1. Training/Validation/Testing accuracy

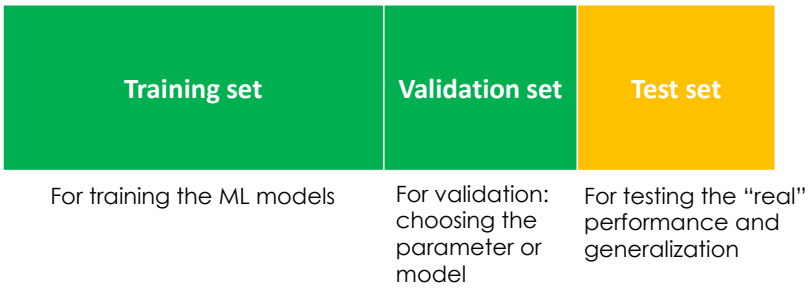

Fig.1 Training/Validation/Test Partition.

Let's say, we use the training set to train two Random Forest classifiers,

- C1 with 100 trees
- C2 with 200 trees

We can then apply C1 and C2 on the training set, and compute the accuracy of C1 and C2, which is known as the **Training Accuracy**. In other words, we check the performance of C1 and C2 on the same dataset which is used for training them.

We apply C1 and C2 on the validation set, and compare their accuracy. In this case, we are checking the accuracy of C1 and C2 on a dataset that is different from the training one. The accuracy of C1 and C2 on the validation set is known as **Validation Accuracy**.

Suppose C1 has a higher validation accuracy than C2, we will select C1 for deployment and use C1 for testing. When we apply C1 on the test set, the accuracy is known as **Testing (or Test) Accuracy.**

2. Is benchmarking on the test set considered as validation?

No, it is not validation. As mentioned during lecture, validation is involved in the training stage and for the sake of parameter selection. Benchmarking on the test set is to compare the generalization capability (or the real performance) of trained models.

<u>Imagine the following scenario:</u>

Xinchao trained a classifier, and he had two parameters to choose. Hence, he trained two models (C1, C2) using the training set. He then applied C1 and C2 on the validation set, and found that C1 works better than C2. Hence, Xinchao decided to use C1 for deployment (or testing).

Vincent trained another type of classifier, and he had three parameters to choose. Hence, trained three models (C3, C4, C5) using the training set. He then applied C3, C4, and C5 on the validation set, and found that C4 works better than C3 and C5. Hence, Vincent decided to use C4 for deployment (or testing).

Then the question is, given the same test set, whose classifier is better? Xinchao's or Vincent's?

To answer this question, we will check the accuracy of Xinchao's C1 and Vincent's C4 on the test set. If Vincent's C4 yields a better accuracy, well, Vincent's classifier beats Xinchao's. In other words, we are comparing the real performance but not doing any parameter selection here.

3. **Can we use the False Negative/False Positive/Confusion Matrix for validation (i.e., for selecting parameters)?**

Yes, we can.

For example, C1 is the Random Forest with 100 trees and C2 is the Random Forest with 200 trees.

If our focus is to select a parameter with a lower False Positive, we can compare C1 and C2 using the metric of False Positive on the validation set, and choose the one with a lower False Positive. The same goes for other metrics.

4. **DET/ROC/AUC Curves**

Don't worry about how to exactly derive these curves -- we won't ask you such questions in the final exam. As long as you know that we *can* derive AUC from ROC, and ROC from DET, it would be great.

From Page 25 to 31, the message is that, if we change the threshold on Y, we will end up having different FN/FP/TP/TN.

What if we want a metric that is invariant of the threshold on Y? Well, we use AUC, which is a single number that accounts for all the thresholds.