Bird Recognition in the City of Peacetopia (Case Study)

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

- [x] True
- [ ] False

Note: The goal is to have one metric that focuses the development effort and increases iteration velocity. 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.

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
98%	9 sec	9MB

Test Accuracy	Runtime	Memory size
99%	13 sec	9MB

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

- [ ] 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.
- [x] Accuracy is an optimizing metric; running time and memory size

are a satisficing metrics.
- [ ] 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.

4. **Structuring your data**

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	3,000,000	1,000,000

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

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

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
- [x] True

Note: 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:

- [x] The test set no longer reflects the distribution of data (security cameras) you most care about.
- [ ] 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.
- [ ] The 1,000,000 citizens' data images do not have a consistent x—>y mapping as the rest of the data (similar to the New York City/Detroit housing prices example from the lecture).

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

- 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 this shows your variance is higher than your bias.
- [ ] 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.
- [x] No, because there is insufficient information to tell.

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

- [ ] The average of regular citizens of Peacetopia (1.2%).
- [ ] The average performance of all their ornithologists (0.5%).

- [x] The performance of their best ornithologist (0.3%).
- [ ] The average of all the numbers above (0.66%).

Note: The best human performance is closest to Bayes' error.

9. A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error. True/False?

- [ ] False
- [x] True

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?

- [x] Address bias first through a larger model to get closest to human level error.
- [ ] Get a bigger training set to reduce variance.
- [x] Decrease regularization to boost smaller signals.
- [ ] Decrease variance via regularization so training and dev sets have similar performance.

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

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

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

- [ ] With only 0.05% further progress to make, you should quickly be able to close the remaining gap to 0%.
- [x] If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤ 0.05.
- [x] It is now harder to measure avoidable bias, thus progress will be slower going forward.
- [ ] This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.

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?

- [x] 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.
- [ ] Pick false negative rate as the new metric, and use this new metric to drive all further development.
- [ ] Expand your model size to account for more corner cases.

Note: 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?

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

Note: 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.)

- [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.

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

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

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

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