机器学习的本质就是确保模型具有泛化能力。要保证这一点，数据被分为训练集和测试集，**测试集是绝对不能用来做训练的**。

课程有个地方需要格外说明一下。整个 Cross Validation[11] 这一个章节的 6 个视频。都是在讲如何更加合理的利用训练数据。这些视频中提到的测试集(Testing Set)，其实又可以称为验证集（Validation Set），它也是训练数据的一部分。

- 视频 1 讲了为什么需要 Cross Validation，什么是 K-Fold Cross Validation。
- 视频 2，3 讲了 K-Fold Validation 的具体应用和局限性。
- 视频 4，5 和 6 讲了 K-Fold 的另一个具体应用，也就是 K-Fold 作为 `GridSearchCV` 的一个可选参数，能够帮我们方便（自动化）地找一些算法的最佳参数，这些算法包括 Naive Bayes，SVM 和 Decision Tree 等。

这几个视频的整个逻辑是这样的：

1. 如果拿全部训练数据做训练，等于我们盲目的相信所有的训练数据，这样很容易出现过拟合。
2. 所以我们可以对训练数据做 Cross Validation，K-fold 是 Cross Validation 的一种方法。这样在去应用到真正的测试数据之前，我们就已经能够判断我们的模型是不是有泛化能力了。
3. K-fold 也不是万能的。模型与数据的排列顺序有很大相关性，例如 100 封 email，1-10 出自 Rob，10-20 出自 Anne...90-100 出自 Mike。分成十份，任何基于其中九份的训练都会完全遗漏一个人。

这个视频[13] 在 P2 部分，但是非常有助于对交叉验证的理解。