ML-SD Modeling: How Machine Learning Can Support Scientific Discovery Learning for K-12 STEM Education

Xiaofei Zhou,[1] **Hanjia Lyu,** [1] **Jiebo Luo,** [1] **Zhen Bai** [1]

[1] Department of Computer Science, University of Rochester
xzhou50@cs.rochester.edu, hlyu5@ur.rochester.edu, jluo@cs.rochester.edu, zhen.bai@rochester.edu

Abstract

The importance of machine learning (ML) in scientific discovery is growing. In order to prepare the next generation for a future dominated by data and artificial intelligence, we need to study how ML can improve K-12 students' scientific discovery in STEM learning and how to assist K-12 teachers in designing ML-based scientific discovery (SD) learning activities. This study proposes research ideas and provides initial findings on the relationship between different ML components and young learners' scientific investigation behaviors. Results show that cluster analysis is promising for supporting pattern interpretation and scientific communication behaviors. The levels of cognitive complexity are associated with different ML-powered SD and corresponding learning support is needed. The next steps include a further co-design study between K-12 STEM teachers and ML experts and a plan for collecting and analyzing data to further understand the connection between ML and SD.

Introduction

As Machine Learning (ML) becomes more prevalent in scientific investigation, it is crucial to introduce it as a new tool for K-12 students' scientific discovery (SD). However, few empirical efforts contributed to supporting K-12 students in conducting scientific investigations, discovering causality, and making arguments with authentic multidimensional data (Kuhn 2016; Kuhn et al. 2017; Kuhn, Ramsey, and Arvidsson 2015). The current school curriculum, limited to bivariate data, may pose a challenge for K-12 students to interpret ML-generated data (Association et al. 2010). With the learning environment we created, our preliminary evaluation indicated that young students can be more equipped with knowledge and skills of a global understanding of a group of data, which is fundamental to interpreting multidimensional data (Ben-Zvi and Arcavi 2001). However, it is under-explored how different ML components can inspire different scientific discovery behaviors during young students' learning.

Existing work either establishes connections between machine learning methods and scientific discovery in the science community's professional practices (Gil et al. 2014), or conducted a descriptive analysis with a few ML-powered SD learning activities created by K-12 STEM teachers (Zhou et al. 2021). To gain a set of more reliable and generalizable connections between ML components and young learners' SD learning behaviors, however, a larger amount of data and modeling are needed. Thus, we raise one research question:

1. How to identify potential connections between ML components and SD behaviors to create adaptive feedback to scaffold novice learners to go through the ML-empowered SD learning processes?

To address this research question, we model the connection between ML components and SD learning behaviors using data from 25 young learners' interactions with an accessible ML-powered SD environment. Our proposed ML-SD authoring system for K-12 teachers allows them to create their own ML-powered SD activities and explore data with ML methods. Initial results show how ML components such as cluster analysis and outlier analysis can enhance pattern interpretation and scientific communication. We also analyze the relationship between learners' performance and language use in scientific discussions, offering insights for effective scaffolding for K-12 students to carry out ML-powered scientific discovery.

Related Work

Integrating ML and K-12 STEM Education

There are emerging research efforts to explore the opportunities of making ML concepts and methods accessible for K-12 students (Evangelista, Blesio, and Benatti 2018; Lin et al. 2020; Wan et al. 2020; Zimmermann-Niefield et al. 2019). One study shows that data visualization supports students with limited computing knowledge to gain a basic understanding of cluster analysis (Wan et al. 2020). Further, it indicates the potential of applying ML methods for pattern interpretation by pattern generation. Zimmermann-Niefield et al. (2019) facilitates youth to train and test ML models of their athletic activities. It shows that ML enhances science learning by aligning ML modeling with modeling scientific phenomena, an essential practice of science recommended in curriculum standards (States 2013). Design guidelines have been extracted from existing research about introducing ML in K-12 STEM contexts, such as unveiling complex ML concepts step by step (Evangelista, Blesio, and Benatti 2018; Lin et al. 2020) and visualizing ML

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Table 1: ML components in the system for scientific investigation.

ML component	ML sub-component for individual tasks
Similarity computation	T1 - Intra-group similarity comparison
	T2 - Intra-group variation comparison
	T3 - Inter-group variation comparison
Centroid	T4 - Centroid
Outlier	T5 - Outlier analysis
K-value selection	T6 - K-value selection for k-means clustering
Cluster analysis	T7 - Cluster analysis with k-means clustering

models for explainability (Essinger and Rosen 2011; Wan et al. 2020; Zimmermann-Niefield et al. 2019). A recent literature review on ML learning and teaching in K-12 (Sanusi et al. 2022) reveals the emerging future research directions include (1) more ML resources are needed for K-12 and informal settings; (2) further research is needed on integrating ML into non-computing subject areas; (3) most studies concentrate on pedagogical development, with limited focus on teacher professional development; (4) future research should examine the societal and ethical implications of ML. These findings align with the fundamental research motivation that we propose to make ML more accessible tools for kids to conduct discovery in diverse subject domains and for teachers to develop corresponding professional skills.

Research shows that ML approaches empower data-driven discovery by enabling hypothesis generation, iterative experimentation with different parameters, and pattern recognition by gradually revealing more refined parameters (McAbee, Landis, and Burke 2017; Muller et al. 2016). Various ML techniques have been proposed to automate SD (Langley 2000). For example, k-means clustering, an unsupervised ML algorithm, is used to discover laws by grouping similar objects (Essinger and Rosen 2011; Evangelista, Blesio, and Benatti 2018), identify dependencies of attributes (Skapa et al. 2012), and form taxonomies (Wang, Nie, and Huang 2014). Such methods, however, are applied in science at a professional level (Gil et al. 2014; Kitano 2016) and thus are inappropriate for K-12 teachers and students with limited CS/ML backgrounds. This points out a demand for designing an ML-powered SD learning environment in K-12 contexts.

Connections between ML and SD

Inquiry-based learning (IBL) is a well-recognized educational strategy in today's K-12 classrooms to facilitate scientific discovery (SD) learning (Pedaste et al. 2015; Furtak et al. 2012; De Jong, Sotiriou, and Gillet 2014; Gormally et al. 2009). To engage students in SD, the main SD stages involve hypothesis generation, investigation, and discussion (Pedaste et al. 2015). Hypothesis generation is the process of generating a testable hypothesis. Investigation refers to the process of analyzing data to answer questions or test hypotheses, which involves two main activities: a) planning and conducting exploration activities or experimentation; b) analyzing and interpreting data. Discussion, the last one, is to communicate with others to present findings and collect feedback.

From existing ML-supported science research, we identified common connections in how different ML methods can be used to support different SD stages. First, ML brings new opportunities for scientists to generate hypotheses by automatically extracting patterns from large datasets (Gil et al. 2014). Second, pairwise similarity comparison (i.e., comparing two data points) supports initial observations of contrastive data points, which can trigger contrastive explanations that support students' abductive reasoning for hypothesis generation and investigation (Folger and Stein 2017). Third, clustering and classification can serve as more data-driven methods for exploration or experimentation (Romesburg 2004; Gil et al. 2014). Existing work also identifies common ML-SD connections in learning activities designed by K-12 teachers (Zhou et al. 2021). For example, teachers preferred hypothesis generation by ML methods with a small amount of data rather than prior knowledge. Teachers also designed iterative investigations from small to large datasets.

System Design: ML-Powered SD Learning Environment

ML Components

The main ML components (Table 1) involved in the system include (1) similarity computation, (2) centroid, (3) outlier, (4) k-value selection, and (5) cluster analysis. With similarity computation, learners are supported in examining shared patterns within a group of data points (T1: intra-group similarity comparison), analyzing variations of individual features within a group (T2: intra-group variation comparison), and comparing the overall variation of all data attributes between two groups of data points (T3: inter-group variation comparison). The ML component of centroids enables learners to generate a centroid for a cluster of data points and compare patterns between different centroids. Outlier analysis allows learners to examine outliers for individual clusters. With the ML components of k-value selection and k-means clustering results, learners can study the impact of k-values on clustering results and examine potential relationships among data attributes through cluster analysis. The corresponding data visualization and interaction design is depicted in Figure 1.

Learning Activity Design

Four scientific investigation learning activities are designed in the system with the ML components described above.

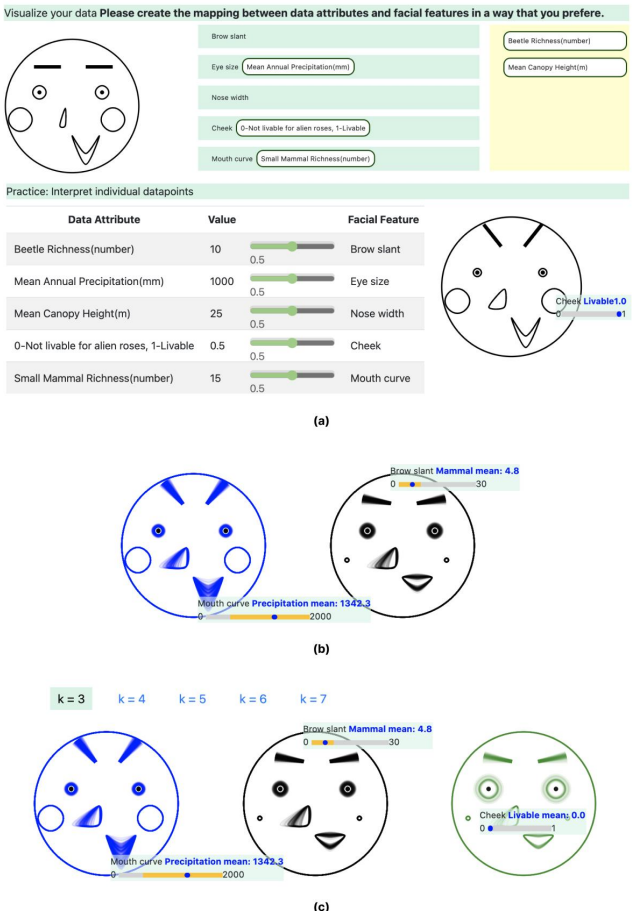

Figure 1: ML-Powered SD learning environment with a set of scientific investigation activities: (a) getting familiar with the data visualization and interaction; (b) interpreting the classification result by applying the ML component - similarity computation; (c) applying ML components related to k-means clustering, including outlier analysis, k-value selection, and cluster analysis.

The first learning activity is designed with a dataset about adult income. Learners are guided to investigate the potential factors influencing a person's income level. In the second learning activity, learners conduct a scientific investigation on what ecological features make a field site livable for an alien rose. The third learning activity, involving a breast cancer dataset, asks learners to discover the patterns of malignant cells and the patterns of benign cells. The fourth activity integrates a dataset related to TV shows and their target audience. Learners apply ML components to investigate what features are related to high-rating TV shows and what features are more likely to form low-rating shows.

Research Method

Participants

This research study enrolled 25 participants through teachers and parents in compliance with the approved Institutional Review Board guidelines. The gender composition of the participants was comprised of 10 women, 14 men, and one participant who chose not to disclose their gender. The participants' grade levels ranged from 7 to 11, with a mean of 9.76 and a standard deviation of 1.16. The pre-study survey revealed that 12 participants had no previous exposure to AI, five had exposure to AI-related consumer products or films, and eight had participated in robotics clubs or possessed basic coding knowledge.

Procedure and Data Collection

The study took place online via Zoom and each study session lasted about 1.5 hours, facilitated by one researcher. During the study, each participant went through four learning activities with the same ML components but different visualization designs. Therefore, the sample size is $25 \times 4 = 100$. To address the issue of a participant's exposure to four activities with different visualization designs affecting the learning outcome, we employed partial counterbalancing and calculated the number of various combinations through the Latin square technique (Gravetter and Forzano 2018).

During ML-powered SD learning, participants are asked to think aloud (Van Someren, Barnard, and Sandberg 1994) while working on individual investigation tasks. Through such think aloud method, we are able to collect audio recording data describing learners' cognitive process of investigating data patterns by applying different ML components.

Data Analysis

Data Annotation A coding scheme along with corpus examples (Table 2) is developed to analyze the investigation behaviors during scientific discovery learning based on the existing literature review on SD learning (Pedaste et al. 2015). The codes mainly focus on the investigation phase during a cycle of inquiry-based learning.

Preliminary Analysis of the ML-SD Connection To have a preliminary view of the association between different ML components and SD learning behaviors, we calculated the occurrence of ML components for individual SD behaviors. This reveals potential patterns in how different ML components are connected with investigation behaviors. To obtain robust findings, we exclude the SD behaviors with less than 10 observations. As a result, behaviors include observation and orientation, exploration and experimentation, analysis, pattern interpretation, and reflection.

Linguistic Analysis To have a further understanding of when the learning support will be most needed, we investigate the difficulty levels for different parts of such ML-powered investigation by analyzing the cognitive complexity that learners experience while applying different ML components for investigation. We apply LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker 2010) to the transcripts of the participants during the experiment. LIWC is a lexicon-based textual data analysis framework that measures the psychological states, sentiment, and linguistic patterns of the authors by counting the words of different categories. It outputs multiple linguistic variables

Table 2: A preliminary coding scheme used to annotate the video and audio recordings of students applying different ML components for scientific investigation.

SD Behaviors in Investigation Phase	Definition	Definition in the study	Corpus Examples
Observation and orientation (Topic)	Behaviors in relation to gaining interest and obtaining background information about a topic at hand.	Learning about the scientific context to be investigated, such as the meaning of data attributes.	"So tell me the definition of adhesion (a data attribute in the Breast Cancer dataset)."
Observation and orientation (Tech)	Behaviors in relation to addressing learning challenges of technology design.	Learning the operation and the function of SmileyDiscovery in supporting scientific investigation, such as the mechanisms behind the connections between data attributes and visual features.	"(The participant is trying to figure out how data attributes and facial features are connected) Does it [data attributes] go anywhere [any facial feature]?"
Exploration and experimentation	Behaviors in relation to collecting data or constructing models.	Conducting observation by applying specific functions of SmileyDiscovery, such as adjusting the pointer, selecting the k value, overlaying glyphs, and canceling the overlay.	Clicking different k values to investigate the changes in different clustering results; shifting between generating centroids and taking back centroids to investigate differences between two data points or clusters.
Analysis	Behaviors in relation to analyzing data and presenting evidence.	Explaining data using visual features of SmileyDiscovery, such as their shapes, positions, and distribution (variation v.s. concentrate)	"...some of these (data points) still like have that variation"
Pattern interpretation	Behaviors in relation to making meaning out of data or models.	Synthesizing or integrating different pieces of information from data analysis to answer a target question.	"I would say low precipitation, and then, um, few small animals."
Reflection	Behaviors in relation to students reconstructing their understanding of the topic or the data after receiving feedback or evaluation.	Responding to the hints, evaluations, or critiques provided by the researcher.	"Uh, because those with higher income."

such as sentiment, cognitive processes, and so on. It can output the linguistic score that indicates the cognitive complexity involved in certain activities (Li and Lee 2019; Biel et al. 2013; Haulcy and Glass 2021). More specifically, we first concatenate the text of each participant between two ML components, and then we apply LIWC. Further, we perform multiple ANOVA experiments to test whether or not there is a significant difference in the linguistic scores among different SD behaviors. The significant level is set as 0.05.

Preliminary Results of ML-SD Modeling

ML components are associated with different SD behaviors. The distributions of the SD behaviors and the ML components across different categories are depicted in Figure 2 and Figure 3, respectively. It is noteworthy that the first column in Figure 3 represents the distributions in the pattern interpretation behaviors during the investigation stage. Each SD behavior is associated with ML components to varying degrees.

First, cluster analysis (T7) is associated with analysis, pattern interpretation, and reflection, the key behaviors during the scientific investigation. The SD behavior triggered by cluster analysis, in particular, exhibits a high proportion of pattern interpretation due to the involvement of a larger number of pattern elements for learners to analyze and integrate. It is also the most frequent ML component that triggers reflection behavior. It shows the need for feedback and evaluation for the investigation supported by cluster analysis. This also indicates that cluster analysis may have a higher potential to support the effective development of the skills for scientific discovery learning and scientific communication. Second, k-value selection for k-means clustering is associated with more than half of the young learners' experimentation behaviors. This suggests that the practice of adjusting ML algorithm parameters and observing its immediate effect on data visualization can encourage experimentation for kids. This can further foster a trial-and-error learning approach. Third, outlier analysis (T5) is observed

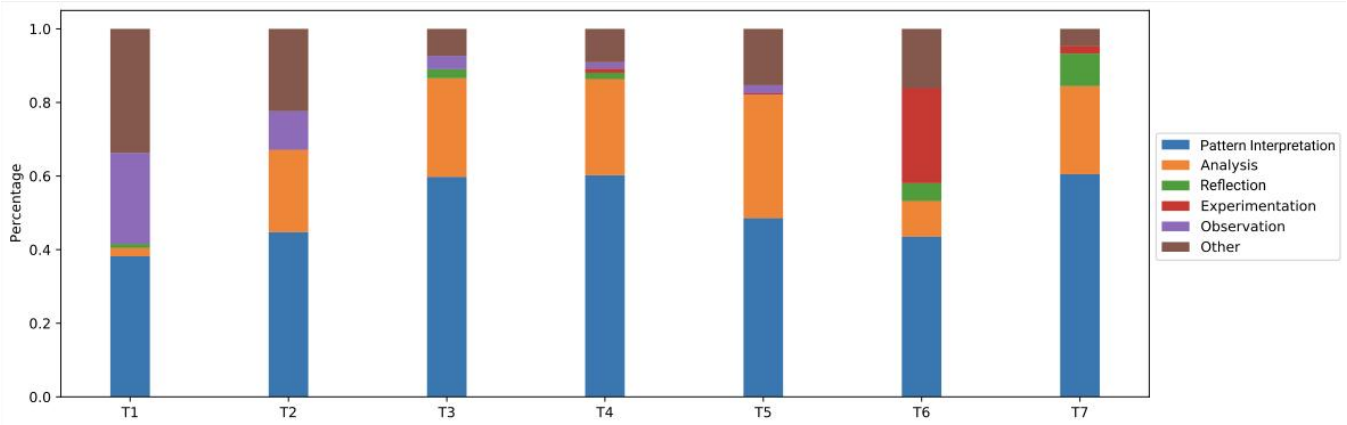

Figure 2: Distribution of the SD behaviors across various ML components: T1 - Intra-group similarity comparison, T2 - Intra-group variation comparison, T3 - Inter-group variation comparison, T4 - Centroid, T5 - Outlier analysis, T6 - K-value selection for k-means clustering, T7 - Cluster analysis with k-means clustering.

to be more likely to trigger analysis and pattern interpretation. This implies that outlier analysis potentially inspires more pattern interpretation behaviors or requires more effort in analyzing data to reach conclusions.

The language reveals different levels of cognitive complexity during various ML components and SD behaviors. According to the results of ANOVA, we find significant differences in the average linguistic scores among different ML components and SD behaviors ($p < .05$). In particular, the cognitive process score is higher during those advanced stages such as reflection and analysis, and lower during the initial stages such as observation. This suggests that the introduction of the new ML components didn't make the initial interaction more cognitively challenging than data analysis and pattern interpretation. The cognitive process score represents the frequency of words such as "cause", "know", and "ought", which suggests the active process of reappraisal. An increasing usage may be indicative of a higher level of cognitive complexity (Tausczik and Pennebaker 2010). This is consistent with the dynamic mechanism of SD behaviors, where synthesizing different pieces of evidence from data is a sign of a more comprehensive understanding (Pedaste et al. 2015).

The ML components, ranked in order of decreasing cognitive process scores, are T6, T2, T7, T3, T5, T1, and T4. It is noteworthy that, despite being a more open-ended and advanced task that involves a combination of ML components, cluster analysis (T7) exhibits a lower cognitive complexity compared to k-value evaluation (T6) and intra-group variation comparison (T2) with similarity computation. The reason for this may stem from our design strategy of incrementally introducing complexity. By placing cluster analysis as the final task in the learning activity, a strong foundation is provided for young learners to gradually gain a basic understanding of how to apply various ML sub-components in scientific investigations. The findings suggest a need for further visualization and interaction design to scaffold the analysis of variations among different attributes within a group of data points, as well as the evaluation of the overall clustering results.

Future Work

Further Co-Design Study

For machine learning educators and researchers, we would like to highlight a limitation in our previous study where teachers co-designed ML-powered SD learning activities. The study used an online platform for creating diagrams but teachers could not refine their designs based on real-time outputs from ML components. This hindered in-depth discussions on ML's role in K-12 STEM education and led to missing details in the final design results. To address this, we suggest conducting a follow-up study where teachers design and implement ML-powered SD learning activities using an ML-SD authoring system, allowing them to interact with intermediate results generated by ML components.

According to teachers' feedback from the prior study, some of them still did not feel prepared enough to develop ML-powered SD learning activities, even when collaborating with other teachers. As a result, the further co-design study should support more elaborated collaboration between teachers and ML experts.

Data Collection and Analysis for Future Modeling

In addition to the video transcripts, we intend to exploit the video itself to further understand the interplay between ML components and SD learning activities. For example, we intend to capture the level of engagement of each participant by analyzing how the participant uses the learning tool of our study.

In terms of modeling, we intend to extend our study in two directions. First, unlike the univariate statistical tests we conduct in the current study, we plan to jointly model the visual features, textual features, and the involvement of different ML components to investigate the SD learning activities. Second, the current experiment design is cross-sectional. We plan to employ models such as LSTM (Long Short-Term

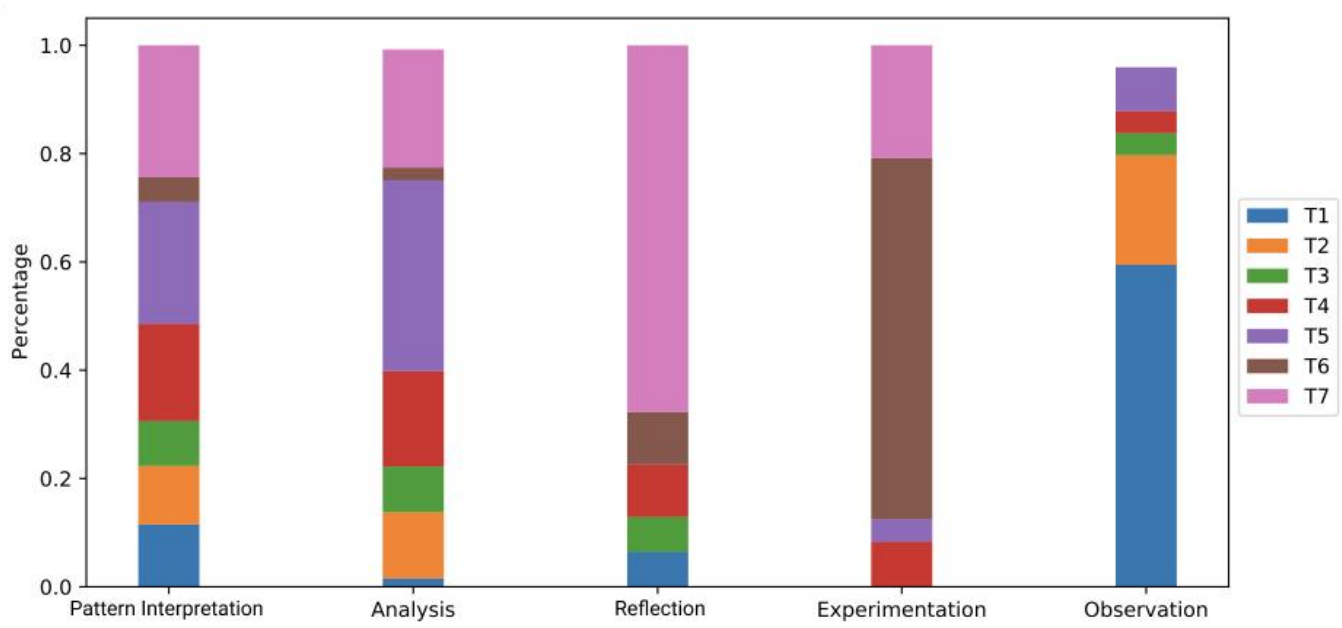

Figure 3: Distribution of the ML components across various SD behaviors.

Memory) (Hochreiter and Schmidhuber 1997) to incorporate the temporal patterns because of the dynamic characteristics of SD learning activities (Pedaste et al. 2015).

Acknowledgement

This work was supported in part by the National Science Foundation (RETTL program award No. 2225227).

References

Association, N. G.; et al. 2010. Common core state standards. *Washington, DC*.

Ben-Zvi, D.; and Arcavi, A. 2001. Junior high school students' construction of global views of data and data representations. *Educational studies in mathematics*, 45(1): 35–65.

Biel, J.-I.; Tsiminaki, V.; Dines, J.; and Gatica-Perez, D. 2013. Hi YouTube! Personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 119–126.

De Jong, T.; Sotiriou, S.; and Gillet, D. 2014. Innovations in STEM education: the Go-Lab federation of online labs. *Smart Learning Environments*, 1(1): 1–16.

Essinger, S. D.; and Rosen, G. L. 2011. An introduction to machine learning for students in secondary education. In *2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, 243–248. IEEE.

Evangelista, I.; Blesio, G.; and Benatti, E. 2018. Why Are We Not Teaching Machine Learning at High School? A Proposal. In *2018 World Engineering Education Forum-Global Engineering Deans Council (WEEF-GEDC)*, 1–6. IEEE.

Folger, R.; and Stein, C. 2017. Abduction 101: Reasoning processes to aid discovery. *Human Resource Management Review*, 27(2): 306–315.

Furtak, E. M.; Seidel, T.; Iverson, H.; and Briggs, D. C. 2012. Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of educational research*, 82(3): 300–329.

Gil, Y.; Greaves, M.; Hendler, J.; and Hirsh, H. 2014. Amplify scientific discovery with artificial intelligence. *Science*, 346(6206): 171–172.

Gormally, C.; Brickman, P.; Hallar, B.; and Armstrong, N. 2009. Effects of inquiry-based learning on students' science literacy skills and confidence. *International journal for the scholarship of teaching and learning*, 3(2): n2.

Gravetter, F. J.; and Forzano, L.-A. B. 2018. *Research methods for the behavioral sciences*. Cengage learning.

Haulcy, R.; and Glass, J. 2021. Classifying Alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11: 624137.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Kitano, H. 2016. Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery. *AI magazine*, 37(1): 39–49.

Kuhn, D. 2016. What do young science students need to learn about variables? *Science Education*, 100(2): 392–403.

Kuhn, D.; Arvidsson, T. S.; Lesperance, R.; and Corprew, R. 2017. Can engaging in science practices promote deep understanding of them? *Science Education*, 101(2): 232–250.

Kuhn, D.; Ramsey, S.; and Arvidsson, T. S. 2015. Developing multivariable thinkers. *Cognitive Development*, 35: 92–110.

Langley, P. 2000. The computational support of scientific discovery. *International Journal of Human-Computer Studies*, 53(3): 393–410.

Li, J.-L.; and Lee, C.-C. 2019. Attentive to Individual: A Multimodal Emotion Recognition Network with Personalized Attention Profile. In *Interspeech*, 211–215.

Lin, P.; Van Brummelen, J.; Lukin, G.; Williams, R.; and Breazeal, C. 2020. Zhorai: Designing a Conversational Agent for Children to Explore Machine Learning Concepts. In *AAAI*, 13381–13388.

McAbee, S. T.; Landis, R. S.; and Burke, M. I. 2017. Inductive reasoning: The promise of big data. *Human Resource Management Review*, 27(2): 277–290.

Muller, M.; Guha, S.; Baumer, E. P.; Mimno, D.; and Shami, N. S. 2016. Machine learning and grounded theory method: Convergence, divergence, and combination. In *Proceedings of the 19th International Conference on Supporting Group Work*, 3–8.

Pedaste, M.; Mäeots, M.; Siiman, L. A.; De Jong, T.; Van Riesen, S. A.; Kamp, E. T.; Manoli, C. C.; Zacharia, Z. C.; and Tsourlidaki, E. 2015. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational research review*, 14: 47–61.

Romesburg, C. 2004. *Cluster analysis for researchers*. Lulu.

Sanusi, I. T.; Oyelere, S. S.; Vartiainen, H.; Suhonen, J.; and Tukiainen, M. 2022. A systematic review of teaching and learning machine learning in K-12 education. *Education and Information Technologies*, 1–31.

Skapa, J.; Dvorsky, M.; Michalek, L.; Sebesta, R.; and Blaha, P. 2012. K-mean clustering and correlation analysis in recognition of weather impact on radio signal. In *2012 35th International Conference on Telecommunications and Signal Processing (TSP)*, 316–319. IEEE.

States, N. L. 2013. *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.

Van Someren, M.; Barnard, Y. F.; and Sandberg, J. 1994. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress*, 11: 29–41.

Wan, X.; Zhou, X.; Ye, Z.; Mortensen, C. K.; and Bai, Z. 2020. SmileyCluster: supporting accessible machine learning in K-12 scientific discovery. In *Proceedings of the Interaction Design and Children Conference*, 23–35.

Wang, D.; Nie, F.; and Huang, H. 2014. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track). In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 306–321. Springer.

Zhou, X.; Tang, J.; Michael, D.; Ahmad, S.; and Bai, Z. 2021. "Now, I Want to Teach it for Real!": Introducing Machine Learning as a Scientific Discovery Tool for K-12 Teachers. (In press).

Zimmermann-Niefield, A.; Turner, M.; Murphy, B.; Kane, S. K.; and Shapiro, R. B. 2019. Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, 121–132.