Distributing Epistemic Functions and Tasks - A Framework for Augmenting Human Analytic Power With Machine Learning in Science Education Research

Marcus Kubsch
IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Christina Krist
University of Illinois at Urbana-Champaign

Joshua M. Rosenberg
University of Tennessee, Knoxville

Author Note

All Authors contributed equally to this manuscript.Correspondence concerning this article should be addressed to Marcus Kubsch. Email: kubsch@leibniz-ipn.de

Abstract

Machine learning has become commonplace in educational research and science education research, especially to support assessment efforts. Such applications of machine learning have shown their promise in replicating and scaling human-driven codes of students' work. Despite this promise, we and other scholars argue that machine learning has not yet achieved its transformational potential. We argue that this is because our field is currently lacking frameworks for supporting creative, principled, and critical endeavors to use machine learning in science education research. To offer considerations for science education researchers' use of ML, we present a framework, Distributing Epistemic Functions and Tasks (DEFT), that highlights the functions and tasks that pertain to generating knowledge that can be carried out by either trained researchers or machine learning algorithms. Such considerations are critical decisions that should occur alongside those about, for instance, the type of data or algorithm used. We apply this framework to two cases, one that exemplifies the cutting-edge use of machine learning in science education research and another that offers a wholly different means of using machine learning and human-driven inquiry together. We conclude with strategies for researchers to adopt machine learning and call for the field to rethink how we prepare science education researchers in an era of great advances in computational power and access to machine learning methods.

Machine learning (ML) is now commonplace in our day-to-day lives and, increasingly, in science education research and assessment. However, the most visible applications of ML published in science education research journals have focused on developing systems for automated assessment (Zhai et al., 2020a). As Zhai et al. argue, this line of research–while providing glimpses at the potential of ML in science education research–has not yet had a transformative impact. We agree with this perspective and argue that it is representative of a broader symptom of applications of ML not only in science education research but across disciplines: an emphasis on teaching machines how to replicate human behavior towards goals of automation.

Teaching machines to replicate human behavior is potentially powerful. Doing so allows us to automate performances, such as the grading of short written responses by students in response to prompts on worksheets or assessments. Historically, this accomplishment was met with great skepticism from funding agencies and other scholars. Based on our experiences at conferences including NARST, this skepticism about whether computers can ever tell us anything meaningful about human-generated text is still present. Despite this skepticism, automating performances in science education research has been the basis for valuable applications like personalized learning or for freeing teachers' time from routine grading tasks to making key teaching-related decisions. At the same time, we argue that a sole emphasis on replicating human performances is limiting and comes with an opportunity cost in terms of how ML could be used in transformational ways for science education research.

Instead of exploring the full range of ML applications, research focused on using ML in ways that automate existing performances can only bring about a change in the quantity of the research efforts undertaken, not their quality (Nelson, 2020; Sherin, 2013). In other words, through these applications of ML, we effectively gain more of the same kinds of insights, perhaps with greater statistical confidence. This leaves a range of applications of ML unexplored, namely applications that aim at producing *qualitatively* new insights[1]. In consequence, it is hardly surprising that Zhai et al. (2020a) found no transformative impact of ML on science assessment. Moreover, replicating human performance through the use of ML still requires substantial expertise and work on the part of humans (see, for example, the detailed methods descriptions in Maestrales et al., 2020)– and they may only work for specific types or subsets of assessments of students' learning or performance (e.g., Zhai et al., 2020a).

While we are enthusiastic and passionate about the potentials of ML for science education research, we are concerned that the science education research community is most visibly taking up ML in a way that inhibits its transformative potential. Our stance is that the field is currently lacking methodological frameworks for supporting creative, principled, and critical endeavors to use ML in science education research. Different from current frameworks that provide incisive guidance for how to use ML in science education assessment (Zhai et al., 2020a), we argue that we need a framework for a broader range of science education inquiry, one that encompasses goals in addition to assessment. We hope that this framework could also provide leverage for scholars to argue for the significance of publishing the varied ways that ML can contribute to science education research methods and scholarship.

[1] Thanks to an anonymous reviewer for pointing out that this apparent lack of exploration of more innovative applications of ML may reflect a publication bias: there may be innovative uses of ML embedded within larger programs of research that are poised to produce these new insights. These findings, however, may be published in such a way that the use of ML is not highlighted as a primary contribution of the research; or, given the challenges associated with publishing methodological innovations, it may be that this work primarily "lives" in conference publications or other less formal venues. We think both of these cases support our argument: that the most visible uses of ML in science education currently are those focused on automating assessment.

We aim to offer a vision for our work and the work of others in the emergent community of scholars advancing ML in science education research. To do so, we present a framework–*Distributing Epistemic Functions and Tasks* (DEFT)–that can help us to think about how ML can be used in our field. To leverage the full potential of ML, we argue that we need to more clearly articulate the work and its epistemic function that both humans and computers are doing during any kind of analytical process to support creative, principled, and critical uses of ML in science education research.

A Primer on Machine Learning

To rethink how we can use ML in science education research, we first briefly describe what ML is and how it comes to its results. We point the interested reader to James et al. (2013) for a more extensive review.

ML can be thought of as a subset of the broader aims of developing *artificial intelligence*. Artificial intelligence (AI) aims at enabling machines to exhibit performances and behavior that is considered "intelligent," such as accurately predicting outcomes in novel situations or responding adaptively to changing circumstances (Russell, 2021; see also James et al., 2013; Breiman, 2001). ML efforts are focused on how computers acquire new information or knowledge. There are three principal approaches for this, depending on what is to be learned: 1) supervised ML for learning relations and classifications, 2) unsupervised ML for learning structures and patterns, and 3) reinforcement learning for learning the rules governing a system.

In supervised learning, a computer is *trained* through the use of already-coded data. In essence, the machine learns a function, $f(x) = y$ that accurately predicts y for any x. Image classification is a typical example of supervised learning.

In unsupervised learning, a computer is used to discover codes (codes as in the codes that are provided to the computer during supervised ML applications). Importantly, unsupervised applications are carried out in an exploratory mode: the nature or even the number of such codes does not need to be known *a priori*. Recommender systems, e.g., for what to watch next on a streaming platform based on previous views, are typical applications of unsupervised learning.

In reinforcement learning, a computer is allowed to act in a variety of ways in a defined environment, e.g., make a move in a virtual chess game. Each way of acting has a certain probability to occur. Depending on the success of the action (winning vs. losing a chess piece), the probability of the respective action to occur again in the future is either increased or decreased. Computer programs that play board games or computer games are typical examples. As reinforcement learning requires very large amounts of data, it is still rarely used in educational research.

Machine Learning in Science Education Research

In the past several years, there has been a coalescence of work carried out in science education on how ML can impact our research and assessment efforts. Central to this coalescing of research have been critical essays (Cheuk, 2021), frameworks (Zhai et al., 2020a), review papers (Zhai et al., 2020b), and meta-analyses (Zhai et al., 2021) on the use of ML in science education contexts. We begin our review of this research with what has been identified as the predominant use: as a part of supervised ML applications.

Supervised ML to Replicate and Scale Human Coding

We know from recent review papers that science education research that uses ML has largely relied on *supervised ML approaches* (Zhai et al., 2020a,b, 2021), largely in efforts to replicate and scale human codes. Several scholars have advanced the use of supervised ML approaches for science education assessment (e.g., Jescovitch et al., 2021; Maestrales et al.,

2021; Nehm & Haertig, 2011; Nehm et al., 2011; Shiroda et al., 2021; Zhai et al., 2020b). Common to all of these approaches is the use of training and testing data as part of supervised ML. Another systematic review by Zhai et al. (2020b) also found that around 85% of reviewed studies that used any type of ML approach used supervised approaches and that automatically assigning scores and classifying and predicting student responses were the most common *reason* for using supervised ML. While the *type* of data used in such supervised ML applications to replicate human codes may vary from constructed response questions that often take the form of prompts for students to respond in relatively short, freely-written responses to longer essays (Zhai et al., 2021b), this type of ML is predominant in science education research, a conclusion also reached and tacit in the inclusion criteria for Zhai et al.'s (2020a) review (namely, only papers using supervised ML were included in their systematic review of how ML was used in science education assessment).

As an example of such an approach, Maestrales et al. (2021) examined the written responses from almost 7,000 high school students from the states of California and Michigan in the United States to four items drawn from the National Assessment of Education Progress (NAEP). First, the authors established the degree of human-human agreement for the use of coding frames for each of the four items; the coding frames had three levels: a) correct, b) incorrect, and c) correct in a way that also draws on what the authors termed "multi-dimensional" reasoning – reasoning that draws not only on an understanding of disciplinary core ideas but also science and engineering practices and crosscutting concepts. After multiple rounds of training and coding, human-human agreement for the four items was sound, with Cohen's Kappa values ranging from .64 to .80. After establishing reliability, the authors coded the remaining responses and then used these responses to train an algorithm using the Automated Analysis of Constructed Responses (*AACR*; AACR, 2021) web-based system. As a part of this training, the authors evaluated the degree to which the AACR system could predict a code (correct, incorrect, or correct in a way that draws on multidimensional reasoning), using the human codes as the reference group (or the "ground truth"), but not providing this known code to the algorithm. The authors found that the AACR algorithm was able to predict the human codes soundly, with Cohen's Kappa values ranging from .69 to .81. The conclusion from this work was that such a system could be used to code other responses at scale.

Such applications of supervised ML – to automate assessment scoring – are common and potentially powerful. As noted earlier, such uses can be and have been used to automatically code a large number of student responses, classify student responses to inform research efforts and make predictions as a part of technology-based tools and systems that can "adapt" the teaching strategies embedded in them to help students to learn about scientific ideas (Zhai et al., 2020a).

At the same time, there are some reasons why supervised ML may not be the panacea it first may appear to be. First, the time-intensity of the training process limits the wide use and applicability of supervised ML assessments (Zhai et al., 2021; this is especially salient as many assessments likely do not readily generalize across content areas at all (Nehm et al., 2011) – or do not generalize as well across them (Gobert et al., 2013).

Further, there is reason to believe that many supervised ML applications may be unsuitable for sizable groups of students. Cheuk (2021) argues that the use of ML can fail to provide valid inferences about students from non-dominant linguistic, racial, or ethnic backgrounds. Such concerns are difficult to ameliorate when supervised ML can only replicate the inputs provided to it—inputs that may reflect bias in how items are constructed, data is analyzed, and inferences in the process of coding are made. Further, while measures against bias exist in principle, e.g., universal design for learning guidelines for item construction (Rose et al., 2018) or differential item fucntioning (DIF) analyses in test

construction, it often remains unclear how exactly they have to be implemented (Kitto & Knight, 2019; Corbett-Davies & Goel, 2018) to achieve the goal of fair and valid inferences.

Another reason concerns the data used: Zhai and colleagues (2020a) note that one reason why science education research and assessment efforts have not been transformed by ML is that scholars have primarily used conventional data collection methods – particularly constructed response items that are intended to collect students' free-written responses. But, they note that some scholars have used "innovative data collected from high-tech systems, such as eye-tracking systems or fMRI" (p. 1451). Moreover, Zhai and colleagues point out that the few studies of educational games and simulations "have transformed the approach of data collection, as traditionally we acquire data by rating students' products" (p. 1450). We consider such data and how data has a bearing on how ML is used in the next section.

Using Digital Data and Other New Sources of Data

Another use of ML in science education research and assessment centers on the analysis of data that is difficult for humans to code, such as digital trace data— "detailed records of social interaction" (Welser et al., 2008, p. 116) recorded via digital technologies, such as simulations and games (Dickler et al., 2019; Gerard & Linn, 2016; Gobert et al., 2012, 2013, 2015; Li et al., 2017; Liu et al., 2016). Such data have also been used widely outside of science education contexts (e.g., Baker et al., 2010; Shute, 2011).

A key difference between using digital data vs. using supervised ML to replicate human codes is that much of this digital data is not readily interpretable by humans. While humans can code constructed responses from students or students' essays, it is not as tractable for humans to make sense of the "log" files that educational technology tools can generate, especially when the data records fine-grained interactions with as many as 70 or more unique variables (e.g., Gobert et al., 2013). In this context, ML offers a means through which analysts can use abundant digital data streams generated by simulations (Xing et al., 2020), games (Dunleavy et al., 2009), intelligent tutoring systems (Graesser et al., 2012), and conversational agents ("chatbots"; Johnson et al., 2000; Xu et al., 2021) to explain and predict students' outcomes, such as their involvement in science practices (Gobert et al., 2013; Spikol et al., 2018).

As an example of the use of digital data for science assessment and research, Gobert et al. (2013) undertook a process of qualitatively coding students' actions within microworlds, simulations for exploring and understanding phase changes and density. Gobert and colleagues used "clips": logs of multiple ways students interacted with the simulation, such as changing the level of heat and the amount of substance in separate steps or opening and closing a list of hypotheses in the phase change microworld. Logs of such actions were then labeled with a code that characterized an aggregate activity, such as *designing a controlled experiment.* The qualitative codes along with the digital log data for particular features (e.g., for how many times students changed the value for a variable, such as the variable for heat and for how many times students ran a trial or experiment within the microworld) were then used in a supervised ML approach to code the "clips." The predictive accuracy of the supervised ML approach was validated and deemed sufficiently accurate for two purposes. First, Gobert and colleagues describe how their approach can serve as a performance assessment that may prove to have greater validity than traditional (constructed response and especially multiple choice) assessments. Second, this work can lead to the development of intelligent tutoring systems or conversational agents that support learners to be able to, for instance, design controlled experiments more readily.

The above-described work is complex and works like it are lauded by Zhai and colleagues (2020a) as more transformational. Moreover, such simulations and games could be equipped with further means of data collection, such as classroom or computer-based

video recording tools (e.g., Bosch et al., 2015), to collect an even more varied collocation of data. Thus, new data sources such as those that are generated by digital technologies *could* contribute to the transformation of science education research and assessment.

At the same time, however, new data sources could be used in traditional ways: the work of Gobert et al. (2013) uses new data sources but toward the relatively well-established aims of assessing students' capacity to engage in science practices or inquiry with greater validity or to be able to use these new data sources to provide timely scaffolds. What might wholly different *functions* of ML be like? And, how might such functions' distinctive features extend beyond the nature or even the complexity of the data sources used? We consider these questions further in the next section.

Triangulation of Analyses

Qualitative analyses have played a role in the research described above, but in both cases, qualitative methods were used to generate data that served as inputs for subsequent ML analyses. There are a few instances of other ways that ML and qualitative analyses have been coupled in science education research (e.g., Anderson et al. 2020; Beggrow et al. 2014; Rosenberg & Krist, 2021; Sherin 2013). In these uses of ML, ML and qualitative methods are seen as two means to achieve the broader epistemic end of the research, i.e., the understanding of a particular focal construct. Each approach – ML and qualitative methods – provides unique, complementary insights: doing so involves using "human-based and computational methods in tandem, in a manner that increases our confidence in both" (Sherin, 2013, p. 602). We can think of such analyses as triangulating inferences from data.

Sherin's (2013) work illustrates triangulating ML and qualitative methods. With the intent of understanding students' conceptions of the Earth's seasons and their causes, Sherin used unsupervised ML methods - namely, a latent semantic analysis technique that involves combining natural language processing of students' transcribed responses to interview questions, leveraging vector space representations of their responses, and then clustering the vector space representations to identify groups of common conceptions of the seasons. Sherin then compared these groups that emerged from the unsupervised learning procedure to those generated through an original qualitative analysis of the same student responses to the interview questions. Sherin reported a high degree of convergent validity that resulted from interpreting the combination of the ML and qualitative analyses. Notably, no training nor testing data were involved in this analysis; instead, the use of ML offered a different, data-based portrait of what students thought and said. Others have built upon Sherin's work by ordering the use of unsupervised ML, qualitative analysis, and supervised ML through the adoption of a *computational grounded theory* approach (Rosenberg & Krist, 2021). Given the inclusion criteria used in Zhai et al.'s (2020a) systematic review (only considering studies that used supervised ML for science education assessment), the work of Sherin (2013) would not have been included – and, thus, its transformational potential for offering analysts new ways to understand and triangulate their findings using traditional methods may be under-appreciated.

Moving Beyond Automation With Machine Learning

As the previous section has described, the primary usage of ML in science education has been in assessment with a focus on automation and efficiency. With ML systems matching the performance levels of humans, it is not hard to see why there has been a focus on automation, as automation is a straightforward application with concrete benefits.

However, it is important to note that how ML systems arrive at these performances does not resemble how humans arrive at them as human and artificial cognition work differently, despite what commonly-used terms like *artificial intelligence* or *neural network*

may suggest (Clark et al., 2016; Schölkopf, 2019; Russel, 2020). In consequence, tasks that are easy for humans can be very difficult for computers and vice versa (Schölkopf, 2019; McElreath, 2021). In the context of automation, this can become an issue because parts of a task to be automated that are trivial for humans may be very hard ML problems.

This raises the question of whether there is a different way to use ML more flexibly than in the context of automation – a way in which we can profit from how humans and machines can work in integrated, complementary ways to come to their results. We present here one such framework that describes a series of dimensions along which various computational, quantitative, and qualitative analytic approaches might be considered and combined for science education research. Loosely inspired by the theoretical perspectives of distributed cognition (Hutchins, 1995), network epistemologies (e.g., Zollman, 2013) and hybrid-augmented intelligence approaches in engineering (e.g., Zheng et al., 2017), we argue for considering humans and machines as part of a single analytical system and strategically considering the roles that each can play. In other words, how can we combine and sequence uses of humans and ML in a complementary, augmentative way? What a "complementary way" is critically depends on the task at hand. In this paper, we will focus on the research task of *analyzing research data*.

A Framework for Rethinking the Roles of Human and Machines in Science Education Research

The framework we propose revolves around how different tasks and their respective epistemic functions in an analysis that uses ML are distributed between the human analyst and the machine. We call this framework DEFT (Distributing Epistemic Functions and Tasks). DEFT consists of three key questions: a) What are the tasks? b) Who is doing what task?, and c) Who *should* be doing what task, based on their relative strengths and weaknesses? We present this framework as a working model. As such, it is intentionally abstracted (and thus oversimplified) in some regards in order to highlight certain points; and it should continue to be refined as more work is done. In this section, we will demonstrate how one might use DEFT to rethink the roles of humans and machines in analyzing research data.

We begin by asking the question a) "*What are the tasks*?" Figure 1 shows an idealized (i.e., intentionally over-simplified) process of analyzing research data that describes three activities that we might consider as the key tasks in analyzing research data: *setting goals*, *assigning codes or numbers to data*, and *looking for relations and drawing inferences*. We acknowledge that analyzing research data is much messier in the "real world" than Figure 1 suggests; we begin with these key tasks as a strategic essentialization of the process to help us illustrate the kind of thinking that we hope the DEFT framework could support.

PLACE FIGURE 1 HERE

1. Goal-setting. The first step in this endeavor is usually to set goals, i.e., determining what one wants to measure or identifying and motivating why that matters. Once the goal of an analysis is set, working with the actual data begins. *2. Assigning codes or numbers to data.* This is a task that requires establishing a connection between the data and what one wants to measure or identify, i.e., the principled assignment of numbers or codes to observations. *3. Looking for relations and drawing inferences.* Now, with numbers or codes at hand, one looks for relations in the numbers or codes, often to draw inferences and describe and explain the phenomenon under study. As suggested by the gray arrow in Figure 1, this whole process is iterative.

We now turn to an in-depth discussion of questions b) *Who – computer or human – is doing what task?* and c) *Who should be doing what task based on their relative strengths and weaknesses?* As we consider these questions, we recognize the potentially artificial nature of

attempting to cleanly delineate what humans do vs. what computers do, especially given the sociomaterial entanglement of scientific work. We view this entanglement from a perspective based broadly in distributed cognition (e.g., Hutchins, 1995): while each individual component of a distributed system has a role, that function only makes sense in the context of the system. In addition, these roles are defined by differences in the type of information available to each individual: ML tools are able to "see" aspects of or patterns in data that are rendered incomprehensible to humans, whereas humans are able to consider aspects of theory and context that can be quite difficult to algorithmatize in a robust way. This distributed delineation of information access amongst a network of cognitive agents is argued to be an important characteristic for effective knowledge-building (e.g., Zollman, 2010). In this spirit, we aim to analytically delineate who is and should be doing various tasks, based on their information access and capabilities, in order to support reflective awareness of and decision-making about the possibilities for taking on roles within the system.

Activity 1: Setting Goals

Currently – to the best of our knowledge – goal-setting in all paradigms of research is predominantly done by humans. This is hardly surprising, as analyses of research data are carried out to answer questions that are being asked by humans. One may wonder how computers could support this task, which seems so intrinsically linked to human curiosity, creativity, and theoretical sensitivity. However, recent research and scholarly work highlight how the questions that are being asked and the analyses that are being done to answer them are not only the results of human curiosity. They are also informed by cultural and societal norms, values, and ideologies (Benjamin, 2018; Crawford, 2021). In the past, ideologies such as eugenics, phrenology, and racism have informed the goals of research (ibid.). Here, computers could support humans by spotting potentially discriminatory questions and goals, suggest complementary analyses, such as a subanalysis that considers how the effects of an intervention vary for different demographics or even highlight gaps in the current literature. A recent study by Odden et al. (2021) that used ML to examine how the research topics in *Science Education* changed over the last 100 years – and how this is in part driven by intellectual cross-pollination – can be seen as the first step in this direction.

Activity 2: Assigning Codes or Numbers to Data

Before discussing *who* is typically assigning codes or numbers to data and what the relative strengths and weaknesses of humans and computers are in doing so, it is helpful to deconstruct the task of assigning codes or numbers to data based on its epistemic function and the level of inference. The *epistemic function* can be characterized by how well-defined the coding categories are a-priori, i.e., do we know what we are looking for because there is an existing survey, coding manual, or rubric? Or, does part of the task involve developing emergent categories and codes (i.e., a grounded approach)? This distinction is similar to the difference between inductive and deductive analysis in qualitative content analysis (e.g., Kuckartz, 2014) and supervised vs. unsupervised ML techniques. The second characteristic, high vs. low inference, concerns how well-defined the coding of the data is. Mappings, i.e., the assignments of codes to data, might be well-defined and require little inference. For example, when we are looking for the usage of technical language, the sequences of letters that signify key terms are clear. In contrast, mappings might be less well-defined and require large amounts of inference for constructs such as students' epistemic beliefs. Such constructs include a range of definitions and operationalizations in the literature. Figure 2 presents these dimensions as four quadrants where each quadrant serves a different epistemic function. We describe typical examples of assigning codes or numbers to data that might fall within each of these quadrants next.

PLACE FIGURE 2 HERE

Low inference, a priori. A typical example of a task in this scenario would be to analyze a test that asks for short written responses to questions about factual knowledge, such as selecting one's birth year or stating Newton's 2nd law. The mapping between the data and the code is low-inference, and the code is determined *a priori*. Thus, it effectively requires following simple rules which both humans and computers can carry out with little training. Here, the obvious strength of the computer is that it follows simple rules flawlessly and relentlessly. An example of such an analysis can be seen in the work of Li et al. (2017) who used a supervised computational approach: hard-coded rules to automatically code students' explanations. Whether these strengths can be effectively deployed mostly depends on how easily the data can be accessed by the computer and how much data there is, i.e., whether scale effects warrant the initial workload of setting up the automated coding system. However, as Zhai et al. (2020a) note in the context of science assessment, besides automation, there is little new insight that the usage of computers offers compared to that achieved by human raters. And while automation at scale does have the potential to allow us to answer research questions that were not answerable without the use of large amounts of data, in the low inference, a priori quadrant, ML has a replicating (rather than augmenting) function.

Low inference, grounded. Using a grounded approach in a low-inference coding task involves *identifying patterns,* specifically patterns in which the general pattern maps closely to the data itself. For example, an analyst might sort student responses by whether they talk about cooking with their family as *enjoyable* to them or whether they describe it as a *chore*. These categories were not determined a priori, but they can easily be inferred by the presence of certain words: *fun, exciting, love* vs. *boring, hard, hate,* etc. Unsupervised computational methods like sentiment analysis (Feldman, 2013) are suited to conducting this kind of pattern recognition. In conducting this type of analysis, humans and machines bring potentially complementary aspects. Humans bring strengths in recognizing patterns by bringing additional knowledge about the data to the task in ways that computers do not. Humans know the social, historical, and cultural context, which can help to reveal mappings that are not in the data accessible by the computer. Furthermore, as unsupervised methods usually require relatively large amounts of data to identify patterns reliably, patterns from underrepresented groups may be missed. For instance, Cheuk (2021) describes cases in which computers struggle with vernacular expressions. Further, unsupervised approaches such as clustering algorithms only group data based on statistical patterns but do not provide interpretations of what the groups represent. This task can only be accomplished by humans.

At the same time, computers may offer some advantages concerning low inference, grounded coding. The social, historical, and cultural context that humans bring to a task can also unconsciously influence how they approach the data. In this way, unsupervised computational methods can find patterns that humans do not expect (e.g., Russel, 2021; Rosenberg & Krist, 2021) or may unconsciously suppress. Another strength of the computer comes from its ability to consider large amounts of data, i.e., given enough data, computers can find patterns that humans may never recognize given the limitations of human memory. Hope and Witmore (2014) provided an example for this when they discovered how the patterns in which the word "the" occurs in Shakespeare's Macbeth is a major source of the uncanny tone of the tale; there may be similar patterns to be explored in science education contexts, such as whether patterns in sentence structure of contributions to whole-class discussions impact the how influential a particular student's contributions are during argumentation. In sum, for low inference, grounded tasks, humans, and unsupervised computational approaches can complement each other. The size of the computer's memory and its insensitivity to context can complement human context-sensitivity and suggest new patterns in data. However, as any pattern identified by a computer has no "natural" meaning,

it requires human interpretation. In this way, the resulting analytical system can be seen as mostly augmenting human memory capacity and adding a layer of less contextualized information processing.

High inference, a priori. A high inference, *a priori* task is one in which the mapping between the data and the abstracted categories can be defined *a priori* but is (comparatively) hard to define in and of itself. Coding a defined range of students' non-normative ideas represents a good example of this kind of task. While the range of students' non-normative ideas about, e.g., energy (Lancor, 2015), is well described in the literature and coding manuals do exist, the same idea can be expressed in numerous different ways and using a range of different metaphors. Further, the context will influence how the idea is expressed. In consequence, the patterns that need to be identified in high inference, supervised learning are often more abstract and less clear cut than those in the low inference, supervised setting. While both humans and computers can carry out this task, more training is required (for both humans and computers!) than if the coding were lower inference.

Maestrales et al. (2021) provide an example of this type of analysis when they use supervised ML to score multidimensional (i.e., NGSS-aligned) chemistry and physics assessments. They demonstrate that carefully calibrated machine scoring algorithms can replicate human codes with good to excellent scoring accuracy. However, as Nehm & Härtig (2012) note, the economic advantage remains questionable unless the same set of items can be used with very large samples and a representative sample of the population which is to be tested can be acquired to train the machine scoring algorithms. Further, as Ruha Benjamin (2019) and Kate Crawford (2021) have noted, economic advantages resulting from ML systems are often generated by exploiting indigenous, minority, or economically underprivileged groups. This exploitation ranges from the destruction of preserved land through mining for minerals and metals needed in computers and mobile devices to the profiteering of cheap labor for spotting and correcting mistakes that ML systems make (Crawford, 2021). Overall, for high inference, supervised ML, the effort required to train computers can severely limit the cases where using computers for coding is faster, more accurate, and cheaper than relying on human coders.

Lastly, Gebru et al. (2021) describe how the algorithms often used for high-inference ML are effectively black box systems, which makes it hard to judge the validity of the codes. This validity essentially hinges on the computer reproducing human codes, positioning those codes as a "gold standard." This positioning has been increasingly questioned by methodologists (e.g., Benjamin, 2019; Cheuk, 2021; Nelson, 2020). In sum, similar to the function of ML in the low inference, a priori quadrant, the use of ML in the high inference, a prior quadrant is mostly replicating (rather than augmenting) human performances, albeit with the intensification of ethical and methodological issues tied to the usage of computers.

High inference, grounded. High inference grounded analysis also requires engaging in pattern recognition. As discussed before, when humans look for patterns that are not defined a priori, they cannot escape that the patterns they will find will be influenced by their sociocultural background and knowledge about the world. This is both a feature and a bug as it can help to recognize patterns that are accessible to a computer but it can also (inadvertently) suppress patterns that a computer can identify based on statistical patterns. Further, computers can detect patterns in data that are not accessible to humans due to limited working memory capabilities. For example, computers can analyze log data from online learning platforms to find patterns that indicate learning behavior (e.g., Zhang et al., 2017).

Given the challenges, this type of coding is rare relative to the others. Sherin's (2013) work provides a notable example. Drawing on the potential differences in how humans and computers might approach high-inference pattern-seeking, Sherin (2013) used unsupervised ML in the analysis of student interview data about seasons. The goal of this application of

ML was to see whether a computational algorithm would identify the same instances of conceptual shifts in how students were explaining seasons as a human coder had. Rather than aiming to replicate the human to automate coding, the goal was to compare between the human and computer's coding as convergent evidence for the validity of the claims made from the data. Other work by Sherin elaborates on and expands how this convergent validity evidence can be generated through using qualitative and computational methods in concert (Sherin, 2015; Sherin et al., 2018).

Activity 3: Looking for Relations and Drawing Inferences

Once data have been assigned codes or numbers, the typical next steps are looking for relations between the codes or numbers, interpreting them, and drawing inferences about them.

In quantitative research, this work is traditionally shared by humans and computers, functioning as a complementary system where humans specify relations based on substantive theory and use computational statistical tools to quantify the relations, test hypotheses, and draw inferences about generalizability. A typical example of this would be to set up a regression model based on substantive theory and use statistical software to calculate the regression coefficients and conduct statistical tests of the model. The result of such an analysis is a descriptive and potentially explanatory model of the phenomenon under study. However, as Breiman (2001) has argued, advances in ML provide an alternative. Instead of letting humans specify the relations between the codes or numbers in the form of a statistical model such as a regression, humans may just specify which codes or numbers x is supposed to be related to codes or numbers y and let the computer find a function $f(x) = y$ that describes the relation. The result is a model, often uninterpretable to humans, that can predict y given x with a certain accuracy. This approach is typical of many supervised ML use cases.

The predictive accuracy of such supervised ML models is usually superior to the more traditional statistical model that aims at providing an interpretable and mechanistic model. In essence, predictive accuracy is traded for interpretability and causality (Breimann, 2001; see also Pearl & McKenzie, 2018, Schölkopf, 2019). However, both interpretability and causality are crucial qualities if one wants to intervene in the phenomenon under study. In sum, when it comes to looking for relations and drawing inferences there are two distinct ways to approach this as they serve different goals: explanation and prediction. In both cases, humans and computers complement each other. They produce an outcome neither could produce on its own. The only decision that humans have to make is which outcome they are aiming for – explanatory or predictive models.

In (grounded) qualitative research, the work of looking for relations and drawing inferences is traditionally done by humans. We address two considerations that continually guide qualitative analyses: (a) whether a pattern or theme is meant to be representative or anomalous, as both have value (Glaser, 1965); and (b) in what ways the positionality of the researcher is influencing the analysis (Lincoln, 1995). In terms of representativeness, commercially-available qualitative analysis software (e.g., NVivo, MAXQDA) increasingly provide a suite of tools to assist in consolidating, organizing, and visualizing qualitatively-analyzed data. These tools range from quite simple, such as displaying the total count of each code across a set of documents, to more complex, such as network-type visualizations of code co-occurrences. The goal of such tools is not to produce statistical outputs, but rather to assist the researcher in conducting second-stage coding or building qualitative claims. In other words, these high-level patterns, such as the number of times a particular code occurred, are often an important starting point in supporting qualitative analysis, even though the counts of codes themselves are typically not the claim being made. Rather than the *end* (i.e., a claim), these patterns are the *beginning* of a qualitative analytic process. It stands to reason, then, that unsupervised approaches such as topic modeling or anomaly detection could serve a similar

role: they are not themselves generating a claim about the data, but they are surfacing a pattern from qualitative data that an analyst could then interrogate more deeply.

In terms of researcher positionality, unsupervised approaches may play a similar role. Rather than being used to *confirm* the human coding as accurate when compared to some computational output, instead, *discrepancies* between human coding and computationally-detected patterns may themselves be an important starting point for inquiry. These investigations may shed light onto biases or ideologically-laden patterns in the data; they may also reveal important researcher standpoints that led to the identification of particular codes, connections, or themes. Articulating and theorizing around these discrepancies may itself prove to be a generative knowledge-building activity for the field.

Applying the Framework: Two Cases from the Literature

To demonstrate the value of the DEFT framework, we will use it to characterize two recent publications: Maestrales et al. (2021), which presents a use of ML to score multi-dimensional science assessments, and Nelson (2020), which presents a use of ML to study how geographically-bound distinctions in political logics persisted across time during first and second wave feminist movements in the US. We have selected these papers to showcase both *where we are* and *where we could go* in terms of utilizing ML in science education research. Specifically, we view Maestrales et al. (2021)'s work as reflecting the "state of the art" in using ML in science assessment. We describe how these authors are doing sophisticated work in one area of the framework we presented above. Then, because a comparable publication does not currently exist in science education, we present Nelson's (2020) work from sociology to stretch our collective imaginations and help us envision how we might apply her methodological framework to further expand the uses of ML in science education research.

Maestrales et al. (2021): Machine Learning in Science Assessment

Maestrales et al. (2021) provide a cutting-edge example of how ML is typically used in science education assessment. Such a use aligns with the first category of a research study using ML in science education we reviewed – those using supervised ML to replicate and scale human coding – generally, a common use of ML (Zhai et al., 2020a, 2020b).

Maestrales et al. (2021) examine "the use and accuracy of an ML text analysis protocol as an alternative to human scoring of constructed response items." The authors describe their overall approach and the development of the automated scoring algorithm in a flowchart of their workflow (Figure 3 presents an adapted version).

PALCE FIGURE 3 HERE

What are the tasks? As seen in Figure 3, the nature of the analytic work focuses on *assigning codes or numbers to data.* More specifically, their work can be characterized to belong in the *high inference*, *a priori / supervised* quadrant of Figure 3, as the coding rubric existed before they started scoring. Scoring three-dimensional student performances from the text can certainly be considered a *high inference* task, as it involves evaluating students' use of disciplinary core ideas, science & engineering practices, and crosscutting concepts in an integrated way.

Who is doing what task? Again, referring to Figure 3 provides a straightforward answer to this question: first, humans score students' answers using a-priori defined rubrics, next ML is used to replicate that process reliably.

Who should be doing what task based on their relative strengths and weaknesses? In this paper, Maestrales et al. utilize humans' interpretive power to generate the initial scoring rubrics. They then train computers on those data and ultimately reproduce humans' coding for three of their items at scale. In this sense, they are leveraging the strengths of

humans and of computers. At the same time, based on the challenges described in their paper, the original human coding *needed* to be low-inference for the computer to be reliably trained to reproduce it. And, as discussed above, the advantage of scalability is limited in this context: the computer can score an infinite number of student responses to three precise items, but no new items can be scored within this coding system.

We posit that considering the other three quadrants – those other than the *high inference*, *a priori / supervised* quadrant – in the framework presented above can offer additional methodological approaches that leverage the strengths of humans and computers in slightly different ways.

Nelson (2020): Integrating ML Through Computational Grounded Theory. In this methodological paper, Nelson used ML techniques to investigate the political logics in first- and second-wave feminist movements, primarily using documents produced by political organizations in New York and Chicago during each era. To do so, she developed a methodological framework she termed *Computational Grounded Theory* (CGT). Such use is in alignment with the research that involves the triangulation of analyses (e.g., Sherin, 2013), rather than replicating and scaling human coding. In this way, it provides a contrast with the work of Maestrales et al. (2021) and the other ML-based studies in science education research and assessment that predominantly use supervised ML.

In CGT, one first leverages the power of computational techniques, especially unsupervised ML techniques, for pattern detection in large datasets—those of a size and scope that may prohibit human-driven analyses from the outset. Then, one leverages the interpretative power of humans to add quality and depth to the quantity and breadth of the first step. Finally, one uses computers to test the reliability and generalizability of the human refined pattern detection and interpretation from the second step. Nelson also provides a figure describing her workflow; an adapted version of it can be found in Figure 4.

What are the tasks? In each of Nelson's three steps, assigning codes or numbers to data is followed by looking for relations and drawing inferences, which inform the next step leading up to the final sociological conclusions. This process is inherently cyclical, and the nature of the analytic work varies slightly at each stage.

PLACE FIGURE 4 HERE

In step one, Nelson utilized *unsupervised* computational methods to identify both *low-inference* patterns (via lexical analysis, i.e., basic natural language processing techniques) and *high-inference* patterns (via topic modeling). In step two – the pattern refinement step – the analytic work remains *grounded* (and primarily done by the human analyst) but is predominantly *high-inference*. In this step, the human analyst uses a standard qualitative method – content analysis via guided deep reading – to interrogate the patterns that were identified using the computational techniques in step 1, though only utilizing a small subset of the dataset. In the third step, pattern confirmation, the analytic work can be characterized as *high-inference, supervised / a priori*. Nelson strategically selected supervised computational techniques that would examine the generality of the claims identified in Step 2 across the entire data corpus.

Who is doing what task? Figure 4 shows that in step one the work is primarily done by computers, in step two almost exclusively by humans, and in step three again primarily by computers.

Who should be doing what task based on their relative strengths and weaknesses?

This framework demonstrates Nelson's (2020) careful consideration of when and how to integrate computational methods to complement the work of the human analyst. Specifically, computational patterns generated in Step 1 were then interrogated qualitatively through standard sociological methods (e.g., content analysis and deep reading). These claims

were inspired by a computationally-detected pattern, but generated through the careful work of a theoretically-sensitive (human) scholar. Finally, Step 3 utilizes (supervised) computational tools to verify that these claims hold across the entire data corpus. By strategically sequencing methods, Nelson's approach explores usages of ML in all but one of the four quadrants in Figure 3, using computation in a way where it complements the role of the human analyst, aiming at "preserving the superior abilities to interpret text holistically provided by humans but incorporating the formal rigor, reliability, and reproducibility of computer-assisted methods." (Nelson, 2020, p.8).

In sum, our analysis – using the DEFT framework – has shown two very different ways of using ML in data analysis. Maestrales et al. (2021) present state-of-the-art usages of ML aiming at replicating the work of human analysts. Nelson (2020) presents a strategic sequencing of different usages of ML aiming at complementing the work of the human analyst. We see promise in following a similar strategy—of thinking *across* quadrants rather than simply within quadrants—to inform the design of research in science education.

Implications for Science Education Research: Looking Forward

While we are convinced that science education research can profit from carefully embracing the developments in ML, we caution the field from missing out on the true potential of this technology. We envision that science education research can and should more creatively explore roles for ML in science education research beyond automating assessment. After all, even an analytic approach aiming for fully automated coding of assessment items involves a significant amount of human analysis and effort (Maestrales et al., 2020). Without carefully considering the roles that humans play in computational analyses, we leave unexamined two critical issues: where and how bias may be introduced into the analysis, and whether and how the approaches are working together.

The first issue is about ethics and transparency. If we are not explicit about what humans are doing in an analysis, or how human and algorithmic analyses are chained together, we are obfuscating potential sources of bias. The goal is not necessarily to remove all sources of bias from analysis (an impossible task!), but rather to be transparent about them such that the impact of inevitable biases can then be discussed alongside the research findings. The second issue is about the nature of the methodological integration we are aiming for as science education researchers. Are we aiming for *replicating* human performances for the sake of automaticity and efficiency, or are we aiming for *augmenting* human performances to generate new insights? In either case, we argue that we need to think carefully about how the roles – of replicating and augmenting can be put to use strategically for our goals. As the framework we developed shows, solely focusing on automation leaves a whole realm of possibly insightful and valuable uses of ML unexplored – the aforementioned opportunity cost of strictly using ML in relatively narrow ways.

We conclude with some implications for the field that have to do with how we might need to re-shape both our methodological thinking and the methodological training we provide for emerging researchers.

Reshaping our Methodological Thinking and Assumptions

In terms of re-shaping our methodological thinking, we need to move away from viewing computers as "output machines" that, if properly trained, can do our work for us. We have seen both emerging and established researchers enamored by the allure of computational techniques the first time they are introduced to them – usually motivated by the sense that if they can only pick the right tool, then a computer could do some of the difficult analytic work for them. We have also seen anxiety from emerging qualitative researchers rooted in a fear that they are wasting time doing analyses themselves when there exists a machine that could do it for them. Unfortunately, efficiency-motivated uses of ML often feed into these allures or

anxieties: they present the goal of such analyses as having computers score your assessments for you, relieving one of the slogs of coding hundreds of assessments, and that they do so with "less bias" – a phrase that often is interpreted to mean that computer scoring is more accurately reflecting "the truth" than would the same scoring done by a human coder.

We think that both of these notions can and need to be dispelled. Yes, computer algorithms can score assessments with a good degree of accuracy – after a substantial degree (e.g., years) of human code development and revision. And yes, computers can quickly identify patterns in text-based qualitative data, such as the most common words, or words that tend to occur together. But these outputs are rarely an answer to a research question that holds much qualitative weight in terms of its contribution to the field. Much like statistical software does not do rigorous quantitative analysis automatically – quantitative analyses involve many decisions on the part of the analyst (Schweinsburg et al., 2021) – computational methods do not replace the difficult work of careful analysts. Given the ways that bias shows up throughout the research process, from question and instrument development to data collection to analysis and interpretation, thoughtful and critical analysts are more essential than ever in doing computational work.

We offer a few considerations that we have begun adopting in our own work as a starting point for using the framework presented to make methodological choices and designs. By labeling them considerations, we mean to emphasize that these are not established best practices or design guidelines (Edelson et al., 2002) but instead are strategies for researchers to think about and potentially adopt when using ML in their research.

Consideration 1: Conceive of the analytic work required when using ML in science education research in terms of how established codes are and the level of inference involved in applying them. While neither piece of this consideration is particularly new (e.g., researchers often consider whether to adopt an existing coding scheme, or whether to create a new one inductively), in our research we have begun talking about these pieces together as we are weighing analytic decisions. Some beginning rules of thumb for us are (1) while *established, low-inference* coding may be easy to do using computational tools, it is often not answering the question we care about; and (2) *inductive, high-inference* coding is difficult all around, and will likely need to be "stepped back" a bit for either humans or computers to conduct it effectively. However, if we need such coding to answer our research questions, it is a good candidate for doing further, deep thinking about how we might strategically sequence methods together.

Consideration 2: Use computational output as an additional source of data in need of further analysis, rather than as a finding on its own. This consideration marks the biggest shift in our thinking and has been essential for us to make progress in thinking about sequencing techniques. It also has freed us from feeling pressure to develop or utilize sophisticated algorithms for their own sake: often simple, straightforward, and off-the-shelf tools like proper noun counters or basic topic modeling can go a long way toward achieving meaningful research goals – if thoughtfully utilized. Most importantly, looking at computational outputs (e.g., topics generated) as mid-point data rather than end-point findings has created space for us to think critically about potential sources of bias, and to then consider how to address those biases with further (human-driven) methodological tools and care in how we conclude from those data. In quantitative analyses, ML could provide additional tools for data exploration, especially in cases where full-powered analyses are not feasible. Similarly, thinking about using simple computational tools to confirm *portions* of patterns undergirding qualitative findings, such as Nelson's (2020) use of a metric of abstractness vs. concreteness of terms, can lend additional confidence to claims without needing to be the foundation of the claim itself.

Consideration 3: Leverage computational tools to help reveal biases in data and in coders. One of the strongest critiques of ML, across all uses, is that it amplifies biases already inherent in data. These biases include asymmetries in which data are collected, and therefore represented in any training set (e.g., Black Americans are overrepresented in terms of arrests, convictions, and time spent in prison), which then lead to biased predictions based on those datasets (e.g., all else equal, simply being Black leads to a higher risk assessment index (RAI) score which predicts the likelihood of committing a crime again; RAIs are used in sentencing). Science education is not immune from these critiques, as we know that schooling and standardized assessment are racially biased. Disaggregating data by gender, race and other demographic variables is often an important recommendation to reveal such potential biases. Although not free of its own issues arising from the collection of very sensitve personal data (e,g., Drachsler & Greller, 2016), this recommendation can also be applied to ML-based analyses. For example, in a topic modeling analysis for open-ended science responses, which topics are most common for white students vs. Black students? Or, which topics are most common in lab report essays written in response to the same prompts written by students at predominantly white vs. minority-serving institutions? (Of course, using critical frameworks to interpret such patterns in terms of the systemic factors causing them rather than interpreting them as individual deficits is essential). These questions could be research questions in and of themselves; see, for example, Ha & Nehm (2016) who studied the effect of spelling mistakes made by English-language learners on classification accuracy of an automated scoring system and did not find meaningful differences. We encourage thinking about those questions of bias and fairness again as preliminary analytic steps to strategically *reveal* and quantify biases in our datasets as first steps towards engaging them. Altogether, this is not to say to abandon ML for the issue of potential bias; rather we urge the field to explicitly engage with the multiple ways that bias may be introduced or amplified when using ML in science education research so that we can tap into the potentials of ML with clear-eyed goals of reducing unintended harms.

We have also begun exploring such uses to examine our research positionalities and biases as well. For instance, if qualitative researchers are writing analytic memos, are there differences in terms of words or topics identified between researchers? What might these differences mean in terms of each scholar's positionality, and how might that impact the development and interpretation of claims? While such a practice is most commonly associated with qualitative research, the same potential pitfalls exist in quantitative work (e.g., Schweinsberg et al., 2021). In other words, rather than attempting to *eliminate* bias, we could adopt a qualitative ethos of *articulating and embracing* sources of variation and how those are impacting the analytic process as criterion of rigor (Lincoln, 1995). In fact, modest versions of such variations have been theorized to lead to more robust scientific conclusions (Zollman, 2010). Such calls have also been made for quantitative psychology researchers: Research findings may be more robust if researchers first work to understand where variation exists using robust analytic methods prior to predicting or attempting to intervene (Rosenberg et al., 2021; Yarkoni, 2019).

Re-thinking Methodological Training in Science Education Research

Across our conversations, we [the authors] have been struck by the interdisciplinarity of our own methods training, coupled with our inclinations to explore methods broadly. A few observations stand out to us from our reflections on our (relatively recent) training. First, we think we were all privileged to be exposed to the *best-case* versions of the use of qualitative *and* quantitative methods, often in integrated ways – rather than the "straw man" version of either. While specific faculty members may have had their preferences, we received training in ways that moved beyond the quantitative-qualitative divide and emphasized critically considering the values and limitations of each – including how to

produce high-quality versions of each type of research. We hope we are not anomalous, and that this is a feature of graduate programs across the globe in science education. At the same time, we recognize that there may be great variations in doctoral student training because of disparities in resources, program structures, or individual preferences.

Second, none of us are "experts" in *all* methodologies, but we know enough about each to be able to engage productively with each other about them. In addition, we respect each others' varied methodological and technical expertise. This mutual respect has been the foundation of several lively conversations about what rigor in science education research is; what the goals of science education (research) are; and even what the value of ML is. We include this point to highlight that we think that multidisciplinary expertise is important, but that without shared language or a shared sense of purpose, we believe it is difficult to creatively, and innovatively, advance any methodological approach. While many scholars may need to develop a grasp on programming or some fundamental computer science principles to be able to engage in conversation with those more technically proficient, the reverse is also true: those with strong technical expertise may also need to saturate themselves in methodological approaches like grounded theory, or deeply interrogate the many applications of educational theories like social constructivism to have truly generative interdisciplinary collaborations.

Third, the inclusion of computational methods as part of a robust science education research toolkit requires thinking carefully about how doctoral training may need to be approached. We note that we each had a mentor or mentors in relatively close proximity who were carefully, critically, and creatively examining emerging computational methods. Additionally, we found peer-group colleagues who were interested in collaborating on exploratory projects "on the side." So, despite our not receiving extensive formal training, we also did not find ourselves in complete isolation trying to make progress on our own. However, we would hope to see resources and support for utilizing computational methods become a standard part of doctoral training. In the past few years, we have seen an increase in one-time support such as pre-conference workshops or online training sessions being offered. Sustained support is critical, including training, workshops, and graduate courses that go beyond, for instance, simply offering tutorials for using ML, and instead couple the skill development with serious consideration of when, why, and how to integrate ML into an analytic workflow.

Conclusion

ML applications are increasingly commonplace in science education research and assessment efforts. Such applications are exciting, but we are not alone in expressing some hesitancy regarding how ML is predominantly used. Most uses of ML have sought to replicate (and scale) human coding, often successfully (Zhai et al., 2020b). Despite these successes, we have argued in this paper that a core issue is that we lack a broader frame with which we can think about the use of ML. Considering the epistemic function of analysis in terms of the degree of inference made in coding and the provenance of the codes (*a priori* or grounded) as well as the tasks that go into analyzing data, we offer the DEFT framework as a way to consider how we are using ML – and how we are not, presently, using ML. Different from a framework for how ML has been used for science education assessment – one highlighting the nature of the construct, function, and automaticity of assessment (Zhai et al., 2020a) – ours focuses on how knowledge is generated through human-driven efforts and the use of ML in science education *research*.

In reviewing two exemplary studies – Maestrales et al.'s (2021) and Nelson's (2020) – we used the framework to draw out what an entirely different use of ML may reveal for science education researchers. Notably, Nelson is a scholar from a different field – sociology – suggesting that science education researchers may gain insights not only from adopting ML

strategies from computer scientists and engineers but also scholars in allied fields who are also seeking to use ML in ways that make progress on the goals that are most important to them as substantive researchers. In selecting and contrasting these two studies, we emphasize the flexible and in-progress, rather than the normative and settled, nature of the framework and also this paper: There is no one way to evoke creative and critical uses of ML in science education research, but like for any truly transformational technological advance, there are likely many and we as researchers need to explore which best support our causes.

References

Anderson, D. J., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). (advance online publication). Evaluating content-related validity evidence using a text-based, machine learning procedure. Educational Measurement: Issues and Practice. https://doi.org/10.1111/emip.12314.

Automated Analysis of Constructed Responses (AACR). (2021). https://beyondmultiplechoice.org/

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies, 68(4), 223-241.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? Journal of Science education and Technology, 23(1), 160-182.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., ... & Zhao, W. (2015, March). Automatic detection of learning-centered affective states in the wild. In Proceedings of the 20th international conference on intelligent user interfaces (pp. 379-388).

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199–231. https://doi.org/10.1214/ss/1009213726

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. Science Education, sce.21671. https://doi.org/10.1002/sce.21671

Clark, A. (2016). Surfing uncertainty: Prediction, action, and the embodied mind. Oxford University Press.

Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. ArXiv:1808.00023 [Cs]. http://arxiv.org/abs/1808.00023

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Dickler, R., Li, H., & Gobert, J. (2019). *A data-driven approach for automated assessment of scientific explanations in science inquiry*. In 12th International Conference on Educational Data Mining, EDM 2019 (pp. 536-539). International Educational Data Mining Society.

Drachsler, H., & Greller, W. (2016). Privacy and analytics: It's a DELICATE issue a checklist for trusted learning analytics. Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16, 89–98. https://doi.org/10.1145/2883851.2883893

Dunleavy, M., Dede, C., & Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. Journal of science Education and Technology, 18(1), 7-22.

Edelson, D. C. (2002). Design research: What we learn when we engage in design. The Journal of the Learning sciences, 11(1), 105-121.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89. https://doi.org/10.1145/2436256.2436274

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92.

Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. Journal of Science Teacher Education, 27(1), 111–129.

Gerard, L., Linn, M. C., & Berkeley, U. C. (2022). Computer-based guidance to support students' revision of their science explanations. Computers & Education, 176, 104351.

Glaser, B. G. (2002). Conceptualization: On theory and theorizing using grounded theory. International journal of qualitative methods, 1(2), 23-38.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. Journal of Educational Data Mining, 4(1), 104-143.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. Journal of the Learning Sciences, 22(4), 521-563.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. Educational Psychologist, 50(1), 43-57.

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. APA educational psychology handbook, Vol 3: Application to learning and teaching., 451-473.

Ha, M., & Nehm, R. H. (2016). The Impact of Misspelled Words on Automated Computer Scoring: A Case Study of Scientific Explanations. Journal of Science Education and Technology, 25(3), 358–374. https://doi.org/10.1007/s10956-015-9598-9

Hope, J., & Witmore, M. (2014). The language of Macbeth. Macbeth: The State of Play. London: Bloomsbury (Arden), 183–208.

Hutchins, E. (1995). How a cockpit remembers its speeds. Cognitive science, 19(3), 265-288.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of ML performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. Journal of Science Education and Technology, 30(2), 150-167.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial intelligence in education, 11(1), 47-78.

Kitto, K., & Knight, S. (2019). Practical ethics for building learning analytics. British Journal of Educational Technology, 50(6), 2855–2870. https://doi.org/10.1111/bjet.12868

Kuckartz, U., & McWhertor, A. (2014). Qualitative text analysis: A guide to methods, practice & using software. SAGE.

Lancor, R. (2015). An Analysis of Metaphors Used by Students to Describe Energy in an Interdisciplinary General Science Course. International Journal of Science Education, 37(5–6), 876–902. https://doi.org/10.1080/09500693.2015.1025309

Li, H., Gobert, J., & Dickler, R. (2017). Automated Assessment for Scientific Explanations in On-Line Science Inquiry. International Educational Data Mining Society.

Lincoln, Y. S. (1995). Emerging criteria for quality in qualitative and interpretive research. *Qualitative inquiry*, *1*(3), 275-289.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. Journal of Research in Science Teaching, 53(2), 215-233.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using Machine Learning to Score Multi-Dimensional Assessments of Chemistry and Physics. Journal of Science Education and Technology, 30(2), 239-254.

McElreath, R. (2021). Regression, Fire, and Dangerous Things. Elements of Evolutionary Anthropology. https://elevanth.org/blog/2021/06/21/regression-fire-and-dangerous-things-2-3/

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. Journal of Science Education and Technology, 21(1), 56-73.

Nehm, R. H., Ha, M., & Mayfield, E. (2011). Transforming biology assessment with ML: automated scoring of written evolutionary explanations. Journal of Science Education and Technology, 21(1), 183-196.

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework. Sociological Methods & Research, 49(1), 3–42. https://doi.org/10.1177/0049124117729703

Odden, T. O. B., Marin, A., & Rudolph, J. L. (2021). How has Science Education changed over the last 100 years? An analysis using natural language processing. Science Education, 105(4), 653–680. https://doi.org/10.1002/sce.21623

Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect (First edition). Basic Books.

Rosenberg, J. M., & Krist, C. (2020). Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations. Journal of Science Education and Technology. https://doi.org/10.1007/s10956-020-09862-4

Rosenberg, J., Beymer, P. N., Phun, V., & Schmidt, J. (2020, December 29). How does situational engagement vary between learners, situations, and classrooms? Findings from the use of intensive longitudinal methods and cross-classified, multi-level models. OSF Preprint. https://doi.org/10.31219/osf.io/pj2v8

Rose, D. H., Robinson, K. H., Hall, T. E., Coyne, P., Jackson, R. M., Stahl, W. M., & Wilcauskas, S. L. (2018). Accurate and Informative for All: Universal Design for Learning (UDL) and the Future of Assessment. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), Handbook of Accessible Instruction and Testing Practices (pp. 167–180). Springer International Publishing. https://doi.org/10.1007/978-3-319-71126-3_11

Russell, S. J. (2020). Human compatible: Artificial intelligence and the problem of control. Penguin Books.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C., Van Assen, M. A., ... & Schulte-Mecklenbeck, M. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organizational Behavior and Human Decision Processes.

Schölkopf, B. (2019). Causality for Machine Learning. ArXiv:1911.10500 [Cs, Stat]. http://arxiv.org/abs/1911.10500

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences, 22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Sherin, B. L. (2015). *Computational analysis and the importance of interactional detail.* In edited A. A. diSessa, M. Levin, & N. J. S. Brown (Eds.), Knowledge and Interaction (pp. 445-462). Routledge.

Sherin, B., Kersting, N. B., & Berland, M. (2018). Learning analytics in support of qualitative analysis. Proceedings of International Conference of the Learning Sciences, ICLS, 1, 464-471.

Shiroda, M., Uhl, J. D., Urban-Lurain, M., & Haudek, K. C. (2021). Comparison of computer scoring model performance for short text responses across undergraduate institutional types. *Journal of Science Education and Technology,* 1-12.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. Computer games and instruction, 55(2), 503-524.

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. Journal of Computer Assisted Learning, 34(4), 366–377. https://doi.org/10.1111/jcal.12263

Welser, H. T., Smith, M., Fisher, D., & Gleave, E. (2008). Distilling digital traces: Computational social science approaches to studying the internet. In N. Fielding, R. M. Lee, & G. Blank, The SAGE handbook of online research methods (pp. 116–141). Thousand Oaks: SAGE.

Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., & Xie, C. (2021). Automatic assessment of students' engineering design performance using a Bayesian network model. Journal of Educational Computing Research, 59(2), 230-256.

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. Computers & Education, 161, 104059.

Yarkoni, T. (2019). The generalizability crisis. *Behavioral and Brain Sciences,* 1-37.

Zhai, X., C Haudek, K., Shi, L., H Nehm, R., & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of ML-based science assessment. Journal of Research in Science Teaching, 57(9), 1430-1459.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020b). Applying machine learning in science assessment: A systematic review. Studies in Science Education, 56(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Zhai, X., Haudek, K. C., Stuhlsatz, M. A., & Wilson, C. (2020c). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. Studies in Educational Evaluation, 67, 100916.

Zhai, X., Shi, L., & Nehm, R. H. (2021). A meta-analysis of ML-based science assessments: factors impacting machine-human score agreements. Journal of Science Education and Technology, 30(3), 361-379.

Zhang, N., Biswas, G., & Dong, Y. (2017). Characterizing Students' Learning Behaviors Using Unsupervised Learning Methods. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), Artificial Intelligence in Education (pp. 430–441). Springer International Publishing.

Zheng, N. N., Liu, Z. Y., Ren, P. J., Ma, Y. Q., Chen, S. T., Yu, S. Y., ... & Wang, F. Y. (2017). Hybrid-augmented intelligence: collaboration and cognition. Frontiers of Information Technology & Electronic Engineering, 18(2), 153-179.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. Erkenntnis, 72(1), 17-35.

Zollman, K. J. (2013). Network epistemology: Communication in epistemic communities. Philosophy Compass, 8(1), 15-27.