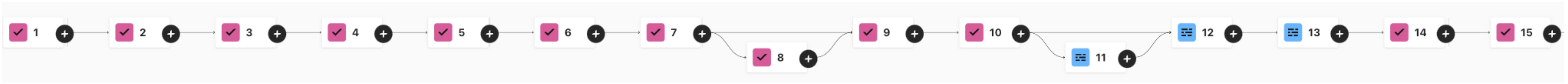
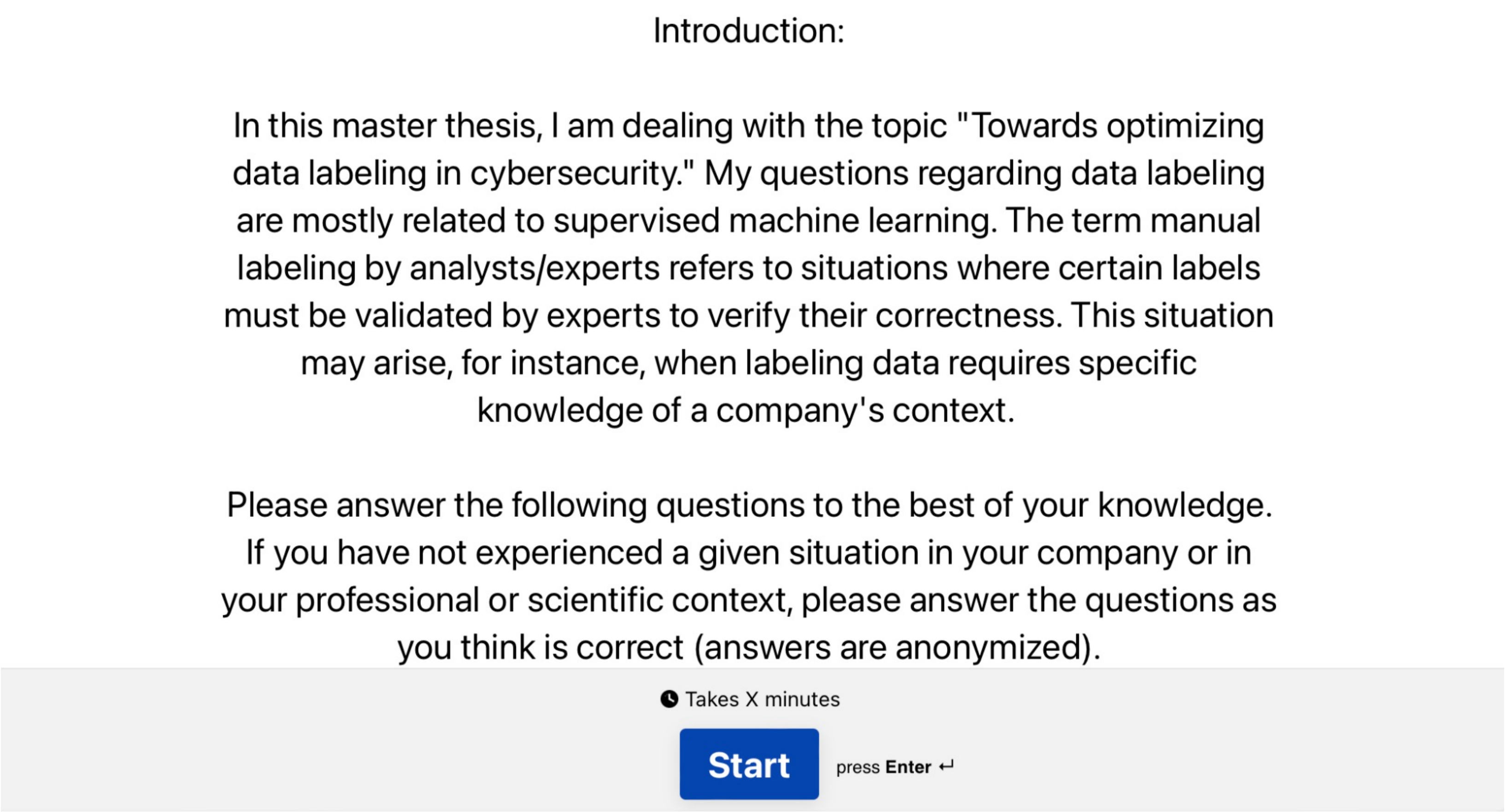
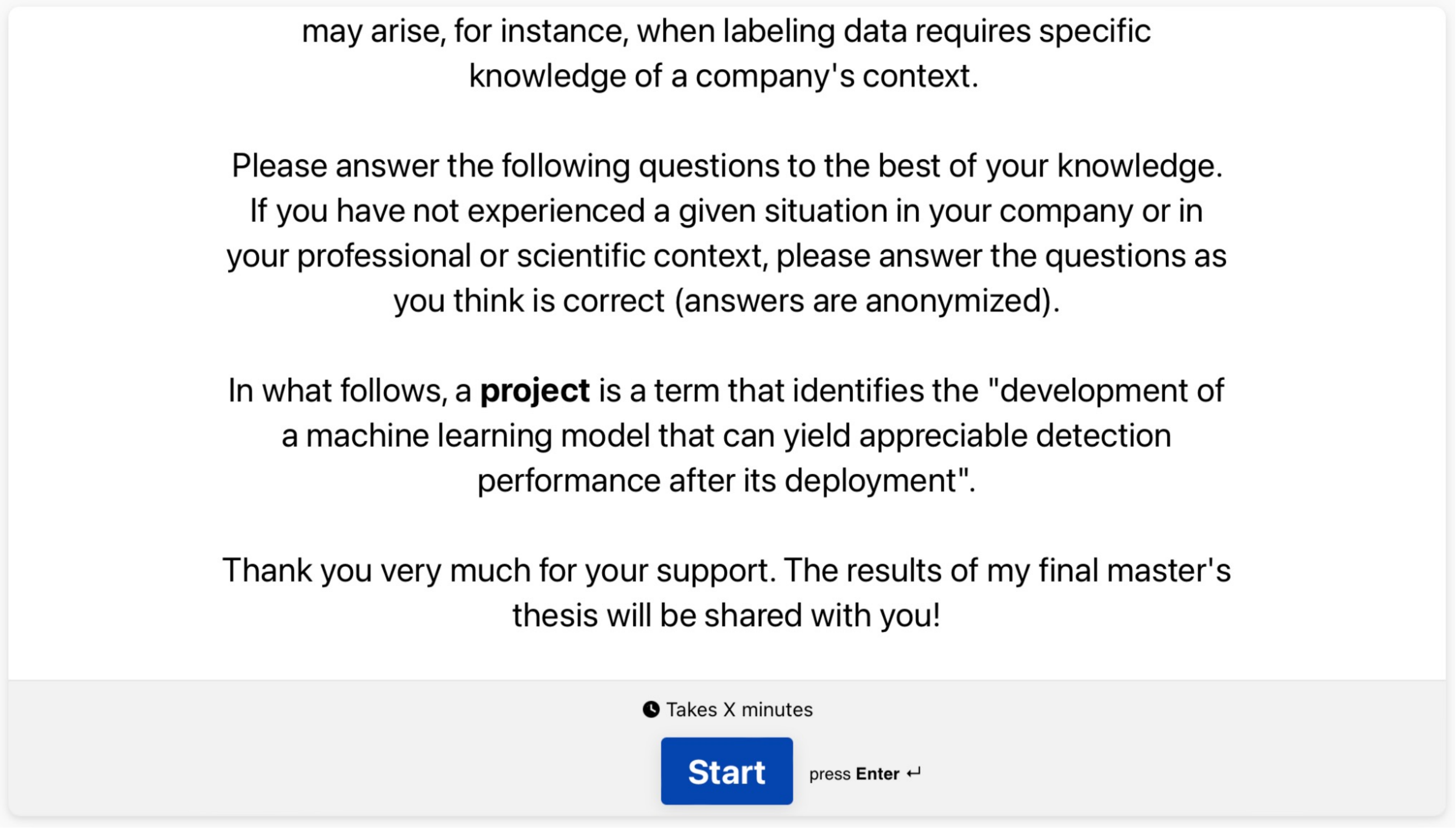

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



Introduction:

In this master thesis, I am dealing with the topic "Towards optimizing data labeling in cybersecurity." My questions regarding data labeling are mostly related to supervised machine learning. The term manual labeling by analysts/experts refers to situations where certain labels must be validated by experts to verify their correctness. This situation may arise, for instance, when labeling data requires specific knowledge of a company's context.

Please answer the following questions to the best of your knowledge. If you have not experienced a given situation in your company or in your professional or scientific context, please answer the questions as you think is correct (answers are anonymized).

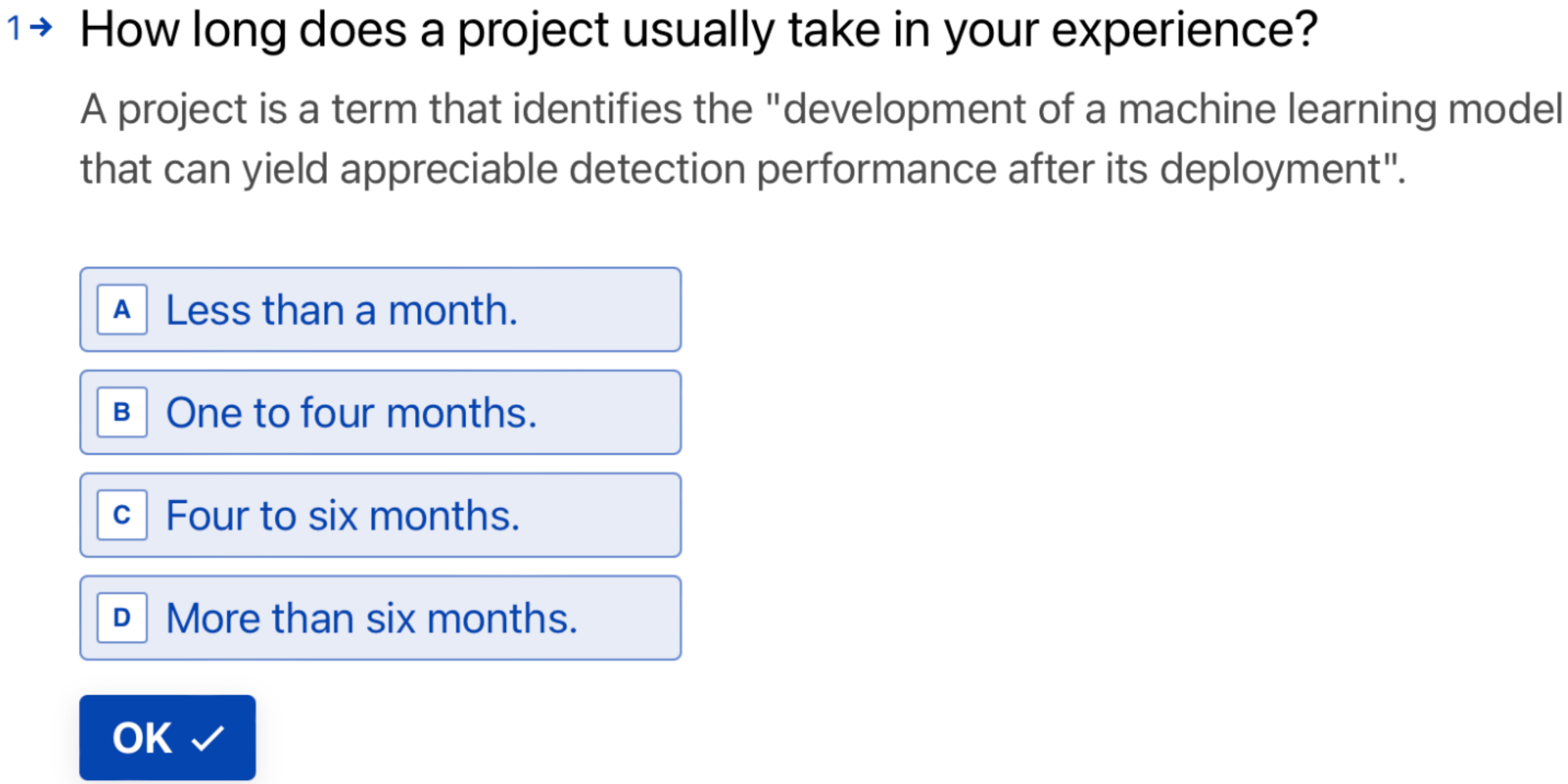


may arise, for instance, when labeling data requires specific knowledge of a company's context.

Please answer the following questions to the best of your knowledge. If you have not experienced a given situation in your company or in your professional or scientific context, please answer the questions as you think is correct (answers are anonymized).

In what follows, a **project** is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

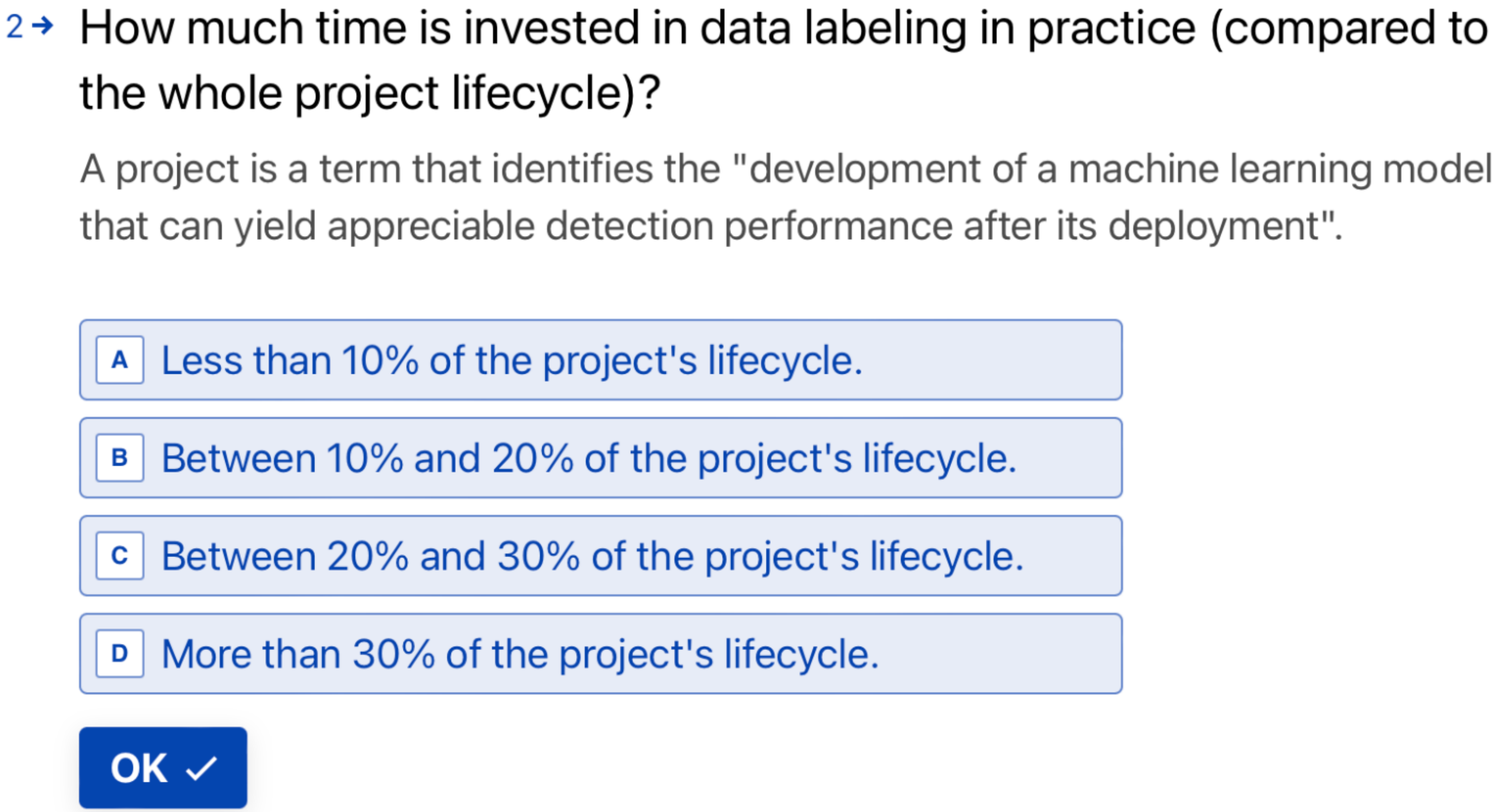
Thank you very much for your support. The results of my final master's thesis will be shared with you!



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

A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

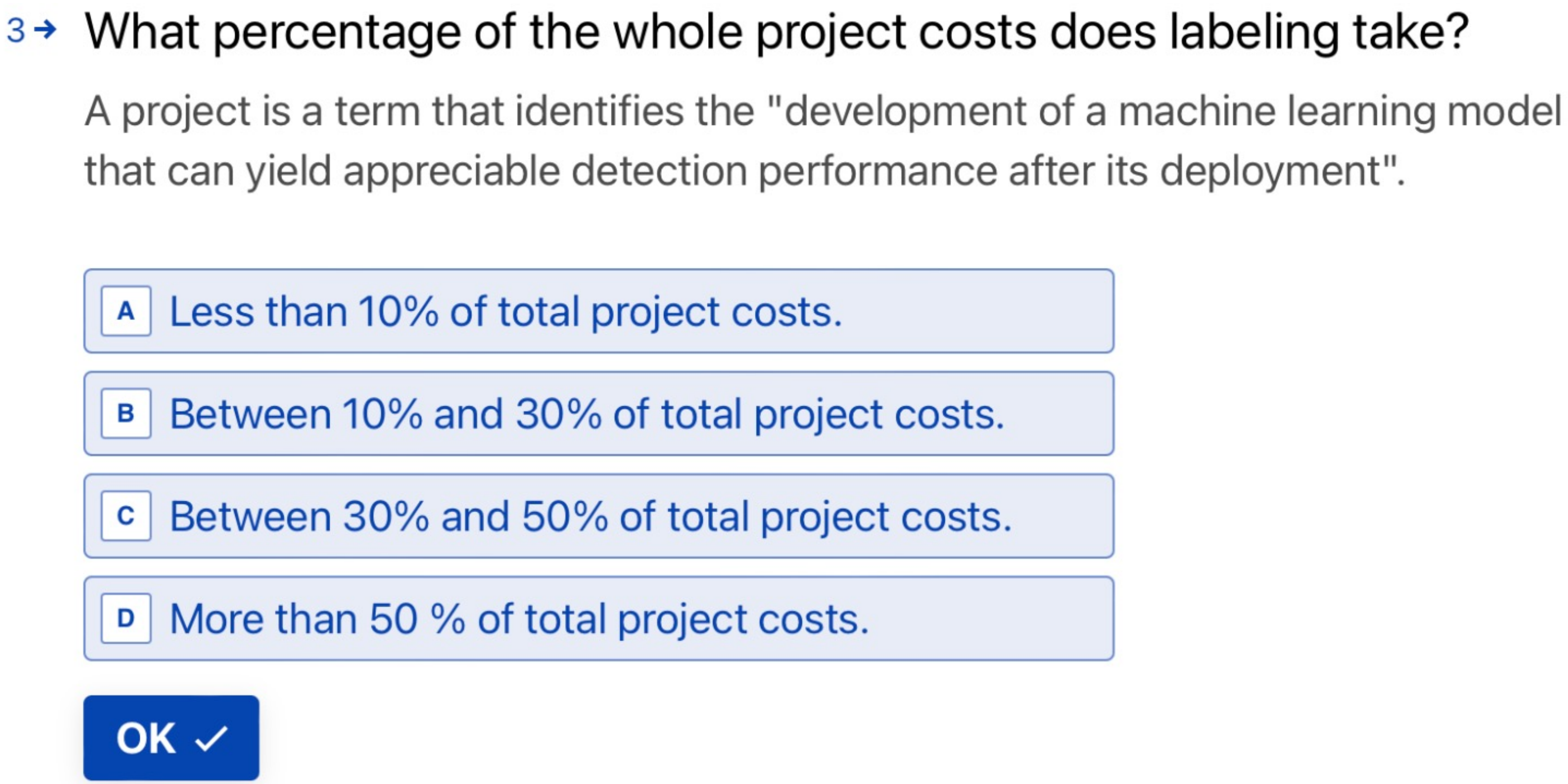
- A Less than a month.
- B One to four months.
- C Four to six months.
- D More than six months.



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

A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

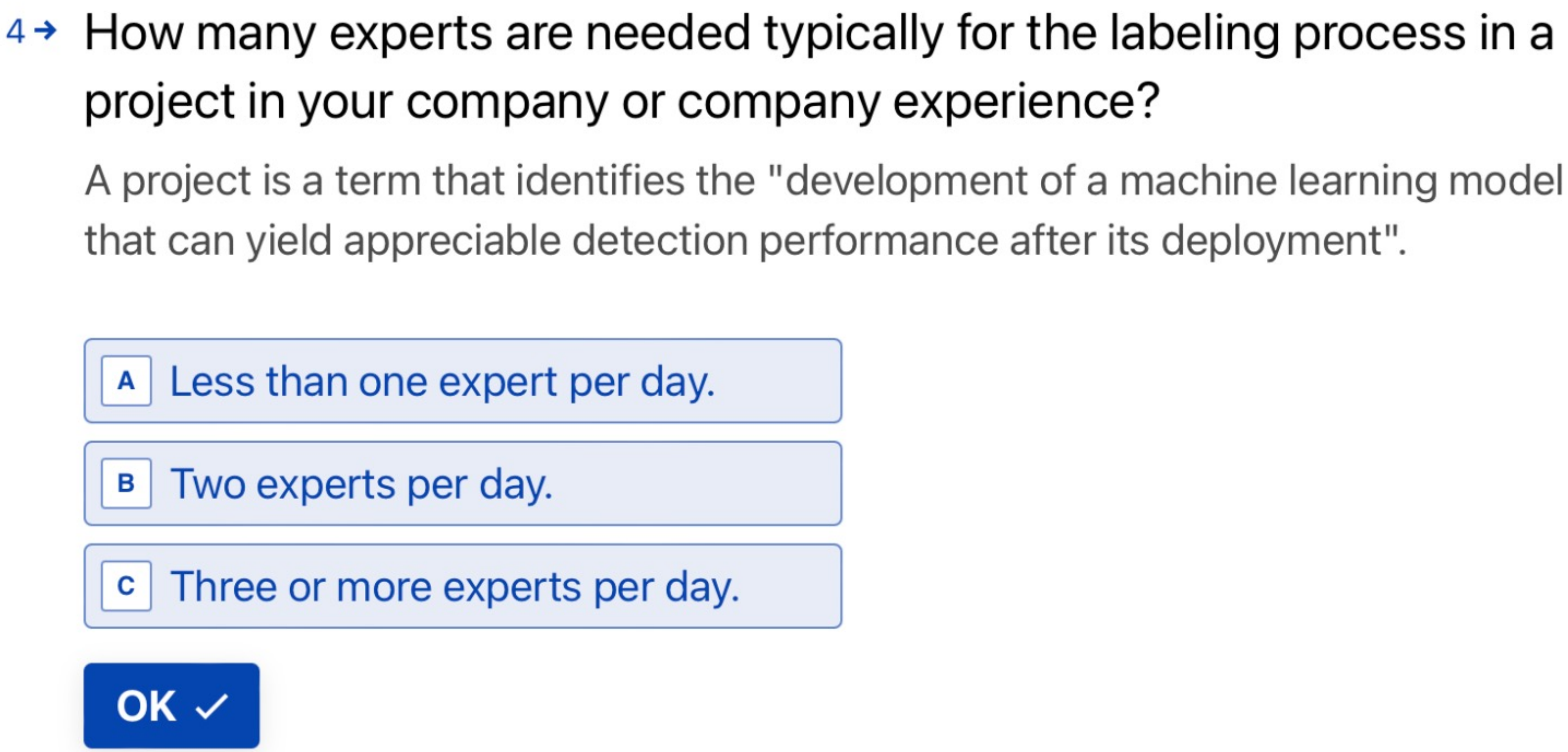
- A Less than 10% of the project's lifecycle.
- B Between 10% and 20% of the project's lifecycle.
- C Between 20% and 30% of the project's lifecycle.
- D More than 30% of the project's lifecycle.



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

A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

- A Less than 10% of total project costs.
- B Between 10% and 30% of total project costs.
- C Between 30% and 50% of total project costs.
- D More than 50 % of total project costs.



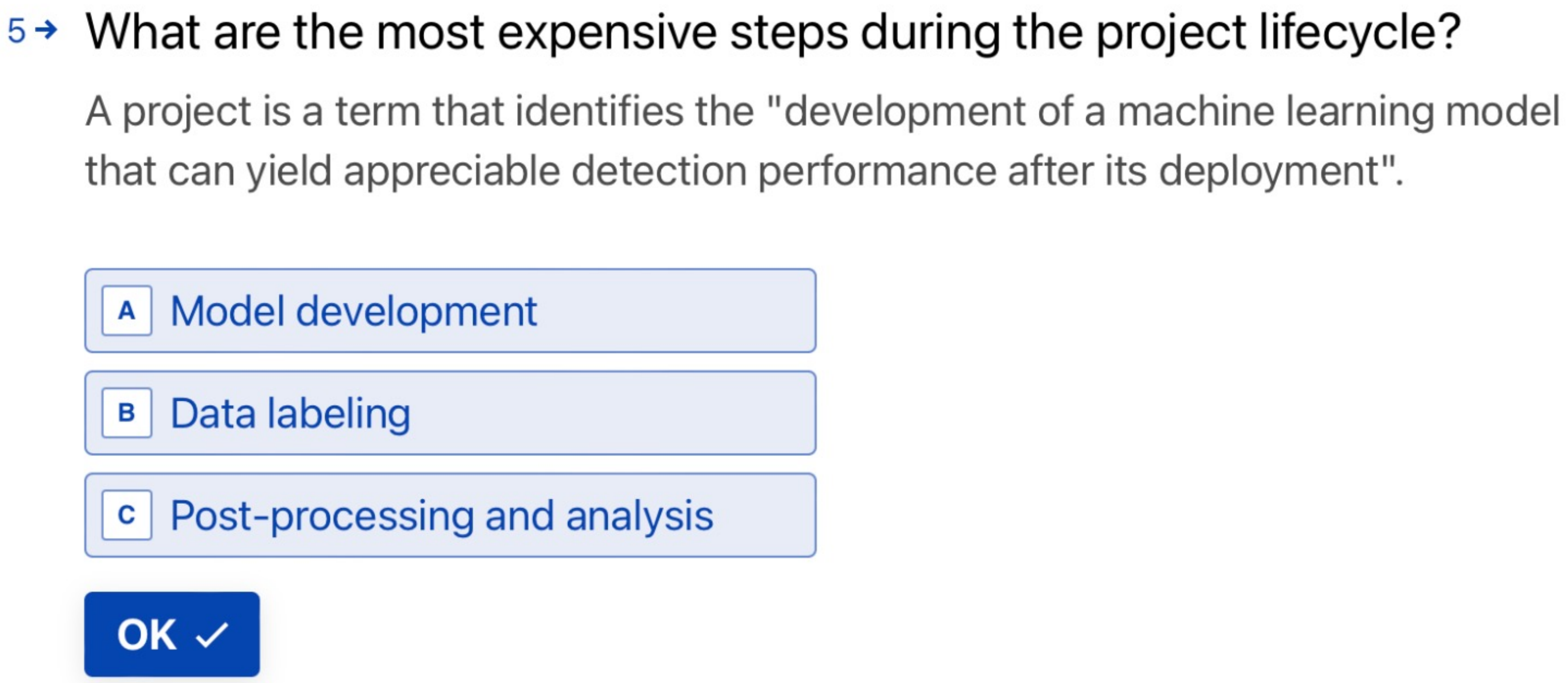
How many experts are needed typically for the labeling process in a project in your company or company experience?

A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

A Less than one expert per day.

B Two experts per day.

C Three or more experts per day.



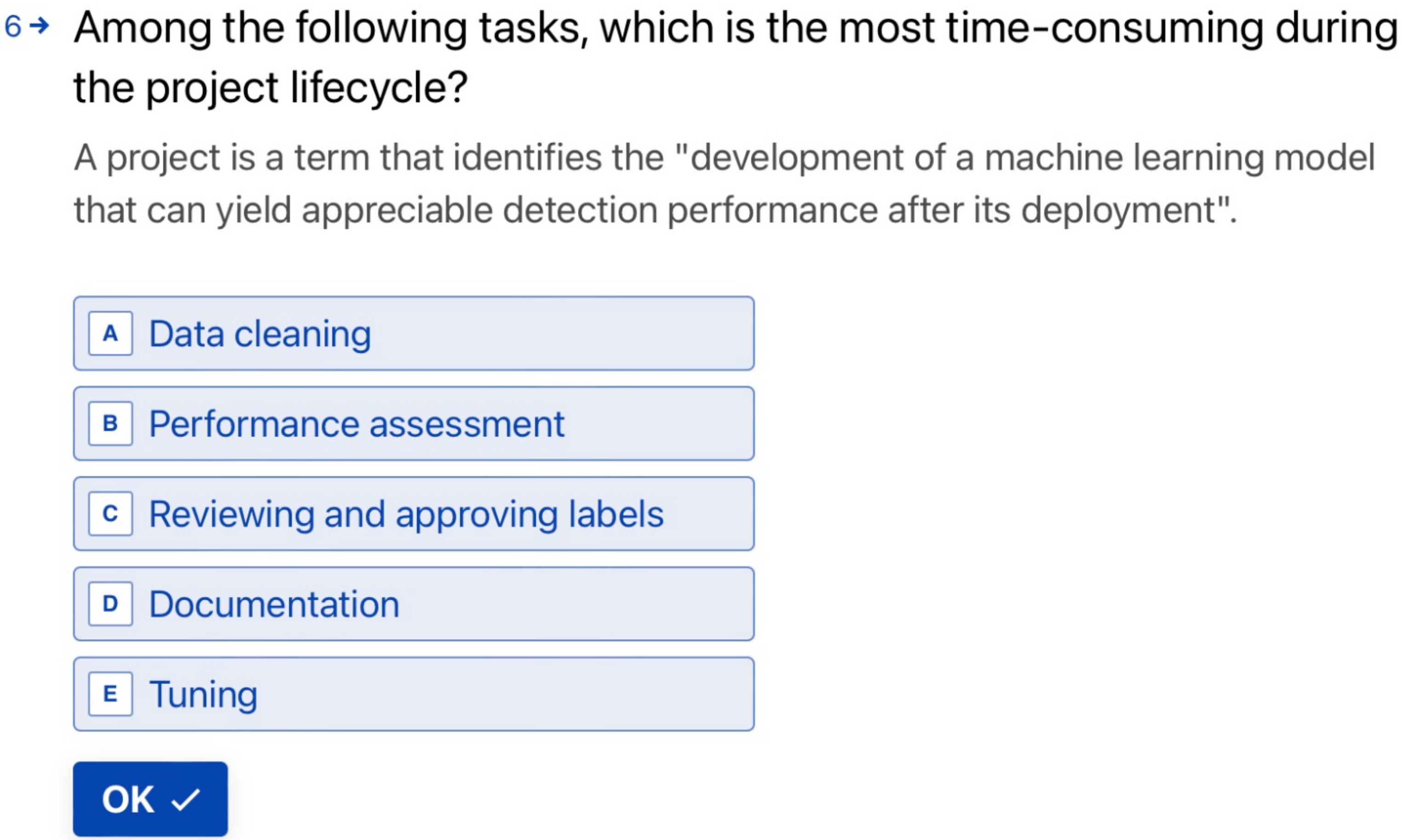
5 → What are the most expensive steps during the project lifecycle?

A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

A Model development

B Data labeling

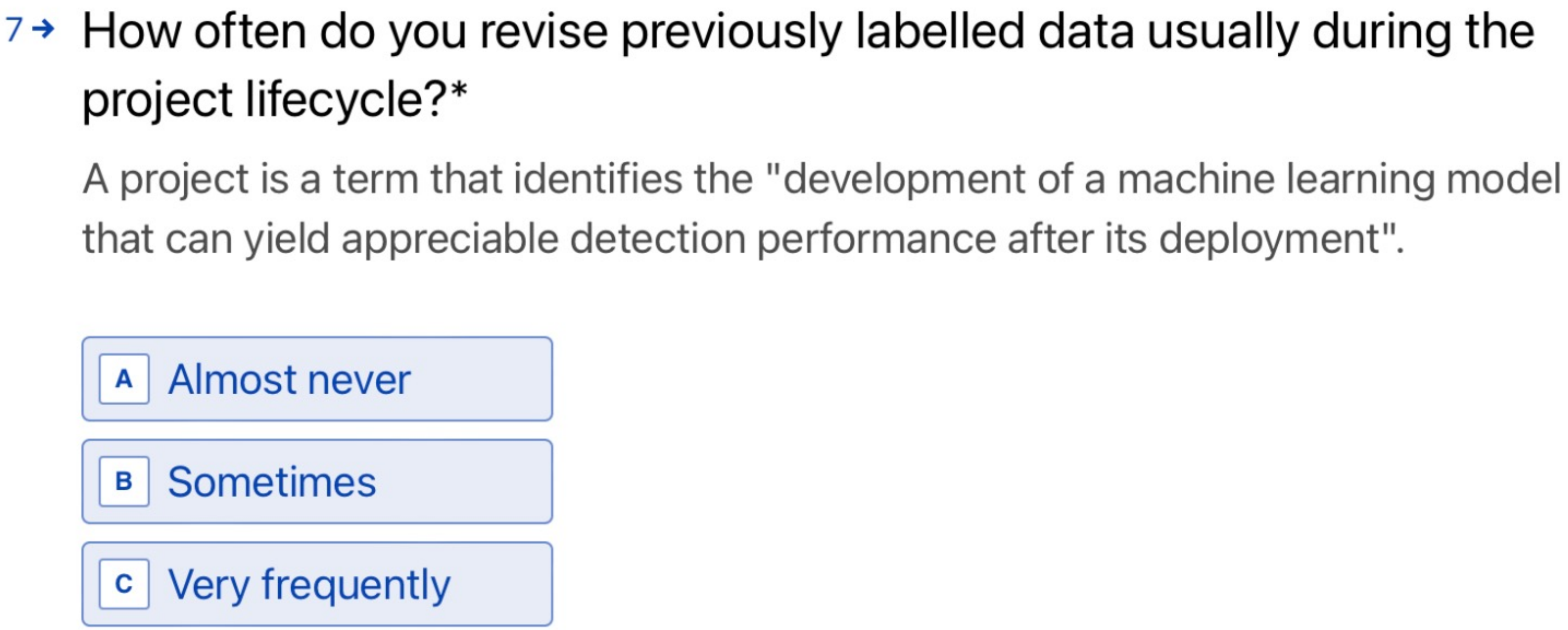
C Post-processing and analysis



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

A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

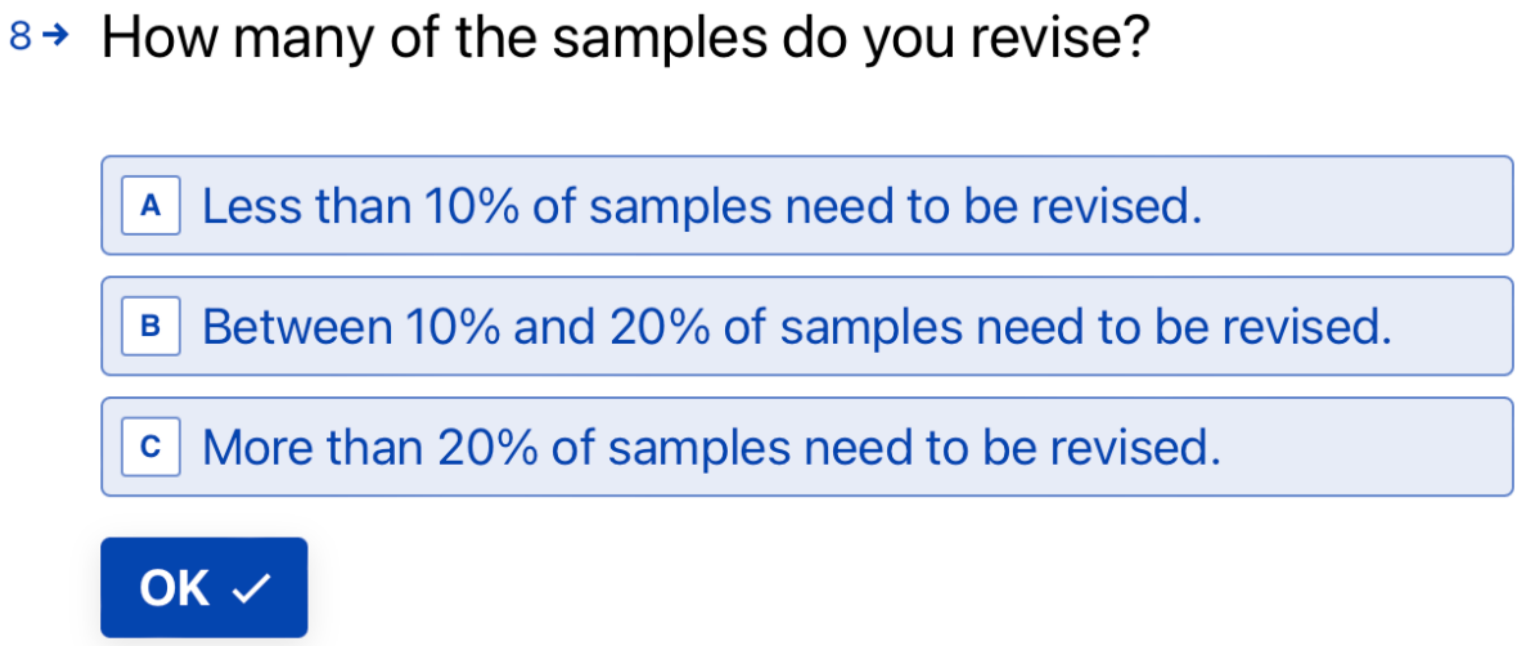
- A Data cleaning
- B Performance assessment
- C Reviewing and approving labels
- D Documentation
- E Tuning



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

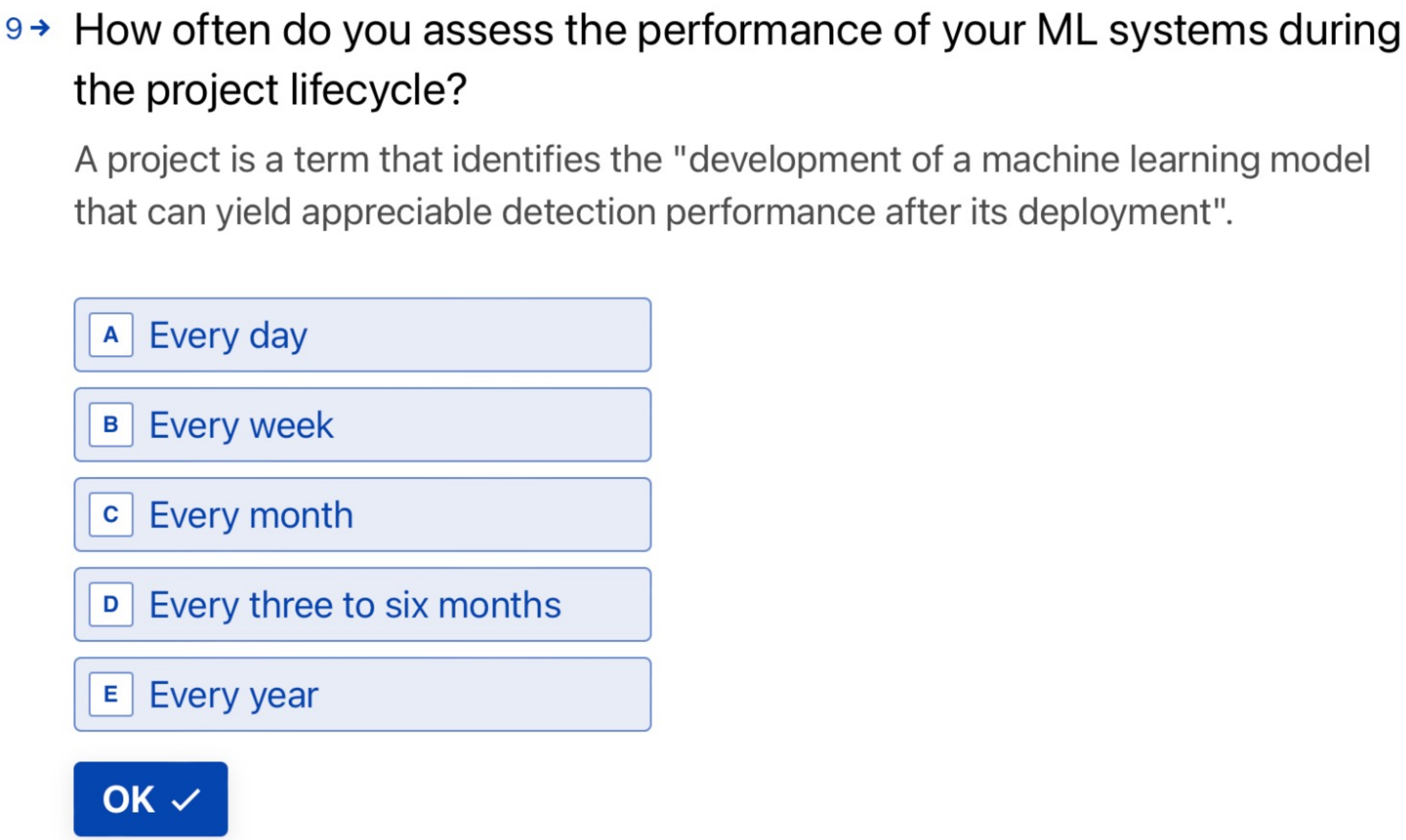
A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

- A Almost never
- B Sometimes
- C Very frequently



8 → How many of the samples do you revise?

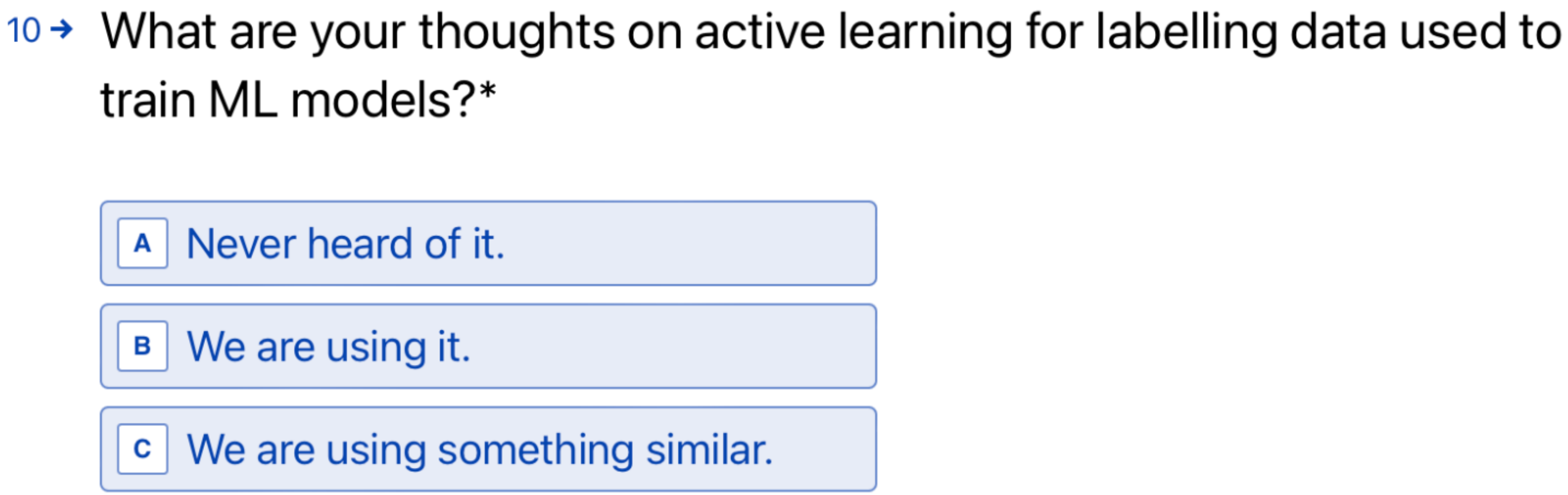
- A Less than 10% of samples need to be revised.
- B Between 10% and 20% of samples need to be revised.
- C More than 20% of samples need to be revised.



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

A project is a term that identifies the "development of a machine learning model that can yield appreciable detection performance after its deployment".

- A Every day
- B Every week
- C Every month
- D Every three to six months
- E Every year

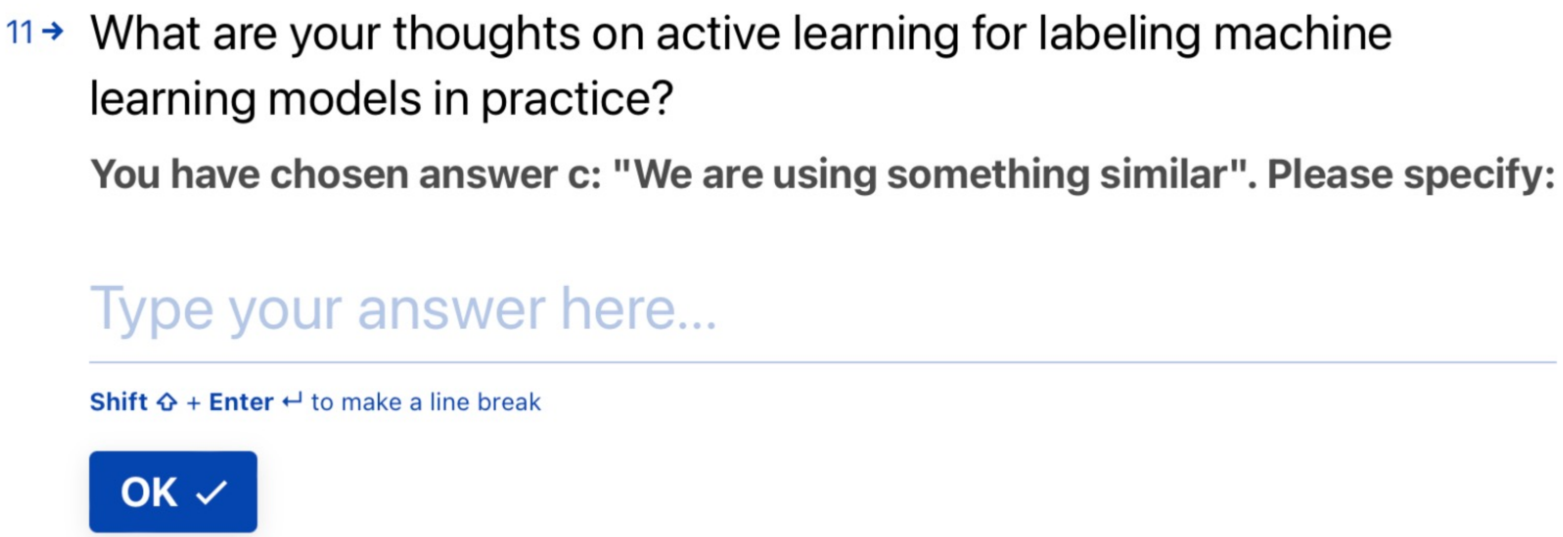


10 → What are your thoughts on active learning for labelling data used to train ML models?*

A Never heard of it.

B We are using it.

C We are using something similar.



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

**You have chosen answer c: "We are using something similar". Please specify:**

Type your answer here...

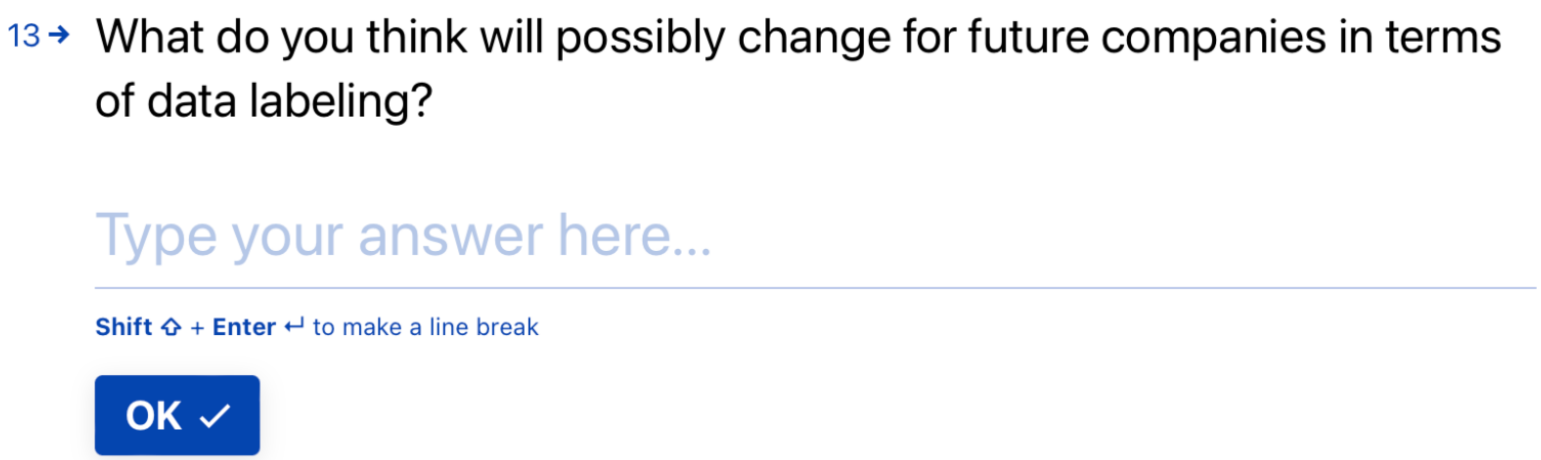
**Shift ⇧** + **Enter ↵** to make a line break

12 → How can data labeling be improved in practice?

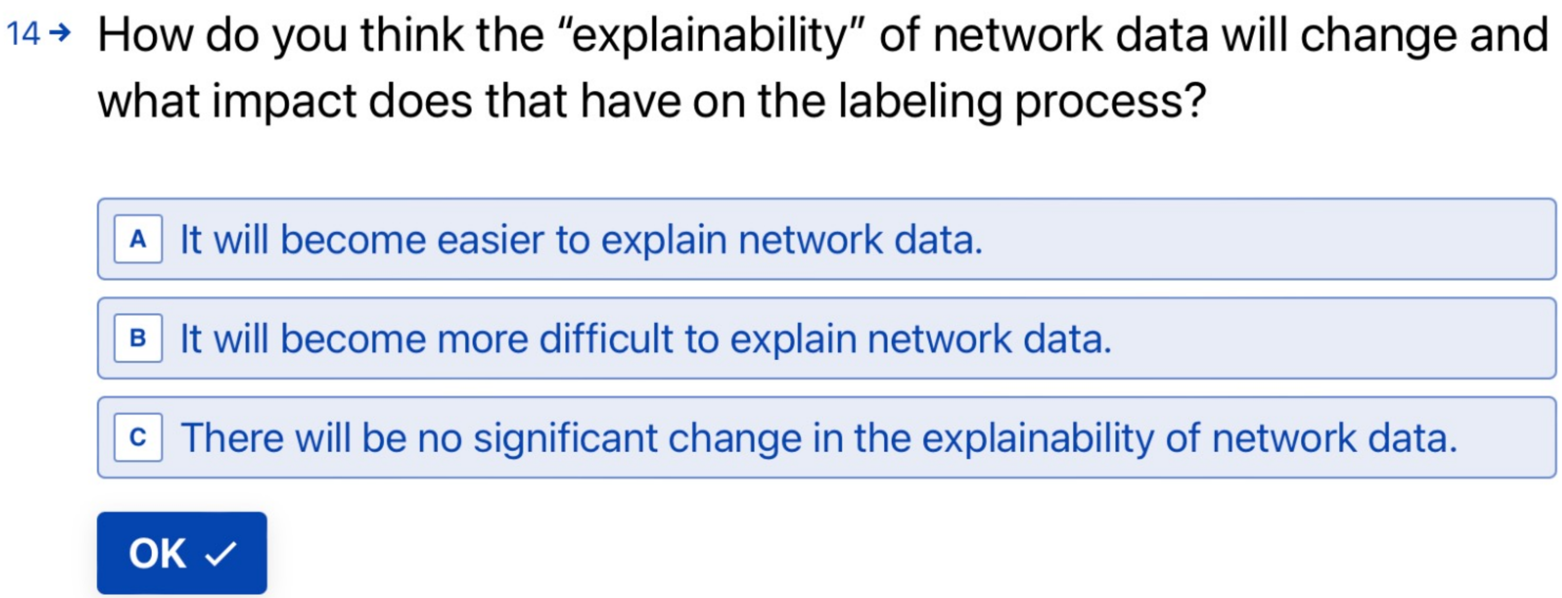
Type your answer here...

**Shift ⇧** + **Enter ↵** to make a line break

**OK ✓**

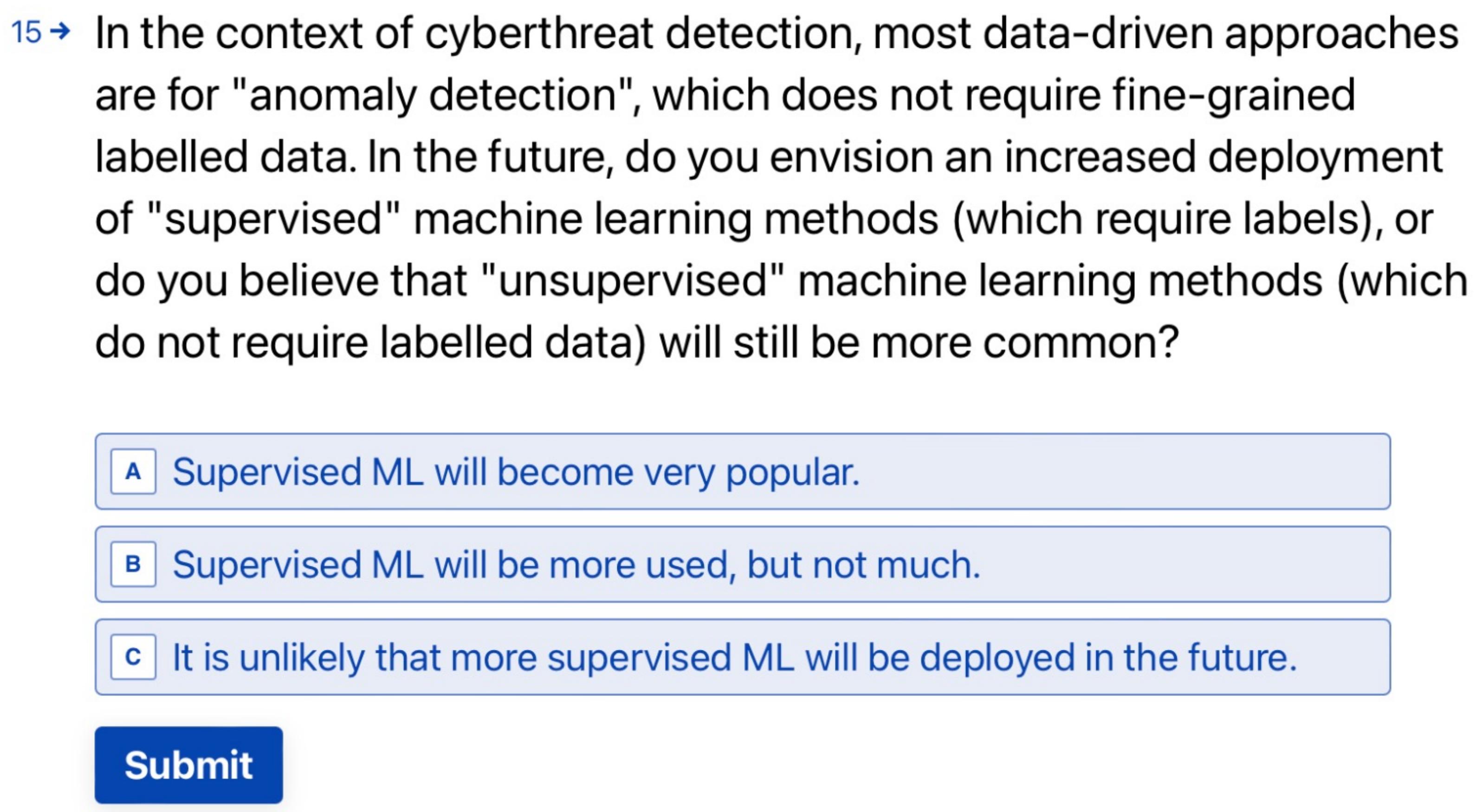


13 → What do you think will possibly change for future companies in terms of data labeling?



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

- A It will become easier to explain network data.
- B It will become more difficult to explain network data.
- C There will be no significant change in the explainability of network data.



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?

A Supervised ML will become very popular.

B Supervised ML will be more used, but not much.

C It is unlikely that more supervised ML will be deployed in the future.