How to Teach Unsupervised Machine Learning with Analogies

A Study on the Effectiveness of Analogies in Teaching Unsupervised Machine Learning

V.J. (Vincent) Ruijgrok[1]

Supervisor(s): Gosia Migut[1]**, Ilinca Rențea**[1]**, Yuri Noviello**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2025

Name of the student: V.J. (Vincent) Ruijgrok
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Ilinca Rențea, David Tax

Abstract

Unsupervised machine learning is a complex and abstract topic, posing challenges for student comprehension. Considering the considerable growth of relevance the topic of machine learning has seen in the past years, teaching it effectively has become ever-so important. Analogy-based teaching approaches offer a potential solution by mapping unfamiliar machine learning concepts to familiar real-world ideas. This paper investigates how analogies can improve the understanding of unsupervised learning, a rather relevant field within machine learning. Contributions include a collection of analogies for teaching unsupervised ML, an expert-based evaluation of these analogies' quality, and a student-centred assessment of analogy-based teaching. The findings from the expert evaluation show a consensus on the effectiveness of several analogies and highlight which analogies might be less effective. The findings from the student assessment suggest that the analogical explanations are more effective than 'generic' explanations and suggest that students have a higher satisfaction while learning through analogies. We conclude that well-crafted analogies can enhance student understanding in unsupervised machine learning. The study's insights can guide educators in integrating analogies to make unsupervised learning more accessible.

1 Introduction

Machine Learning (ML) has become a major part of modern society, with applications ranging from healthcare and finance to entertainment and education [1]. As its practical relevance grows, so does the need for effective teaching methods that help students, not only understand, but also apply complex ML concepts [2].

Using analogies in science education helps students form mental models by linking abstract concepts to familiar, tangible experiences [3]. Within machine learning education, the use of analogies is only beginning to be systematically explored [4]. Traditional ML instruction tends to be maths-heavy and abstract, which can alienate newcomers [4]. Analogies offer a promising way to demystify concepts by comparing them to real-life processes. There is a lack of formal studies assessing how such analogies impact learning [5]. Saxena et al. [5] note that the role of analogies in computer science and ML teaching has been under-investigated in the literature.

This research will focus on the concepts in unsupervised machine learning. This sub-field was chosen primarily because of its relevance: it is used in healthcare [6, 7], urban studies [8], and finance [9–11]. Additionally, unsupervised learning algorithms, like clustering, can struggle with interpretability [12], signifying the need for proper and effective teaching methods.

This research aims to bridge the gap by collecting, evaluating, and organizing useful metaphors for key ML concepts in unsupervised learning. The main research question is:

> *How does the use of analogies in teaching unsupervised learning affect the knowledge gain measured by the answering of theoretical questions?*

To support this, these three sub questions are considered:

- What are key concepts for unsupervised learning?
- Which analogies can be used to explain these key concepts, according to expert teachers in the field?
- What is the effect on students' quiz results for the analogies, in comparison to standard explanations?

1.1 Contribution and Structure

The main contribution of this research are

- a set of evaluated analogies
- a mapping for each analogy to a core concept in unsupervised learning
- empirical data showing their impact on student understanding

These results provide insights for educators looking to improve ML teaching and lay a foundation for future work on systematic analogy use in machine learning education.

In Figure 1 the organisation of this research is visualised. This research is organized into three main phases: (1) analogy generation, (2) expert reviews, and (3) student survey.

The rest of this paper is structured as follows. Section 2 contains the background to this research and references to related literature. Section 3 describes Phase (1), where the relevant concepts within unsupervised learning were collected and an analogy was generated for each concept. Section 4 describes Phase (2), where these analogies were reviewed by several experts on several evaluation criteria. Section 5 describes Phase (3), where the best analogies were selected from the results in Phase (2) and the difference in knowledge gain between the students who received an analogy-based explanation and the students who received a "generic" explanation was measured. Additionally, Section 6 addresses responsible research aspects, such as ethical concerns and transparency goals. Section 7 provides a discussion of the results and Section 8 concludes the paper and outlines directions for future work.

2 Background

Analogy definition

The Cambridge Dictionary defines an analogy as "*a comparison between things that have similar features, often used to help explain a principle or idea*"[1]. Bhavya et al. [13] have defined an analogy to be a set of three things:

1. Target concept: the concept that is being explained
2. Source concept: the concept the target is being compared to
3. Mapping: the links between features of the target concept and the source concept

[1] https://dictionary.cambridge.org/dictionary/english/analogy

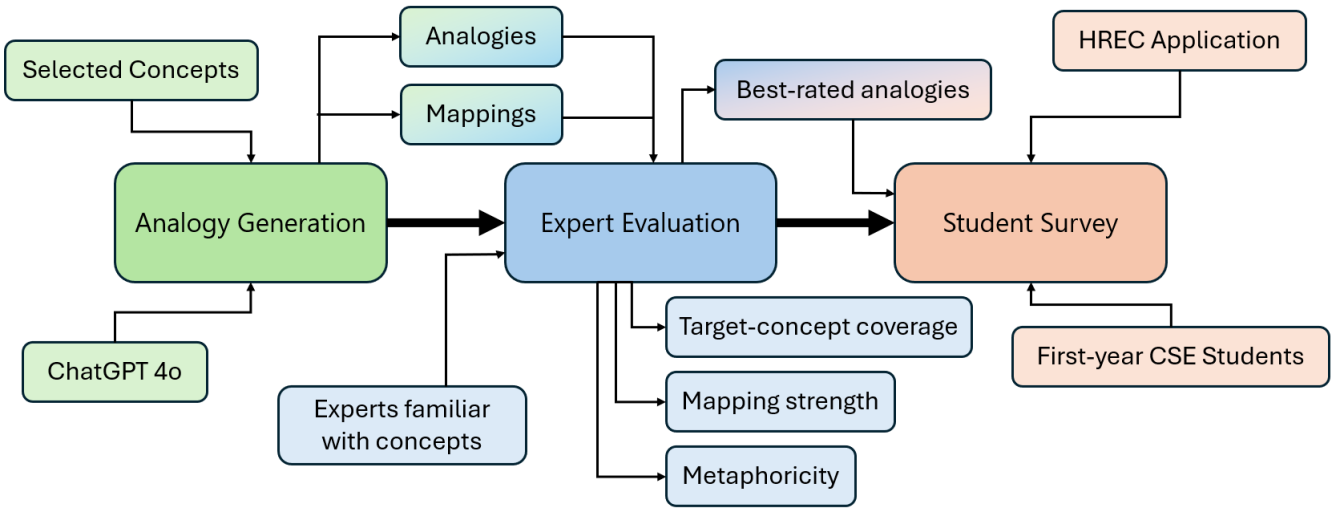

Figure 1: An overview of this research's methodology

Take the analogy "*an array is like a stack of boxes*". The array is the target concept and the stack of boxes is the source concept. The mapping can be defined as: the array is like the stack of boxes; an array element is like a box; an element index is like the number on each box and the element value is like the concents of the box [2].

Analogies in Science Education

Prior research confirms multiple benefits of analogy-based learning. Effective analogies can motivate and engage learners, provide vivid mental models, and act as cognitive scaffolding for new knowledge [3]. They help students visualize abstract processes and can even address misconceptions by seeing the problem in a new light. Crucially, analogies can reduce cognitive load when learning complex material: by drawing parallels to familiar situations, students expend less mental effort decoding pure abstractions [14].

Saxena et. al [5] note that computer science concepts are highly abstract and therefore hard to grasp, and that analogy-driven instruction can significantly ease this burden on learners' working memory. In their work, they developed several analogies for computer science topics and observed improved student engagement and retention. Experiments by Saxena et al. [5] showed that using analogies in teaching led to measurable improvements in student learning outcomes, validating the efficacy of analogy-based techniques. While these are promising prior results, their research was based on a single class, a single instructor, and two concepts (operating systems and module cohesion) and therefore might not be representative of topics within machine learning.

Analogies in Machine Learning Education

An extensive search for literature on the use of analogies in ML resulted in a single paper from Pendyala [4]. They also state in their paper that "From a literature survey and to the best of the author's knowledge, this work is the first of its kind.", signifying the research gap here.

Pendyala started filling the research gap for analogies in ML education with their 2022 paper [4]. In this paper, they stated numerous analogies for different topcis within ML. The analogies are generated "based on human intuition and ingenuity". Additionally, the analogies are not evaluated.

Analogies are a proven education strategy in other related fields (like software engineering). Analogies for ML concepts have been generated before, but their effectiveness has not been measured. That is the research gap this research tries to fill.

3 Analogy Generation

To identify the core topics within unsupervised machine learning, the lecture notes and slides of the introductory machine learning course for second year CS majors at Delft University of Technology were used[3]. The ten biggest, most-emphasized concepts were chosen to be relevant for this research: unsupervised learning, clustering, intra-cluster cohesion, inter-cluster cohesion, K-means clustering, single linkage, complete linkage, average linkage, agglomerative dendrogram, and divisive dendrogram.

Method

According to Shao et. al [15], using Large Language Models (LLMs) to generate analogies for concepts, can result in analogies that are effective. However, it is also stated by Shao et al. [15] that teachers or experts should review the results of an LLM-response. Because the author did not feel that their own ingenuity would lead to sound analogies, it was chosen to use this method. An LLM would generate the analogies and experts would then review them first in Phase (2).

The prompt for the LLM (ChatGPT 4o) contained the research background and motivation, as well as an early version of the research questions and the list of concepts. Finally, it

[2] https://notionalmachines.github.io/nms/ArrayAsStackOfBoxes.html

[3] https://ml-teaching-analogies.github.io/cse2510-notes.pdf

prompted the model to come up with a definition of the given concepts, an analogy for that concept and a mapping, where features of the concepts are linked to features of the analogies. The prompt can be found in Appendix B.

Results

The resulting analogies can be found in Appendix A. The definitions in the results were checked against the definitions in the slides and lecture notes of the machine learning course and no mistakes were found. Below, as an example, the analogy for **complete linkage** and for **clustering** will be given.

Complete Linkage
If two cities want to ensure their farthest apart homes are still within reach, complete linkage measures the distance between the two farthest houses before connecting the cities.

- Clusters = Cities
- Data points = Houses
- Linkage distance = Maximum distance between any two houses across the two cities
- Result = Clusters only merge when all elements are close enough

Clustering
Sorting your socks after doing laundry: you don't have labels, but you pair socks based on their colour, size, and material to form neat clusters of matching socks.

- Data points = Individual socks
- Similarity = Matching in colour/size/material
- Clusters = Piles of similar socks
- Clustering process = You sorting socks based on similarities

4 Expert Evaluation

To evaluate the quality and appropriateness of the generated analogies before testing them with students, an expert review was conducted.

4.1 Method

It was decided that experts were people who had at least passed one ML course in a CS bachelor, since they should be familiar with all these concepts and thus can review the quality of the analogies.

Procedure
The survey was constructed in two parts: first, all participants answer the question *What is your Machine Learning knowledge level? Select the one that fits your situation best.*. The given options were: 'I have passed the Machine Learning course in a Computer Science Bachelor', 'I have TA'd the Machine Learning course in a Computer Science Bachelor', 'I have passed a Master course on the topic of Machine Learning', 'I am a lecturer/professor in a Machine Learning course', and 'None of the above'.

The main body of the survey consisted of randomly-ordered questions, with one analogy per question. The question first defined the concept, then gave a textual analogy and lastly gave the mapping of the concept features to the target features. The participants were then asked to rate the analogy and mapping on three evaluation criteria.

Evaluation Criteria
The chosen criteria were: target-concept coverage, mapping strength and metaphoricity, each of which could be ranked with 1 (low), 2 (mid), and 3 (high). These criteria were based on an improved version of the criteria as designed by Bhavya et al. [13], as described on their research's website [4]:

- **Target-concept coverage** describes whether the target (the analogy) covers the concept.
- **Mapping strength** describes the logical soundness of the mappings of features of the concepts to features of the target.
- **Metaphoricity** describes the conceptual distance between the target and the concept, that is, how closely related the target and concept are.

Participants
The participants were approached through the personal network of the author. At the start of the survey, only the ML proficiency was asked, so there is no data on where the students came from or which study they had followed. However, it is likely that the participants of the study were students or teachers at Delft University of Technology, most of them doing the bachelor Computer Science and Engineering or a Computer Science master, since those are the people that were approached.

In total, the survey saw 16 respondents, 10 of whom had passed a ML course in any CS Bachelor programme, 3 had been a Teaching Assistant for such a course, 2 had passed an ML course in any Master programme and 1 was a lecturer or professor in an ML course. 1 responder indicated to not have passed any ML courses, so their ML proficiency remains undetermined and their response was not taken into account during the analysis.

Survey
The expert feedback form was designed to be independently answered by each expert to avoid group influence. Experts submitted their responses anonymously via an online questionnaire to ensure candid feedback. Because the survey also contained analogies for other studies, the number of analogies (50) was deemed too large for one participant to evaluate. Therefore, participants were instructed to submit their response when they did not want to continue any more. The result of this is that not all analogies have been rated by the same number of experts.

4.2 Results

Figure 2 summarizes the expert evaluation of each analogy across the three criteria of target-concept coverage, mapping strength, and metaphoricity. Each analogy was rated by multiple experts (N); mean scores as well as standard deviations were calculated per criterion.

[4] https://sites.google.com/illinois.edu/analogyeval24/analogy-evaluation-criteria

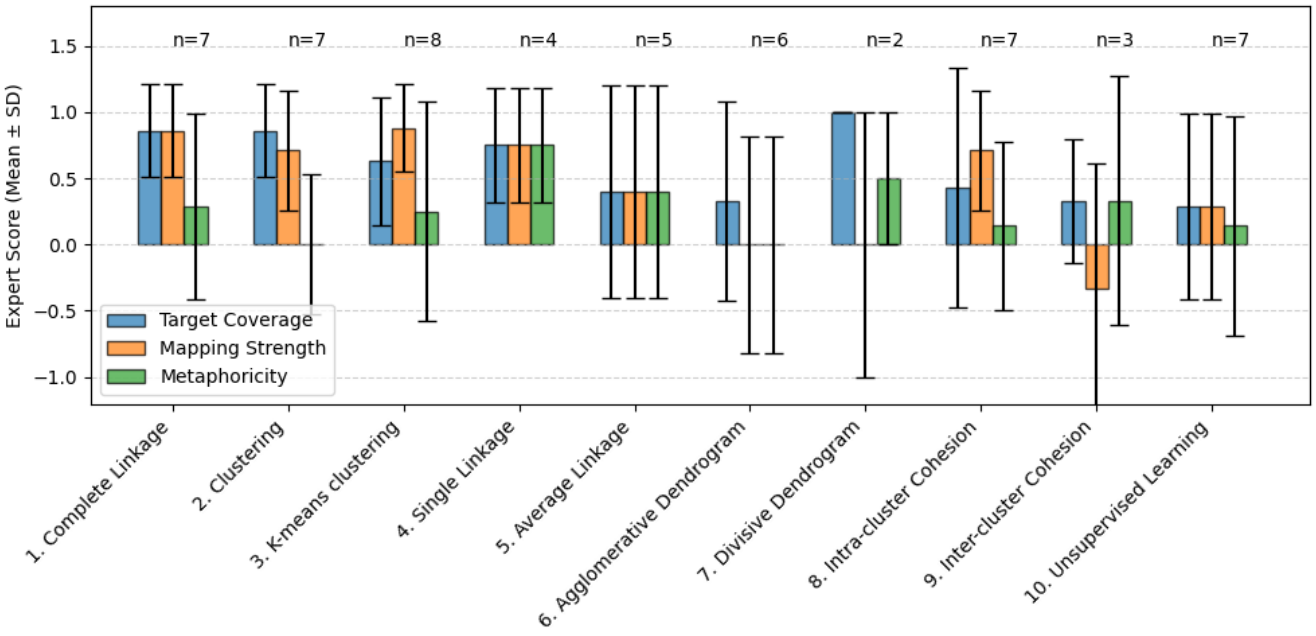

Figure 2: Expert evaluation results per analogy on three criteria: target-concept coverage, mapping strength, and metaphoricity (mean ± SD).

The evaluation identified a top tier of three analogies that outperformed the others. These four highest-rated analogies (Analogies 1–4 in Figure 2) each achieved a similarly higher rating than the other six. Standard deviations for their scores tend to be smaller than the other analogies, reflecting a stronger consensus among experts regarding their quality.

The remaining analogies formed a lower tier with moderate to lower performance overall. In general, this group's mean ratings on coverage and mapping strength were lower. Several of these analogies struggled with target-concept coverage: for example, unsupervised learning and inter-cluster cohesion. Likewise, mapping strength was a weakness for some in this tier, for example divisive dendrogram.

It should be noted that all the inter-rater agreement values are low, this could be because of the subjectivity of learning and analogies, as mentioned by He et al. [16].

Some analogies in this group were quite metaphorical (for instance, divisive dendrogram had a high metaphoricity mean), but this was not enough to overcome shortcomings mapping strength.

Inter-rater Agreement

For the evaluation of the agreement among experts, the Krippendorf's Alpha value was used. It was calculated with the online K-Alpha calculator[5], which has been validated by Marzi et al. [17]. The results are presented in Figure 3.

The analogies can be grouped into three groups:

1. **Analogies 1-3**: the experts have at least some agreement on the quality of these analogies. They tend to agree that these analogies are of a higher quality, according to the previous section.
2. **Analogies 4-7, 10**: the experts do not agree on the quality of these analogies.
3. **Analogies 8-9**: the experts have at least some agreement on the quality of these analogies. They tend to agree that these analogies are of a lower quality, according to the previous section.

Considering this evaluation, analogies 1, 2, and 3 will be further evaluated in the student evaluation.

5 Student Survey

To evaluate the impact of the best-rated analogies on student learning, an A/B study with first-year computer science student participants was conducted.

5.1 Method

The survey's design and set-up was based on the research by Dagher et al. [3] and Saxena et al. [5]. Participants were alternately assigned to the two groups. All participants were first-year university students in the programme Computer Science and Engineering at Delft University of Technology, who had little or no prior experience with machine learning, ensuring that the concepts presented were unfamiliar. They were approached by the author during one of the university-organised lab-sessions, where students can collaborate on homework and ask teaching assistants questions. They were

The study was conducted online using a structured survey form for each group, hosted through university-hosted Microsoft Forms to ensure data security. The form consisted of three parts, one for each concept (clustering, K-means clustering, and complete linkage). For each concept, a learning goal was determined by the author on the level of 'Understanding' in Bloom's taxonomy [18]:

[5]https://k-alpha.org

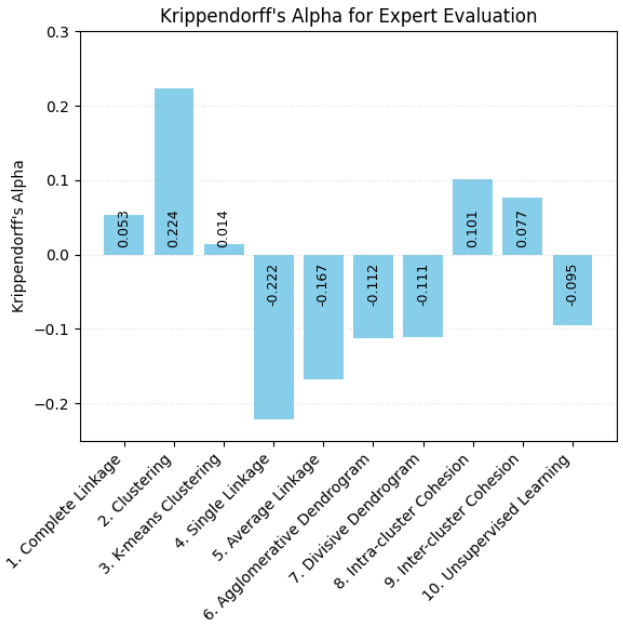

Figure 3: The Krippendorf's Alpha value for the expert evaluation of the different analogies. The values are on a range of [-1, 1].

- **Clustering**: Students understand how clustering groups similar data points based on features without prior labels.
- **K-means clustering**: Students understand how the K-means algorithm assigns data to K clusters based on proximity to cluster centers.
- **Complete linkage**: Students understand how the distance between clusters is defined using the maximum distance between their points.

For each learning goal, a pre-test and post-test was designed by the author with two questions with three multiple choice options and a fourth for 'I don't know' each. Group A's survey included analogy-enriched explanations for each concept, whereas Group B's survey presented equivalent explanations covering the same content without the use of analogies. This design isolates the effect of the analogical framing while keeping other instructional content consistent. The length and structure of the explanations were kept similar between groups, differing primarily in the presence or absence of the analogy. The complete surveys can be found in Appendix C.

The process (pre-test - explanation - post-test) was repeated for the three concepts in sequence for each participant. Finally, after completing all three concept sections, participants were presented with a non-cognitive survey to capture their perceptions of the learning experience, the Reduced Instructional Materials Motivation Survey. This survey, designed by Loorbach et al. [19], includes 12 Likert-scale statements covering attention, relevance, confidence, and satisfaction, according to the ARCS model by Keller [20]. Participants rated each statement on a 5-point scale from "Not True" to "Very True" All participants in both groups responded to the same set of statements.

5.2 Results

In total, 38 students participated in the study. 18 joined group A and 20 joined group B.

Cognitive Survey Results

Figure 5 summarizes the mean pre-test and post-test scores (proportion of correct answers) for each concept by group. Both the analogy group and the control group showed improvement from pre-test to post-test on all three concepts.

The knowledge gain is defined as in Equation 1.

$$\text{knowledge gain} = \text{score}_{\text{post}} - \text{score}_{\text{pre}} \tag{1}$$

For clustering, the mean gain was +0.50 for Group A and +0.38 for Group B. On K-means clustering, the gain was +0.53 for Group A and +0.63 for Group B. On complete linkage, Group A achieved a gain of +0.72, compared to +0.48 for Group B. These values indicate that the analogy group (A) gained more on the clustering and complete linkage concepts, whereas the control group (B) gained slightly more on the K-means concept. It should be noted that due to the high post-test scores and relatively small sample size, the differences in final scores between groups were modest for clustering and K-means.

Non-Cognitive Survey Results

Figure 4 summarises the results from the non-cognitive survey. The responses showed differences in how the two groups perceived the learning experience.

The RIMMS statements can be grouped into four groups according to the ARCS model [19,20]: Attention (Statements 1-3), Relevance (Statements 4-6), Confidence (Statements 7-9), and Satisfaction (Statements 10-12).

For the statements on attention, the control group (B) scored marginally higher. There is no clear patterns for the statements on relevance: statement 4 is marked considerably higher by Group A, while Statement 5 is marked higher by Group B. Group A did score consistently higher on all statements in the Confidence and Satisfaction groups.

The results from the cognitive and non-cognitive evaluation have been tested for statistical significance with the tool provided by Sapio Research[6]. None of the results were found to be significant.

6 Responsible Research

It is important during this research to adhere to principles of ethical and responsible research, especially given that the work involves human participants and pedagogical interventions.

6.1 Ethical Approval and Consent

Prior to executing the research, the complete application to the Human Research Ethics Committee at Delft University of Technology was filled out and contained no unmitigated risks and a sensible data management plan.

[6] https://sapioresearch.com/significant-difference-calculator

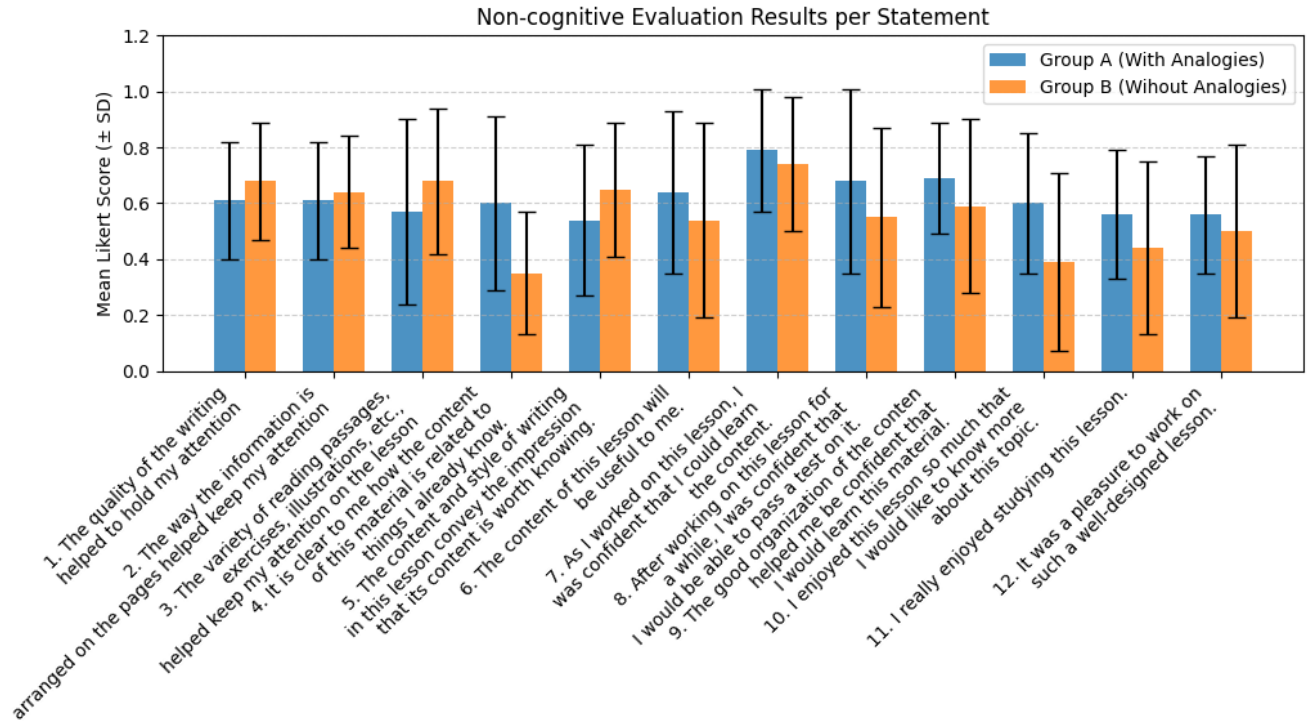

Figure 4: Non-coginitive student evaluation results, with the mean of the results of a 5-point Likert scale (0 = Not True, 1 = Very True).

All expert and student participation is entirely voluntary. Informed consent was obtained from each expert who participated in the analogy review. They consented to have their responses included in our analysis, under the condition of anonymity, by participation. Similarly, for the student survey, informed consent will be obtained from all student participants. They will indicate consent by proceeding. Students can freely choose to participate or not. It is also made clear that they can skip any question or withdraw at any point without penalty.

It is crucial to protect participant identity and data confidentiality. The expert feedback and student test results were collected anonymously: experts submitted their reviews via a Microsoft Form that did not collect names or identifiable information. In any publications or reports, only aggregated results are used.

6.2 Reproducibility and Transparency

To support reproducibility, the process and materials are described in this paper. The set of analogies developed and the prompt for the LLM that generated them, along with the expert review questionnaire and the student survey, are included in the appendix of this paper for transparency. By sharing these resources, open science practices are followed, enabling follow-up studies.

6.3 Addressing Bias and Validity

It is could be that that, as researchers, there is a bias towards the effectiveness of the methodologies. To mitigate this, the independent expert reviewers was incorporated to critically evaluate the analogies, providing an external check. The student survey is designed to capture authentic student reactions, whether positive or negative. Additionally, consider the limitations of the sample: the expert panel, while knowledgeable, is small in number; their views might not represent all possible instructors. Similarly, the student sample is from one institution and programme.

6.4 Generative AI Usage

To adhere strictly to all ethical concerns regarding the use of generative AI as ones own work, the prompt for the LLM used in this research is provided in Appendix B. Generative AI was not used literally for the writing of this paper, all the writing was done by the author himself.

7 Discussion

7.1 Expert Review

Across all analogies, certain patterns emerge in the data. First, there appears to be a positive correlation between target-concept coverage and mapping strength. Analogies that covered the concept more completely tended also to have stronger mappings: experts gave high coverage and high mapping scores together for the top-tier analogies, whereas analogies with limited coverage generally received only middling mapping scores as well. This pattern suggests that when an analogy thoroughly addresses a concept, it likely does so by establishing many clear, strong correspondences. This finding is consistent with the literature on analogical reasoning, which emphasizes that an analogy's pedagogical power

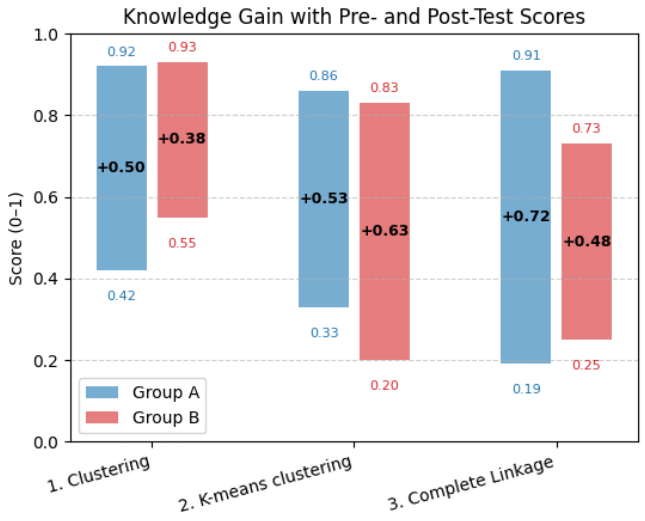

Figure 5: Student evaluation results, with the mean of the pre-test and post-test (between 0 and 1) and the knowledge gain (the difference between the pre-test and post-test means).

lies in the overlap of relational structure rather than merely in shared surface features [3].

Second, the role of metaphoricity in expert ratings seems to be more nuanced. While all of the top-tier analogies were indeed highly metaphorical, metaphoricity by itself did not elevate an analogy into the top tier unless it was accompanied by solid coverage and mapping. In fact, the data indicate that the most effective analogies struck a balance: they were sufficiently creative to engage (high metaphoricity) and systematically aligned with the concept's elements (high mapping), without sacrificing completeness of content. Finally, differences in rating dispersion highlight the consistency of expert judgments. The top four analogies not only have higher mean scores, but also lower standard deviations, implying strong agreement among experts that these analogies are effective.

The expert panel's insights provide a valuable guide for educators and curriculum designers: invest effort in developing analogies that maximize concept coverage and mapping strength.

7.2 Student evaluation

The results of the student survey provide signals that incorporating analogies into machine learning explanations can enhance student learning outcomes and engagement. In the study, participants who learned with the help of analogical explanations (Group A) showed equal or greater improvements in understanding the concepts compared to those who received standard explanations without analogies (Group B). This finding aligns with the theory that analogies are particularly helpful for grasping complex or abstract topics by mapping them onto more familiar experiences [14] and is in line with the results from previous research [3, 5].

It is worth discussing the one concept (K-means clustering) where the control group saw a slightly larger gain than the analogy group. In this case, both groups achieved high post-test scores (over 83% correctness on average), but Group B's improvement from a lower baseline was greater. The analogy used for K-means might not have provided a significant additional boost over the plain description, or it may have introduced extra information that some students found tangential.

Beyond test scores, an important outcome of this study is the impact of analogies on student motivation and engagement. Group A participants reported higher enjoyment and interest in the lesson compared to Group B. They were more likely to agree that they enjoyed the lesson and wanted to learn more about the topic. This suggests that the use of relatable analogies made the learning experience more engaging or satisfying for students. This result aligns with motivational theory in instructional design: analogies and examples that connect academic content to a student's real-life experiences can increase the perceived relevance of the material and spark interest [4].

Limitations

While the results are encouraging, consider several limitations of this study. First, the sample size (N=38) was small, and participants were not a fully random sample of all learners (they were volunteers from a university subject pool). Thus, the findings should be interpreted with caution and may not generalize to all populations.

Second, the learning outcomes were measured immediately after the lesson; we did not assess long-term retention of the concepts. It would be valuable to see if the analogical learning advantage persists over time.

Third, our design focused on three specific concepts in unsupervised learning. Different domains or more complex tasks might yield different results; analogies that work for introductory concepts might need adaptation for more advanced topics.

Fourth, none of the results were statistically significant, which does not mean the analogies were not useful; their usefulness just cannot be concluded conclusively from this study.

Despite these limitations, our controlled comparison provides suggestions supporting the efficacy of analogies in teaching machine learning, complementing the expert insights from our earlier evaluation.

8 Conclusions and Future Work

In this work, the use of real-world analogies as a pedagogical strategy for teaching foundational machine learning concepts to novices was explored. It began by conducting an expert review to review LLM-generated analogical explanations for key concepts. Following the expert feedback, a student study was executed to empirically test the impact of these analogies on learning outcomes and student motivation. The results indicate that knowledge gain might increase when learning unsupervised ML through analogies, thereby answering the research question of what the effect of analogies is on knowledge gain for unsupervised ML education.

Moreover, the analogy group students reported greater enjoyment, interest, and perceived relevance of the lesson, suggesting that analogies also contribute positively to the learning experience beyond just factual understanding. These findings align with prior research that highlights the benefits of analogical teaching in science and technology education.

The full list of analogies and evaluation results from, among others, this research, can be found at our site: https://ml-teaching-analogies.github.io/.

In conclusion, incorporating analogies into machine learning instruction appears to be a promising approach to improve both cognitive and affective outcomes for beginners. Educators can leverage analogies to reduce the initial intimidation of abstract concepts, thereby lowering cognitive barriers and increasing student motivation. Future work could continue to develop a library of vetted analogies for a wider range of machine learning topics, and to investigate their impact in larger and more diverse learner populations. Additionally, examining long-term retention and understanding potential pitfalls (such as analogical misconceptions) will be important for creating best practices for analogy-based teaching. Ultimately, we hope that our findings encourage instructors to thoughtfully integrate analogies in their curricula, as a means to make machine learning and other complex STEM subjects more accessible and engaging for all learners.

References

[1] R. Benjamin Shapiro and Rebecca Fiebrink. Introduction to the special section: Launching an agenda for research on learning machine learning. *ACM Transactions on Computing Education*, 19(4):1–6, October 2019.

[2] Orit Hazzan and Koby Mike. *The Pedagogical Challenge of Machine Learning Education*, page 199–208. Springer International Publishing, 2023.

[3] Zoubeida R. Dagher. Review of studies on the effectiveness of instructional analogies in science education. *Science Education*, 79(3):295–312, June 1995.

[4] Vishnu S. Pendyala. *Relating Machine Learning to the Real-World: Analogies to Enhance Learning Comprehension*, page 127–139. Springer International Publishing, 2022.

[5] Pawan Saxena, Sanjay Kumar Singh, and Gopal Gupta. Achieving effective learning outcomes through the use of analogies in teaching computer science. *Mathematics*, 11(15):3340, July 2023.

[6] Christina M. Eckhardt, Sophia J. Madjarova, Riley J. Williams, Mattheu Ollivier, Jón Karlsson, Ayoosh Pareek, and Benedict U. Nwachukwu. Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31(2):376–381, November 2022.

[7] Christian Lopez, Scott Tucker, Tarik Salameh, and Conrad Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. *Journal of Biomedical Informatics*, 85:30–39, September 2018.

[8] Jing Wang and Filip Biljecki. Unsupervised machine learning in urban studies: A systematic review of applications. *Cities*, 129:103925, October 2022.

[9] Daniel Hoang and Kevin Wiegratz. Machine learning methods in finance: Recent applications and prospects. *European Financial Management*, 29(5):1657–1701, February 2023.

[10] James Ming Chen and Charalampos Agiropoulos. Hints of earlier and other creation: Unsupervised machine learning in financial time-series analysis. In *ITISE 2023*, ITISE 2023, page 42. MDPI, July 2023.

[11] Peter Chew. Unsupervised-learning financial reconciliation: a robust, accurate approach inspired by machine translation. In *Proceedings of the First ACM International Conference on AI in Finance*, ICAIF '20, page 1–12. ACM, October 2020.

[12] Mohammed Tuays Almuqati, Fatimah Sidi, Siti Nurulain Mohd Rum, Maslina Zolkepli, and Iskandar Ishak. Challenges in supervised and unsupervised learning: A comprehensive overview. *International Journal on Advanced Science, Engineering and Information Technology*, 14(4):1449–1455, August 2024.

[13] Bhavya Bhavya, Chris Palaguachi, Yang Zhou, Suma Bhat, and ChengXiang Zhai. Long-form analogy evaluation challenge. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, page 1–16. Association for Computational Linguistics, 2024.

[14] Lindsey E. Richland and Janice Hansen. Reducing cognitive load in learning by analogy. *International Journal of Psychological Studies*, 5(4), November 2013.

[15] Zekai Shao, Siyu Yuan, Lin Gao, Yixuan He, Deqing Yang, and Siming Chen. Unlocking scientific concepts: How effective are llm-generated analogies for student understanding and classroom practice? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–19. ACM, April 2025.

[16] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):89–101, October 2022.

[17] Giacomo Marzi, Marco Balzano, and Davide Marchiori. K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545, June 2024.

[18] David R. Krathwohl. A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218, November 2002.

[19] Nicole Loorbach, Oscar Peters, Joyce Karreman, and Michaël Steehouder. Validation of the instructional materials motivation survey (¡scp¿imms¡/scp¿) in a self-directed instructional setting aimed at working with technology. *British Journal of Educational Technology*, 46(1):204–218, February 2014.

[20] John M. Keller. Development and use of the arcs model of instructional design. *Journal of Instructional Development*, 10(3):2–10, September 1987.

A Analogies

A.1 Unsupervised Learning

Definition: Unsupervised learning is a type of machine learning where the algorithm is given data without any labeled outcomes and tries to find patterns or structure in it.

Analogy: Imagine entering a party where you don't know anyone, and no one is wearing name tags. You observe how people talk, dress, and behave to figure out who might be friends, who might be colleagues, and who seems to be alone.

Mapping:

- Data = People at the party
- No labels = No name tags or prior info
- Finding structure = Noticing who talks to whom or has similar behavior
- Algorithm = You, the observer
- Outcome = Informal groupings based on observed patterns

A.2 Clustering

Definition: Clustering is the process of grouping data points so that those in the same group (cluster) are more similar to each other than to those in other groups.

Analogy: Sorting your socks after doing laundry: you don't have labels, but you pair socks based on their color, size, and material to form neat clusters of matching socks.

Mapping:

- Data points = Individual socks
- Similarity = Matching in color/size/material
- Clusters = Piles of similar socks
- Clustering process = You sorting socks based on similarities

A.3 Intra-cluster Cohesion

Definition: Intra-cluster cohesion refers to how similar or close the members of a single cluster are to each other.

Analogy: Think of a tight-knit group of friends sitting together at lunch. They talk easily, share inside jokes, and understand each other well.

Mapping:

- Cluster = Group of friends
- Data points = Individual friends
- Cohesion = How well the friends interact and bond
- High cohesion = Everyone talks closely and shares interests
- Low cohesion = Some friends feel left out or don't get along well

A.4 Inter-cluster Cohesion (Distance)

Definition: Inter-cluster cohesion (or separation) refers to how different or far apart different clusters are from each other.

Analogy: Different school clubs meeting in separate classrooms — the drama club, the chess club, and the soccer team. Each club stays in its room and rarely mixes with others.

Mapping:

- Clusters = Different clubs
- Data points = Students in each club
- Inter-cluster distance = Physical and interest-based distance between clubs
- High separation = Clubs in different buildings with very different interests
- Low separation = Clubs that meet nearby and share members

A.5 K-means Clustering

Definition: K-means is a clustering algorithm that divides data into a fixed number (K) of clusters by minimizing the distance between data points and the center (mean) of their assigned cluster.

Analogy: Imagine assigning students to study groups based on proximity to a few fixed tables in a library. Each student joins the table they are closest to, and the table becomes the "center" of that group.

Mapping:

- Data points = Students
- Cluster centers = Tables
- Clusters = Groups around each table
- Distance = How far each student is from each table
- K = Number of tables (clusters)
- Mean = Average location of students at the table

A.6 Hierarchical Clustering

Definition: Hierarchical clustering builds a tree of clusters, where data points are merged or split based on similarity to form a hierarchy of groupings.

Analogy: Organizing your family tree: start with individuals, then group siblings, then families, then extended families, building up to an ancestry chart.

Mapping:

- Data points = Individual people
- Clusters = Family groupings
- Tree = Family tree structure
- Merging/splitting = Grouping relatives together or separating family lines
- Hierarchy = Levels from individual to extended family

A.7 Single Linkage

Definition: Single linkage is a method in hierarchical clustering where the distance between two clusters is defined as the shortest distance between any two points in each cluster.

Analogy: Imagine connecting two cities by roads — single linkage means choosing the shortest possible road connecting any two houses from each city.

Mapping:

- Clusters = Cities
- Data points = Houses
- Linkage distance = Length of the shortest road between any house in one city to any house in another
- Result = Clusters merge when even a single short connection exists

A.8 Complete Linkage

Definition: Complete linkage defines the distance between two clusters as the longest distance between any two points in each cluster.

Analogy: If two cities want to ensure their farthest apart homes are still within reach, complete linkage measures the distance between the two farthest houses before connecting the cities.

Mapping:

- Clusters = Cities
- Data points = Houses
- Linkage distance = Maximum distance between any two houses across the two cities
- Result = Clusters only merge when all elements are close enough

A.9 Average Linkage

Definition: Average linkage computes the distance between two clusters as the average distance between all pairs of points in each cluster.

Analogy: Think of measuring the average travel time between every person in one city visiting every person in another city.

Mapping:

- Clusters = Cities
- Data points = People
- Linkage distance = Average travel time between all cross-city pairs
- Result = Clusters merge based on overall similarity, not extremes

A.10 Agglomerative Dendrogram

Definition: An agglomerative dendrogram is a tree diagram representing the process of hierarchical clustering from the bottom up, starting with individual points and merging them into clusters.

Analogy: Imagine building a LEGO tower from many small pieces. You first snap two blocks together, then combine them with others, until you have one big structure. The diagram of how you combined the pieces is the dendrogram.

Mapping:

- Data points = LEGO blocks
- Clusters = Groups of blocks
- Merging = Snapping together pieces
- Dendrogram = Blueprint showing the building steps from pieces to whole
- Bottom-up = Start from small pieces, build up to full model

A.11 Divisive Dendrogram

Definition: A divisive dendrogram represents a hierarchical clustering process that starts with all data in one cluster and splits it into smaller clusters recursively.

Analogy: Think of cutting a cake: you start with the whole cake and make cuts to divide it into smaller slices, then cut those slices again to make even smaller pieces.

Mapping:

- Full data = Whole cake
- Clusters = Slices or pieces
- Splitting = Cutting the cake
- Dendrogram = Diagram showing where and when each cut was made
- Top-down = Start from the whole, divide into parts

B Prompt

The used prompt, as described in Section 3.

Background and motivation

”We need to lear how to teach Machine Learning” - this quote of Amy J. Ko from her insightful blog post [1] has been circulating in the Machine Learning education community sinds 2017. There is a lot of empirical knowledge on how to structure and effectively teach Machine Learning, but very few structured approaches are proposed, evaluated, and compared, for both CS and nonCS majors. But why do we care?

Machine learning is being used extensively in industry and academia. If we want Machine Learning experts to be good at what they do, the effective and efficient teaching of core concepts in this field becomes of high importance. Many graduate curricula of STEM education include a large number of ML courses, thereby preparing the ML engineers for professional work. The growth of the number of ML courses is also visible in the undergraduate programs, aiming on one hand to prepare students for applying ML in the industrial context and on the other hand to prepare students for the master programmes.

In this project we would like to investigate what is the influence on students learning when using metaphors and analogies in Machine Learning teaching.

This is a well-established approach to make abstract topics more digestible. It can help learners relate new information to familiar concepts, making it easier to develop a mental model of how a system works. For example, in introductory programming, a variable is often explained as a labeled ”box” that stores data, and recursive functions as a ”set of nesting dolls”. These metaphors simplify abstract ideas by linking them to real-world objects and experiences, enabling learners to understand the logic behind them before consolidating with the technical details. Previous works collected and studied popular analogies for programming education: Notional Machines [4]. The use of metaphors and analogies in Machine Learning education remains under-explored. While some instructors may incorporate informal metaphors to describe ML concepts (e.g., explaining neural networks as ”brain-inspired systems” or gradient descent as ”rolling down a hill”), there is no structured repository of metaphors and analogies specifically designed to enhance ML instruction. Our project aims to build a collection of the evaluated (useful for students and teachers) metaphors/analogies for Machine Learning concepts.

Research Questions for the Sub-Projects

1. Can concepts in Machine Learning be taught effectively using metaphors?
2. For which concepts can the meaningful analogies be generated?
3. Which analogies are useful in Machine Learning education?

We would like all the students to answer both sub-research questions for a specific list of concepts in Machine Learning that we will divide at the beginning of the project (e.g. PCA, generative vs discriminative models, ML pipeline, clustering, etc.).

References

[1] Amy J. Ko: “We need to learn how to teach Machine Learning”. https://medium.com/bits-and-behavior/weneedto-learn-how-to-teach-machine-learning-acc78bac3ff8

[2] Shapiro, R. B., and Fiebrink, R. (2019). Introduction to the special section: Launching an agenda for research on learning machine learning. ACM Transactions on Computing Education (TOCE), 19(4), 1-6.

[3] Hazzan, O., and Mike, K. (2023). The Pedagogical Challenge of Machine Learning Education. In Guide to Teaching Data Science: An Interdisciplinary Approach (pp. 199-208). Cham: Springer International Publishing.

[4] Fincher, S. et. al. 2020. Notional Machines in Computing Education: The Education of Attention. In Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education (ITiCSE-WGR '20).

Consider this list of concepts, think of a GOOD analogy that helps students with very limited knowledge and experience in the machine learning field to understater and grasp the concepts. Make sure the analogies are not co-dependent, that is, reading one explanation with an analogy should have you understand that concept without having to read others. Write a paragraph for each concept where you explain it with the use of the generated analogy.

- Unsupervised learning
- Clustering
- Intra-cluster cohesion
- Inter-cluster cohesion
- K-means clustering
- Hierarchical clustering
- Single linkage
- Complete linkage
- Average linkage
- Agglomerative dendrogram
- Divisive dendrogram

For each concept construct the following things:

- A definition of the concept (that is used as a basis for the analogy)
- The analogy itself, textually explained
- The mapping: where each feature of the concept is mapped to a feature in the analogy, try to make the mapping as complete as possible, where a lot of features are mapped, prefereably all features for a specific concept

C Student survey

By responding to this survey, you agree with our informed constent, you can read it here: https://we-need-to-learn-how-to-teach-ml.github.io/consent.pdf.

This short survey checks how well you understand a few machine learning concepts. You will get some questions first to assess your prior knowledge, then you will read an explanation of a topic and answer some more questions. From this, we evaluate your knowledge gain. There are three concepts, each takes just a couple of minutes. Choose "I don't know" if you're unsure, that is more effective for us than you guessing.

Clustering

Pre-test

Recall that your are not meant to be able to answer these questions. Only give an answer when you are fairly certain of your choice, otherwise, answer "I don't know".

1. What best describes the goal of clustering?
   A. To assign labels to known categories
   B. To divide data into random groups
   C. To group similar items together without knowing labels
   D. I don't know
2. In clustering, what defines the similarity between data points?
   A. Their order in the dataset
   B. Their values or features
   C. Their color on a chart
   D. I don't know

Explanation with Analogy (Group A)

Clustering is like sorting your socks after doing laundry. Imagine you've dumped all your socks (these are your data points) into a pile, but none of them have labels. You don't know which ones came as a pair — but you can still sort them. You start grouping them based on color, size, and material — these are your features, and they help you decide which socks are similar. The more alike two socks are, the more likely they go into the same pile — which represents a cluster. So:

Data points = Individual socks
Similarity = Matching in color/size/material
Clusters = Piles of similar socks
Clustering process = You manually sorting socks based on similarity

In machine learning, clustering does the same thing: it forms groups (clusters) of data points based on how similar they are, even though the data has no labels.

Explanation without Analogy (Group B)
Clustering is a machine learning technique that organizes data points into groups, or clusters, based on how similar they are in terms of certain features or values. Importantly, it does this without using pre-labeled categories. The algorithm measures similarity using mathematical distance (like Euclidean distance) and assigns points to the cluster where they fit best. The end result is a structure in the data that reflects natural groupings or patterns.

Post-test
Only give an answer when you are fairly certain of your choice, otherwise, answer "I don't know".
3. Which of the following is most similar to how clustering works?
   A. Grouping identical images by filename
   B. Grouping animals by their habitat
   C. Grouping objects with shared features but no labels
   D. I don't know
4. What is the main requirement for points to be in the same cluster?
   A. They must have unique features
   B. They must appear next to each other
   C. They must be more similar to each other than to other points
   D. I don't know

K-means Clustering

Pre-test
Recall that your are not meant to be able to answer these questions. Only give an answer when you are fairly certain of your choice, otherwise, answer "I don't know".
5. What is the purpose of the "K" in K-means clustering?
   A. It's a random variable
   B. It's the number of clusters to create
   C. It's the name of the algorithm creator
   D. I don't know
6. How does K-means decide where clusters are?
   A. By randomly picking points
   B. By averaging nearby points' positions
   C. By choosing the largest group
   D. I don't know

Explanation with Analogy (Group A)
K-means is like organizing students into study groups around tables in a library. Imagine there are a fixed number of tables — these represent your K cluster centers. Each student (your data point) walks to the nearest table — this is based on distance, like how we measure similarity in K-means. Once all students are seated, each table calculates the average position of its nearby students — this becomes the new center of the group. The tables may then shift position, and students re-evaluate which one is closest. This process repeats until no one changes tables — just like K-means stabilizing the clusters.
   Data points = Students
   Cluster centers = Tables
   Clusters = Groups around each table
   Distance = How far each student is from a table
   K = Number of tables (clusters)
   Mean = Average location of students at a table (new center)

Explanation without Analogy (Group B)
K-means clustering is an iterative algorithm that groups data into K predefined clusters. Each data point is assigned to the nearest cluster center, often measured using Euclidean distance. After assignment, the mean of the data points in each cluster is calculated, and that becomes the new cluster center. This process repeats until the cluster centers stabilize — meaning points stop changing clusters. The goal is to minimize the total distance between points and their respective centers, forming compact and well-separated groups.

Post-test
Only give an answer when you are fairly certain of your choice, otherwise, answer "I don't know".
7. In K-means clustering, how is the center of a cluster determined?
   A. By the data point closest to the edge
   B. By choosing a random member
   C. By averaging the positions of all points in that cluster

D. I don't know

8. What happens when K is increased in a K-means clustering task?
   A. You get fewer, broader clusters
   B. You get more, finer clusters
   C. It makes clustering impossible
   D. I don't know

Complete Linkage

Pre-test

Recall that your are not meant to be able to answer these questions. Only give an answer when you are fairly certain of your choice, otherwise, answer "I don't know".

9. What does "complete linkage" measure between two clusters?
   A. The shortest distance between any two points
   B. The average of all distances
   C. The longest distance between any two points in each cluster
   D. I don't know
10. Why does complete linkage often prevent merging clusters that are spread out?
   A. It only merges clusters with many points
   B. It requires the farthest points in both clusters to be close
   C. It ignores the distances between points
   D. I don't know

Explanation with Analogy (Group A)

Complete linkage is like deciding whether two cities should be connected by checking how far apart their most distant homes are. Each city is a cluster, and the houses are the data points within them. Instead of connecting the cities just because some homes are close, the rule says: only connect them if even the farthest house in City A is still close to the farthest house in City B. This ensures the two cities are close in every possible way — not just at one edge. In clustering, this means clusters are only merged when every data point in one is not too far from any point in the other — preserving tight, compact clusters.

Clusters = Cities
Data points = Houses
Linkage distance = Maximum distance between any two houses from the two cities
Result = Cities (clusters) only merge if even the farthest houses are acceptably close

Explanation without Analogy (Group B)

Complete linkage is a method used in hierarchical clustering to determine the distance between two clusters. Unlike other methods that use the nearest point or the average distance, complete linkage focuses on the maximum distance between any pair of points — one from each cluster. This approach ensures that merged clusters are tightly packed, as the largest possible internal distance remains small. It's particularly useful when you want to avoid long, stretched-out clusters.

Post-test

Only give an answer when you are fairly certain of your choice, otherwise, answer "I don't know".

11. When using complete linkage, when are two clusters considered close?
   A. When their closest points are near
   B. When all their points are near
   C. When their farthest points are still relatively close
   D. I don't know
12. What feature of complete linkage helps avoid combining very distant points?
   A. It considers the maximum distance between the two clusters
   B. It always merges the smallest clusters first
   C. It picks the cluster with the most similar average values
   D. I don't know

What did you think of the explanations?

Please rate how true each statement is for you after reading the explanations and answering the questions in this lesson. Not all statements might equally refer to your situation. Please fill in the questions to the best of your ability.

1. The quality of the writing helped to hold my attention
2. The way the information is arranged on the pages helped keep my attention
3. The variety of reading passages, exercises, illustrations, etc., helped keep my attention on the lesson

4. It is clear to me how the content of this material is related to things I already know.
5. The content and style of writing in this lesson convey the impression that its content is worth knowing.
6. The content of this lesson will be useful to me.
7. As I worked on this lesson, I was confident that I could learn the content.
8. After working on this lesson for a while, I was confident that I would be able to pass a test on it.
9. The good organization of the content helped me be confident that I would learn this material.
10. I enjoyed this lesson so much that I would like to know more about this topic.
11. I really enjoyed studying this lesson.
12. It was a pleasure to work on such a well-designed lesson.

The statements could be answered with the options: Not True, Slightly True, Moderately True, Mostly True, and Very True.