Bird Recognition in the City of Peacetopia (Case Study)

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 focus on whichever criterion is important and then eliminate the other two.
- (selected) Suggest to them that they define which criterion is most important. Then, set thresholds for the other two.

Expand

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

3. Which of the following best answers why it is important to identify optimizing and satisficing metrics? 1 / 1 point

- Identifying the optimizing metric informs the team which models they should try first.
- It isn't. All metrics must be met for the model to be acceptable.
- (selected) 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. **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
9,500,000	250,000	250,000

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

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

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

Expand

Correct
Yes.

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. Which of the following is the best use of that additional data? 1 / 1 point

- Split it among train/dev/test equally.
- Add it to the training set.
- Add it to the dev set to evaluate how well the model generalizes across a broader set.
- Do not use the data. It will change the distribution of any set it is added to.

Expand

Correct
Yes. It is not a problem to have different training and dev distributions. Different dev and test distributions would be an issue.

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"?

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

Expand

✓ Correct

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

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

Expand

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

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.) 0 / 1 point

- [x] Pushing to even higher accuracy will be slow because you will not be able to easily identify sources of bias.

  ✓ **Correct**
  Yes. Exceeding human performance means you are close to Bayes error.

- [ ] This result is not possible since it should not be possible to surpass human-level performance.
- [x] There is still avoidable bias.

  ! **This should not be selected**
  No. Exceeding human performance makes the identification of avoidable bias very challenging.

- [x] The model has recognized emergent features that humans cannot. (Chess and Go for example)

  ✓ **Correct**
  Yes. When Google beat the world Go champion, it was recognized that it was making deeper moves than humans.

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? 1 / 1 point

- ( ) Pick false negative rate as the new metric, and use this new metric to drive all further development.
- ( ) 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.
- (•) Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

Expand

✓ **Correct**

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.) 1 / 1 point

- [x] Needing two weeks to train will limit the speed at which you can iterate.

  ✓ Correct

- [x] 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

- [ ] 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.
- [x] Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.

  ✓ Correct

Expand

Correct
Great, you got all the right answers.

2. The city revises 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."

Given models with different accuracies, runtimes, and memory sizes, how would you choose one?

- ( ) Create one metric by combining the three metrics and choose the best performing model.
- ( ) Take the model with the smallest runtime because that will provide the most overhead to increase accuracy.
- (•) Find the subset of models that meet the runtime and memory criteria. Then, choose the highest accuracy.
- ( ) Accuracy is an optimizing metric, therefore the most accurate model is the best choice.

Expand

Correct
Yes. Once you meet the runtime and memory thresholds, accuracy should be maximized.

3. Which of the following best answers why it is important to identify optimizing and satisficing metrics? 1 / 1 point

- ○ Knowing the metrics provides input for efficient project planning.
- ○ 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.
- ○ It isn't. All metrics must be met for the model to be acceptable.

Expand

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

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? 1 / 1 point

- ○ 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 this shows your variance is higher than your bias.
- ◉ No, because you do not know what the human performance level is.

Expand

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

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"?

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

Expand

✓ Correct

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

- ○ 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.
- ○ 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.

Expand

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

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.)

- [ ] This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.
- [x] It is now harder to measure avoidable bias, thus progress will be slower going forward.

✓ Correct

- [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

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

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. 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.
- ( ) Apply regularization to minimize the false negative rate.
- ( ) Pick false negative rate as the new metric, and use this new metric to drive all further development.
- ( ) Ask your team to take into account both accuracy and false negative rate during development.

Expand

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

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

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

Expand

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

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

- ☐ 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.
- ☑ 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**

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

✓ **Correct**

Expand

Correct
Great, you got all the right answers.

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
98%	9 sec	9MB

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

Expand

Correct
Correct! This model has the highest test accuracy, the prominent criteria you are looking for, compared with other models, and also has a runtime <10 seconds and memory size < 10MB.

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.
- 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.
- 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.

Expand

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

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

- (●) The performance of their best ornithologist (0.3%).
- ( ) The average of regular citizens of Peacetopia (1.2%).
- ( ) The average performance of all their ornithologists (0.5%).
- ( ) The average of all the numbers above (0.66%).

Expand

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 be better than human-level performance but it can never be 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 never be better than human-level performance nor better than Bayes error.
- ( ) A learning algorithm's performance can be better than human-level performance and better than Bayes error.

Expand

Correct

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?

- ○ Port the code to the target devices to evaluate if your model meets or exceeds the satisficing metrics.
- ◉ 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.
- ○ Evaluate the test set to determine the magnitude of the variance.

Expand

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

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)

- ☐ 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.

Expand

Correct
Great, you got all the right answers.

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 / 1

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.
- ○ Try data augmentation/data synthesis to get more images of the new type of bird.
- ○ Put the 1,000 images into the training set so as to try to do better on these birds.

Expand

✓ Correct

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.)

- [x] This significantly impacts iteration speed.

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

- [ ] Reducing the model complexity will allow the use of the larger data set but preserve accuracy.

- [x] 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.

Expand

Correct
Great, you got all the right answers.

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: 1 / 1 point

- ○ The training set will not be as accurate because of the different distributions.
- ○ The additional data would significantly slow down training time.
- ○ The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data.
- ◉ If we add the images to the test set then it won't reflect the distribution of data expected in production.

Expand

✓ 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.

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 performance of the average citizen of Peacetopia.
- ○ The performance of the head of the City Council.
- ◉ The best performance of a specialist (ornithologist) or possibly a group of specialists.
- ○ The performance of their volunteer amateur ornithologists.

Expand

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