DLS/C3/W1/ QUIZ

Bird Recognition in the City of Peacetopia (Case Study)

1. Problem Statement

This example is adapted from a real production application, but with details disguised to protect confidentiality.

You are a famous researcher in the City of Peacetopia. The people of Peacetopia have a common characteristic: they are afraid of birds. To save them, you have **to build an algorithm that will detect any bird flying over Peacetopia** and alert the population.

The City Council gives you a dataset of 10,000,000 images of the sky above Peacetopia, taken from the city's security cameras. They are labeled:

- y = 0: There is no bird on the image
- y = 1: There is a bird on the image

Your goal is to build an algorithm able to classify new images taken by security cameras from Peacetopia.

There are a lot of decisions to make:

- What is the evaluation metric?
- How do you structure your data into train/dev/test sets?

Metric of success

The City Council tells you that they want an algorithm that

1. Has high accuracy.
2. Runs quickly and takes only a short time to classify a new image.
3. Can fit in a small amount of memory, so that it can run in a small processor that the city will attach to many different security cameras.

Note: Having three evaluation metrics makes it harder for you to quickly choose between two different algorithms, and will slow down the speed with which your team can iterate. True/False?

● True

○ False

✓ **Correct**

2. After further discussions, the city narrows down its criteria to:

- "We **need** an algorithm that can let us know a bird is flying over Peacetopia as accurately as possible."
- "We want the trained model to take no more than 10 sec to classify a new image."
- "We want the model to fit in 10MB of memory."

If you had the three following models, which one would you choose?

○

Test Accuracy	Runtime	Memory size
97%	3 sec	2MB

○

Test Accuracy	Runtime	Memory size
97%	1 sec	3MB

○

Test Accuracy	Runtime	Memory size
99%	13 sec	9MB

●

Test Accuracy	Runtime	Memory size
98%	9 sec	9MB

✓ Correct

Correct! As soon as the runtime is less than 10 seconds you're good. So, you may simply maximize the test accuracy after you make sure the runtime is <10 seconds.

3. Which of the following best answers why it is important to identify optimizing and satisficing metrics?

○ It isn't. All metrics must be met for the model to be acceptable.

○ Identifying the optimizing metric informs the team which models they should try first.

● Identifying the metric types sets thresholds for satisficing metrics. This provides explicit evaluation criteria.

○ Knowing the metrics provides input for efficient project planning.

Expand

✓ Correct

Yes. Thresholds are essential for evaluation of key use case constraints.

4. With 10,000,000 data points, what is the best option for train/dev/test splits?

○ train - 33.3%, dev - 33.3%, test - 33.3%

○ train - 60%, dev - 10%, test - 30%

● train - 95%, dev - 2.5%, test - 2.5%

○ train - 60%, dev - 30%, test - 10%

Expand

✓ **Correct**
Yes. The size of the data set allows for bias and variance evaluation with smaller data sets.

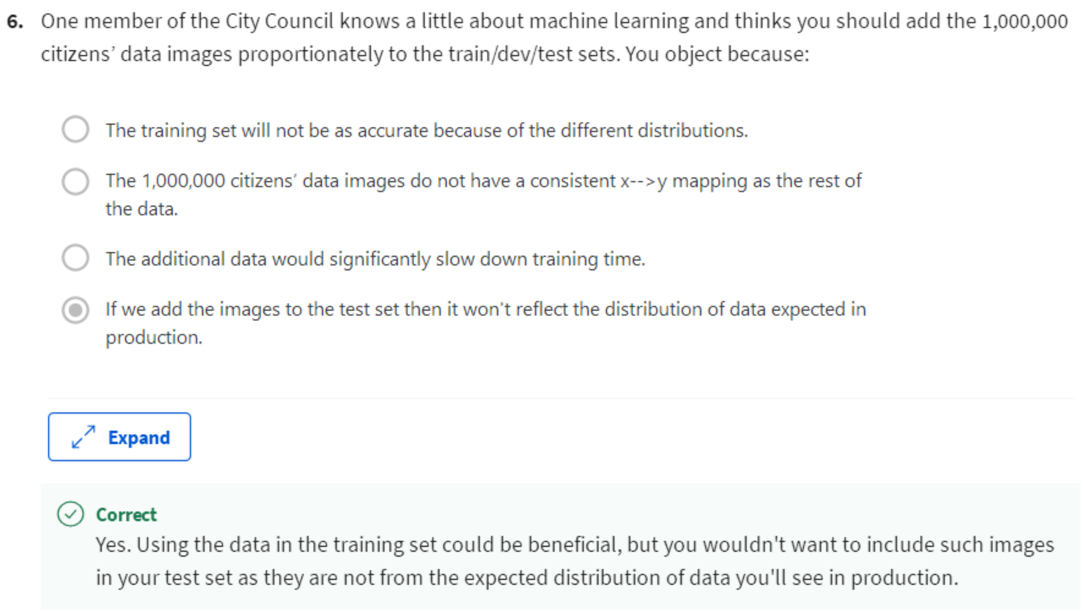
5. Now that you've set up your train/dev/test sets, the City Council comes across another 1,000,000 images from social media and offers them to you. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm. You should add the citizens' data to the training set. True/False?

○ False

● True

Expand

✓ **Correct**
Yes. This will cause the training and dev/test set distributions to become different, however as long as dev/test distributions are the same you are aiming at the same target.



6. One member of the City Council knows a little about machine learning and thinks you should add the 1,000,000 citizens' data images proportionately to the train/dev/test sets. You object because:

○ The training set will not be as accurate because of the different distributions.

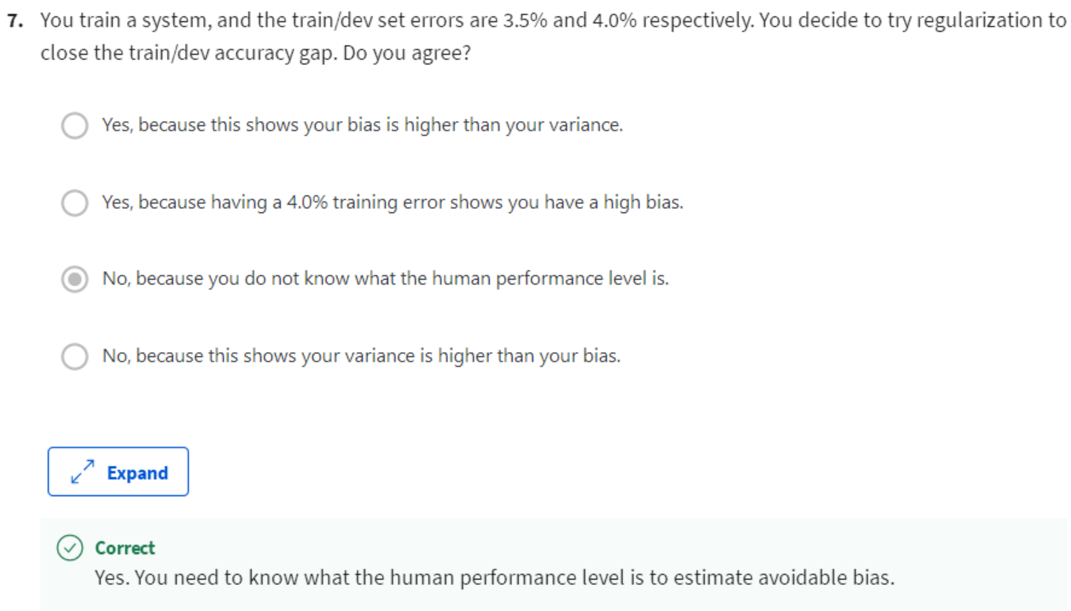
○ The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data.

○ The additional data would significantly slow down training time.

● If we add the images to the test set then it won't reflect the distribution of data expected in production.

✓ **Correct**

Yes. Using the data in the training set could be beneficial, but you wouldn't want to include such images in your test set as they are not from the expected distribution of data you'll see in production.



7. You train a system, and the train/dev set errors are 3.5% and 4.0% respectively. You decide to try regularization to close the train/dev accuracy gap. Do you agree?

○ Yes, because this shows your bias is higher than your variance.

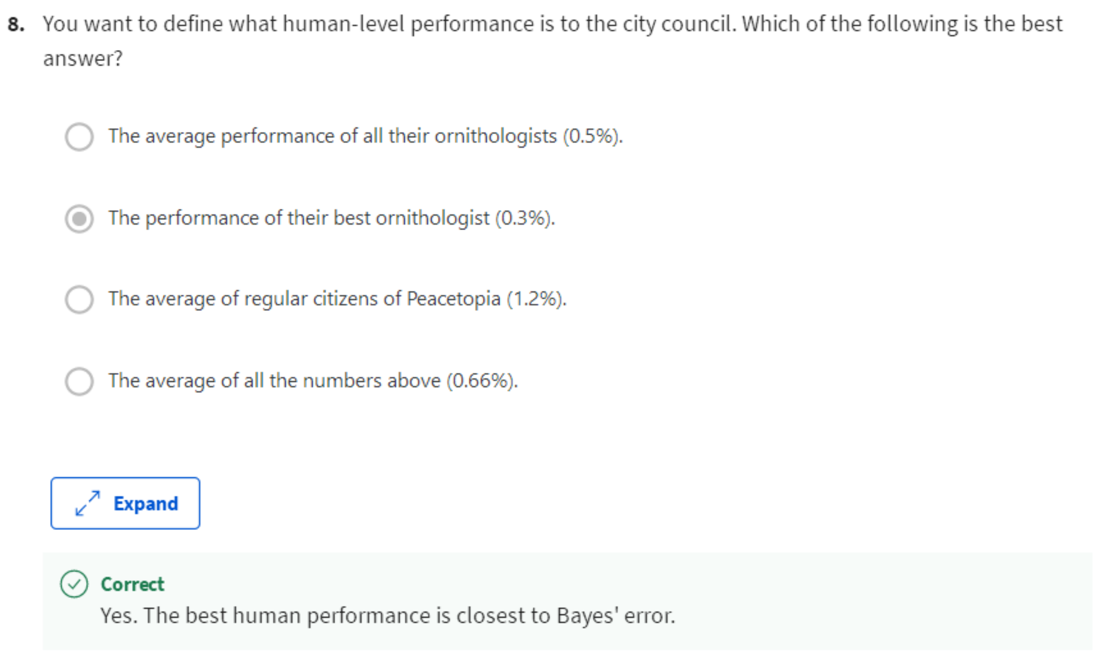
○ Yes, because having a 4.0% training error shows you have a high bias.

● No, because you do not know what the human performance level is.

○ No, because this shows your variance is higher than your bias.

✓ **Correct**

Yes. You need to know what the human performance level is to estimate avoidable bias.



8. You want to define what human-level performance is to the city council. Which of the following is the best answer?

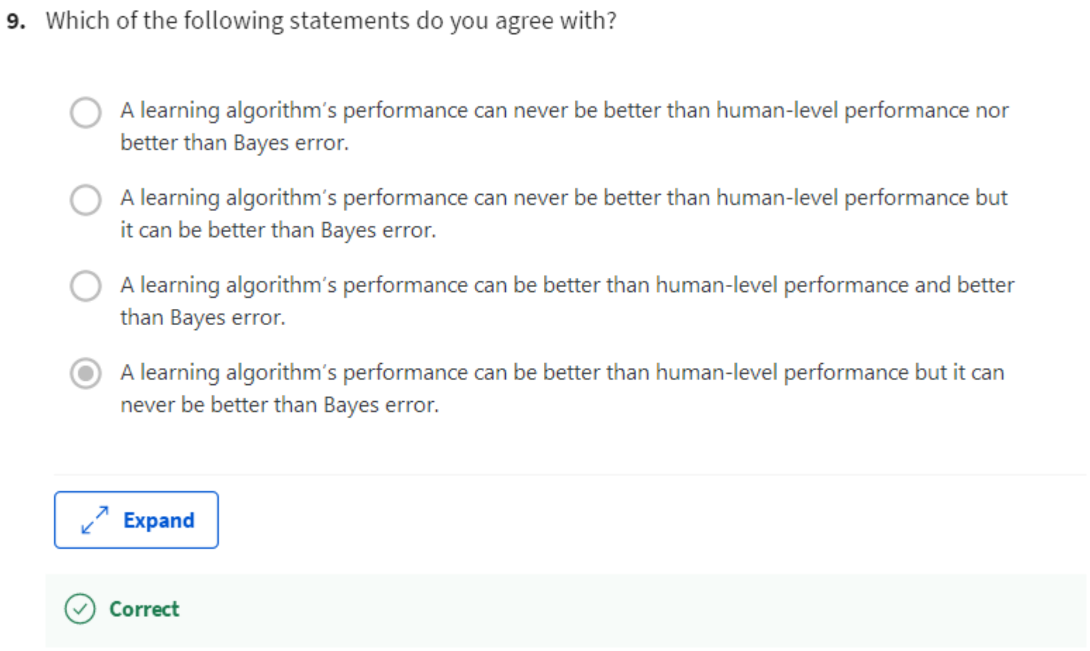
○ The average performance of all their ornithologists (0.5%).

● The performance of their best ornithologist (0.3%).

○ The average of regular citizens of Peacetopia (1.2%).

○ The average of all the numbers above (0.66%).

✓ **Correct**
Yes. The best human performance is closest to Bayes' error.



9. Which of the following statements do you agree with?

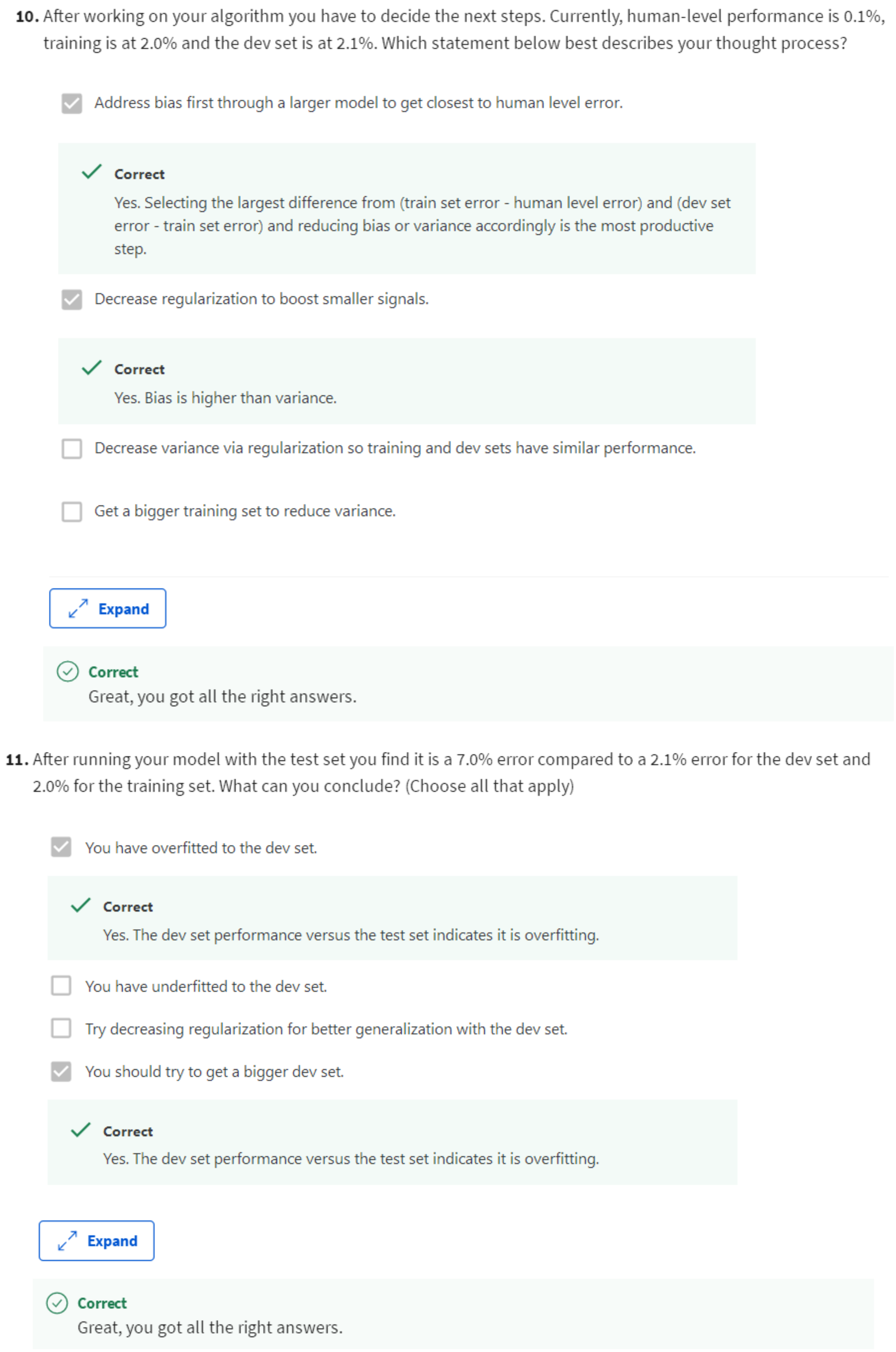
○ A learning algorithm's performance can never be better than human-level performance nor better than Bayes error.

○ A learning algorithm's performance can never be better than human-level performance but it can be better than Bayes error.

○ A learning algorithm's performance can be better than human-level performance and better than Bayes error.

● A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error.

✓ **Correct**



10. After working on your algorithm you have to decide the next steps. Currently, human-level performance is 0.1%, training is at 2.0% and the dev set is at 2.1%. Which statement below best describes your thought process?

☑ Address bias first through a larger model to get closest to human level error.

✓ **Correct**
Yes. Selecting the largest difference from (train set error - human level error) and (dev set error - train set error) and reducing bias or variance accordingly is the most productive step.

☑ Decrease regularization to boost smaller signals.

✓ **Correct**
Yes. Bias is higher than variance.

□ Decrease variance via regularization so training and dev sets have similar performance.

□ Get a bigger training set to reduce variance.

✓ **Correct**
Great, you got all the right answers.

11. After running your model with the test set you find it is a 7.0% error compared to a 2.1% error for the dev set and 2.0% for the training set. What can you conclude? (Choose all that apply)

☑ You have overfitted to the dev set.

✓ **Correct**
Yes. The dev set performance versus the test set indicates it is overfitting.

□ You have underfitted to the dev set.

□ Try decreasing regularization for better generalization with the dev set.

☑ You should try to get a bigger dev set.

✓ **Correct**
Yes. The dev set performance versus the test set indicates it is overfitting.

✓ **Correct**
Great, you got all the right answers.

12. After working on this project for a year, you finally achieve:

Human-level performance	0.10%
Training set error	0.05%
Dev set error	0.05%

What can you conclude? (Check all that apply.)

□ This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.

☑ It is now harder to measure avoidable bias, thus progress will be slower going forward.

✓ **Correct**

□ With only 0.05% further progress to make, you should quickly be able to close the remaining gap to 0%

☑ If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤ 0.05

✓ **Correct**

Expand

Correct
Great, you got all the right answers.

13. It turns out Peacetopia has hired one of your competitors to build a system as well. Your system and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! However, when Peacetopia tries out your and your competitor's systems, they conclude they actually like your competitor's system better, because even though you have higher overall accuracy, you have more false negatives (failing to raise an alarm when a bird is in the air). What should you do?

○ Ask your team to take into account both accuracy and false negative rate during development.

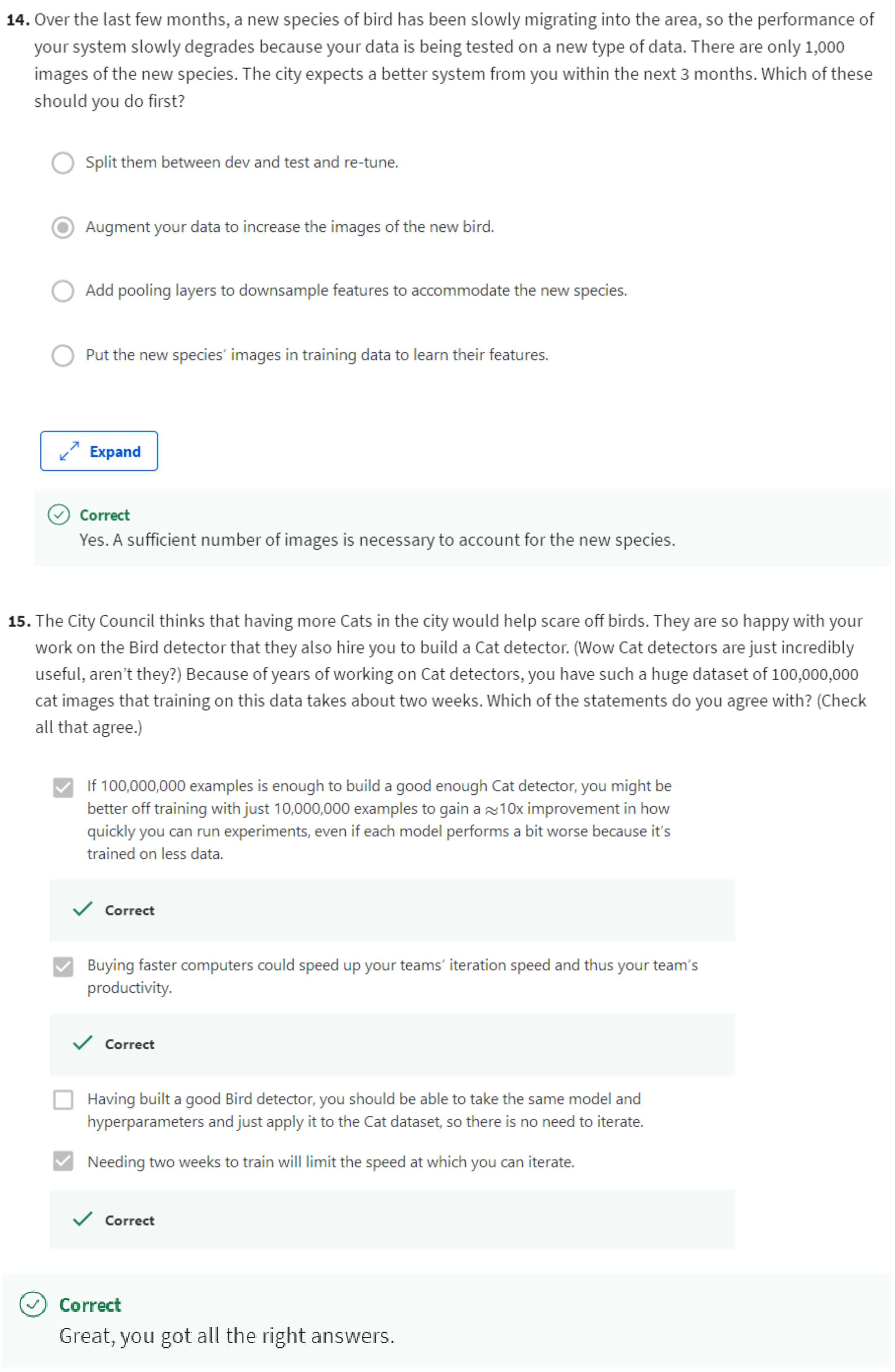
● Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

○ Look at all the models you've developed during the development process and find the one with the lowest false negative error rate.

○ Pick false negative rate as the new metric, and use this new metric to drive all further development.

Expand

Correct



14. Over the last few months, a new species of bird has been slowly migrating into the area, so the performance of your system slowly degrades because your data is being tested on a new type of data. There are only 1,000 images of the new species. The city expects a better system from you within the next 3 months. Which of these should you do first?

○ Split them between dev and test and re-tune.

● Augment your data to increase the images of the new bird.

○ Add pooling layers to downsample features to accommodate the new species.

○ Put the new species' images in training data to learn their features.

Correct

Yes. A sufficient number of images is necessary to account for the new species.

15. The City Council thinks that having more Cats in the city would help scare off birds. They are so happy with your work on the Bird detector that they also hire you to build a Cat detector. (Wow Cat detectors are just incredibly useful, aren't they?) Because of years of working on Cat detectors, you have such a huge dataset of 100,000,000 cat images that training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

☑ If 100,000,000 examples is enough to build a good enough Cat detector, you might be better off training with just 10,000,000 examples to gain a $\approx$10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it's trained on less data.

✓ **Correct**

☑ Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.

✓ **Correct**

□ Having built a good Bird detector, you should be able to take the same model and hyperparameters and just apply it to the Cat dataset, so there is no need to iterate.

☑ Needing two weeks to train will limit the speed at which you can iterate.

✓ **Correct**

Correct

Great, you got all the right answers.