1. **Problem Statement**

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 that they want an algorithm that

Has high accuracy.

Runs quickly and takes only a short time to classify a new image.

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.

Note: 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. True/False?

True

1. 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: 98%
* Runtime: 9 sec
* Memory size: 9MB

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

Accuracy is an optimizing metric; running time and memory size are satisfying metrics.

1. **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: 9.500.000
* Dev: 250.000
* Test: 250.000

1. 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?

Add it to the training set.

1. 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:

If we add the images to the test set then it won’t reflect the distribution of data expected in production.

1. 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 there is insufficient information to tell.

1. 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%).

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

True.

1. 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:

* Human-level-performaance: 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.
* Try decreasing regularization.

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

* You should try to get a bigger dev set.
* You have overfitted to the dev set.

1. 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 true? (Check all that apply.)

* All or almost all of the avoidable bias has been accounted for.
* You are close to Bayes error and possible overfitting.

1. 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?

Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

1. 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.

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?

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.

1. 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.)

* Buying faster computers could speed up your teams’ iteration speed and thus your team’s productivity.
* Needing two weeks to train will limit the speed t which you can iterate.
* If 100.000.000 examples is enough to build good enough Cat detector, you might be better off training with just 10.000.000 examples to gain a improvement in how quickly you can run experiments, even if each model performs a bit worse because it’s trained on less data.