Analogies for Machine Learning Loss Functions: An Empirical Study on Understanding and Motivation

Ahmet Arif Özmen

Supervisors: Gosia Migut, Ilinca Renţea, Yuri Noviello

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Ahmet Arif Özmen
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Ilinca Renţea, Yuri Noviello, David Tax

Abstract

This study examines the effect of analogies on conceptual understanding of machine learning (ML) loss functions, and the motivation to learn in first-year bachelor computer science students. For a set of 10 ML loss functions, analogies were generated and evaluated by 15 experts. 3 of these analogies were subsequently tested with 22 students. The results show no conclusive evidence for improvement in understanding and motivation to learn. The study outlines a general strategy for evaluation of analogies on student understanding and motivation. The study further provides 10 expert-rated analogies, 3 of which have been tested with students.

1 Introduction

Machine learning (ML) has become an indispensable part of modern civilization. This fact can easily be seen in the acceleration of the amount of compute needed for ML, which has quadrupled between 2010 and 2022. After 2022 this rapid acceleration has subsided, but it has nonetheless not stopped growing [23]. The field of machine learning is also becoming ever more prevalent in many other fields besides computer science. Sarker [21] outlines many real-world applications of ML models, such as cybersecurity, healthcare, e-commerce and agriculture, to name a few.

These statistics show us that machine learning is in great demand currently. Not only in the field of computer science and programming, but across nearly every sector. Given the wide-spread application of machine learning techniques, the impact they will have on society is just as wide-spread. Therefore it is highly important that people implementing these algorithms have a good understanding of their strengths, but also the potential drawbacks these methods bring along. This strongly underlines the importance of not only machine learning itself, but by extension the education of this field for up-and-coming ML-experts as well.

ML education

Despite the importance of machine learning education, there is very little research and literature on it. In their article titled "*We need to learn how to teach machine learning*" [1], which forms an important part of the background for this thesis, the author Amy J. Ko puts a spotlight on this issue. They argue that the knowledge about teaching ML itself is still lacking. Fiebrink [8] also shares this same notion, and argues that the subject of teaching machine learning to any group of people, is an underexplored topic.

To solve this issue, Ko argues that the *pedagogical content knowledge* (PCK) in the field of machine learning has to be explored. PCK is the knowledge that is required for a person to teach that subject matter to someone else [24].

Analogies in ML education

In the exploration of the PCK of the machine learning field, analogies come forward as an important category[1]. Analogies are a way to link two different concepts together by similarity. An example for an analogy would be: "The CPU is the brain of the machine. It takes input data, processes it and produces outputs." [10]. Here two different concepts, the CPU and the brain, are compared to one another due to their resemblance in processing inputs, giving us an analogy.

Analogies are a way for humans to relate two different concepts by way of similarity. In the analogy above, a CPU is explained by using the brain as a metaphor. The two different concepts are related to each other because they have a resemblance in processing inputs and producing outputs. This can make analogies especially useful in teaching abstract concepts in computer science and ML, as the difficult to grasp concepts in these fields, can be related to more concrete and amenable topics. Besides improving understanding, analogies can also help improve students' motivation to learn [20].

Despite there existing a good amount of research on analogies for computer science education, the same cannot be said for ML education. Looking through the existing literature landscape, only one such example could be found. Pendyala [19] gives in their paper a set of analogies and examples, used in their own teaching of machine learning concepts to their students.

Research aim

The importance of machine learning, and the lack of research into its education, indicate that more work is invaluable in this field. This research contributes to the field by introducing and evaluating a number of analogies for ML loss functions. Loss functions are the mechanism by which ML algorithms are evaluated to be accurate or not, therefore it is of great importance that they are understood by students of machine learning.

The aim of this research is to answer the following question: *How does the use of analogies in explaining loss functions of machine learning algorithms affect the conceptual understanding and motivation to learn in Computer Science students?*
This goal can be subdivided into the following sub-questions:

- Which analogies can be used to enhance the explanation of these particular machine learning loss functions?
- Is the conceptual understanding in Computer Science students positively influenced by the use of these analogies?
- Is the motivation to learn in Computer Science students positively influenced by the use of these analogies?

Structure

The paper starts in Section 2 with an exploration of the research field's background. Section 3 defines what an analogy is, and how it differs from a metaphor. Section 4 outlines the methodology used for the research. Section 5 shortly describes the findings, with a more detailed discussion of these findings and their limitations in Section 6. In Section 7 the ethical considerations in designing and conducting the research are discussed. Finally, Section 8 contains the concluding remarks, with suggestions for future work.

[1] Amy J. Ko. We need to learn how to teach machine learning. 8 2017

2 Background

Research on the field of machine learning education tends to focus mostly on introducing it into K-12 schools. However, there exists a distinct lack of research for the education of machine learning to students in higher education. If instead computer science and programming are considered in general, then a more comprehensive set of research exists on the topic.

Analogies in computer science

Fincher et al. [9] look at a general definition of analogies, called a *Notional Machine* (NM). The authors define them as *"a pedagogic device to assist the understanding of some aspect of programs or programming"*. This research is particularly noteworthy, as the researchers classify analogies as a category of NMs. An example for such an NM, that uses an analogy to explain a programming concept, is the representation of arrays as a row of parking spaces. In this analogy, a set of correspondences between the programming concept and the analogy itself exists to relate the abstract concept, to a more easily interpretable concrete concept. Figure 1 shows the example, along with the correspondences between two concept domains. There is much conceptual overlap between analogies and notional machines. This paper therefore forms a strong basis for further research on analogies in the field of computer science in general, but also for this paper's focus of machine learning.

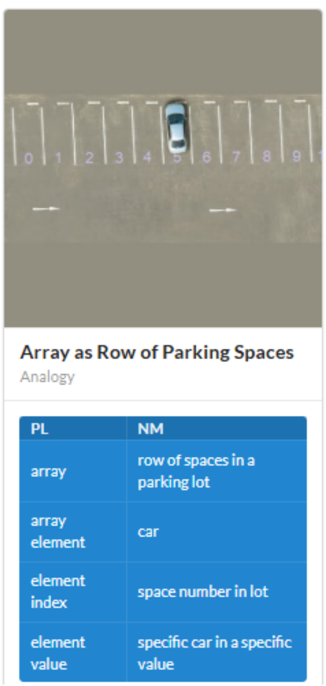

Figure 1: Notional machine example from [9]. The column "PL" (programming language) denotes the conceptual elements of an array, the column "NM" (notional machine) are the analogy's conceptual elements.

There is also research on evaluating existing analogies, or introducing new ones. These works examine examples of analogies that can be used for teaching, among other things, recursion [6], algorithms [10], the switch-statement [17], parallel computing [18], and design patterns [1]. Some other works focus more on examining the effectiveness of analogies in teaching programming and computer science concepts. Cao et al. [4] measure the value added by analogies for long-term and short-term knowledge retention and find evidence for analogies being beneficial for short-term retention, but no conclusive evidence for long-term retention. Alongside improvement in knowledge retention, there is also evidence showing that analogies help improve semantic knowledge in programming education [16]. Finally, Saxena et al. [22] perform an experimental analysis of the use of analogies in a university lecture. An A/B test was done, where both groups got a lecture on operating system scheduling, but the experimental group was taught using analogies. The research reveals a positive learning outcome for the experimental group. The results mentioned in these studies above, strongly suggest that similar positive outcomes may be achievable by employing analogies in teaching machine learning, as ML contains – similar to programming and computing – a plethora of abstract concepts.

Analogies in machine learning

The same breadth of research on analogies in ML education does not exist. A search through the literature landscape yielded only one example of a study that specifically pertains to analogies for machine learning concepts. It is a paper by Pendyala [19], who in their research give a number of analogies that can be applied to certain machine learning concepts. They also attempt to provide analogies for the concept of a loss function. The paper however doesn't contain any analogies for specific loss functions.

In their exploration of loss functions, the author doesn't provide analogies directly, but rather they give some explanations for the semantics of loss functions. In these explanations, sometimes they use an analogy and sometimes they use an example. Here it is important to first define the differences between examples and analogies. An example, refers to a pattern that should or shouldn't be imitated [2]. An analogy, on the other hand, refers to a similarity between two things that are otherwise not related to each other [3].

The author exemplifies a wrong choice of a loss function, using height and weight as incorrect measures when using squared error for classification. This is an example, as they provide a pattern that shouldn't be imitated; height and weight should not be used with squared loss. They also give an analogy for loss functions, as the difference between the sale price and the cost price of products sold by a business. This is an analogy, because ML loss functions and business profit are different concepts, but share a resemblance in this regard. This confusion as to what exactly constitutes an analogy, calls for a proper definition of the term. Section 3 provides definitions for some terms related to this study's work.

[2]https://www.merriam-webster.com/dictionary/example
[3]https://www.merriam-webster.com/dictionary/analogy

3 Definitions of terms

To properly define and understand analogies, firstly the terms *target and source concepts*, and their *mapping* has to be defined. The source concept is the concept that is already known or assumed to be known. In the analogy of "arrays are like a row of parking spaces", given in figure 1, the source concept is the row of parking spaces. This is general knowledge that is assumed to be known. The target concept, then, is the concept being taught: the array. The set of similarities between elements of these two concepts, are the *mapping* between them. For example: The array element is likened to a car, and the array element's index to the parking lot number. The Merriam Webster dictionary defines the word "analogy" as a similarity between two things that are otherwise not related to each other[4]. In other words, one or more *mapping(s)* between the *source* and *target* domains.

A metaphor on the other hand, makes a more direct comparison, where the two concepts aren't likened to one another, but equated to one another[5]. "Love is a journey" is an example for this. Love is not exactly the same thing as a journey, but they are equated in this metaphor to point out that love involves a process that one has to go through. Note that there is no explicit mapping between the source and target concepts, the similarity between them is implied.

We can turn this metaphor into a simile[6], by changing the phrasing to include "like", or "as": "Love is like a journey". Now there is an explicit mapping stated between "love" and "a journey". The individual elements of the two concepts are not explicitly explained, however.

Finally, if we add explicit mappings between the concepts' respective elements, we obtain an analogy: "Love is like a journey. Partners move through challenges and work toward goals, just as a journey involves facing obstacles, and navigating a path toward a destination."[7] Analogies therefore always contain a figure of speech, like a metaphor or a simile.

Since this study makes explicit the mappings between source and target domains, we speak of analogies. These analogies are used as a pedagogic device to enhance learning in ML, just like notional machines are used as a pedagogic device to enhance learning in programming.

4 Methodology

The research conducted in this study consisted of four phases. First the ML concepts were chosen, and the analogies for them were generated. Then the analogies were evaluated by experts and tested with students. Figure 2 shows the research pipeline. Each of these phases is described in detail in their corresponding subsections below.

4.1 ML concepts

In accordance with the main question for this study, loss functions of machine learning algorithms had to be chosen. In order to keep the concepts relevant to a university machine learning course, the curriculum of the 2024-2025 CSE2510 Machine Learning course at TU Delft was referenced. Using this, a list of 10 ML loss functions was compiled. These loss functions cover different domains within ML, like supervised and unsupervised learning, and dimensionality reduction.

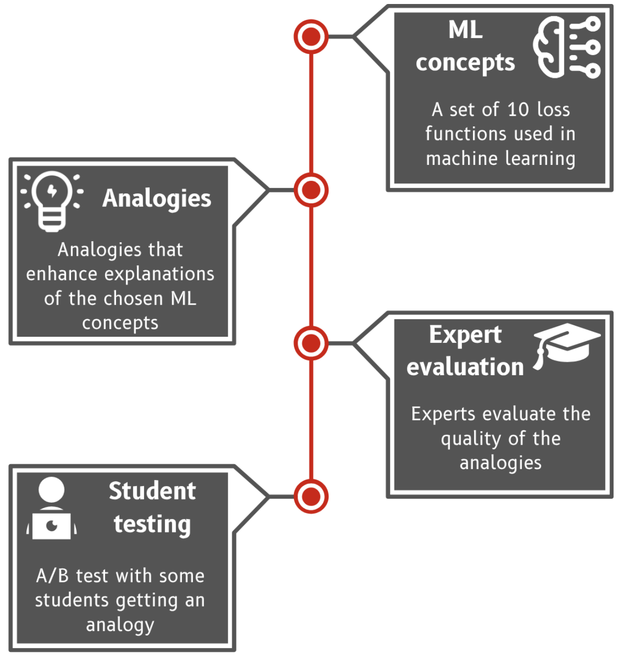

Figure 2: Visualization of the research process.

1) Misclassification error
2) Loglikelihood
3) Mean Squared Error
4) Absolute error
5) Hinge loss
6) Gini index
7) Manhattan distance
8) Cross-entropy
9) Reconstruction error
10) Kullback-Leibler divergence

4.2 Analogies

The process of finding appropriate analogies for the chosen concepts requires linking two different domains to one another, with correct mappings between the concepts. This means that knowledge of the target domain, and a broad knowledge of different source domains is necessary; not every source domain will be suitable for a useful analogy. Therefore, creativity will be required for this task of finding analogies.

In a study by Koivisto & Grassini [13], which compares creativity in humans with that of AI chatbots, the authors describe convergent thinking as specific or deep thinking, and divergent thinking as broad or creative thinking. The authors found that the AI chatbots generally outperformed humans in divergent thinking, but the high-performing humans did better than the AI chatbots. They conclude that this is explained by the large amount of low-quality ideas that humans can produce in contrast to AI chatbots.

Since finding analogies requires creative thought, it was deemed appropriate to use an AI language model to gener-

[4] https://www.merriam-webster.com/dictionary/analogy
[5] https://www.merriam-webster.com/dictionary/metaphor
[6] https://www.merriam-webster.com/dictionary/simile
[7] Yuri Noviello. Introduction to "We need to learn how to teach machine learning". *Delft University of Technology*, 4 2025

ate the analogies for this study. All of the analogies were generated using ChatGPT 4o. The same prompt template was used for each analogy. In order to make the analogies useful in a lecture setting, the model was prompted to optimize the analogies for a general audience, avoiding specific or niche topics. To get the best possible output from the model, prompting techniques as outlined by Zamfirescu-Pereira et al. [25] were used. A detailed log of the prompt template used, and the responses from the model are available in Appendix A.

Below is an example of an analogy generated for Reconstruction error:

You saw a person and you're now describing them to a sketch artist who hasn't seen them. The sketch artist draws a portrait based on your description. Once finished, you compare the sketch to the real person. The more it differs, the higher the reconstruction error

- *The real person's face → Original input*
- *Your verbal description → Encoded representation (compressed form)*
- *The sketch drawn from your description → Reconstructed output*
- *Comparing the real face with the sketch → Calculating reconstruction error*
- *A good likeness → Low reconstruction error*
- *A bad likeness → High reconstruction error*

4.3 Expert evaluation

Participants

The target audience for the expert evaluations consisted of:

- bachelor and master university students in computer science, having completed a course on machine learning,
- teaching assistants for a machine learning course,
- lecturers for a machine learning course.

Participants from different educational backgrounds will have different mental models of the concepts when rating the analogies. This way, the aim is to make sure that any bias present in one group of experts is mitigated by the inclusion of other groups. The differing levels of expertise will also help in balancing how critical ratings are, as raters with higher expertise may be too critical while raters with lower expertise may be too uncritical of the analogies presented.

Survey

For the expert evaluations of the analogies, a survey was conducted. It contained a brief explanation of the machine learning concept, then the analogy to be rated. The experts rated the quality of the analogies on a three-point balanced Likert scale (low, medium, high), in the following three categories:

1. **Target concept coverage**: How well the analogy covers the elements of the ML loss function.
2. **Mapping strength**: The logical soundness and consistency of the mapping between concepts.
3. **Metaphoricity**: Conceptual distance between the source and the target concept.

These categories were taken from research by Bhavya et al. [2], where the quality assessment of textual analogies is done using these criteria[8].

The survey was a combined questionnaire containing all the analogies being studied by the research project peer group. The reviewers were able to quit the survey at any point. By randomizing the question order it was ensured that every analogy was rated, albeit not uniformly. Appendix B contains the full expert evaluation survey.

Analysis of survey results

The analysis of the survey results is based on research by Zumrawi & Macfayden [26]. The authors propose a combination of the *Interpolated Median* (IM) and *Percent Favourable* (PF) metrics for the interpretation of ordinal survey data from a balanced Likert scale. The expert evaluation survey data, fit exactly in this category of data.

The PF is the percentage of responses that were higher than the neutral rating. It is given by the following ratio:

$$PF = \frac{N_{\text{favourable}}}{N_{\text{total}}} \cdot 100\%$$

Where $N_{\text{favourable}}$ is the amount of favourable ratings, and N_{total} is the total amount of ratings.

The IM is the dataset's median value, but adjusted by addition of a number in the range $[-0.5, 0.5]$ to better represent the distribution of the ratings. It's given by:

$$IM = \begin{cases} M + \frac{N_g - N_s}{2 \cdot N_e} & \text{if } N_e > 0 \\ M & \text{if } N_e = 0 \end{cases}$$

Where M is the median, and N_s, N_e, N_g are the ratings that are less than, equal to, or greater than M respectively.

Krippendorff's alpha [14] was used as a metric for inter-rater agreement. Bhavya et al. [2] use this alpha metric to measure inter-rater agreement as part of their automatic analogy evaluation. To calculate this alpha value, an online tool called ReCal was used [9].

4.4 Student tests

Participants

The target participants for the student tests consisted of university students majoring in computer science, with no prior education on machine learning. The aim is to measure the difference in knowledge gain between a group of students receiving an explanation with analogies, and a control group receiving no analogies. By comparing the knowledge gain between the experimental and control group, the effect that analogies have on conceptual understanding can be measured.

Survey

For the student survey, three analogies were tested. In order to measure knowledge gain, a pre- and post-test structure was used. The student received a pre-test, measuring their pre-existing level of knowledge. After the pre-test, a small explanation of the concept was given, which included

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

[9] Deen Freelon. ReCal OIR, 2013

an analogy in the experimental group. Finally, a post-test was conducted to measure the knowledge gain on the concept. Each question was multiple-choice and they were the same in both the control and experimental groups. Figure 3 shows a visual representation of the experiment process. The pre- and post-test questions all contained the option of "I don't know/understand". The students were encouraged to select this option if they didn't know the answer, ensuring that guessing on the questions did not occur. Below is an example from one of the test questions asked in the survey. The full survey is available in Appendix D.

Two machine learning models make 5 predictions:

- *Model A makes a small error of 1, consistently on every prediction.*
- *Model B makes 4 predictions flawlessly (0 error), but one prediction with a large error of 5.*

Which model will have a higher Mean Squared Error?

A *Model A will have a higher MSE*

B *Model B will have a higher MSE*

C *Both will have the same MSE*

D *I don't know/understand*

The pre and post-questions were different, but both devised from the same learning objective. This process of constructive alignment of the assessment material to learning objectives, was outlined by Biggs & Tang [3]. This way both questions were made to measure the same learning outcome.

The learning objectives were produced using Bloom's Taxonomy, which provides a hierarchy of different levels of cognition. These start at the simplest level, with remembering, then understanding, applying, analyzing, evaluating and creating. These levels are used to guide learning objectives into the correct level of cognition that needs to be measured [10].

For the questions, the second level (understanding), was chosen to determine the learning objectives, as this study aims to measure conceptual understanding. The learning objectives were devised by referencing University of Arkansas' guide on creating learning objectives [11]. Inspiration was also taken from examples in AAFP's guide on the subject[12].

Along with a cognitive test focused on understanding, a survey to measure the non-cognitive aspect of motivation was also conducted. To this end, the Reduced Instructional Materials Motivation Survey (RIMMS) was used. This survey assesses instructional materials according to the ARCS-model, which represents a student's motivation to learn using four categories; Attention, Relevance, Confidence and Satisfaction [15].

Analysis of survey results

For the analysis of knowledge gain, the same method used by Delucchi, and Jordan et al. [5, 12], in their research into

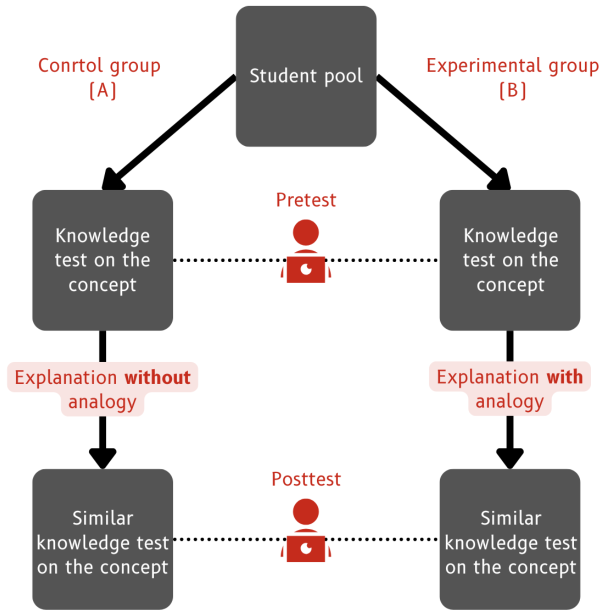

Figure 3: Visualization of the student test experiment.

student knowledge gain, was used. The difference in percentages of correct answers between the pre and post-test was used as a measure of knowledge gain.

The difference in knowledge gain between the control and experimental groups was analyzed for statistical significance using the Mann-Whitney U statistic, as a normally distributed knowledge gain cannot be expected with the questions having a correct or incorrect, i.e. binary outcome. For the RIMMS survey on the other hand, Welch's t-test was used as the answers are on a 5-point Likert scale, resulting in a normally distributed dataset.

5 Findings

5.1 Expert evaluation

16 participants filled in the expert survey. Some were rated more often than others, however. One of the participants indicated that they didn't have any ML experience. This participant was removed from the dataset, resulting in 15 samples.

Figure 4 shows the Percent Favourable (PF) and Interpolated Median (IM) values as blue bars and a red line respectively. For the IM, the values between 1-3 represent the evaluations from low to high respectively.

As visible in figure 4, the PF and IM values are congruent with one another. The rankings from highest to lowest will not change if only one metric is considered. This shows that both metrics give the same quality ranking to the analogies.

Figure 5 shows Krippendorff's Alpha for each of the analogies. This metric has a range between $[-1, 1]$, where 1 indicates total agreement, and -1 total disagreement. The metric could not be computed for the analogy on Manhattan distance because there was no variability in the ratings; every rating on

[10] https://www.ru.nl/en/staff/lecturers/designing-education/designing-courses/formulating-learning-objectives

[11] https://tips.uark.edu/using-blooms-taxonomy/

[12] https://www.aafp.org/dam/AAFP/documents/cme/faculty_development/LearningObjectivesGuidelines.pdf

the analogy was "high". With the exception of Misclassification error and Manhattan distance, each analogy had some level of disagreement between raters.

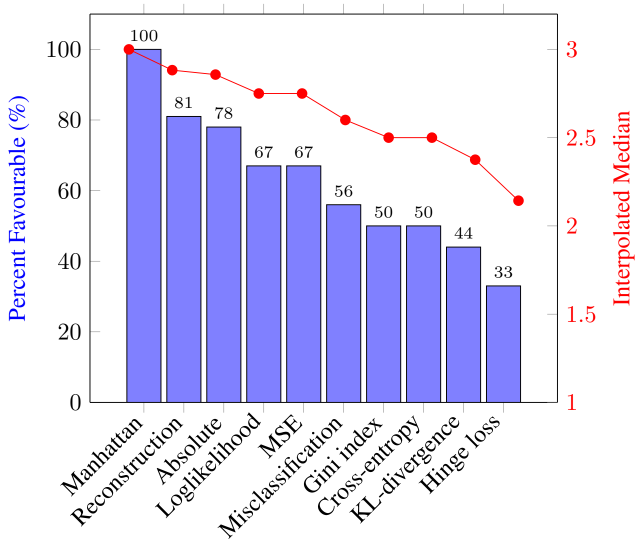

Figure 4: Bar and line chart showing evaluation metrics across each analogy. The Percent Favourable metric is plotted as blue bars on the left Y-axis. The Interpolated Median is plotted as a red line on the right Y-axis.

5.2 Student tests

For the student A/B test, 12 participants filled in the control survey and 10 participants filled in the experimental survey. All participants in the survey were first-year bachelor computer science students from the Dutch universities TU Delft and VU Amsterdam.

Figure 6 shows the knowledge gains between the pre and post-tests for each of the analogies. This metric is the difference in percentage of correct answers between the pre and post-tests. The knowledge gain is lower across the experimental group in contrast to the control group. For Manhattan distance, the knowledge gain in the control group was more than twice the knowledge gain in the experimental group.

Table 1 shows the Mann-Whitney U test results between the control and experimental groups. The critical value of 29, corresponds to $n_1 = 12$ and $n_2 = 10$ with $\alpha = 0.05$[13]. None of the analogies showed a statistically significant difference between the control and experimental groups on knowledge gain.

Figure 7 shows the results of the Reduced Instructional Materials Motivation Survey (RIMMS). On all four of the domains, the control group outperforms the experimental group. For each of the statements in the questionnaire, the participants indicated how true they believe it to be, ranging from not true to very true. These are represented in the figure as numbers from 0 to 4.

[13]Z. Bobbitt. Mann-Whitney U test. Statology, Jul. 1, 2022. Available at: https://www.statology.org/mann-whitney-u-test/

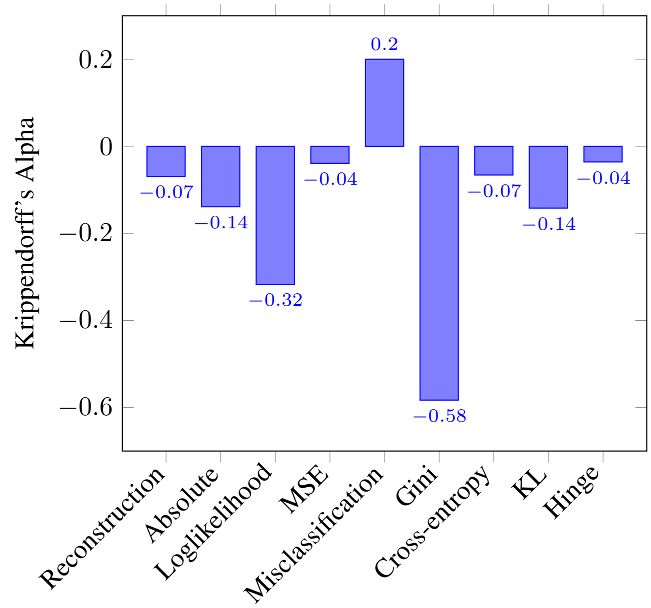

Figure 5: Bar chart showing Krippendorff's Alpha across each analogy. Manhattan distance had full agreement, therefore no alpha value could be computed.

Table 1: Mann–Whitney U test results comparing control and experimental groups knowledge gains.

Analogy	U statistic	critical value
MSE	174,6	29
Reconstruction error	174,6	29
Manhattan distance	173,9	29

Table 2 shows the Welch's t-test[14] results between the control and experimental groups. The critical value for $\alpha = 0.05$ is given by 2.00 for all categories. No statistically significant difference was found in any of the categories.

Table 2: Welch's t-test results for the RIMMS survey.

ARCS domain	t statistic	critical value
Attention	1.34	2.00
Relevance	0.78	2.00
Confidence	0.98	2.00
Satisfaction	1.19	2.00

6 Discussion

Expert evaluation

The amount of expert ratings varied per analogy, most getting three ratings, but some getting 6-7 ratings as well. The options with the highest PF and IM values, namely, Manhattan distance and Reconstruction error were chosen for student

[14]Z. Bobbitt. Welch's t-test. Statology, Dec. 20, 2020. Available at: https://www.statology.org/welchs-t-test/

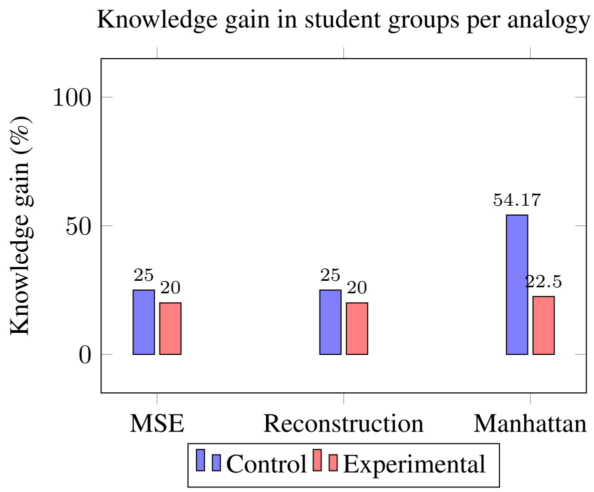

Figure 6: Bar chart showing the percent knowledge gain between pre- and post-tests across each analogy.

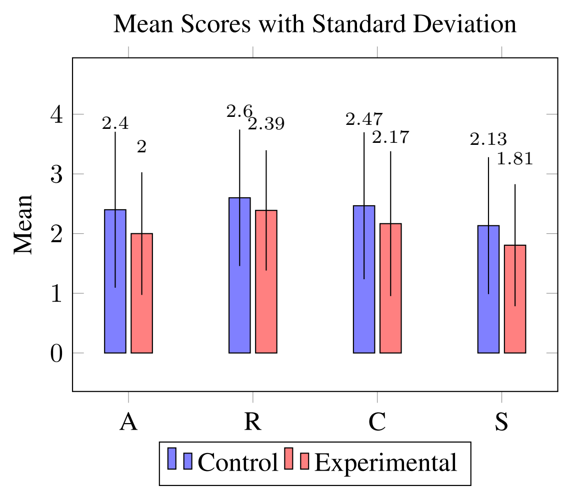

Figure 7: Mean scores and standard deviations across the categories Attention, Relevance, Confidence, and Satisfaction.

testing. Due to Absolute error's conceptual similarity to Manhattan distance, and Loglikelihood's high inter-rater disagreement, they were not used in the student test. Mean Squared Error was chosen as the third, next-best analogy. This selection procedure introduces a survivorship bias in the results for the students tests however, which should be avoided in future research to get a more balanced result.

An interesting pattern in the results, is the generally low inter-rater agreement. This is likely not due to the small sample size, as the analogies that were rated 2 or 3 times, were the ones with the most extreme, rather than the lowest, alpha values. The positive 0.2 value for Misclassification error, with 3 raters, is an example for this. Krippendorff's Alpha also accounts for small sample sizes in general, and is robust against disagreements by chance [2, 14].

A reason for the low agreement between raters, can be the subjectivity inherent to rating the quality of analogies. A study by He et al. [11] on analogical explanations similarly found low agreement between experts. The authors conclude that this is likely due to the subjective nature of expert ratings.

Another factor in the low agreement could be the differences in educational background of the experts. Narrowing down the inclusion criteria for the participants may increase inter-rater agreement. This, however, will also reduce the diversity in the participant pool, introducing a stronger bias.

Future implementations of this method for expert evaluation should take this tendency of low inter-rater agreement into consideration by collecting a sizable amount of expert ratings.

Student tests

The results of the student tests show no significant increase in knowledge gain and motivation. This could be explained by the small sample size of the tests. An example of the effects of the sample size, is the high difference in knowledge gain between control and experimental groups for Manhattan distance. In the control, 7 out of 12 students gave "I don't know / understand" as the pre-test answer, then 6 of them gave the correct answer in the post-test. This alone results in a 50% knowledge gain. In the experimental group however, 6 out of 10 students already knew the correct pre-test answer, meaning that the knowledge gain was capped at 40%. Further research into the effects of analogies on knowledge gain should be conducted in order to examine whether they improve conceptual understanding. The methodology of this study provides a good framework for future studies to be built upon.

There is a general trend of slightly higher knowledge gain in the control group, which is also mirrored by the RIMMS survey. These trends could be caused by sampling noise, as the sample sizes were quite small. Another explanation could be that the introduction of analogies in the explanations was adding extra cognitive load, reducing the capability of students to correctly answer the post-test question. Especially considering that the students voluntarily took the survey during self-study hours, this becomes a plausible reason. Finally, it's also possible that regardless of the high expert ratings on the analogies, they were not sufficiently high in quality to enhance understanding in the students. Future research taking these limitations into consideration could reveal the effects of analogies on understanding and motivation.

It is also important to mention that the setup of the student tests measures only the short-term knowledge gain in students. The pre-test, explanation, and post-test all immediately follow one another. So the results of the research cannot be generalized to long-term knowledge gain. This mirrors the results found by Cao et al. on long-term knowledge gain [4].

The results of the student tests do not confirm that the use of analogies provide an improvement in understanding of ML loss functions, nor in the motivation to learn. A more extensive study with more participants would need to be performed to assess whether analogies provide a better understanding.

7 Research ethics

Replicability and reproducibility

Research is considered replicable, when the same results can be replicated using the described research setup. While reproducibility refers to the possibility of reaching the same results using the data and methods provided in the research paper [15].

Based on these definitions, the research described in this paper is reproducible. All of the data that was available to the author, is made public in the appendix. Using the evaluation metrics described in the paper, the results can be reproduced.

Using the methodology described in this paper, this research can be replicated. Section 4 explains in detail: all of the data collection methods, the used model and prompt for generating the analogies, the survey questions and target audience for both surveys, and the data evaluation metrics.

While there is no guarantee that metric computation tools used, such as ReCal [16], will remain available in the future. It nonetheless remains possible to compute the necessary metrics manually, albeit less conveniently.

Data usage and privacy

During the research process, data from human subjects were collected. Special care was taken to preserve the subjects' privacy. All surveys were conducted anonymously, with no personally identifiable information being collected.

TU Delft requires thesis authors to submit an application to the Human Research Ethics Committee (HREC) regarding data collection procedures during the research. These HREC applications also contain a Data Management Plan (DMP)[17], which has been created and submitted for the EEMCS datasteward and responsible professor to verify. The completed HREC application was submitted by the project's responsible professor.

FAIR and open data

The data collected during this research follow the FAIR data policies [15].

Findable and Accessible data: The data of this research is both findable and accessible in the paper's appendix. The thesis paper itself is accessible to the public through TU Delft Repository, which contains student theses like this one.

Interoperable and Reusable data: The data has no interoperability risks. It consists of numeric data represented in tables, which can trivially be reused as needed.

Bias

Analogies are a way to scaffold understanding for students, and connect a familiar source domain to an unfamiliar target domain. Since familiarity in domains will vary between people, the effectiveness of the analogy can also vary. Bias is therefore an important factor to consider when choosing an analogy for teaching.

The analogies tested in this study were generated using ChatGPT 4o, which has a bias towards Western views [18]. This has to be taken into account when using the output as teaching material. The analogy for cross-entropy loss makes use of the topic of horse betting for example. A student with a Western cultural background may resonate more with this analogy, as sports betting regulations are generally more relaxed in Western countries [7].

Knowledge-based biases can also exist in analogies. In generating analogies for this research, the model was prompted to avoid niche topics. The generated analogies can be reasonably expected to be understood by any university student. However, some students may still acquire a deeper understanding of the concept being taught, due to familiarity with the analogy.

To prevent unfair advantages for students, it is important to take these biases into consideration. A fully unbiased analogy may not be achievable, but taking these biases into account is important to generate analogies that are useful to as many students as possible, if not every student.

Artifical Intelligence

Artificial Intelligence (AI) tools, specifically Large Language Models have been used during the research process for the following:

- Producing the analogies in the second phase of the study. A full log is available in Appendix A.
- LaTeX formatting of the paper, using ChatGPT 4o.
- Performing literature searches, using Perplexity AI's "Academic" feature.

No AI-tools have been used for writing or reviewing any content of this paper. It is entirely written by the author himself.

8 Conclusion and Future Work

In this study, a four-phase approach was used to produce and measure the quality of analogies that can be used to enhance student understanding of machine learning loss functions. The aim of the study was to quantitatively analyze these analogies and assess whether they improved students' conceptual understanding and motivation to learn. The research question could not be definitively answered, as the results show no conclusive evidence for improvement or decline in student understanding and motivation to learn.

Analogies can be a useful pedagogic tool to assist teachers in turning the abstract into something more concrete for their students. While the results in this study do not confirm a positive effect on understanding and motivation in teaching ML loss functions, there is nonetheless still a strong indication for the possibility of a positive result, as has been shown in other fields besides machine learning.

With the growing importance of machine learning in every field, so does the importance of its education grow with it. Little research exists on the topic of analogies in machine learning education, and this study provides a model framework which can be replicated or adapted by future studies on this topic. Besides this, it also provides 10 expert-rated analogies, and 3 student-tested analogies for some common ML loss functions.

As future research we recommend studies on the effects of analogies on long-term knowledge retention of ML loss functions, as these concepts form the basis upon which more in-depth ML knowledge will be built.

[15] Andrea Gammon. Responsible Research for EWI research projects. 4 2025

[16] Deen Freelon. ReCal OIR, 2013

[17] https://www.tudelft.nl/en/about-tu-delft/strategy/integrity-policy/human-research-ethics/

[18] https://help.openai.com/en/articles/8313359-is-chatgpt-biased

References

[1] Joseph Bergin. Teaching polymorphism with elementary design patterns. In *Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, pages 167–169, New York, NY, USA, 10 2003. ACM. doi:10.1145/949344.949393.

[2] Bhavya Bhavya, Chris Palaguachi, Yang Zhou, Suma Bhat, and Chengxiang Zhai. Long-Form Analogy Evaluation Challenge. Technical report, Tokyo, Japan, 9 2024.

[3] John Biggs and Catherine Tang. Constructive Alignment: An Outcomes-Based Approach to Teaching Anatomy. In *Teaching Anatomy*, pages 23–30. Springer International Publishing, Cham, 2020. doi:10.1007/978-3-030-43283-6{_}3.

[4] Yingjun Cao, Leo Porter, and Daniel Zingaro. Examining the Value of Analogies in Introductory Computing. In *Proceedings of the 2016 ACM Conference on International Computing Education Research*, pages 231–239, New York, NY, USA, 8 2016. ACM. doi:10.1145/2960310.2960313.

[5] Michael Delucchi. Measuring Student Learning in Social Statistics. *Teaching Sociology*, 42(3):231–239, 7 2014. doi:10.1177/0092055X14527909.

[6] Jefferey Edgington. Teaching and viewing recursion as delegation. *Journal of Computing Sciences in Colleges*, 23(1):241–246, 10 2007.

[7] Repairer Etuk, Tiange Xu, Brett Abarbanel, Marc N Potenza, and Shane W Kraus. Sports betting around the world: A systematic review. *Journal of behavioral addictions*, 11(3):689–715, 9 2022. doi:10.1556/2006.2022.00064.

[8] Rebecca Fiebrink. Machine Learning Education for Artists, Musicians, and Other Creative Practitioners. *ACM Transactions on Computing Education*, 19(4):1–32, 12 2019. doi:10.1145/3294008.

[9] Sally Fincher, Johan Jeuring, Craig S. Miller, Peter Donaldson, Benedict Du Boulay, Matthias Hauswirth, Arto Hellas, Felienne Hermans, Colleen Lewis, Andreas Mühling, Janice L. Pearce, and Andrew Petersen. Notional Machines in Computing Education: The Education of Attention. In *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, pages 21–50. Association for Computing Machinery, 6 2020. doi:10.1145/3437800.3439202.

[10] Michal Forišek and Monika Steinová. Metaphors and analogies for teaching algorithms. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education*, pages 15–20, New York, NY, USA, 2 2012. ACM. doi:10.1145/2157136.2157147.

[11] 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, 10 2022. doi:10.1609/hcomp.v10i1.21990.

[12] Rebecca C. Jordan, Steven A. Gray, David V. Howe, Wesley R. Brooks, and Joan G. Ehrenfeld. Knowledge Gain and Behavioral Change in Citizen-Science Programs. *Conservation Biology*, 25(6):1148–1154, 12 2011. doi:10.1111/j.1523-1739.2011.01745.x.

[13] Mika Koivisto and Simone Grassini. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1):13601, 9 2023. doi:10.1038/s41598-023-40858-3.

[14] Klaus Krippendorff. Computing Krippendorff's Alpha Reliability, 1 2011. URL: https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf.

[15] 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, 1 2015. doi:10.1111/bjet.12138.

[16] Katherine N. Macfarlane and Barbee T. Mynatt. A study of an advance organizer as a technique for teaching computer programming concepts. In *Proceedings of the nineteenth SIGCSE technical symposium on Computer science education - SIGCSE '88*, pages 240–243, New York, New York, USA, 1988. ACM Press. doi:10.1145/52964.53024.

[17] John Mason. Teaching by analogy. In *Working group reports from ITiCSE on Innovation and technology in computer science education*, pages 105–107, New York, NY, USA, 6 2004. ACM. doi:10.1145/1044550.1041670.

[18] Henry Neeman, Lloyd Lee, Julia Mullen, and Gerard Newman. Analogies for teaching parallel computing to inexperienced programmers. *ACM SIGCSE Bulletin*, 38(4):64–67, 12 2006. doi:10.1145/1189136.1189172.

[19] Vishnu S. Pendyala. Relating Machine Learning to the Real-World: Analogies to Enhance Learning Comprehension. volume 1572, pages 127–139. 5 2022. doi:10.1007/978-3-031-05767-0{_}11.

[20] Fugui Qi, Yizhu Chen, Zhao Li, Tao Lei, Jianqi Wang, and Guohua Lu. Exploring the Use of Associative Analogy Teaching in BME Computer Courses in Medical University. *Advances in Medical Education and Practice*, Volume 16:483–492, 4 2025. doi:10.2147/AMEP.S436882.

[21] Iqbal H. Sarker. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3):160, 5 2021. doi:10.1007/s42979-021-00592-x.

[22] 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, 7 2023. doi:10.3390/math11153340.

[23] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute Trends Across Three Eras of Machine Learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 7 2022. doi:10.1109/IJCNN55064.2022.9891914.

[24] Lee S. Shulman. Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2):4, 2 1986. doi:10.2307/1175860.

[25] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, New York, NY, USA, 4 2023. ACM. doi:10.1145/3544548.3581388.

[26] Abdel Azim Zumrawi and Leah P. Macfadyen. Proposed metrics for summarizing student evaluation of teaching data from balanced Likert scale surveys. *Cogent Education*, 10(2), 12 2023. doi:10.1080/2331186X.2023.2254665.

A Large Language Model prompt log

Below is a log of the prompt used, and the responses from the AI language model. OpenAI's ChatGPT 4o was used.

The prompt

I have a class of Computer Science university students that I'm teaching Machine Learning to. I want to make some machine learning concepts easier to understand by making use of analogies.

- Please provide me with an analogy for the loss function: <Name of loss function>.
- Make sure the analogy is general enough to be understood by most.
- Try to make the mapping between source and target concept as complete as possible.

See the example below:
Example analogy for: Gradient descent
Imagine you're on a foggy mountain. Your car is at the base of the mountain. What is a good strategy to get back to your car? You find the steepest decline and take steps to go downwards until you find your car at the lowest point. Here the mapping between source and target (analogy and concept) is given by:

- The fog is analogous to the fact that the gradient descent algorithm does not know the optimum.
- The steepest decline is analogous to the gradient.
- The steps you take as a person is analogous to the step size in gradient descent.
- A valley in the mountain is analogous to a local optimum.
- The base of the mountain is analogous to the global optimum.

A.1 Response: Misclassification error

Analogy: You're a grocery worker tasked with sorting fruits into labeled crates: apples, oranges, and bananas. For each crate, you pick the most frequent fruit and label the crate accordingly. If a crate labeled "apple" still contains some oranges and bananas, these are considered sorting mistakes.
Mapping:

- Crate of fruits → A node in the decision tree.
- Types of fruits → Class labels of data points.
- Label on the crate = majority fruit → Predicted class for that node.
- Fruits that don't match the label → Misclassified data points.
- Fraction of non-majority fruits → Misclassification loss.

A.2 Response: Loglikelihood

Analogy: Imagine you're a detective trying to identify the criminal from a group of suspects. You rank each suspect based on how likely you think they committed the crime. When the true criminal is near the top of your list (assigned high probability), you're considered a good detective. If they're low on your list, your performance is judged harshly.
Mapping:

- Each suspect → A possible class label.
- The actual criminal → The true label.
- Probability assigned to each suspect → The model's predicted class probabilities.
- High rank (high probability) for the real criminal → Loglikelihood is high (loss is low).
- Low rank for the real criminal → Loglikelihood is low (loss is high).
- Penalizing bad rankings → Loglikelihood loss increases quickly when true label gets low predicted probability.

A.3 Response: Mean Squared Error

Analogy: Imagine a group of archers practicing their aim by shooting arrows at a target. After each round, the coach evaluates how accurate the archers were by looking at how far each arrow landed from the bullseye.
Mapping:

- Each arrow → A prediction made by the model.
- Bullseye (center of target) → The true value or label.
- Distance between arrow and bullseye → The error (difference between prediction and true value).
- Worse points for missing the bullseye by a large margin → Squaring the distance
- Tallying the points of all arrows → Calculating the mean squared error across all predictions.

A.4 Response: Absolute error

Analogy: You're managing a student club. You estimate how much money you'll spend for each event, but afterward you compare it to how much you actually spent. Whether you overspend or underspend, you just care about how far off you were, and every euro of error matters the same.
Mapping:

- Estimated budget for an event → Predicted value
- Actual money spent → True value
- Budget error (overspent or underspent) → Difference between prediction and truth
- Taking the absolute value → You're only interested in how far off, not whether it was high or low

A.5 Response: Hinge loss

Analogy: Imagine a train track with a safety zone marked on both sides. Any object inside this safety zone is a danger and must be removed — even if it's not directly on the track. Only objects outside this margin are considered safe.
Mapping:

- The train track → The decision boundary
- The safety zone → The margin required for classification confidence
- Objects near or on the track → Data points near or on the wrong side of the margin

- Removing those objects → Penalizing predictions with hinge loss
- Objects far away from the safety zone → Confident, correct predictions (no hinge loss)
- The need to keep the margin clear → Encouraging classifiers to not just be correct, but confidently correct

A.6 Response: Gini index

Analogy: Imagine you're doing laundry and want to separate your clothes into baskets based on color before washing. You aim to have one basket per color (e.g., all-white clothes in one, all-black in another). If a basket has a perfect mix of different colors, you'll have trouble deciding which cycle to run — and the chance of ruining clothes is high. If a basket contains clothes of only one color, everything's fine.
Mapping:

- Each basket of laundry → A node in a decision tree.
- Colors of clothes in the basket → Class labels of samples.
- A mixed basket (e.g., white + red + black...) → High Gini impurity.
- A basket with only one color → Gini impurity = 0 (pure).
- The more mixed the basket, the more uncertain you are → Higher uncertainty of classification.

A.7 Response: Manhattan distance

Analogy: Imagine you're a delivery person in a city like Manhattan, where the streets are laid out in a perfect grid. You need to go from your location to the customer's house. However, you can't walk diagonally—only north/south or east/west along the streets. You try to minimize the total number of blocks you walk.
Mapping:

- Current location → Model's prediction.
- Destination (customer's house) → True value or label.
- Blocks walked north/south/east/west → Absolute differences in each dimension.
- Total number of blocks walked → Manhattan distance = sum of absolute errors.

A.8 Response: Cross-entropy

Analogy: You're betting on a horse race with multiple horses. You don't just pick one horse—you distribute your money (belief) across several based on how likely you think each will win. If the winning horse was the one you bet heavily on, you earn a lot. If you bet heavily on a horse that lost, you lose big.
Mapping:

- Each horse → A class label.
- The actual winning horse → The true label.
- Your bet distribution → The predicted probabilities over classes.
- Betting most of your money on the winner → Low cross-entropy loss (high confidence in the correct class).
- Betting on the wrong horse(s) → High cross-entropy loss (high confidence in incorrect labels).

A.9 Response: Reconstruction error

Analogy: You saw a person and you're now describing them to a sketch artist who hasn't seen them. The sketch artist draws a portrait based on your description. Once finished, you compare the sketch to the real person. The more it differs, the higher the reconstruction error.
Mapping:

- The real person's face → Original input
- Your verbal description → Encoded representation (compressed form)
- The sketch drawn from your description → Reconstructed output
- Comparing the real face with the sketch → Calculating reconstruction error
- A good likeness → Low reconstruction error
- A bad likeness → High reconstruction error

A.10 Response: Kullback-Leibler divergence

Analogy: Imagine you're planning a picnic, and you check two weather apps: App A (your "true" trusted source) and App B (your new model). Both apps predict the probability of different weather conditions (sunny, cloudy, rainy).
Mapping:

- The true weather forecast (App A) → True probability distribution (P).
- The predicted weather forecast (App B) → The model's predicted distribution (Q).
- Chance of rain given by App B differs from App A → High KL-divergence.
- Chance of rain is the same on both apps → KL-divergence is zero.

B Expert survey questions

Below are all the questions that were asked in the expert survey. For each loss function, the experts rated three categories on a 3-point Likert scale, as shown below.
Target concept coverage: Low / Mid / High
Mapping strength: Low / Mid / High
Metaphoricity: Low / Mid / High

B.1 Misclassification error

Recall the definition of factors:

Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition

Misclassification loss (or classification error) measures how many data points are wrongly classified in a node. For each node, you assign the majority class label. The loss is the fraction of data points in the node that do not belong to the majority class.

Analogy

You're a grocery worker tasked with sorting fruits into labeled crates: apples, oranges, and bananas. For each crate, you pick the most frequent fruit and label the crate accordingly. If a crate labeled "apple" still contains some oranges and bananas, these are considered sorting mistakes.

Mapping

- Crate of fruits $\rightarrow$ A node in the decision tree.
- Types of fruits $\rightarrow$ Class labels of data points.
- Label on the crate = majority fruit $\rightarrow$ Predicted class for that node.
- Fruits that don't match the label $\rightarrow$ Misclassified data points.
- Fraction of non-majority fruits $\rightarrow$ Misclassification loss.

B.2 Loglikelihood

Recall the definition of factors:

Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition

Log-likelihood measures how probable the observed data is under the model's predicted probability distribution. In classification:

- The model predicts probabilities for each class.
- The log-likelihood is the log of the probability assigned to the correct class.
- The goal is to maximize this value, meaning the model is assigning high probability to the correct labels.

Analogy

Imagine you're a detective trying to identify the criminal from a group of suspects. You rank each suspect based on how likely you think they committed the crime. When the true criminal is near the top of your list (assigned high probability), you're considered a good detective. If they're low on your list, your performance is judged harshly.

Mapping

- Each suspect $\rightarrow$ A possible class label.
- The actual criminal $\rightarrow$ The true label.
- Probability assigned to each suspect $\rightarrow$ The model's predicted class probabilities.
- High rank (high probability) for the real criminal $\rightarrow$ Loglikelihood is high (loss is low).
- Low rank for the real criminal $\rightarrow$ Loglikelihood is low (loss is high).
- Penalizing bad rankings $\rightarrow$ Loglikelihood loss increases quickly when true label gets low predicted probability.

B.3 Mean Squared Error

Recall the definition of factors:

Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition

The Mean Squared Error (MSE) is a metric used to measure how well a machine learning model's predictions match the actual outcomes. It works by:

- Calculating the difference between each predicted value and the actual value (the error).
- Squaring these errors (to penalize larger errors more).
- Taking the average of all the squared errors.
- A lower MSE means the model's predictions are closer to the real values, while a higher MSE means the model is making larger mistakes.

Analogy
Imagine a group of archers practicing their aim by shooting arrows at a target. After each round, the coach evaluates how accurate the archers were by looking at how far each arrow landed from the bullseye.

Mapping
- Each arrow $\rightarrow$ A prediction made by the model.
- Bullseye (center of target) $\rightarrow$ The true value or label.
- Distance between arrow and bullseye $\rightarrow$ The error (difference between prediction and true value).
- Worse points for missing the bullseye by a large margin $\rightarrow$ Squaring the distance
- Tallying the points of all arrows $\rightarrow$ Calculating the mean squared error across all predictions.

B.4 Absolute error

Recall the definition of factors:
Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition
Absolute loss measures the absolute difference between the predicted value and the actual value: $|y - \hat{y}|$. It treats all errors equally, regardless of size.

Analogy
You're managing a student club. You estimate how much money you'll spend for each event, but afterward you compare it to how much you actually spent. Whether you overspend or underspend, you just care about how far off you were, and every euro of error matters the same.

Mapping
- Estimated budget for an event $\rightarrow$ Predicted value
- Actual money spent $\rightarrow$ True value
- Budget error (overspent or underspent) $\rightarrow$ Difference between prediction and truth
- Taking the absolute value $\rightarrow$ You're only interested in how far off, not whether it was high or low.

B.5 Hinge loss

Recall the definition of factors:
Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition
Hinge loss is used primarily for maximum-margin classification, such as in Support Vector Machines (SVMs).
- It penalizes predictions that are not only incorrect but also too close to the decision boundary.
- Even correct predictions incur a loss if they are not confidently correct (i.e., margin ¡ 1).
- The loss is 0 when the prediction is correct and far enough from the decision boundary.

Analogy
Imagine a train track with a safety zone marked on both sides. Any object inside this safety zone is a danger and must be removed — even if it's not directly on the track. Only objects outside this margin are considered safe.

Mapping
- The train track $\rightarrow$ The decision boundary
- The safety zone $\rightarrow$ The margin required for classification confidence
- Objects near or on the track $\rightarrow$ Data points near or on the wrong side of the margin
- Removing those objects $\rightarrow$ Penalizing predictions with hinge loss
- Objects far away from the safety zone $\rightarrow$ Confident, correct predictions (no hinge loss)
- The need to keep the margin clear $\rightarrow$ Encouraging classifiers to not just be correct, but confidently correct

B.6 Gini index

Recall the definition of factors:
Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition
Gini loss measures how mixed the classes are in a group of data points. It is used to decide how to split data in a decision tree.
- For each node, you calculate the probability (p) of each class.
- For each class, multiply the chance of picking that class by the chance of not picking it, then add up these values for all classes. This is the Gini index, also called "impurity".
- Lower Gini impurity means the node is more "pure" (mostly one class).

- Decision trees prefer splits that result in lower Gini impurity in the child nodes.

Analogy
Imagine you're doing laundry and want to separate your clothes into baskets based on color before washing. You aim to have one basket per color (e.g., all-white clothes in one, all-black in another). If a basket has a perfect mix of different colors, you'll have trouble deciding which cycle to run — and the chance of ruining clothes is high. If a basket contains clothes of only one color, everything's fine.

Mapping
- Each basket of laundry → A node in a decision tree.
- Colors of clothes in the basket → Class labels of samples.
- A mixed basket (e.g., white + red + black...) → High Gini impurity.
- A basket with only one color → Gini impurity = 0 (pure).
- The more mixed the basket, the more uncertain you are → Higher uncertainty of classification.

B.7 Manhattan distance

Recall the definition of factors:
Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition
Manhattan distance is a way to measure how far two points are from each other by only moving along horizontal and vertical lines (like a grid).

- You take the absolute difference between each predicted value and the actual value.
- You sum up all those absolute differences.

Analogy
Imagine you're a delivery person in a city like Manhattan, where the streets are laid out in a perfect grid. You need to go from your location to the customer's house. However, you can't walk diagonally—only north/south or east/west along the streets. You try to minimize the total number of blocks you walk.

Mapping
- Current location → Model's prediction.
- Destination (customer's house) → True value or label.
- Blocks walked north/south/east/west → Absolute differences in each dimension.
- Total number of blocks walked → Manhattan distance = sum of absolute errors.

B.8 Cross-entropy

Recall the definition of factors:
Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition
Cross-entropy loss measures the difference between two probability distributions: the true labels and the model's predicted probabilities. It is used for classification tasks.

- For each data point, it calculates how far off the predicted probability is from the correct class.
- It assigns a high penalty when the model is confident but wrong.
- The loss is lowest when the model assigns high probability to the correct class.

Analogy
You're betting on a horse race with multiple horses. You don't just pick one horse—you distribute your money (belief) across several based on how likely you think each will win. If the winning horse was the one you bet heavily on, you earn a lot. If you bet heavily on a horse that lost, you lose big.

Mapping
- Each horse → A class label.
- The actual winning horse → The true label.
- Your bet distribution → The predicted probabilities over classes.
- Betting most of your money on the winner → Low cross-entropy loss (high confidence in the correct class).
- Betting on the wrong horse(s) → High cross-entropy loss (high confidence in incorrect labels).

B.9 Reconstruction error

Recall the definition of factors:
Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition

Reconstruction error is used in unsupervised learning to measure how well the model can recreate the original input.

- It compares the input and its reconstruction.
- Lower reconstruction error means the model is good at capturing the structure of the data.

Analogy

You saw a person and you're now describing them to a sketch artist who hasn't seen them. The sketch artist draws a portrait based on your description. Once finished, you compare the sketch to the real person. The more it differs, the higher the reconstruction error.

Mapping

- The real person's face → Original input
- Your verbal description → Encoded representation (compressed form)
- The sketch drawn from your description → Reconstructed output
- Comparing the real face with the sketch → Calculating reconstruction error
- A good likeness → Low reconstruction error
- A bad likeness → High reconstruction error

B.10 Kullback-Leibler divergence

Recall the definition of factors:

Target concept coverage: How well the analogy covers the topics in the description.
Mapping strength: The logical soundness and consistency of the correspondence between source and target concepts.
Metaphoricity: Conceptual distance between the source and the target concept.

Definition

KL divergence measures how one probability distribution differs from a reference distribution.

- It's used to quantify how different the predicted distribution is from the true distribution.
- Lower values mean the predicted distribution is closer to the target distribution.

Analogy

Imagine you're planning a picnic, and you check two weather apps: App A (your "true" trusted source) and App B (your new model). Both apps predict the probability of different weather conditions (sunny, cloudy, rainy).

Mapping

- The true weather forecast (App A) → True probability distribution (P).
- The predicted weather forecast (App B) → The model's predicted distribution (Q).
- Chance of rain given by App B differs from App A → High KL-divergence.
- Chance of rain is the same on both apps → KL-divergence is zero.

C Expert evaluation results

This appendix contains the survey data for the expert evaluations. The numbers in each rated category correspond to their rating as follows: Low = 1, Medium = 2, High = 3. Note that expert 1 is not necessarily the same participant across each analogy.

Table 3: Misclassification error expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	3	2	3
Expert 2	3	3	2
Expert 3	3	2	2

Table 4: Gini index expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	3	3	1
Expert 2	2	2	3

Table 5: Loglikelihood expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	3	3	3
Expert 2	3	3	3
Expert 3	1	1	2

Table 6: Manhattan distance expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	3	3	3
Expert 2	3	3	3
Expert 3	3	3	3

Table 7: Mean Squared Error expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	3	3	3
Expert 2	3	3	3
Expert 3	3	3	3
Expert 4	2	2	3
Expert 5	2	2	2
Expert 6	3	2	3

Table 8: Cross-entropy loss expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	2	2	2
Expert 2	3	3	3
Expert 3	3	3	3
Expert 4	2	2	3
Expert 5	3	2	1
Expert 6	2	1	3

Table 9: Kullback-Leibler divergence expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	2	2	2
Expert 2	2	3	3
Expert 3	3	3	1

Table 10: Absolute error expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	3	3	3
Expert 2	3	2	3
Expert 3	3	3	1

Table 11: Reconstruction error expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	3	3	3
Expert 2	3	3	1
Expert 3	3	2	3
Expert 4	3	3	2
Expert 5	3	3	3
Expert 6	3	3	3
Expert 7	2	3	3

Table 12: Hinge loss expert ratings

Participant	Target concept coverage	Mapping strength	Metaphoricity
Expert 1	2	1	2
Expert 2	2	1	3
Expert 3	3	3	2
Expert 4	3	2	1
Expert 5	2	2	3

D A/B test survey questions

Below are the pre- and post-questions, and explanations for each of the analogies that were tested in the student survey.

D.1 Mean Squared Error

Pre-test

Two machine learning models make a prediction:

- Model A has an error of 4.
- Model B has an error of 2.

Which statement is true?

A Model A has an MSE that's twice as large as that of model B.

B **Model A has an MSE that's four times as large as that of model B.**

C Model A has an MSE that's eight times as that of model B.

D Model A and B have the same MSE

E I don't know/understand

Post-test

Two machine learning models make 5 predictions:

- Model A makes a small error of 1, consistently on every prediction.
- Model B makes 4 predictions flawlessly (0 error), but one prediction with a large error of 5.

Which model will have a higher Mean Squared Error?

A Model A will have a higher MSE

B **Model B will have a higher MSE**

C Both will have the same MSE

D I don't know/understand

Explanation without analogy

Mean Squared Error is a way to measure how far off predictions are from the true, real values. For each data point, the distance between the predictions and the true values are squared, then the average (mean) over these squared distances is taken. Because the distances are squared, larger distances yield increasingly larger Mean Squared Error values.

Explanation with analogy

Mean Squared Error is a way to measure how far off predictions are from the true, real values. For each data point, the distance between the predictions and the true values are squared, then the average (mean) over these squared distances is taken. Because the distances are squared, larger distances yield increasingly larger Mean Squared Error values.

Imagine a group of archers practicing their aim by shooting arrows at a target. After each round, the coach evaluates how accurate the archers were by looking at how far each arrow landed from the bullseye.

- Each arrow → A prediction.
- Bullseye (center of target) → The true value.
- Distance between arrow and bullseye → The error (difference between prediction and true value).
- Worse points for missing the bullseye by a large margin → Squaring the distance.
- Calculating the average score of your shots → Calculating the mean squared error across all predictions.

D.2 Reconstruction error

Pre-test

When performing dimensionality reduction, if you increase the number of dimensions in a projection, what will likely happen to the reconstruction error?

A The reconstruction error will increase.

B **The reconstruction error will decrease.**

C The reconstruction error will stay the same.

D I don't know/understand.

Post-test

An image was reduced in dimensionality in two different ways:

- Projection A: the image was projected onto a 262k (262 thousand) dimensional space.
- Projection B: the image was projected onto a 1m (1 million) dimensional space.

Which of the following statements will generally be true? (Select all that apply)

A **The reconstruction error will be higher with projection A.**

B The reconstruction error will be higher with projection B.

C The reconstruction error will be the same in both cases.

D The reconstructed image will be closer to the original in projection A, than in projection B.

E **The reconstructed image will be closer to the original in projection B, than in projection A.**

F The reconstructed image will be the same as the original in both projections.

G I don't know/understand.

Explanation without analogy

Reconstruction error is used in dimensionality reduction. It is a way to measure how much information is lost when data is projected to a lower dimensional subspace, then reconstructed back into it's original dimensions.

Explanation with analogy

Reconstruction error is used in dimensionality reduction. It is a way to measure how much information is lost when data is projected to a lower dimensional subspace, then reconstructed back into it's original dimensions.

You saw a person and you're now describing them to a sketch artist who hasn't seen them. The sketch artist draws a portrait based on your description. Once finished, you compare the sketch to the real person. The more it differs, the higher the reconstruction error

- The real person's face → Original input
- Your verbal description → Lower dimensional projection
- The sketch drawn from your description → Reconstructed output
- Comparing the real face with the sketch → Calculating reconstruction error

D.3 Manhattan distance

Pre-test

For each of the images, please indicate the Manhattan distance value. One point represents the prediction, the other the true value, it doesn't matter which is which. (open image in new tab to enlarge it)

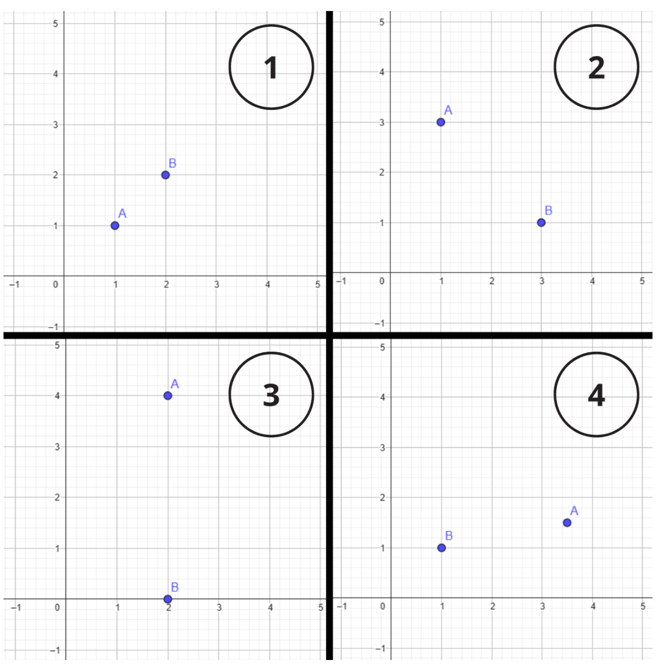

Figure 8: Image accompanying pre-test

Post-test

For each of the images, please indicate the Manhattan distance value. One point represents the prediction, the other the true value, it doesn't matter which is which. (open image in new tab to enlarge it)

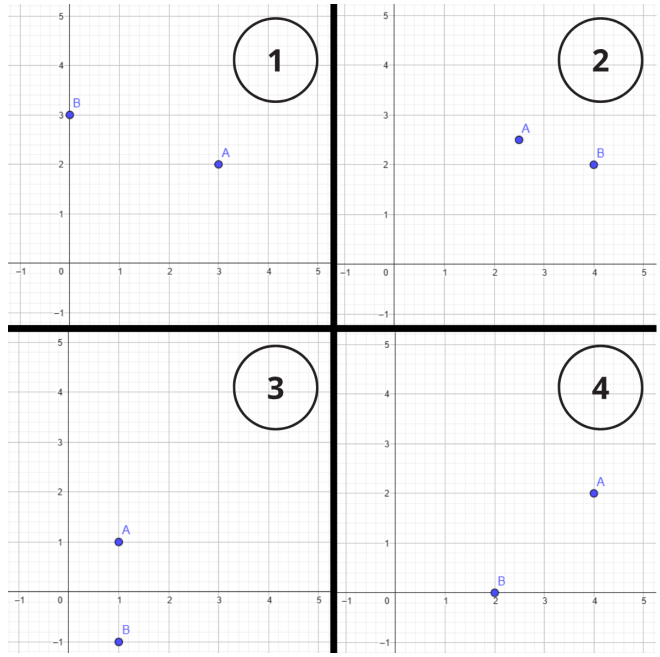

Figure 9: Image accompanying post-test

Explanation without analogy

Manhattan distance is a way to measure how far off predictions are from the true, real values. For each data point, the distance between the predictions and the true values is found by only moving along horizontal and vertical lines. The total distance measure, is the sum of the horizontal distance and the vertical distance.

Explanation with analogy

Manhattan distance is a way to measure how far off predictions are from the true, real values. For each data point, the distance between the predictions and the true values is found by only moving along horizontal and vertical lines. The total distance measure, is the sum of the horizontal distance and the vertical distance.

Imagine you're a delivery person in a city like Manhattan, where the streets are laid out in a perfect grid. You need to go from your location to the customer's house. However, you can't walk diagonally—only north/south or east/west along the streets. You try to minimize the total number of blocks you walk.

- Your current location → Model's prediction.
- Destination (customer's house) → Real value.
- Each block walked north/south/east/west → Amount of error in each direction.
- Total number of blocks walked → Manhattan distance = sum of absolute errors.

D.4 Reduced Instructional Materials Motivation Survey

This part consisted of the 12 questions as outlined in the study by Loorbach et al. [15]. These questions were ordered in a way to ensure each category's questions were not grouped together.

1. The quality of the writing helped to hold my attention.(11A03)
2. The way the information is arranged on the pages helped keep my attention. (17A06)
3. The variety of reading passages, exercises, illustrations, etc, helped keep my attention on the lesson. (28A10)
4. It is clear to me how the content of this material is related to things I already know. (06R01)
5. The content and style of writing in this lesson convey the impression that its content is worth knowing. (23R06)
6. The content of this lesson will be useful to me. (33R09)
7. As I worked on this lesson, I was confident that I could learn the content. (13C05)
8. After working on this lesson for a while, I was confident that I would be able to pass a test on it. (25C07)
9. The good organization of the content helped me be confident that I would learn this material. (35C09)
10. I enjoyed this lesson so much that I would like to know more about this topic. (14S02)
11. I really enjoyed studying this lesson. (21S03)
12. It was a pleasure to work on such a well-designed lesson. (36S06)

E A/B test survey results

E.1 Control group

Table 13: Pre- and post-test answers for Mean Squared Error

Participant	Time to complete survey	Pre-test	Post-test
1	00:09:41	E	B
2	00:00:51	B	C
3	00:08:34	E	B
4	00:31:09	B	B
5	00:05:40	B	C
6	00:02:53	B	C
7	00:03:19	B	B
8	00:06:57	B	B
9	00:07:35	E	B
10	00:09:42	E	D
11	00:14:16	E	B
12	00:05:06	B	B
13	00:10:42	E	B

Table 14: Pre- and post-test answers for Reconstruction error

Participant	Time to complete survey	Pre-test	Post-test
1	00:09:41	D	G
2	00:00:51	C	A-C
3	00:08:34	D	A
4	00:31:09	D	A
5	00:05:40	D	E-A
6	00:02:53	D	A-E
7	00:03:19	D	B-D
8	00:06:57	D	A-E
9	00:07:35	D	G
10	00:09:42	D	G
11	00:14:16	B	A-D
12	00:05:06	D	G
13	00:10:42	D	A-E

Table 15: Pre- and post-test answers for Manhattan distance

Participant	Time to complete survey	Pre-test	Post-test
1	00:09:41	–	3, 2, 2, 3
2	00:00:51	2, 2.5, 2, 2	–
3	00:08:34	2, 4, 4, 4	4, 2, 2, 4
4	00:31:09	2, 4, 4, 3	4, 2, 2, 4
5	00:05:40	–	4, 2, 2, 4
6	00:02:53	2, 4, 4, 3	4, 2, 2, 4
7	00:03:19	1.5, 2.5, 4, 3	3.5, 1.5, 2, 2.5
8	00:06:57	–	4, 2, 2, 4
9	00:07:35	–	4, 2, 2, 4
10	00:09:42	–	4, 2, 2, 4
11	00:14:16	–	4, 2, 2, 4
12	00:05:06	2, 4, 4, 3	4, 2, 2, 4
13	00:10:42	–	4, 2, 2, 4

E.2 Experimental group

Table 16: Pre- and post-test answers for Mean Squared Error

Participant	Time to complete survey	Pre-test	Post-test
1	00:11:38	B	B
2	00:11:27	E	C
3	00:09:44	B	B
4	00:14:22	B	B
5	00:11:06	A	B
6	00:04:30	B	B
7	00:06:54	B	B
8	00:10:34	B	B
9	00:15:52	B	B
10	00:21:10	E	B

Table 17: Pre- and post-test answers for Reconstruction error

Participant	Time to complete survey	Pre-test	Post-test
1	00:11:38	A	B
2	00:11:27	D	A-E
3	00:09:44	A	B-C
4	00:14:22	D	B-D
5	00:11:06	B	A-E
6	00:04:30	A	E-B
7	00:06:54	D	A-E
8	00:10:34	D	G
9	00:15:52	B	A-E
10	00:21:10	D	A-E

Table 18: Pre- and post-test answers for Manhattan distance

Participant	Time to complete survey	Pre-test	Post-test
1	00:11:38	–	–
2	00:11:27	2, 4, 4, 3	4, 2, 2, 4
3	00:09:44	–	4, 2, 2, 4
4	00:14:22	–	4, 2, 2, 4
5	00:11:06	2, 4, 4, 3	4, 2, 2, 4
6	00:04:30	2, 4, 4, 3	4, 2, 2, 4
7	00:06:54	–	3, 1.5, 2, 2.5
8	00:10:34	2, 4, 4, 3	4, 2, 2, 4
9	00:15:52	2, 4, 4, 3	4, 2, 2, 4
10	00:21:10	2, 4, 4, 3	4, 2, 2, 4

Table 19: Motivation survey results: Control group

Participant	**Time**	11A03	17A06	28A10	33R09	06R01	23R06	35C09	25C07	13C05	14S02	21S03	36S06
1	00:09:41	2	3	3	3	2	2	3	1	3	2	2	3
2	00:00:51	1	2	1	2	2	1	2	1	2	3	3	2
3	00:08:34	3	3	3	1	1	4	4	4	4	3	3	3
4	00:31:09	1	2	2	1	3	4	3	0	1	1	2	1
5	00:05:40	1	3	2	2	2	2	2	1	2	1	1	2
6	00:02:53	3	2	3	3	4	4	3	3	3	3	3	3
7	00:03:19	2	3	4	3	3	3	3	2	2	3	3	3
8	00:06:57	1	2	2	1	2	1	2	1	2	0	2	1
9	00:07:35	2	3	3	3	3	2	1	0	3	3	1	1
10	00:09:42	1	0	1	3	1	1	0	0	1	2	2	0
11	00:14:16	3	2	2	1	3	4	2	4	4	1	1	0
12	00:05:06	0	0	0	2	3	1	1	2	2	0	1	1
13	00:10:42	1	2	2	3	3	2	3	3	3	2	3	2

Table 20: Motivation survey results: Experimental group

Participant	**Time**	11A03	17A06	28A10	33R09	06R01	23R06	35C09	25C07	13C05	14S02	21S03	36S06
1	00:11:38	0	0	1	0	3	3	0	1	1	0	0	1
2	00:11:27	3	4	3	3	4	2	3	3	3	2	3	3
3	00:09:44	4	2	3	4	2	3	3	1	3	3	3	3
4	00:14:22	3	3	3	2	3	3	3	3	3	2	3	2
5	00:11:06	3	2	3	2	2	3	3	2	3	2	3	3
6	00:04:30	3	4	4	4	4	4	4	4	4	3	4	4
7	00:06:54	3	3	3	3	3	3	3	3	3	2	2	3
8	00:10:34	1	2	2	3	2	1	2	3	3	1	2	1
9	00:15:52	0	0	0	4	0	0	0	0	0	0	1	0
10	00:21:10	3	4	3	2	3	3	3	4	3	2	3	3