1. **Problem Statement** 0 / 1 point

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 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 are delighted because this list of criteria will speed development and provide guidance on how to evaluate two different algorithms. True/False?

- (●) True:
- ( ) False

⊗ **Incorrect**
No. The goal is to have one metric that focuses the development effort and increases iteration velocity.

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 / 1 point**

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

Expand

Correct
Yes. The thresholds provide a way to evaluate models head to head.

3. Based on the city's requests, which of the following would you say is true? **1 / 1 point**

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

Expand

Correct

4. You propose a 95/2.5%/2.5% for train/dev/test splits to the City Council. They ask for your reasoning. Which of the following best justifies your proposal? **1 / 1 point**

- 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 10M individual samples, 2.5% represents 250k samples, which should be more than enough for dev and testing to evaluate bias and variance.
- The emphasis on the training set provides the most accurate model, supporting the memory and processing satisficing metrics.

Expand

Correct
Yes. The purpose of dev and test sets is fulfilled even with smaller percentages of the data.

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? **1 / 1 point**

- (•) True
- ( ) False

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 to the test set. You object because: **1 / 1 point**

- [x] The test set no longer reflects the distribution of data (security cameras) you most care about.

Correct

- [ ] The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data.
- [ ] A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.
- [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.

Correct

Expand

Correct
Great, you got all the right answers.

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%

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 there is insufficient information to tell.
- ( ) Yes, because having a 4.0% training error shows you have a high bias.
- ( ) No, because this shows your variance is higher than your bias.
- ( ) Yes, because this shows your bias is higher than your variance.

Expand

Correct

8. You ask a few people to label the dataset so as to find out what is human-level performance. You find the following levels of accuracy: **1 / 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 have "human-level performance" be a proxy (or estimate) for Bayes error, how would you define "human-level performance"?

- ○ 0.4% (average of 0.3 and 0.5)
- ◉ 0.3% (accuracy of expert #1)
- ○ 0.75% (average of all four numbers above)
- ○ 0.0% (because it is impossible to do better than this)

Expand

✓ **Correct**

9. Which of the below shows the optimal order of accuracy from worst to best? **0 / 1 point**

- ◉ The learning algorithm's performance -> human-level performance -> Bayes error.
- ○ Human-level performance -> Bayes error -> the learning algorithm's performance.
- ○ The learning algorithm's performance -> Bayes error -> human-level performance.
- ○ Human-level performance -> the learning algorithm's performance -> Bayes error.

Expand

✗ **Incorrect**
No. in an optimal scenario, your algorithm's performance would be better than HLP but it can never be better than BE.

10. Which of the following best expresses how to evaluate the next steps in your project when your results for human-level performance, train, and dev set error are 0.1%, 2.0%, and 2.1% respectively? **1 / 1 point**

- ○ Evaluate the test set to determine the magnitude of the variance.
- ◉ Based on differences between the three levels of performance, prioritize actions to decrease bias and iterate.
- ○ Keep tuning until the train set accuracy is equal to human-level performance because it is the optimizing metric.
- ○ Port the code to the target devices to evaluate if your model meets or exceeds the satisficing metrics.

Expand

✓ **Correct**
Yes. Always choose the area with the biggest opportunity for improvement.

11. You also evaluate your model on the test set, and find the following: **1 / 1 point**

Human-level performance	0.1%
Training set error	2.0%
Dev set error	2.1%
Test set error	7.0%

What does this mean? (Check the two best options.)

- [x] You have overfit to the dev set.

✓ **Correct**

- [ ] You have underfitted to the dev set.
- [ ] You should get a bigger test set.
- [x] You should try to get a bigger dev set.

✓ **Correct**

12. After working on this project for a year, you finally achieve: **1 / 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.)

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

✓ **Correct**

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

⤢ **Expand**

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

13. Your system is now very accurate but has a higher false negative rate than the City Council of Peacetopia would like. What is your best next step? **0 / 1 point**

- ( ) Expand your model size to account for more corner cases.
- (•) Pick false negative rate as the new metric, and use this new metric to drive all further development.
- ( ) Reset your "target" (metric) for the team and tune to it.
- ( ) Look at all the models you've developed during the development process and find the one with the lowest false negative error rate.

⤢ **Expand**

✗ **Incorrect**
No. This choice also points to the incorrect target.

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?

1 / 1 point

- (●) Augment your data to increase the images of the new bird.
- ( ) Split them between dev and test and re-tune.
- ( ) Add pooling layers to downsample features to accommodate the new species.
- ( ) Put the new species' images in training data to learn their features.

Expand

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. 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

- [x] You could consider a tradeoff where you use a subset of the cat data to find reasonable performance with reasonable iteration pacing.

Correct
Yes. This is similar to satisficing metrics where "good enough" determines the size of the data.

- [x] Given a significant budget for cloud GPUs, you could mitigate the training time.

Correct
Yes. More resources will allow you to iterate faster.

- [ ] With the experience gained from the Bird detector you are confident to build a good Cat detector on the first try.
- [x] 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.

Correct
Yes. The 10x size increase adds a small amount of accuracy but takes too much time.