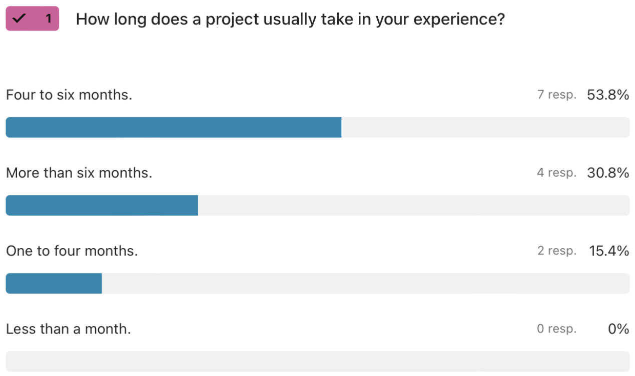
User Study



✓ 1 How long does a project usually take in your experience?

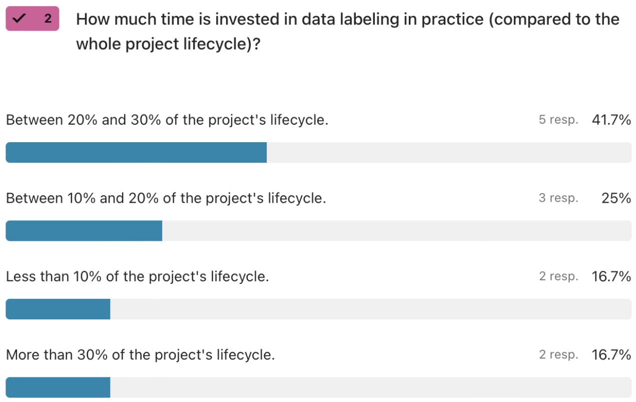
Four to six months. 7 resp. 53.8%

More than six months. 4 resp. 30.8%

One to four months. 2 resp. 15.4%

Less than a month. 0 resp. 0%

The majority of SMEs indicated that the duration of a ”project” ranges between four and six months. The second most common response was that projects last longer than six months. Two SMEs assessed the situation differently, stating that their projects span a time frame of one to four months.



✓ 2 How much time is invested in data labeling in practice (compared to the whole project lifecycle)?

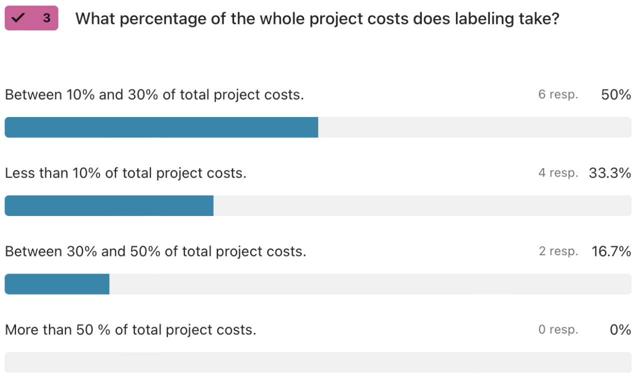
Between 20% and 30% of the project's lifecycle. 5 resp. 41.7%

Between 10% and 20% of the project's lifecycle. 3 resp. 25%

Less than 10% of the project's lifecycle. 2 resp. 16.7%

More than 30% of the project's lifecycle. 2 resp. 16.7%

The responses were more diverse regarding the question of how much time is spent on data labeling compared to the overall project lifecycle. However, the majority of SMEs expressed the opinion that at least 10 percent to 30 percent of the available time is used for data labeling. Two SMEs even indicated that more than 30 percent of the time available in the project lifecycle is dedicated to labeling data.



✓ 3 What percentage of the whole project costs does labeling take?

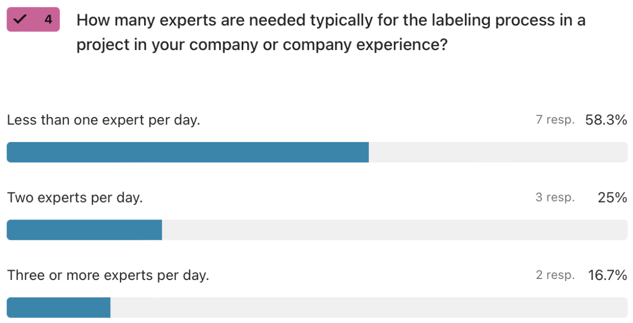
Between 10% and 30% of total project costs. 6 resp. 50%

Less than 10% of total project costs. 4 resp. 33.3%

Between 30% and 50% of total project costs. 2 resp. 16.7%

More than 50 % of total project costs. 0 resp. 0%

From the results of this question, most SMEs assume that 10 to 30 percent of the total project costs are attributable to data labeling, with an additional two SMEs stating that 30 to 50 percent of costs are due to data labeling. Four of the respondents to this question indicated that, in their experience, less than 10 percent of the total project costs are devoted to data labeling.



✓ 4 How many experts are needed typically for the labeling process in a project in your company or company experience?

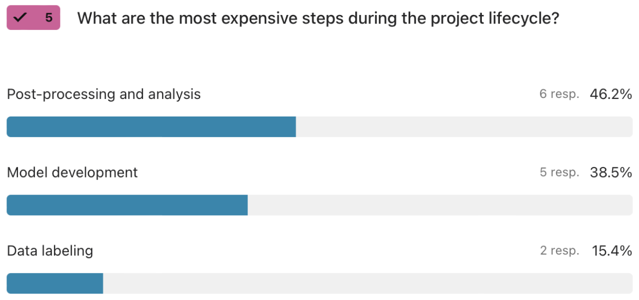
Less than one expert per day. 7 resp. 58.3%

Two experts per day. 3 resp. 25%

Three or more experts per day. 2 resp. 16.7%

The answers to this question were ambivalent. However, it can be stated that experts are needed for the labeling of data. In response to the question of how many experts are practically required for the labeling process in a project, the majority believe that less than one expert per day is needed for labeling. These seven responses from SMEs are countered by five other SME responses stating from their company experience that at least two experts per day are needed. From this population of SMEs, two SMEs in the user study also indicated that three or more experts per day are needed in a project for data labeling.

A definitive answer to what the most costly step within a project lifecycle is could not be determined from this question. Six of the SMEs stated that post-processing and analysis is the most expensive step, and five SMEs suggested that model development is the most costly. Two additional experts indicated that in a project, the most expensive step of the project lifecycle is data labeling.

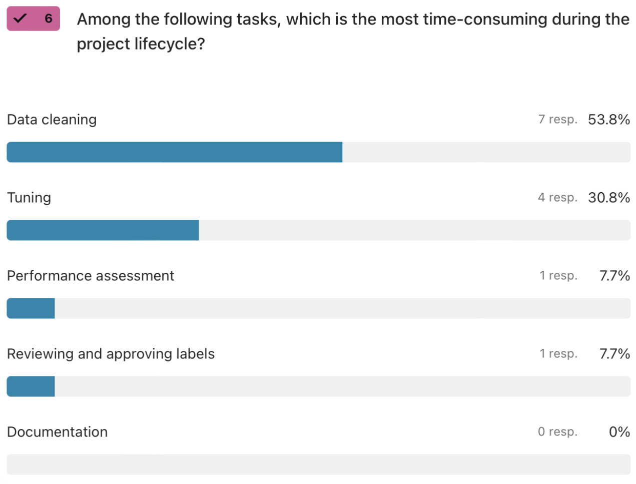


✓ 5 What are the most expensive steps during the project lifecycle?

Post-processing and analysis 6 resp. 46.2%

Model development 5 resp. 38.5%

Data labeling 2 resp. 15.4%


✓ 6 Among the following tasks, which is the most time-consuming during the project lifecycle?

Data cleaning 7 resp. 53.8%

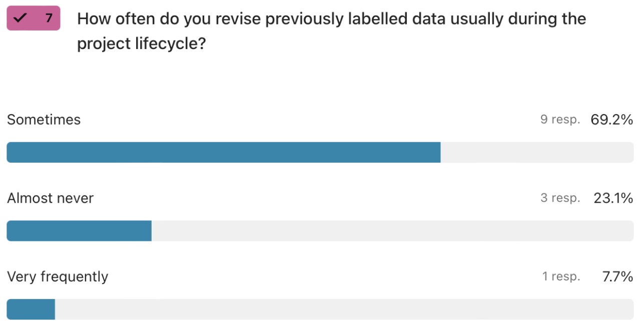
Tuning 4 resp. 30.8%

Performance assessment 1 resp. 7.7%

Reviewing and approving labels 1 resp. 7.7%

Documentation 0 resp. 0%

Based on the responses, the majority of the SMEs believe that data cleaning is the most time-consuming task within a project lifecycle. Tuning was the next most frequently selected option. Both performance assessment and reviewing and approving labels were chosen by one SME each. No respondents identified documentation as the most time-consuming task. In this context, reviewing and approving labels refers to revisiting previously labeled data to confirm its correctness or approve it, which is part of the data labeling process.

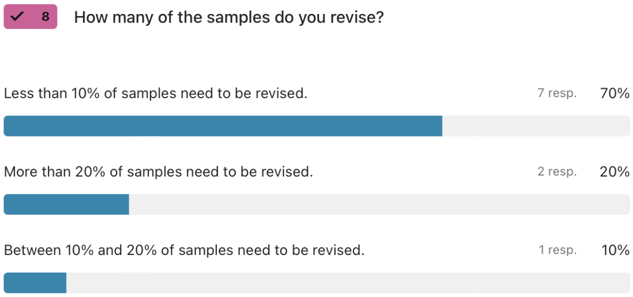

✓ 7 How often do you revise previously labelled data usually during the project lifecycle?

Sometimes 9 resp. 69.2%

Almost never 3 resp. 23.1%

Very frequently 1 resp. 7.7%

The responses to this question indicate that out of thirteen SMEs, ten reported that they at least ”sometimes” have to revise already labeled data within a project lifecycle. One SME even reported that samples need to be revised ”very frequently” during the project lifecycle. On the other hand, three of the user study SMEs stated that samples hardly ever need to be revised. Here again, we see opposed responses, with the majority of SMEs assuming that labels need to be revised.



✓ 8 How many of the samples do you revise?

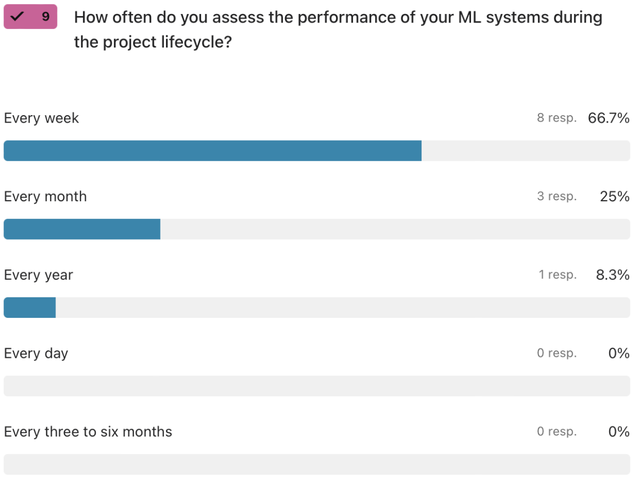
Less than 10% of samples need to be revised. 7 resp. 70%

More than 20% of samples need to be revised. 2 resp. 20%

Between 10% and 20% of samples need to be revised. 1 resp. 10%

This question serves as a follow-up to question seven and provides more details regarding the proportion of samples that are revised when ”sometimes” or ”very often” were selected in the previous response, which was the case for the majority of the participating SMEs. The results indicate that the majority of SMEs revise less than 10 percent of the samples. However, one SME reported revising between 10 percent and 20 percent of the samples. Additionally, two other SMEs stated that they revised more than 20 percent of the samples based on their practical experience in projects.

This question aims to determine how frequently the performance of an ML model needs to be assessed during a project lifecycle. Most of the SMEs indicated that they do this on a weekly basis. Three additional SMEs reported that they assess performance monthly, and one other SME stated that in their practical experience, this assessment is performed annually.



✓ 9 How often do you assess the performance of your ML systems during the project lifecycle?

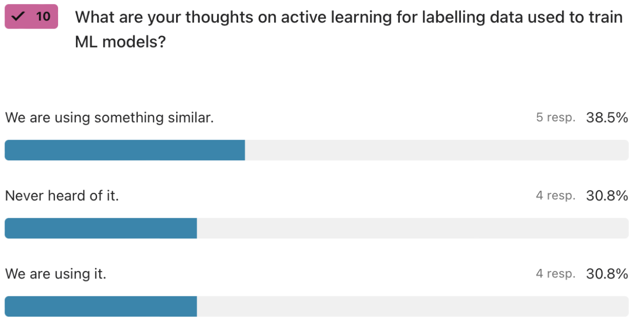
Every week 8 resp. 66.7%

Every month 3 resp. 25%

Every year 1 resp. 8.3%

Every day 0 resp. 0%

Every three to six months 0 resp. 0%



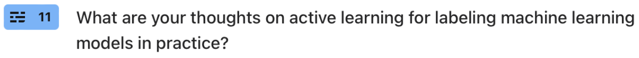
✓ 10 What are your thoughts on active learning for labelling data used to train ML models?

We are using something similar. 5 resp. 38.5%

Never heard of it. 4 resp. 30.8%

We are using it. 4 resp. 30.8%

Based on the responses, the largest group, stated that they are using a method similar to AL for labeling data in their ML models or never heard of AL in the context of data labeling for ML. Only four SMEs stated that they are using AL for data labelling in their ML models.



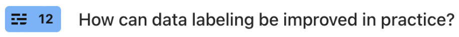
11 What are your thoughts on active learning for labeling machine learning models in practice?

This open-ended query elicited a range of responses, which can be succinctly summarized as follows:

- *Quality and Impact of Labels on Machine Learning Projects:* In the ML projects conducted, labels of average quality are commonly used. Heuristics and auxiliary data sources are utilized to improve these labels. There's a general consensus that if the quality of the labels could be improved, it would positively impact the outcomes of these ML projects. In reality, labels are not typically black-and-white. For data characterized by "weak" labels, an optimal approach would involve under-sampling during the creation of training batches, among other strategies.

- *Prioritizing Training Samples for Effective Resource Use:* Given that labeling resources are finite, there's a need to prioritize the most impactful training samples based on the current model. This approach is key to reducing potential false positives and false negatives. The focus should be on samples at the border of the classes and outliers that significantly alter the decisions of the current model.

- *Interpretation and Implications of Active Learning:* Labelling can often be a tiring process for experts, thus AL can assist in reducing missed positive samples. However, it's worth noting that if an ML solution doesn't actively point to a label, an expert might miss it. Conversely, this can lead to overconfidence, causing experts to only approve or discard labels suggested by ML solutions. This is a significant pitfall that needs to be taken into account. A major advantage of AL for an ML engineer is that it provides insights into how the model behaves at the moment of application. These insights can guide the engineer in modifying the model to produce better predictions, especially as it is frequently evaluated on real-world data.



12 How can data labeling be improved in practice?

In summary, the following responses were collected from this open-ended question:

- *Label Quality and Resource Limitations:* In certain projects, time and budget constraints prevent the improvement of label quality. Ideally, experts in the field would be employed to examine and enhance labels for complex cases. Real-world projects usually involve certain ”certainty” for labels, which should be tracked and utilized for over/under-sampling during the creation of the training dataset.

- *Human-Model Collaboration:* In the labeling process, a model and a human expert often collaborate. However, models do not always offer suggestions that are easily interpretable by humans, leading to increased labeling costs. User Interface and human-readable interpretations significantly impact labeling costs.

- *AL Techniques:* Techniques such as prioritizing and grouping samples to be labeled can simplify data labeling. However, human intervention is still required to maintain the integrity of ground truth in AL.

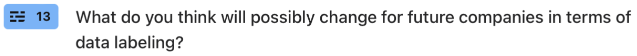
- *Batch Labelling:* Manual labeling of fewer samples and subsequent application of learning techniques to identify similar samples can enhance efficiency. This approach helps avoid the need for experts to switch contexts and understand the concept of a new one for each sample.

- *Labelling Tools:* The use of tools like ”Label Studio” can facilitate the involvement of more people, including non-technical individuals, and make the process less tedious.

- *Label Inconsistency:* Different human experts may produce different labels, leading to discrepancies. Reaching an agreement on the true label can be a time-consuming issue. Addressing this issue can improve data labeling in practice.

- *Early Data Labelling:* Initiating the labeling process as early as possible can help ensure properly characterized datasets. This proactive approach not only benefits ML projects but also enhances overall data management.

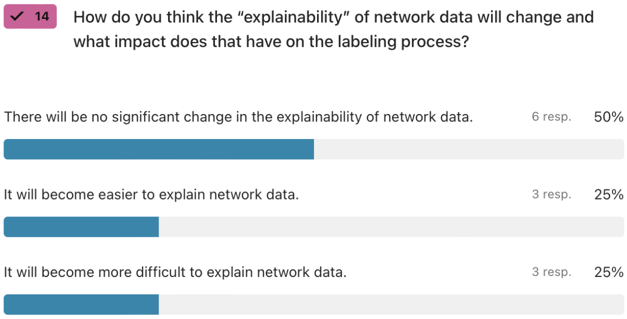
- *Adressing Alert Fatigue:* Several issues like labeling fatigue and inconsistency in labels between different experts plague the data labeling process. Solutions like task rotation and multiple expert reviews can help, but they often lead to reduced efficiency and increased costs.



13 What do you think will possibly change for future companies in terms of data labeling?

This open-ended question provided a summary of the following responses:

- *Automatic Grouping:* Automatic grouping techniques can be used to enhance efficiency in the data labeling process.
- *Third-Party Labelling and AI Enhancement:* As data labeling requirements increase, outsourcing to third-party entities and the development of tools focused on labeling will become more common. The quality of labeling can also be incrementally improved using Artificial Intelligence.
- *Reduced Time for Labelling:* The goal is to use fewer but good-quality labels to optimize the labeling process, reducing the time required for this task.
- *Public Datasets and Quality Labels:* As the cost of data decreases and public datasets become more accessible, the quality of labels and user interactions that yield good labels will become key differentiators for companies. This is due to the fact that better labels generally result in better models.
- *Semi-automated Labelling and Risks:* Many organizations are expected to adopt semi-automated labeling, which combines heuristics with expert labels. This approach enables the labeling of much larger datasets but may decrease label accuracy. Large models like ChatGPT, trained using techniques such as Reinforcement Learning from Human Feedback, appear to work well with this approach. However, these methods often decrease explainability, posing additional risks.
- *The Future of AL:* The future of data labeling will likely be shaped by an increased emphasis on AL, reinforcement learning, and feedback learning. This shift is aimed at optimizing the labeling process by using fewer but high-quality labels, reducing the time required for this task. Concurrently, the use of abstract data embeddings, which may challenge human-interpretable features, is also predicted to rise.



✓ 14 How do you think the “explainability” of network data will change and what impact does that have on the labeling process?

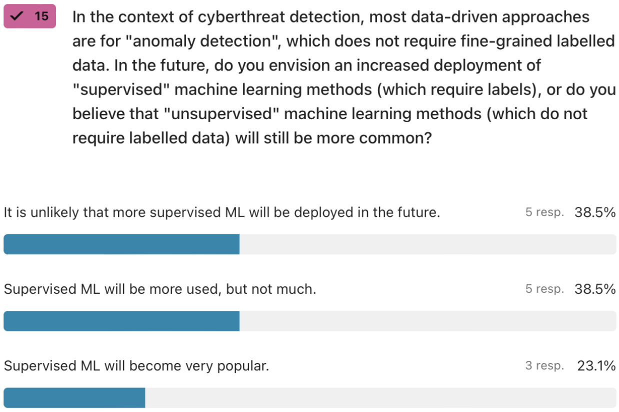
There will be no significant change in the explainability of network data. 6 resp. 50%

It will become easier to explain network data. 3 resp. 25%

It will become more difficult to explain network data. 3 resp. 25%

Based on the responses, 50 percent of the Subject Matter Experts (SMEs) anticipate no significant change in the explainability of network data. Three SMEs expect that the process of explaining network data will become easier, while an equal number of SMEs predict it will become more difficult. Thus, it can be inferred that 75 percent of the SMEs in this user study do not expect an increase in the ease of network data explainability.

While five SMEs consider it unlikely that the use of supervised ML in the future will increase, an equal number foresee a moderate increase in their use. Additionally, three SMEs predict that supervised ML will become highly popular. Despite the varied views, the majority do anticipate an increase in the use of supervised ML in the future.



✓ 15 In the context of cyberthreat detection, most data-driven approaches are for "anomaly detection", which does not require fine-grained labelled data. In the future, do you envision an increased deployment of "supervised" machine learning methods (which require labels), or do you believe that "unsupervised" machine learning methods (which do not require labelled data) will still be more common?

It is unlikely that more supervised ML will be deployed in the future. 5 resp. 38.5%

Supervised ML will be more used, but not much. 5 resp. 38.5%

Supervised ML will become very popular. 3 resp. 23.1%

The range of views presented by the SMEs indicates a necessity for additional analysis. Determining subgroups and emphasizing both correlations and non-correlations within their specific responses might bring about greater clarity. Therefore, this strategy was adopted in the succeeding chapter. This approach serves to highlight patterns and differences in viewpoints, potentially offering insights into the decision-making processes that guided the experts’ responses.