Peter Wulff[1] [1]Pädagogische Hochschule Heidelberg

Machine Learning in Science Education – Realized potentials, expected developments, and fundamental challenges

An important goal for science education researchers and scholars is it to understand and improve processes of science learning and teaching in an evidence-based way (Abell & Lederman, 2007). However, processes of learning and teaching are complex: intra- and interpersonal phenomena interact with each other on multiple levels (vgl.: Hilpert & Marchand, 2018). Moreover, cognitive and non-cognitive constructs of interest in science education are typically complex, e.g., non-linear and dynamic (Stamovlasis, 2016; Zhai, Yin, Pellegrino, Haudek, & Shi, 2020). Hence, science education researchers need sophisticated analysis tools to model relationships and make sense of them. Stochastic data models such as linear models are oftentimes inappropriate, even incapable, to capture these complex relationships (Breiman, 2001). Singer (2019) argues that educational scholars can benefit from adopting data science methods such as machine learning (ML) to explore potentials to understand these complex processes and phenomena. Here we seek to eclectically review realized potentials, expected developments, and fundamental challenges with regards to applying ML in science education research.

What is ML?

ML refers to data-driven, i.e., inductive, problem solving with computers (Marsland, 2015). A widely recognized operational definition for ML states: „A computer program is said to learn from experience E with respect to some classes of task T and performance measure P if its performance can improve with E on T measured by P" (T. Mitchell, 1997). Hence, ML is a form of inductive learning, i.e., learning from examples/experience/data (Nisbet, Elder, & Miner, 2009). Inductive learning moves from the specific to the general (the underlying rules/laws), thus it is sometimes called the inverse of deduction (Domingos, 2015). Inductive learning is consequently associated with an (empirical) risk when extracting rules from examples and generalizing to unseen examples (Vapnik, 1996). Through ML, a machine attains capabilities without being explicitly programmed, but rather through providing of input-output pairs (Géron, 2018). This represents a marked shift from traditional programming, where explicit instruction was necessary to transform an input into an output. Therein lies also the potential of ML, given the unprecedented availability of large datasets in the modern world, especially in the education sector (Baig, Shuib, & Yadegaridehkordi, 2020; Halevy, Norvig, & Pereira, 2009).

As a form of inductive learning, ML has some resemblance with experiential human learning (Kolb, 1984; Marsland, 2015). Experiential learning represents an important form of learning for humans and animals, because it enables them to act in uncertain, novel situations by recalling relevant knowledge from similar experiences. Central categories for experiential learning are memory, adaptation, and generalization (Marsland, 2015). Memory enables recognizing of similar situations. Adaptation enables the flexibility to react differently, depending on outcomes. Finally, similarities and differences are used to form generalizations

across situations. In one form or another, these categories also play an important role for ML. Especially with the advent of deep artificial neural network architectures, activity patterns are stored in the networks that form a sort of associative memory (Engel & van Broeck, 2001). Adaptation is achieved through providing the network in the learning phase a loss signal, which directs it to modify its weights and eventually achieve better generalization capabilities.

Inductive learning, more generally, is considered an important approach also in science for problem solving and scientific inquiry (Rothchild, 2006). From a formal logical point of view, deductive reasoning is limited because it can only work out consequences of what is already known (King et al., 2009). Also from a logical point of view, inductive reasoning is limited, because inferring general rules that "describe every member of a set, one must have information about every member of that set" (Goodfellow, Bengio, & Courville, 2016, p. 113). Hence, forms of reasoning such as induction or abduction are necessary for progress in science. Valiant (1984), from a computer science perspective, outlined important theoretical underpinnings for the possibility of inductive learning and inference. Valiant (1984) introduced the concept of "probably approximately correct" (PAC). It was shown that with a certain success probability an arbitrary function could be learnt from examples. However, as sometimes associated with the "no-free-lunch" theorem (Domingos, 2015), researchers have to provide structure (e.g., constrain hypotheses spaces and inductive biases in algorithms) in order to learn something, i.e., extract knowledge from raw data. Not least the practical success of ML applications, especially the advent of deep learning (i.e., nested artificial neural networks, ANNs), eventually undergirded the acceptance of inductive learning and ML, and informed ML researchers on what ML algorithms capability to solve certain problems.

Potentials of ML

ML is particularly adept to extract information from large and complex datasets (Domingos, 2015; Halevy et al., 2009). Phenomena in nature and society are typically complex and oftentimes highly non-linear, because multiple influencing factors are present which impact the system's behavior on multiple levels in multiple ways (Domingos, 2015; Koopmans & Stamovlasis, 2016). Equally complex are processes of communication and language (Lieberman, Michel, Jackson, Tang, & Nowak, 2007) – and, consequently, processes of learning and teaching (Koopmans & Stamovlasis, 2016), because they intricately rely on natural language as a means of representation and conveying information (Brookes & Etkina, 2007). For example, language is characterized to be compositional (build from elementary units), recursive, and hierarchical (Beule, 2008). Meaning in language emerges from the interplay of words in a sequential order. However, simple stochastic data models, especially linear models, are not well suited to model language: „Complex problems in the real world may require much more expressive hypothesis spaces than can be provided by linear functions" (Nisbet et al., 2009, p. 12). ML-based methods can facilitate better modelling and assessment of language-related processes such as learning and teaching (Goldberg, 2017; Zhai, Haudek, Shi, Nehm, & Urban-Lurain, 2020). With ML, relationships in complex datasets can be inferred and used for problem solving (Rauf, 2021). Especially the learning from examples puts ML at an advantage over stochastic data models, because minimal constraints are posed onto the algorithm (Breiman, 2001; Nisbet et al., 2009).

To model and learn relationships in complex datasets, various learning approaches within the domain of ML are commonly differentiated which offer different potentials and challenges for researchers. Widely employed learning approaches in science education research are supervised and unsupervised ML.[1] Both learning approaches differ not least in the following dimensions: goals, data requirements, algorithms, and learning procedures.

In supervised ML, goals are to classify samples according to categories or score them according to a range of real values (Bishop, 2006). Data samples are required to be annotated (mostly by human raters). Oftentimes, input-output mappings have to be collected and created. Then, a particular learning algorithm is selected. A common distinction is between shallow and deep ML algorithms. Shallow ML algorithms take a set of inputs and essentially produce a combination of these inputs (Marsland, 2015). Examples of shallow ML algorithms are (multinomial) Naïve Bayes, support vector machines (SVM), or (multinomial) logistic regression. Naïve Bayes estimates the probability of a category based on Bayes' rule (prior probabilities and given evidence). Support vector machines classifier/regressor seeks to maximize the decision margin across categories in a higher dimensional space. Logistic regression is essentially a linear model with an activation function (non-linear) (find a more detailed discussion with science education focus here: Wulff et al., 2020; Zhai, Yin, et al., 2020, or for general discussion here: Bishop, 2006; Marsland, 2015).

However, while these shallow algorithms have proven effective in many problems, ML research has seen a surge of interest into deep ML algorithms where initial features are progressively transformed into derived features (Marsland, 2015). Deep ANN architectures have been found to excel in vision and language processing (Goldberg, 2017; M. Mitchell, 2020), and increasingly replace shallow ML architectures. Even simple ANN such as the multilayer perceptron (a kind of hydrogen atom for artificial intelligence research, Engel & van Broeck, 2001) are capable to model arbitrary (smooth) functional relationships ("universal approximation theorem") even with only one layer and non-linear activation (Engel & van Broeck, 2001; Marsland, 2015). It is, then, important to know how to setup the ANN (number of nodes and number of layers). Also, it is desirable to know the amount of training data necessary to train the network. Unfortunately, neither the setup nor the amount of training data can be specified a priori with much certainty (Marsland, 2015). Both depend largely on the problem at hand. Moreover, some problems require different architectures. For example, language processing is inherently sequential and characterized by long-range dependencies (Alvarez-Lacalle, Dorow, Eckmann, & Moses, 2006). Specific ANN architectures such as recurrent neural networks, long short-term memory networks, or transformer architectures have been devised to cope with these requirements (Goldberg, 2017; Vaswani et al., 2017). Pictorial data was found to be processed effectively with convolutions, hence convolutional neural networks were designed. Essentially, these different architectures specify the information flow in the networks with reference to the input and output data. The more complex these network architectures become, the more difficult it is to retrieve

[1] Reinforcement learning and semi-supervised learning are not further considered here, but will likely become increasingly important in educational fields, see: M. Mitchell (2020).

information on the ML model decisions, which raises the “black-box” problem that is addressed in research on explainable AI.

Supervised training shallow and deep ML algorithms requires a loss function (performance indicator) that attributes how far off the actual ML algorithms output is in reference to the gold standard label provided by the annotated training dataset. Given this loss information, an optimization procedure assures that this information is passed through the architecture to update parameters or weights in the ML model to reach a final model. In ANNs oftentimes a gradient-based optimization procedure is utilized. The learning procedure is controlled by hyperparameters (e.g., amount of update for the weights, given a certain loss). These have to be set in advance of the learning. The more hyperparameters there are, the more combinations of hyperparameters have to be tested in order to find an optimal hyperparameter configuration. In supervised learning, cross-validation seeks to illuminate generalizability of the trained ML model. As such, researchers hold back a test dataset from the beginning. This test dataset is then used to estimate to what extent the trained ML model can predict the unseen cases, hence indicate how well the ML model generalizes beyond the training data. Moreover, a validation (sometimes called development) dataset need to be extracted from the training dataset as well, if different hyperparameter configurations are tested. A common challenge with cross-validation is data leakage, where information from the test data leaks into the training regime, causing replicability failures in ML (Kapoor & Narayanan, 2022). Researchers need to make sure this does not happen for otherwise inflated fit statistics might result. Finally, in supervised learning, human-machine agreement is oftentimes used as an evaluation criterion for model performance (assuming the humans agreed in the first place). Depending on the goal (classification, regression) different agreement metrics are available such as precision, recall for classification, and mean squared distance for regression.

Unsupervised ML, on the other hand, seeks to approximate the probability density of data, group or cluster similar samples, or reduce the dimensionality of a dataset (Bishop, 2006). Conveniently, no human annotation of the data samples is required to apply unsupervised ML algorithms. This spares resources and allows the processing of datasets of unwieldy sizes. Unsupervised ML algorithms comprise procedures for probability density estimation (latent Dirichlet allocation, LDA), clustering (e.g., k nearest neighbors, kNN) and dimensionality reduction (latent semantic analysis, LSA). LDA is a generative probabilistic model to extract topics in text documents (Blei, Ng, & Jordan, 2003). KNN is an algorithm that assigns a decision boundary based on the spatially closest neighbors in input space (Marsland, 2015). LSA is a method in natural language processing (NLP). In LSA, the input space is reduced in dimensionality by singular value decomposition of the document-term-matrix, and this lower-dimensional representation is used as a new feature vector where informative similarities can be calculated by algebraic means (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Though data requirements (i.e., annotation efforts) in unsupervised ML are advantageous compared to supervised ML, model validation is oftentimes more difficult. For example, the number of clusters, topics or dimensions have to be chosen by the researcher without much theoretical guidance. This poses the requirements to make ablation studies, e.g., systematically vary the number of clusters and monitor differences and similarities in outputs.

Besides supervised and unsupervised ML, meta learning or transfer learning are other promising learning approaches in ML. Brazdil, van Rijn, Soares, and Vanschoren (2022) differentiate between algorithm selection, hyperparameter optimization, pipeline optimization, and few-shot learning. Few-shot learning became particularly relevant in the context of deep ANNs (Ruder, 2019). ML researchers noticed a phenomenon called "catastrophic forgetting" with ANNs (McCloskey & Cohen, 1989). After training an ANN on, say, a simple calculation problem, it would be possible for the ANN to "forget" (i.e., decrease performance) its capabilities after having been trained on another, slightly different, problem. In transfer learning and few-shot learning, ML models are sought to be trained in a way to generalize across problems, i.e., transfer their knowledge to the new problem. To do so, in a phase called generative pre-training, a backbone ML model is trained to appropriately capture the structure (e.g., correlations) in language or images (Radford, Narasimhan, Salimans, & Sutskever, 2018). It is then the goal to further train (fine tune) the ML model with examples from the new problem (Wang, Yao, Kwok, & Ni, 2020). For language and vision ML models it was shown that prior training on large databases could boost performance on novel tasks and even reduce the data requirements to reach a certain performance (Ruder, 2019).

Applications of ML in science and science education

In the natural sciences, the system AlphaFold can count as a milestone in the application of ML and an exemplary case of how to apply ML (deep learning in particular). AlphaFold is an ML-based system that generates the spatial structure of a protein based on the sequence of amino acids, a problem known as protein folding (Jumper et al., 2021). Some experts judged this problem to be unsolvable without use of ML methods. AlphaFold now reaches accuracies on par with experimental methods, however, without the excess of resource requirements that are necessary to determine the structure via experiments. Training data comprised a structured database in which given amino acid sequences were paired with the respective 3D protein structures. Test data were newly determined (unseen) protein structures; these had to be predicted given the amino acid sequence. Other important applications of ML in science (see Table 1) are in cosmology, quantum physics, materials properties prediction, and elementary particle physics, where new insights could be gained, or calculations and simulations became possible (Carleo et al., 2019; Cranmer et al., 2020; Joss & Müller, 2019; Udrescu & Tegmark, 2020). For example, a supervised ML approach was used to efficiently determine the red shift of distant galaxies based on photometric data (Kind & Brunner, 2013). Spectroscopic analyses, which yield exact red shift values, are resource intense, hence the determination of red shift values based on photometric data is considered a valuable resource for researchers (Carleo et al., 2019). Generalizability in these red shift analyses is considered a challenge, because it is unclear to what extent the training data is representative for later applications in the field. Potentials of transfer learning are considered promising advancements to address these challenges (Carleo et al., 2019; Leistedt, Hogg, Wechsler, & DeRose, 2019). In quantum physics, ML is applied to address the "Quantum Many-Body Problem". In this problem the positional probability density for multiple quantum particles such as electrons should be estimated (Carleo et al., 2019). It could be shown, among others, that ANNs retrieve and store information for quantum entanglement of electrons rather. Also simulations of quantum systems could be improved, e.g., by efficient sampling through ML models (Carleo et al., 2019). ML, and ANNs in particular, have also used to predict boiling points of fluids more

accurately compared to mere linear models such as multiple regression (Joss & Müller, 2019). Finally, ML is of crucial importance in elementary particle physics. Unsupervised and supervised ML approaches are used in conjunction to reduce large datasets of particle collisions (trigger) (Carleo et al., 2019).

In science education, a wealth of studies used ML approaches to answer novel research questions and extend research capabilities (Zhai, Yin, et al., 2020). Applications cover assessment of diverse contents and scientific practices. We examine areas where ML offers potentials and pose challenges along the following recurring themes in the ML-based science education literature:

(1) Extending the inquiry and discovery capabilities with ML
    (1a) Information extraction in complex datasets
    (1b) Model validation
    (1c) Automating assessment and feedback
(2) Extending the research process and capabilities in science education with ML

Table 1: Application of ML in science and science education, grouped by ML approaches used and specific goals.

Approach	Supervised ML		Unsupervised ML		
Goal, Task / Domain	**Regression**	**Classification**	**Discovery**	**Clustering**	**Approximation**
Natural science	Predict red shift on cosmic images (Carleo et al., 2019); Predict boiling points of fluids (Joss & Müller, 2019)	Classify relevant events in particle collisions (zit. in: Carleo et al., 2019)	Extract laws/symmetries/regularities in synthetic or real data (AIFeynman, AI Poincaré; (Y. Liu et al., 2022; Udrescu & Tegmark, 2020)	Clustering of stars; Genetic structures in DNA micro arrays (Hastie, Tibshirani, & Friedman, 2008)	Fast approximation/sampling for simulations (Many-Body quantum systems, elementary particle collisions); Representations of molecule structure (Gómez-Bombarelli, 2017)
Science education	Scoring of argumentation quality in learner responses; scoring of utility value in essays (Zhu et al., 2017; Beigman Klebanov et al., 2017)	Classification of verbs to assess conceptual change in physics (Yan, 2014)	Identify neural activation patterns in learner brain for physics concepts (Mason & Just, 2016)	Grouping of learner responses with reference to their estimation of generality (Rosenberg & Krist, 2021)	Representation of test and terms in word vectors (Sherin, 2013; Wulff et al., 2022)

(1) Extending the inquiry and discovery capabilities with ML

The inquiry process and capabilities lie at the core of scientific disciplines. In empirically oriented disciplines, the data-processing capabilities and the validity of the data models are central concerns around the inquiry and discovery capabilities:

(1a) Information extraction in complex datasets

A widely recognized potential of ML for science education is the capacity to extend assessment by means of analyzing complex data formats such as language-based responses (Zhai, Yin, et al., 2020). Given that competencies are conceptualized as complex, context-dependent dispositions and science education researchers long argued to extend assessment

formats to include knowledge-in-use aspects, ML can potentially enhance effective and efficient analysis of such assessments (Maestrales et al., 2021). In particular, assessment formats such as closed-form questions (e.g., multiple choice items) were criticized to lack capabilities to adequately measure complex, procedural cognitive abilities (Haudek, Prevost, Moscarella, Merrill, & Urban-Lurain, 2012; Martinez, 1999). Moreover, human coding of constructed responses requires resources and is error prone due to expertise differences for the raters, fatigue, and other implicit biases (Zehner, Sälzer, & Goldhammer, 2016). A promise of ML methods is to compensate for some of these drawbacks, because it is a principled, computer-based approach. Early assessments and analysis of constructed responses used rule-based procedures that do not contain ML aspects. Haudek et al. (2012) used lexical analyses to successfully group learners explanations for acid-base behavior in biological systems. Nehm and Härtig (2012) defined rules to extract key concepts in learners responses for evolution, and found that these rules were sufficient for automatically scoring the responses with reference to key concepts. Later studies employed ML-based tools to score and classify responses. O. L. Liu et al. (2014) used a concept-based coding with the help of the program c-rater (similar approach in: Donnelly, Vitale, & Linn, 2015). Maestrales et al. (2021) trained ML algorithms to score students' constructed responses to assess chemistry and physics learning. Mostly, these studies find that human-machine agreement is substantial and thus automated coding of constructed responses is possible with some caveats. For example, Maestrales et al. (2021) reported that classification performance (human-machine agreement) for responses decreased when specific vocabulary in chemistry and physics was present. They hypothesize that the learner responses in the training data had less formal vocabulary. This potentially points to a problem of generalizability, similar to the remarks on representativeness of red-shift data in the physics example above.

While typical constructed responses are rather short (on average 1 to 3 sentences), ML can also be used to analyze entire documents such as papers or essays. Odden, Marin, and Rudolph (2021) analyzed 100 years of research papers in the journal *Science Education*, overall some 5577 papers, with an unsupervised ML approach. This dataset surpasses typical review studies due to resource limitations. They used LDA to find topics in the papers. The authors identify the overarching themes: "science content topics, teaching-focused topics, and student-focused topics," and track the occurrence of these themes over the 100 years in the journal's existence. Using this ML algorithms would also allow the researchers to examine relationships of topic trends with covariates such as societal discourses or journal editors at the time. Beigman Klebanov, Burstein, Harackiewicz, Priniski, and Mulholland (2017) used ML in conjunction with NLP to assess utility-value essays of students in biology. They report that NLP could be used to extract features and ML could be used to accurately score the essays.

To evaluate accuracy of supervised ML models, Cohen's kappa is often used as a measure for chance-corrected agreement between human raters or between human and machine. Typical agreements range from .55 to 1.00 (O. L. Liu, Rios, Heilman, Gerard, & Linn, 2016). However, Cohen's kappa – as a single score – is rather opaque on specific classification problems that might occur with single categories and other performance metrics are also important to consider. For example, precision, recall, F1 (as the average of precision and recall), or area under curve (AUC) yield diagnostic information on the success to single out

specific categories. Also, the confusion matrix yields valuable information where systematic disagreements might occur between human and machine. Further intricate challenges await when more than two raters are present or the categories can overlap with each other. At present, coding in science education research focuses on sentence-level coding units without overlapping or hierarchically nested categories. While this makes coding easier for researchers, this does not necessarily recognize the complex structure of language (compositionality, hierarchy, and recursion) and cognitive processes more generally. Another challenge relates to the robust finding in essay scoring that essay length and essay score are significantly positively correlated (Chodorow & Burstein, 2004). Carpenter, Geden, Rowe, Azevedo, and Lester (2020) and Krüger and Krell (2020) also found this effect in their studies in science education. Researchers who assess constructs with constructed responses need to monitor such correlations. Given that text length is a surface feature of constructed responses, researchers then need to find a way to control for text length and find more informative features that explain quality.

Besides language, other types of complex data such as images, log-data, or eye movement are increasingly analyzed with ML algorithms by science education researchers. Zhai, He, and Krajcik (2022) used graphical representations and constructed responses to assess modelling competencies of students. They found that an ANN could imitate the human coding with substantial agreement. Küchemann, Klein, Becker, Kumari, and Kuhn (2020) showed that eye-tracking data could be used as a feature in ML models to predict success in a kinematics assessment. Interestingly, the ANN performed worse compared to more shallow ML models, which points to the necessity for science education researchers to considerately chose their ML models in reference to the problem at hand. In sum, ML can be used to process a variety of information and even integrate different kinds of information in science education research.

(1b) Model validation

While the capacity of ML models to extract information from complex data is of great value for science education, it is equally important to assure that the models allow for valid inferences (Zhai, Yin, et al., 2020). This is all the more relevant, since ML models such as ANNs can approximate arbitrary smooth function and researchers need to assure that the model picks up on relevant features, where human raters oftentimes lack the capacity to systematically analyse the datasets such as in the case for boiling point prediction of fluids where up to hundreds or thousands of molecule descriptors might be integrated to reach accurate predictions (Joss & Müller, 2019). Validation of ML models is oftentimes more encompassing compared to the procedure known for stochastic data models where (among others) fit indices are calculated and compared (Breiman, 2001).

To validate ML models, science education researchers used several criteria such as (a) human-machine agreement as an indicator for model validity, (b) correlations with covariates, (c) important features for model decisions, and (d) cross-validation to assess generalizability of the ML model. (a) To determine human-machine agreement, ML researchers ground their work in established theoretical frameworks in science education. For example, Krüger and Krell (2020) examined modelling competence according to an established modelling framework that distinguishes five competencies related to modelling. Human-machine

agreement of the ML models ranged from acceptable to substantial. Similarly, studies on argumentation ground their study in respective frameworks on argumentation structure (Zhu et al., 2017). These frameworks are used to guide human annotation/coding to classify the data into distinct categories. (b) Besides human-machine agreement Krüger and Krell (2020) also used covariates to evaluate validity of model decisions. For example, they found that some ML algorithms' decisions correlated with external criterion text lengths which, to some extent, raised validity concerns. (c) Analysis of important decision features can enhance transparency of ML model decisions and thus establish validity. Wulff, Mientus, Nowak, and Borowski (2022) determined attribution scores of their ML model to find relevant features that the model used for classifying physics teachers' sentences into categories that were posited by the reflection-supporting model. They found that certain words were predictive for elements in the reflection-supporting model (see also: Krüger & Krell, 2020). (d) To evaluate generalizability through cross-validation, mostly the dataset is split into training, validation, and test data, and performance of the ML model on unseen test data is assessed (Wulff, Mientus, et al., 2022; Zhai et al., 2022). Researchers conclude that the ML model generalizes well, once performance on the unseen test dataset is substantial. However, less clear are the principles or rules that the ML model learned which allow it to generalize from the training examples, which is an important problem to address in future research.

An important challenge is the question for appropriate gold standards. ML models, most naturally, almost never achieve full agreement, because the human raters disagree in the first place (sometimes substantially so). The root causes for this might lie in the insuffiency of the theoretical framework in the first place. Theories in social sciences lack the mathematical rigor of physics theories, because the phenomena that social scientists seek to model and explain are more complex to begin with (Halevy et al., 2009). How can we be certain that the theories are appropriate? ML might play an important role in improving theory building via data-centered, unsupervised means in the future. However, as of now, guiding ML models on the basis of insufficient theoretical frameworks introduces uncertainty for determining gold standards. Moreover, the theoretical frameworks have to be operationalized and human raters have to be trained appropriately to correctly utilize the theory to classify examples. Questions of human raters' expertise, prior experiences, and situational determinants (e.g., fatigue) need to be addressed to ascertain how valid the coding process is, and, thus, determine to what extend we can expect the ML model to reach full agreement with the human raters. Next, the researchers specify with their choices of ML algorithms the specific hypothesis space to be considered. E.g., some ML algorithms are well versed to cope with small data and non-linearities. Again, these decisions constrain the possibility for perfect agreement with the gold labels, and make it more difficult to assess model validity. Furthermore, while generalizability is tested through performance of the ML model on unseen data, there is oftentimes no formal justification for the particular train-test split of the dataset. Generalizability then means performance expectation, given a randomly sampled data point from the representative sample. Research with the perceptron more rigorously determined generalizability criteria, e.g., through introduction of related but novel reasoning tasks (Engel & van Broeck, 2001). It is unclear what such a more rigorous conceptualization of generalizability would look like in science education research, because the theoretical frameworks are necessarily more fuzzy – given the complexity of the processes and phenomena under study.

Model validation for unsupervised ML can be even more challenging. For example, determining the hyperparameter configuration, e.g., the number of clusters, the number of dimension, or the number of topics, requires trade-offs between interpretability and sparsity. Sherin (2013) systematically varied experimental parameters to find a suitable number of topics in students' transcribed interviews that relate to explanations of the seasons. He contends that no definite solution for this problem of finding a suitable number of topics exists in his case, however, that the systematic variation can provide some insights into topic validity. Typically, human interpretation of topics is necessary, e.g., through providing human raters the most representative words for a topic (Rosenberg & Krist, 2020; Wulff, Mientus, et al., 2022). As in supervised ML, model validation in unsupervised ML involves substantial and critical involvement of human researchers to appropriately set up testing conditions and interpret outputs. Model validation, in consequence, can only function when humans and machines work in tandem (Sherin, 2013).

(1c) Automating assessment and feedback

Once ML models have been found to reach substantial human-machine agreement, automation is a major goal for many researchers. We also saw in the science examples above (Table 1) that automation and efficient analysis are important goals that can advance research capabilities. Zhai, Yin, et al. (2020) identifies automation as a crucial feature of ML in science education, i.e., outsourcing human decision making to machines. Many studies in science education (and science) refer to the resource argument in their motivation and implications. Resources can refer to human raters' time, the costs associated with coding, or merely the availability of human raters. All of these resources are scarce in practice and should be spared if possible. Automation can furthermore enable researchers to more readily answer derived research questions. For example, reliable coding of argumentation elements allows researchers to filter parts of an argumentation and analyze these samples in greater depth – similar to the filtering in elementary particle collision data, such that researchers do not have to sift through the entirety of collision data. Lee et al. (2019) further highlight the elimination of human elements in coding processes and evaluation as a potential benefit of automation through ML.

Automation can be achieved with supervised and unsupervised ML, however, science education researchers engaged with ML mostly employ supervised ML for purposes of automation (Zhai, Yin, et al., 2020). A rough estimate for reliability required to automate coding is a quadratic Cohen's kappa value of .70 or above (Williamson, Xi, & Breyer, 2012). An early example of automation represents the study by Nehm and Härtig (2012). They extracted key concepts with specified rules and implemented EvoGrader as a web-based tool to allow fellow researcher to freely use their coding. They estimate the initial invest as substantial, however, after two years the invest should pay off. On the other hand, O. L. Liu et al. (2014) used a concept-based coding and concluded that substitution of human raters was not possible. Training of the machine would require 10k human ratings and not all concepts (including misconceptions) are documented in the manual appropriately. More generally, the question of dataset size is pertinent to ML research. Ha, Nehm, Urban-Lurain, and Merrill (2011) could show that 500 responses could be sufficient to train a reliable ML model. Similarly, Zehner et al. (2016) found acceptable performance for 249 samples. However, the

amount of data is crucially linked to the complexity of the problem at hand. Deep learning ML algorithms might raise the requirements for sample size. However, with transfer and meta-learning these concerns might be mitigated (see below).

Once a trained and validated ML model allows for automated assessment, individualized and adaptive tutoring systems can be implemented. Adaptive feedback and guidance is an important facilitator for learning, and ML models can be considered valuable parts of these systems to make them more flexible. The ability of ML models to process and analyze language input is among the most central features. For example, Donnelly et al. (2015) used ML models to automatically score thermodynamics essays and adaptively choose guidance sentences for the learners. Interestingly, learners with lower prior knowledge benefitted more compared to other learners. Similarly, Zhu et al. (2017) could show for 16 items for climate change that ML models reach substantial human-machine agreement. The ML model could then be used to adaptively choose a suitable feedback sentence for learners which had a positive impact on post-test scores. Unsupervised ML such as LSA has been used to implement tutoring systems that enact dialogs with learners. Graesser et al. (2004) defined curriculum scripts to successfully guide learners to solve physics problems in the AutoTutor (Person & Graesser, 2002).

(2) *Extending the research process and capabilities in science education with ML*
Besides more specific potentials of ML to extend the inquiry process in science education, ML can also enhance the overall research processes and capabilities in science education research. Science education research is oftentimes empirically grounded and seeks to test hypotheses with evidence gleaned from data. Digitization will all the more facilitate to gather data on learning and teaching processes (Baig et al., 2020), and – as argued above – ML will play a role to extract meaningful information from this data. Besides the data processing and information extracting capabilities of ML to enhance science education research, ML also offers novel capabilities to enhance the overall research process. Once ML models are trained and validated, they can be shared across research contexts and enhance collaboration. While this would also be true for established quantitative and qualitative science education research (e.g., linear regression models or coding manuals could also be shared across research sites), ML can ease collaborative inquiry processes. For example, reuse of coding manuals requires training of new raters who have to interpret the manual just as the previous raters did. This, however, is oftentimes inefficient and error prone.

ML research can incite collaboration and model sharing. In the context of deep learning, ML offers novel potentials to share and collaborate trained ML models and further fine tune them in specific research contexts. In particular, ML researchers showed that pretrained deep learning-based language models can be reused in different contexts (transfer and meta-learning). For example, Carpenter et al. (2020) used pretrained word embeddings to estimate reflective depth of learners‘ responses in a game-based learning environment and found that the pretrained embeddings outperformed other methods. Wulff, Mientus, et al. (2022) showed that pretrained language models could be used to accurately classify preservice physics teachers‘ written reflections. Specific deep learning architectures such as pretrained language models were found to be more performant compared to other deep learning architectures to

classify written reflections. Fine tuning of language models to specific tasks facilitated coding even for small samples (Wulff et al., 2022). The same pretrained language model architectures could then be utilized to inform the process of clustering sentences of preservice physics teachers writing about a physics lesson and extract interpretable topics (Wulff, Buschhüter, et al., 2022). The pretrained language model seemed to be particularly useful to extract robust topics. Fine-tuning pretrained language models even enabled ML models to perform steps of quantitative reasoning (Lewkowycz et al., 2022). In a few-shot learning paradigm through chain-of-thought prompting the ML model learned with reasonable accuracy to solve high school and university problems in mathematics and science. In a representative dataset the model solved one third of the tasks with providing the relevant reasoning behind the solution—either in formal mathematical language and in natural language. To our estimation, this study marks a milestone in utilizing ML to advance science education research and offers exciting new ways for assessing students' problem solving abilities via ML methods.

Given these advancements, some posited that ML can replace human researchers. However, we suspect that the role of the human researchers will remain vital in research processes in science education. While robot scientists such as Adam show impressive progress for automating scientific research in specific applications such as yeast growth (King et al., 2009), it proves difficult to implement general robot scientists, because this would entail many more capabilities (such as implementing experimental setups in reality) than data collection, analysis, and reporting. Science education researchers argued for a human-machine tandem and integration, rather than replacement (Rosenberg & Krist, 2020; Sherin, 2013). Based on his findings, Sherin (2013) argued that the ML model can support the human analyst in a bootstrapping program that can help confirm a larger theoretical and empirical program, i.e., raise confidence in our theories. In this line, Rosenberg and Krist (2020) textured the further path of how human and machine analysts can be integrated. They used unsupervised ML to explore patterns and then human raters validated these patterns to find a robust coding manual for employing supervised ML. These studies highlight potentials for synergy effects between human and machine.

Concluding remarks

ML has offered science education researchers a valuable tool to enhance the inquiry process and research capabilities. Advancements in the field of ML research will continue to provide science education researchers novel potentials to answer their research questions and pose entirely new research questions. We have seen applications in all science disciplines (biology, chemistry, and physics) across different scientific practices (argumentation, explanation, reflection). ML has provided specific potentials to assess complex constructs that can be operationalized, among others, through language-based responses. Given the intricate relationship of language and science learning, ML offers a valuable modelling tool that can enhance assessment, automation, and, more generally, learning and teaching. If such assessment is valid and can be automated, researchers can share their instruments more easily and collaboratively improve them.

Challenges await, however. We outlined that fundamental questions regarding model validity and generalizability remain unsolved. ML is an inductive learning approach and if humans

cannot understand the patterns that the ML algorithm picked up, theory development is hampered. We also touched upon some areas where bias can be introduced into the machine's learning. For example, if large language models are trained on corpora such as the Internet and Wikipedia, and if these corpora are written by specific individuals (e.g., related to gender), this imposes problems of implicit biases. As a matter of fact, language models output similar biases as humans (Caliskan, Bryson, & Narayanan, 2017). Other biases relate to decisions made for the training of ML models (algorithms selection, loss-function selection, hyperparameters) and for reporting the findings (visualizations used).

In consequence, implementing ML models in educational institutions, especially with children who form their identities, requires substantially more research efforts to assure that ML models' decisions and feedback do not implicitly impose harm or disadvantage certain individuals. This is likely true for subjects other than the sciences as well.

References

Abell, S. K., & Lederman, N. G. (2007). Preface. In S. K. Abell & N. Lederman (Eds.), *Handbook of research on science education*. Mawhah, New Jersey: Lawrence Erlbaum Associates Publishers.

Alvarez-Lacalle, E., Dorow, B., Eckmann, J.-P., & Moses, E. (2006). Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(21), 7956–7961. https://doi.org/10.1073/pnas.0510673103

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-00223-0

Beigman Klebanov, B., Burstein, J., Harackiewicz, J. M., Priniski, S. J., & Mulholland, M. (2017). Reflective Writing About the Utility Value of Science as a Tool for Increasing STEM Motivation and Retention – Can AI Help Scale Up? *International Journal of Artificial Intelligence in Education*, *27*(4), 791–818. https://doi.org/10.1007/s40593-017-0141-4

Beule, J. de (2008). Compositionality, Hierarchy and Recursion in Language: A Case Study in Fluid Construction Grammar.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. *Information Science and Statistics*. New York, NY: Springer Science+Business Media LLC. Retrieved from https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022.

Brazdil, P. B., van Rijn, J. N., Soares, C., & Vanschoren, J. (2022). *Metalearning: Applications to automated machine learning and data mining* (Second edition). *Springer eBook Collection*. Cham: Springer. https://doi.org/10.1007/978-3-030-67024-5

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231.

Brookes, D. T., & Etkina, E. (2007). Using conceptual metaphor and functional grammar to explore how language used in physics affects student learning. *Physical Review Special Topics - Physics Education Research*, *3*(1), 771. https://doi.org/10.1103/PhysRevSTPER.3.010105

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.)*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., . . . Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4). https://doi.org/10.1103/RevModPhys.91.045002

Carpenter, D., Geden, M., Rowe, J., Azevedo, R., & Lester, J. (2020). Automated Analysis of Middle School Students' Written Reflections During Game-Based Learning. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 67–78). Cham: Springer International Publishing.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on Toefl essays*. ETS.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., & Ho, S. (2020). Discovering Symbolic Models from Deep Learning with Inductive Biases. *ArXiv*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis.

Domingos, P. (2015). *The Master Algorithm : How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books.

Donnelly, D. F., Vitale, J. M., & Linn, M. C. (2015). Automated Guidance for Thermodynamics Essays: Critiquing Versus Revisiting. *Journal of Science Education and Technology*, *24*(6), 861–874. https://doi.org/10.1007/s10956-015-9569-1

Engel, A., & van Broeck, C. den (2001). *Statistical mechanics of learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139164542

Géron, A. (2018). *Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme* (K. Rother, Trans.). *Animals*. Heidelberg: O'Reilly. Retrieved from http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1303476

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool.

Gómez-Bombarelli, R. (2017). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ArXiv*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, Massachusetts, London, England: MIT Press. Retrieved from http://www.deeplearningbook.org/

Graesser, A. C., Lu, S., Jackson, G. T., Mitchel, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 180–192.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE Life Sciences Education*, *10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 8–12.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE Life Sciences Education*, *11*(3), 283–293. https://doi.org/10.1187/cbe.11-08-0084

Hilpert, J. C., & Marchand, G. C. (2018). Complex Systems Research in Educational Psychology: Aligning Theory and Method. *Educational Psychologist*, *53*(3), 185–202. https://doi.org/10.1080/00461520.2018.1469411

Joss, L., & Müller, E. A. (2019). Machine Learning for Fluid Property Correlations: Classroom Examples with MATLAB. *Journal of Chemical Education*, *96*(4), 697–703. https://doi.org/10.1021/acs.jchemed.8b00692

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kapoor, S., & Narayanan, A. (2022). Leakage and the Reproducibility Crisis in ML-based Science.

Kind, M. C., & Brunner, R. J. (2013). TPZ : Photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, *432*(2), 1483–1501. https://doi.org/10.1093/mnras/stt574

King, R. D., Rowland, J., Aubrey, W., Liakata, M., Markham, M., Soldatova, L. N., . . . Pir, P. (2009). The Robot Scientist Adam. *Computer*, *42*(7), 46–54. https://doi.org/10.1109/MC.2009.270

Kolb, D. (1984). *Experiential Learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.

Koopmans, M., & Stamovlasis, D. (Eds.) (2016). *Complex Dynamical Systems in Education*. Springer.

Krüger, D., & Krell, M. (2020). Maschinelles Lernen mit Aussagen zur Modellkompetenz. *Zeitschrift Für Didaktik Der Naturwissenschaften*, *26*(1), 157–172. https://doi.org/10.1007/s40573-020-00118-7

Küchemann, S., Klein, P., Becker, S., Kumari, N., & Kuhn, J. (2020). Classification of Students' Conceptual Understanding in STEM Education using Their Visual Attention Distributions: A Comparison of Three Machine-Learning Approaches: CSEDU 2020, 36–46. https://doi.org/10.5220/0009359400360046

Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622. https://doi.org/10.1002/sce.21504

Leistedt, B., Hogg, D. W., Wechsler, R. H., & DeRose, J. (2019). Hierarchical modeling and statistical calibration for photometric redshifts. *The Astrophysical Journal*, *881*(1), 80. https://doi.org/10.3847/1538-4357/ab2d29

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., . . . Misra, V. (2022). Solving Quantitative Reasoning Problems with Language Models. *ArXiv*.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*, 713–716.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28. https://doi.org/10.1111/emip.12028

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. https://doi.org/10.1002/tea.21299

Liu, Y., Zhang, L., Wang, W., Zhu, M. [Min], Wang, C., Li, F., . . . Liu, H. (2022). Rotamer-free protein sequence design based on deep learning and self-consistency. *Nature Computational Science*, *2*(7), 451–462. https://doi.org/10.1038/s43588-022-00273-6

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. https://doi.org/10.1007/s10956-020-09895-9

Marsland, S. (2015). *Machine learning: An algorithmic perspective* (Second edition). *Chapman & Hall / CRC machine learning & pattern recognition series*. Boca Raton, FL: CRC Press. Retrieved from http://proquest.tech.safaribooksonline.de/9781466583283

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218.

Mason, R. A., & Just, M. A. (2016). Neural Representations of Physics Concepts. *Psychological Science*, *27*(6), 904–913. https://doi.org/10.1177/0956797616641941

McCloskey, M., & Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Gordon H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 24, pp. 109–165). Academic Press.

Mitchell, M. (2020). *Artificial Intelligence: A guide for thinking humans*. Pelican Books.

Mitchell, T. (1997). *Machine learning*. New York, NY: McGraw-Hill Education.

Nehm, R. H., & Härtig, 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. https://doi.org/10.1007/s10956-011-9282-7

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis & Data Mining Applications*. Elsevier.

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

Person, N., & Graesser, A. C. (2002). Human or Computer? AutoTutor in a Bystander Turing Test. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems* (pp. 821–830). Berlin, Heidelberg: Springer Berlin Heidelberg.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *ArXiv*.

Rauf, I. A. (2021). *Physics of Data Science and Machine Learning*. Boca Raton: CRC Press. https://doi.org/10.1201/9781003206743

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*. Advance online publication. https://doi.org/10.1007/s10956-020-09862-4

Rothchild, I. (2006). Induction, Deduction, and the Scientific Method: An eclectic overview of the practice of science. *SSR*.

Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing: Dissertation*. Ireland: National University of Ireland.

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

Singer, J. D. (2019). Reshaping the Arc of Quantitative Educational Research: It's Time to Broaden Our Paradigm. *Journal of Research on Educational Effectiveness*, *12*(4), 570–593. https://doi.org/10.1080/19345747.2019.1658835

Stamovlasis, D. (2016). Catastrophe Theory: Methodology, Epistemology, and Applications in Learning Science. In M. Koopmans & D. Stamovlasis (Eds.), *Complex Dynamical Systems in Education* (pp. 141–175). Springer.

Udrescu, S.-M., & Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, *6*.

Valiant, L. G. (1984). A theory of the learnable. *Communication of the ACM*, *27*.

Vapnik, V. (1996). Structure of Statistical Learning Theory. In A. Gammerman (Ed.), *Computational Learning and Probabilistic Reasoning* (pp. 3–31). Chichester, New York: John Wiley & Sons.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is All you Need: Conference on Neural Information Processing Systems. *Advances in Neural Information Processing Systems*, 6000–6010.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ArXiv*.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Wulff, P., Buschhüter, D., Nowak, A., Westphal, A., Becker, L., Robalino, H., . . . Borowski, A. (2020). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology*. Advance online publication. https://doi.org/10.1007/s10956-020-09865-1

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning — A Case for Pretrained Language Models-Based Clustering. *Journal of Science Education and Technology*. Advance online publication. https://doi.org/10.1007/s10956-022-09969-w

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2022). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. *International Journal of Artificial Intelligence in Education*. Advance online publication. https://doi.org/10.1007/s40593-022-00290-6

Yan, J. (2014). *A Computer-Based Approach For Identifying Student Conceptual Change: Open Access Theses. 289. https://docs.lib.purdue.edu/openaccesstheses/289.*

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic Coding of Short Text Responses via Clustering in Educational Assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. https://doi.org/10.1177/0013164415590022

Zhai, X., Haudek, K., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

Zhai, X., He, P., & Krajcik, J. S. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). 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

Zhu, M. [Mengxiao], Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668. https://doi.org/10.1080/09500693.2017.1347303