1. This scenario is adapted from a real-world application, with specifics altered for confidentiality. 1 point

As a distinguished researcher in the City of Peacetopia, you face a unique challenge. The citizens are universally afraid of birds, and your task is to develop an algorithm to detect birds flying over the city and alert the populace.

The City Council has provided you with 10,000,000 sky images from Peacetopia's security cameras, labeled as follows:

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

Your objective is to engineer an algorithm that can accurately classify new security camera images. Decisions regarding the evaluation metric and data structuring into train/dev/test sets are critical.

The City Council has outlined their desires for the algorithm:

1. High accuracy.
2. Quick response time for classifying new images.
3. Low memory requirement to function on small processors across various cameras.

True or False: You realize that having a single evaluation metric will expedite development and ease algorithm comparison.

- ○ False
- ◉ True

2. The city asks for your help in further defining the criteria for accuracy, runtime, and memory. How would you suggest they identify the criteria? 1 point

- ◉ Suggest to them that they define which criterion is to be optimized. Then, set thresholds for the other two.
- ○ Suggest to them that they focus on whichever criterion is to be optimized and then eliminate the other two.
- ○ Suggest that they purchase more infrastructure to ensure the model runs quickly and accurately.

3. Based on the context of a city's data analysis project, **which of the following statements is true regarding the metrics used?** 1 point

- ○ Accuracy, running time, and memory size are all satisficing metrics because you have to do sufficiently well on all three for your system to be acceptable.
- ○ Accuracy is a satisficing metric; running time and memory size are an optimizing metric.
- ○ Accuracy, running time, and memory size are all optimizing metrics because you want to do well on all three.
- ◉ Accuracy is an optimizing metric; running time and memory size are satisficing metrics.

4. You propose a 95% / 2.5% / 2.5% for train / dev / test splits to the City Council. They ask for your reasoning. 1 point

Which of the following **best justifies your proposal**, given that the total data set contains 10,000,000 data points?

- ○ The emphasis on the training set provides the most accurate model, supporting the memory and processing efficiency.
- ○ The emphasis on the training set will allow us to iterate faster.
- ○ The most important goal is achieving the highest accuracy, and that can be done by allocating the maximum amount of data to the training set.
- ◉ With a dataset comprising 10,000,000 individual samples, 2.5% represents 250,000 samples, which should be more than enough for dev and testing to evaluate bias and variance.

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 have a different distribution from the images the City Council originally provided, but you think they 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 distribution of the dev and test sets. 1 point

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

6. One member of the City Council wants to add 1,000,000 citizen data images to the test set. Your original data is from security cameras, and you object because: (Choose all that apply) 1 point

- ☐ A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.
- ☑ The test set no longer reflects the distribution of data (security cameras) you most care about.

- [ ] The 1,000,000 citizen data images do not have a consistent input-output relationship as the security camera data.
- [x] 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.

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 error gap. Do you agree? 1 point

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

8. You ask a few people to label a bird species dataset to determine human-level performance. The following error rates were recorded: 1 point

Bird watching expert #1	0.3% error
Bird watching expert #2	0.5% error
Normal person #1 (not a bird watching expert)	1.0% error
Normal person #2 (not a bird watching expert)	1.2% error

If your goal is to use "human-level performance" as an estimate for Bayes error, how would you define "human-level performance" in this scenario?

- ○ 0.75% (Average of all four error rates)
- ○ 0.4% (Average of the two experts' error rates)
- ○ 0.0% (Perfect accuracy, representing an unattainable ideal)
- ◉ 0.3% (The lowest error rate achieved by an expert)

9. **True or False:** A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error. 1 point

- ◉ True.
- ○ False.

10. Which of the following best describes the **most effective next step in your project**, given the following performance metrics? 1 point

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

- ○ Evaluate the test set to determine the variance.
- ◉ Prioritize actions to decrease bias by increasing model complexity, as the training error significantly exceeds human-level performance.
- ○ Continue tuning until the training set error matches human-level performance, focusing solely on the optimizing metric.
- ○ Deploy the model to target devices to evaluate against satisficing metrics.

11. After running your model with the test set, you find the error rate is 7.0% compared to a 2.1% error rate for the dev set and 2.0% for the training set. What can you conclude? (Choose all that apply) 1 point

- [ ] Try decreasing regularization for better generalization with the dev set.
- [x] You have overfitted to the dev set.
- [x] You should try to get a bigger dev set.
- [ ] You have underfitted to the dev set.

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%. Which of the following are likely? (Check all that apply.) 1 point

- [x] Pushing to even higher accuracy will be slow because you will not be able to easily identify sources of bias.
- [x] The model has recognized complex, emergent features that humans may not readily perceive. (Chess and Go, for example).
- [ ] This result is not possible since it should not be possible to surpass human-level performance.
- [ ] There is still avoidable bias.

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! 1 point

Still, when Peacetopia tries out both your system and your competitor's system, 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?

- ○ Look at all the models you've developed during the development process and find the one with the lowest false negative error rate.
- ○ 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.
- ○ Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

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. 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.

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
- ○ Try data augmentation/data synthesis to get more images of the new type of bird.
- ○ Use the data you have to define a new evaluation metric (using a new dev/test set) that accounts for the new species, and use that metric to guide further improvements.
- ◉ Put the 1,000 images into the training set so as to try to do better on these birds.

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. 1 point

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.)

- ☑ You could consider a tradeoff where you use a subset of the cat data to find reasonable performance with reasonable iteration pacing.
- ☐ Accuracy should exceed the City Council's requirements, but the project may take as long as the bird detector because of the two-week training/iteration time.
- ☐ With the experience gained from the Bird detector, you are confident to build a good Cat detector on the first try.
- ☑ Given a significant budget for cloud GPUs, you could mitigate the training time.