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

1 / 1 point

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 the following 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.

You meet with them and ask for just one evaluation metric. True/False?

○ False

● True:

✓ **Correct**
Yes. The goal is to have one metric that focuses the development effort and increases iteration velocity.

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

1 / 1 point

- "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%	1 sec	3MB

○

Test Accuracy	Runtime	Memory size
99%	13 sec	9MB

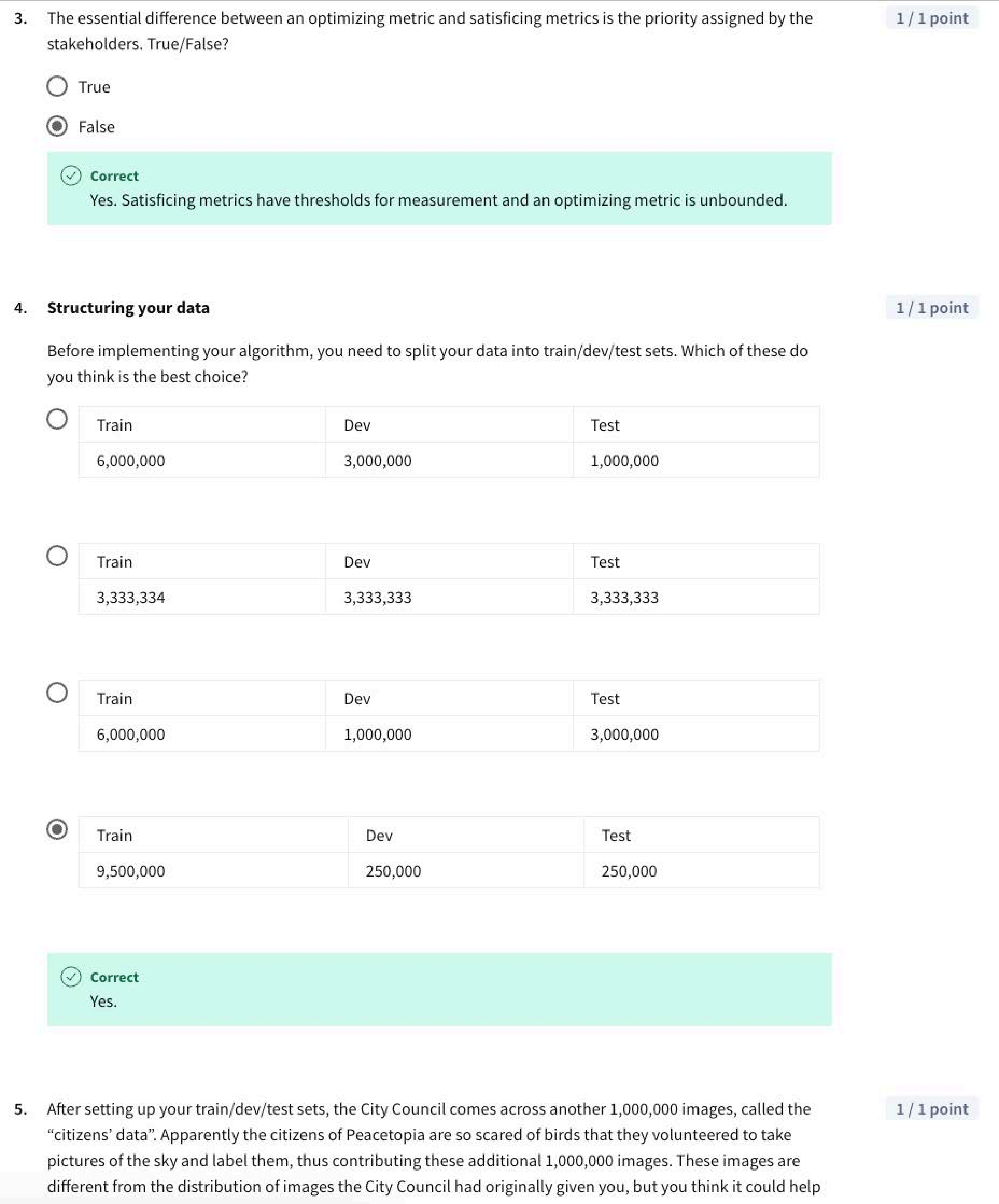
○

Test Accuracy	Runtime	Memory size
97%	3 sec	2MB

●

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. The essential difference between an optimizing metric and satisficing metrics is the priority assigned by the stakeholders. True/False?

1 / 1 point

○ True

● False

Correct

Yes. Satisficing metrics have thresholds for measurement and an optimizing metric is unbounded.

4. **Structuring your data**

1 / 1 point

Before implementing your algorithm, you need to split your data into train/dev/test sets. Which of these do you think is the best choice?

○

Train	Dev	Test
6,000,000	3,000,000	1,000,000

○

Train	Dev	Test
3,333,334	3,333,333	3,333,333

○

Train	Dev	Test
6,000,000	1,000,000	3,000,000

●

Train	Dev	Test
9,500,000	250,000	250,000

Correct

Yes.

5. After setting up your train/dev/test sets, the City Council comes across another 1,000,000 images, called the "citizens' data". Apparently the citizens of Peacetopia are so scared of birds that they volunteered to take pictures of the sky and label them, thus contributing these additional 1,000,000 images. These images are different from the distribution of images the City Council had originally given you, but you think it could help

1 / 1 point

different from the distribution of images the City Council had originally given you, but you think it could help your algorithm.

Notice that adding this additional data to the training set will make the distribution of the training set different from the distributions of the dev and test sets.

Is the following statement true or false?

"You should not add the citizens' data to the training set, because if the training distribution is different from the dev and test sets, then this will not allow the model to perform well on the test set."

○ True

● False

> ✓ **Correct**
> False is correct: Sometimes we'll need to train the model on the data that is available, and its distribution may not be the same as the data that will occur in production. Also, adding training data that differs from the dev set may still help the model improve performance on the dev set. What matters is that the dev and test set have the same distribution.

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 to the dev set. You object because: (Choose all that apply)

1 / 1 point

☑ The dev set no longer reflects the distribution of data (security cameras) you most care about.

> ✓ **Correct**
> Yes. The performance of the model should be evaluated on the same distribution of images it will see in production.

□ A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.

☑ This would cause the dev and test set distributions to become different. This is a bad idea because you're not aiming where you want to hit.

> ✓ **Correct**
> Yes. Adding a different distribution to the dev set will skew bias.

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

7. You train a system, and its errors are as follows (error = 100%-Accuracy):

1 / 1 point

Training set error	4.0%
Dev set error	4.5%

Dev set error	4.5%

This suggests that one good avenue for improving performance is to train a bigger network so as to drive down the 4.0% training error. Do you agree?

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

● No, because there is insufficient information to tell.

○ 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.

✓ **Correct**

8. If your goal is to have “human-level performance” be a proxy (or estimate) for Bayes error, how would you define “human-level performance”?

1 / 1 point

● The best performance of a specialist (ornithologist) or possibly a group of specialists.

○ The performance of their volunteer amateur ornithologists.

○ The performance of the head of the City Council.

○ The performance of the average citizen of Peacetopia.

✓ **Correct**
Yes. This is the peak of human performance in this task.

9. Which of the below shows the optimal order of accuracy from worst to best?

1 / 1 point

○ The learning algorithm’s performance -> human-level performance -> Bayes error.

● Human-level performance -> the learning algorithm’s performance -> Bayes error.

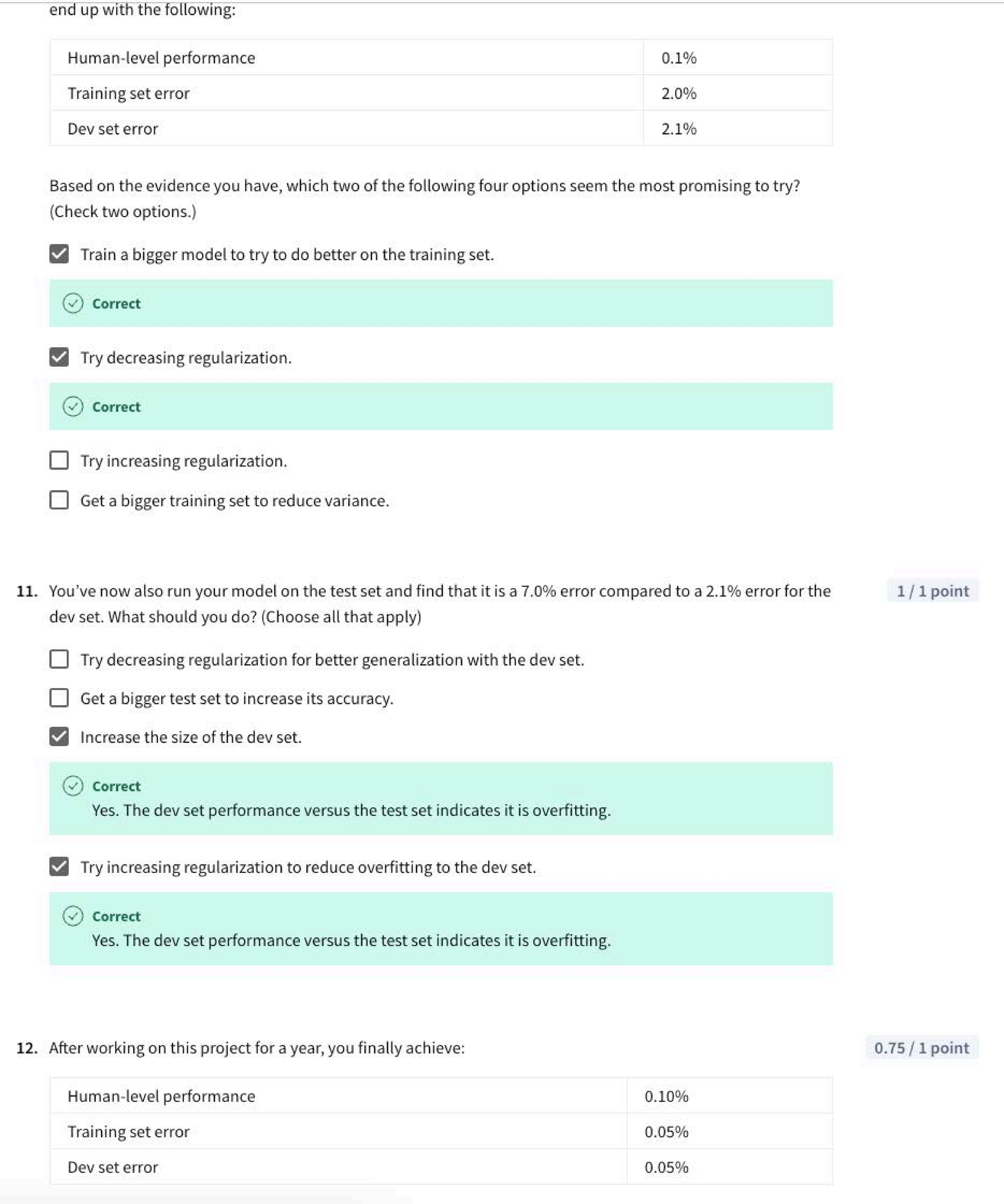
○ Human-level performance -> Bayes error -> the learning algorithm’s performance.

○ The learning algorithm’s performance -> Bayes error -> human-level performance.

✓ **Correct**
Yes. A learning algorithm’s performance can be better than human-level performance but it can never be better than Bayes error.

10. You find that a team of ornithologists debating and discussing an image gets an even better 0.1% performance, so you define that as “human-level performance.” After working further on your algorithm, you end up with the following:

1 / 1 point



end up with the following:

Human-level performance	0.1%
Training set error	2.0%
Dev set error	2.1%

Based on the evidence you have, which two of the following four options seem the most promising to try? (Check two options.)

☑ Train a bigger model to try to do better on the training set.

✓ **Correct**

☑ Try decreasing regularization.

✓ **Correct**

□ Try increasing regularization.

□ Get a bigger training set to reduce variance.

11. You've now also run your model on the test set and find that it is a 7.0% error compared to a 2.1% error for the dev set. What should you do? (Choose all that apply)

1 / 1 point

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

□ Get a bigger test set to increase its accuracy.

☑ Increase the size of the dev set.

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

☑ Try increasing regularization to reduce overfitting to the dev set.

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

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

0.75 / 1 point

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

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

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

Correct

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

Correct

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

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

This should not be selected

13. It turns out Peacetopia has hired one of your competitors to build a system as well. You and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! Still, when Peacetopia tries out both systems, they conclude they 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?

1 / 1 point

● Brainstorm with your team to refine the optimizing metric to include false negatives as they further develop the model.

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

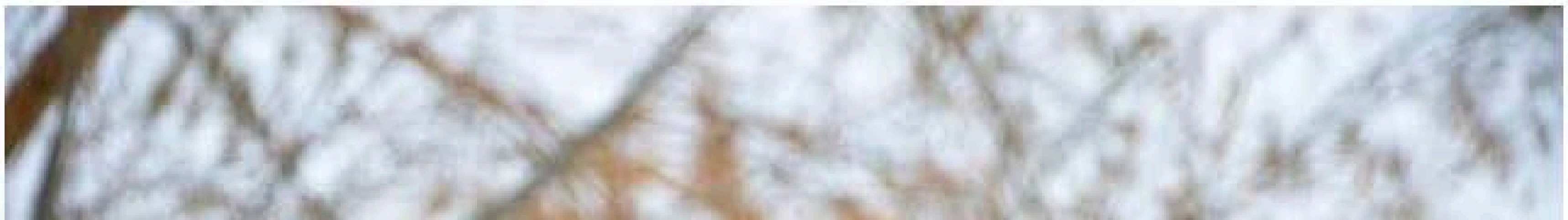
○ Pick false negative rate as the new metric, and use this new metric to drive all further development.

○ Apply regularization to minimize the false negative rate.

Correct
Yes. The target has shifted so an updated metric is required.

14. You've handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the citizens from birds! But 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.

0 / 1 point



You have only 1,000 images of the new species of bird. The city expects a better system from you within the next 3 months. Which of these should you do first?

○ Add the 1,000 images into your dataset and reshuffle into a new train/dev/test split.

○ Use the data you have to define a new evaluation metric (using a new dev/test set) taking into account the new species, and use that to drive further progress for your team.

○ Put the 1,000 images into the training set so as to try to do better on these birds.

● Try data augmentation/data synthesis to get more images of the new type of bird.

> ⊗ **Incorrect**
> The true data distribution is changed. It means you need to adjust your evaluation. Because you evaluate your learning algorithm on dev and test sets, adding more data only to the training set doesn't help the algorithm to perform better.

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. You have a huge dataset of 100,000,000 cat images. Training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

1 / 1 point

☐ Reducing the model complexity will allow the use of the larger data set but preserve accuracy.

☑ Lowering the number of images will reduce training time and likely allow for an acceptable tradeoff between iteration speed and accuracy.

> ✓ **Correct**
> Yes. There is a sweet spot that allows development at a reasonable rate without significant accuracy loss.

☑ This significantly impacts iteration speed.

> ✓ **Correct**
> Yes. This training time is an absolute constraint on iteration.