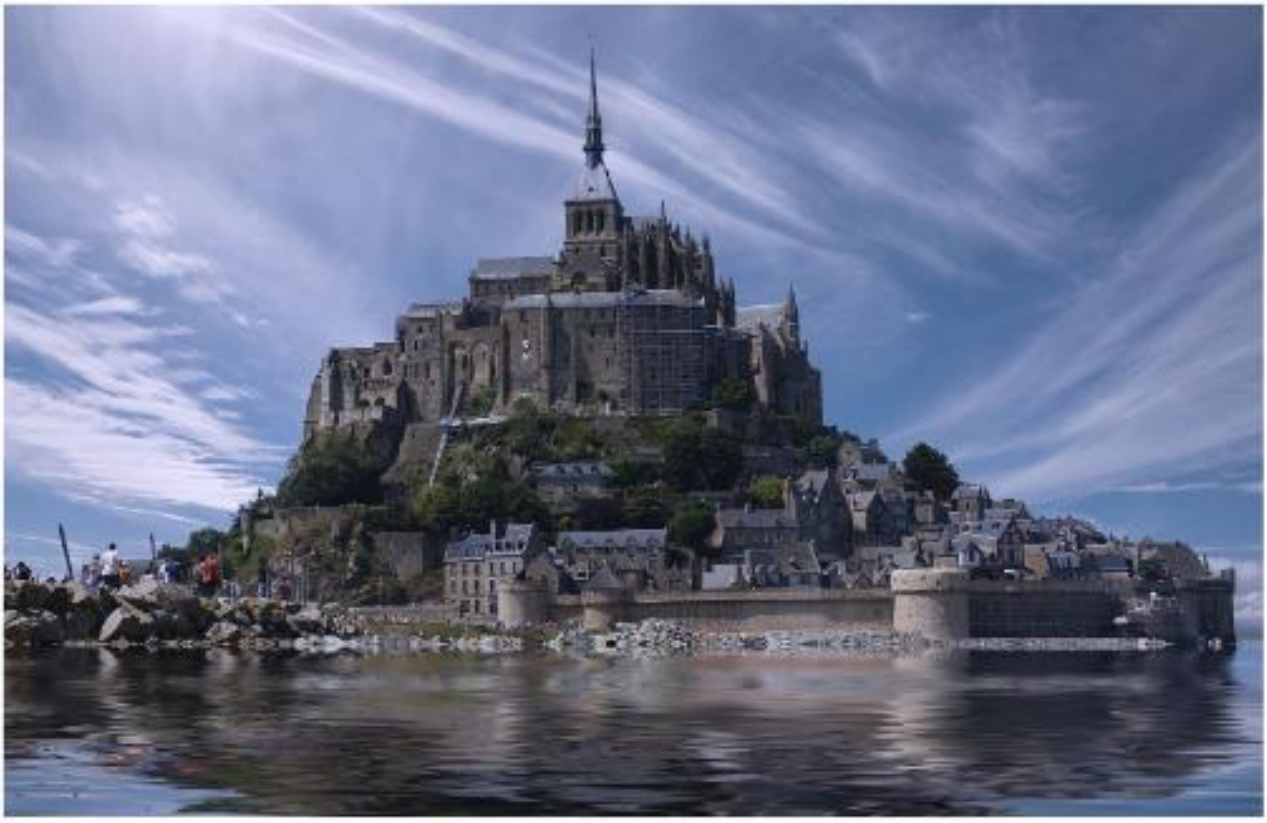
1. This example is adapted from a real production application, but with details disguised to protect confidentiality. 0 / 1 point



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?

- ○ True:
- ◉ False

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?

- ○ 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.
- ○ Create one metric by combining the three metrics and choose the best performing model.

Expand

✓ Correct

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

3. The essential difference between an optimizing metric and satisficing metrics is the priority assigned by the stakeholders. True/False?

1 / 1 point

- ○ True
- ◉ False

Expand

Correct

Yes. Satisficing metrics have thresholds for measurement and an optimizing metric is unbounded.

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
3,333,334	3,333,334	3,333,334

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
9,500,000	250,000	250,000

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. 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 dev set. You object because: (Choose all that apply) 1 / 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.
- [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

Yes. Adding a different distribution to the dev set will skew bias.

- [ ] The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data.
- [x] The dev set no longer reflects the distribution of data (security cameras) you most care about.

✓ Correct

Yes. The performance of the model should be evaluated on the same distribution of images it will see in production.

Expand

✓ Correct

Great, you got all the right answers.

7. Human performance for identifying birds is < 1%, training set error is 5.2% and dev set error is 7.3%. Which of the options below is the best next step? 0 / 1 point

- ○ Get more data or apply regularization to reduce variance.
- ○ Validate the human data set with a sample of your data to ensure the images are of sufficient quality.
- ○ Train a bigger network to drive down the >4.0% training error.
- ◉ Try an ensemble model to reduce bias and variance.

Expand

⊗ Incorrect
No. A best practice is to address the largest gap first.

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

Expand

Correct

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

9. Which of the following statements do you agree with? 1 / 1 point

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

Expand

Correct

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, two of the following four, statements best describe your thought process?

1 / 1 point

- [ ] Get a bigger training set to reduce variance.
- [x] Address bias first through a larger model to get closest to human level error.

✓ **Correct**

Yes. Selecting the largest difference from (train set error - human level error) and (dev set error - train set error) and reducing bias or variance accordingly is the most productive step.

- [ ] Decrease variance via regularization so training and dev sets have similar performance.
- [x] Decrease regularization to boost smaller signals.

✓ **Correct**

Yes. Bias is higher than variance.

Expand

✓ Correct

Great, you got all the right answers.

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)

1 / 1 point

- [x] Try increasing regularization to reduce overfitting to the dev set.

  ✓ Correct
  Yes. The dev set performance versus the test set indicates it is overfitting.

- [x] Increase the size of the dev set.

  ✓ Correct
  Yes. The dev set performance versus the test set indicates it is overfitting.

- [ ] Get a bigger test set to increase its accuracy.
- [ ] Try decreasing regularization for better generalization with the dev set.

Expand

Correct
Great, you got all the right answers.

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] 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%

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. 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

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

Expand

✓ Correct

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?

0 / 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

✗ Incorrect

No. Pooling layers won't reduce the features in a meaningful way to learn 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. (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] 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

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

  ✓ Correct

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

  ✓ 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.

Expand

Correct
Great, you got all the right answers.