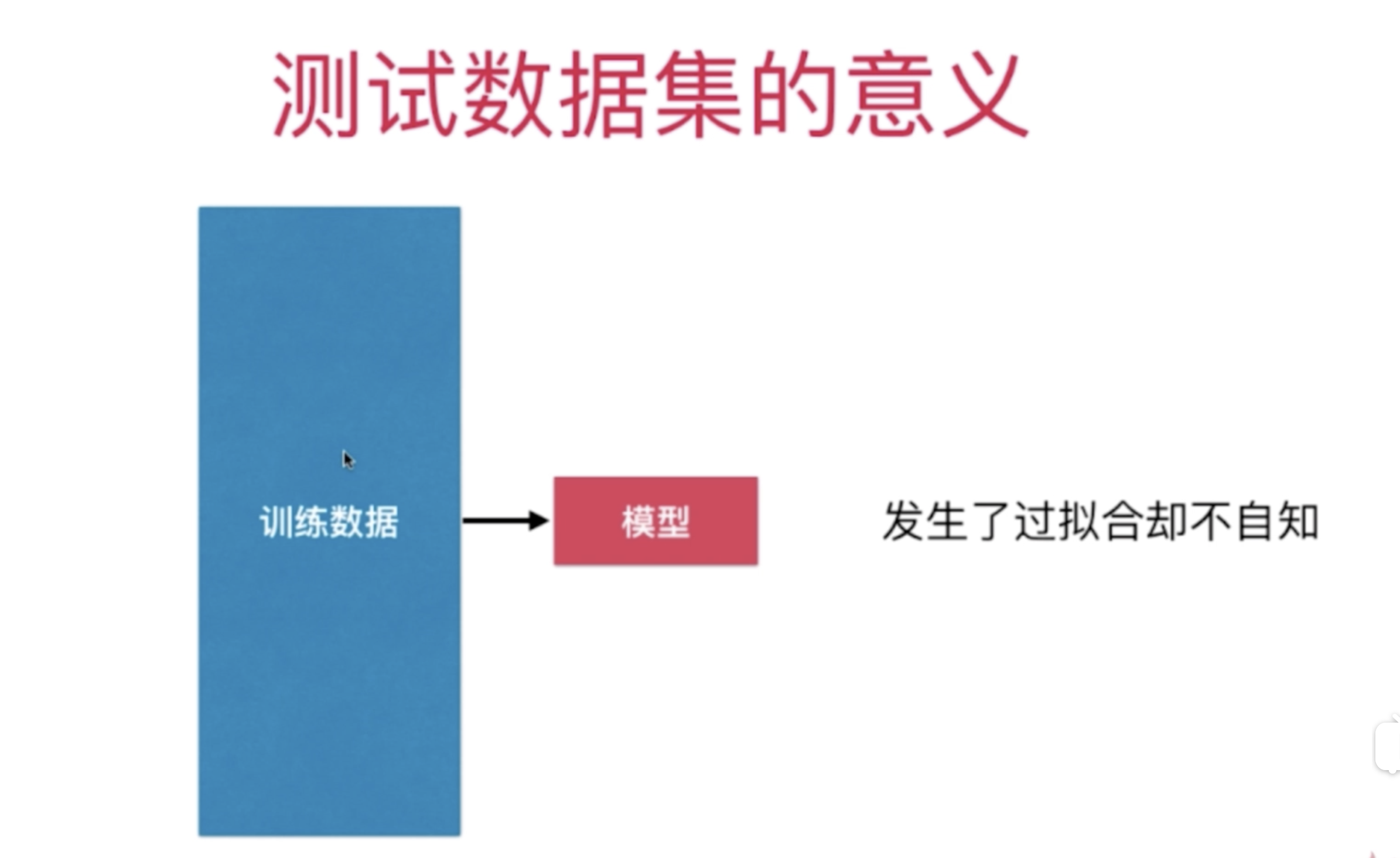
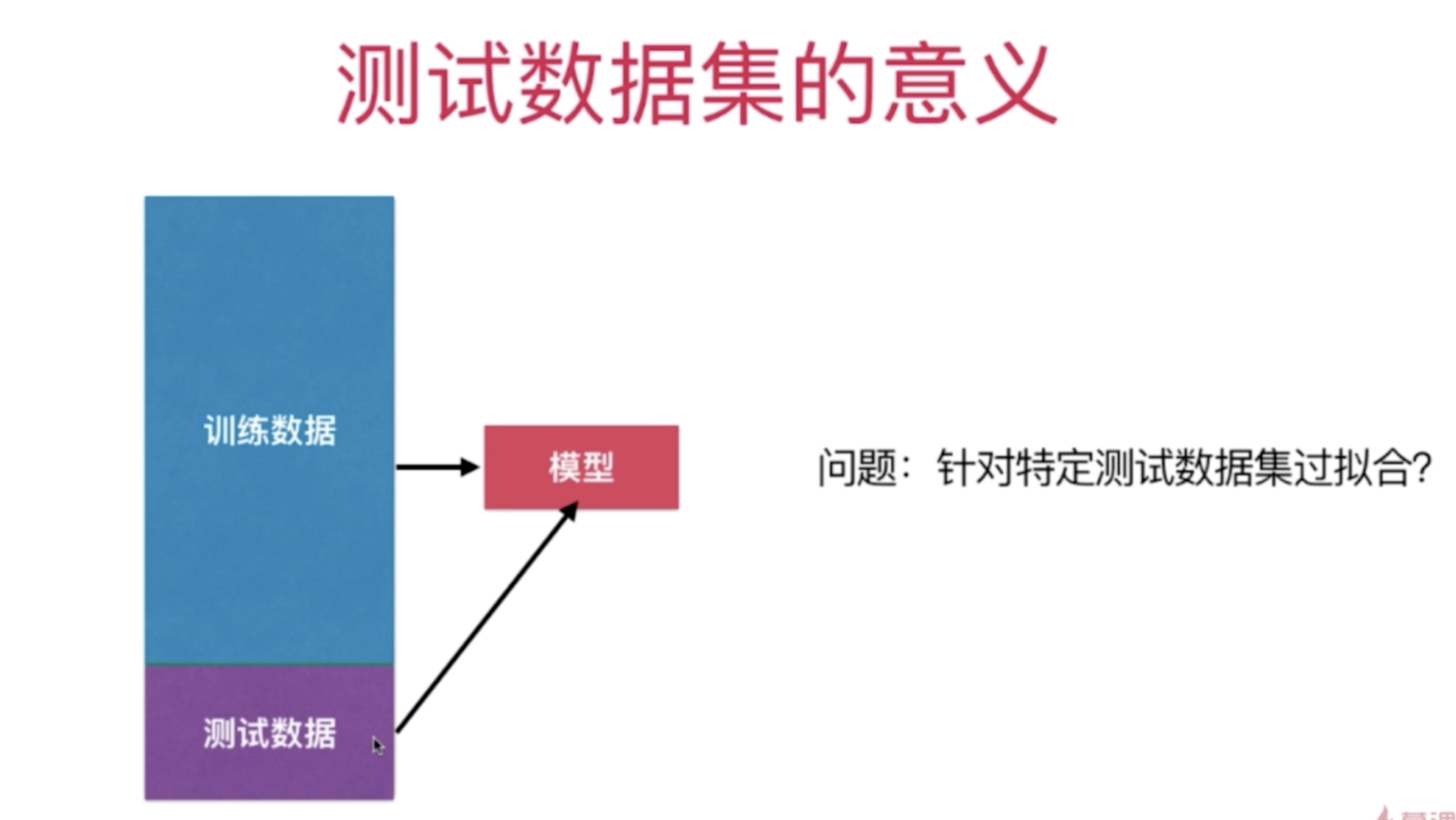
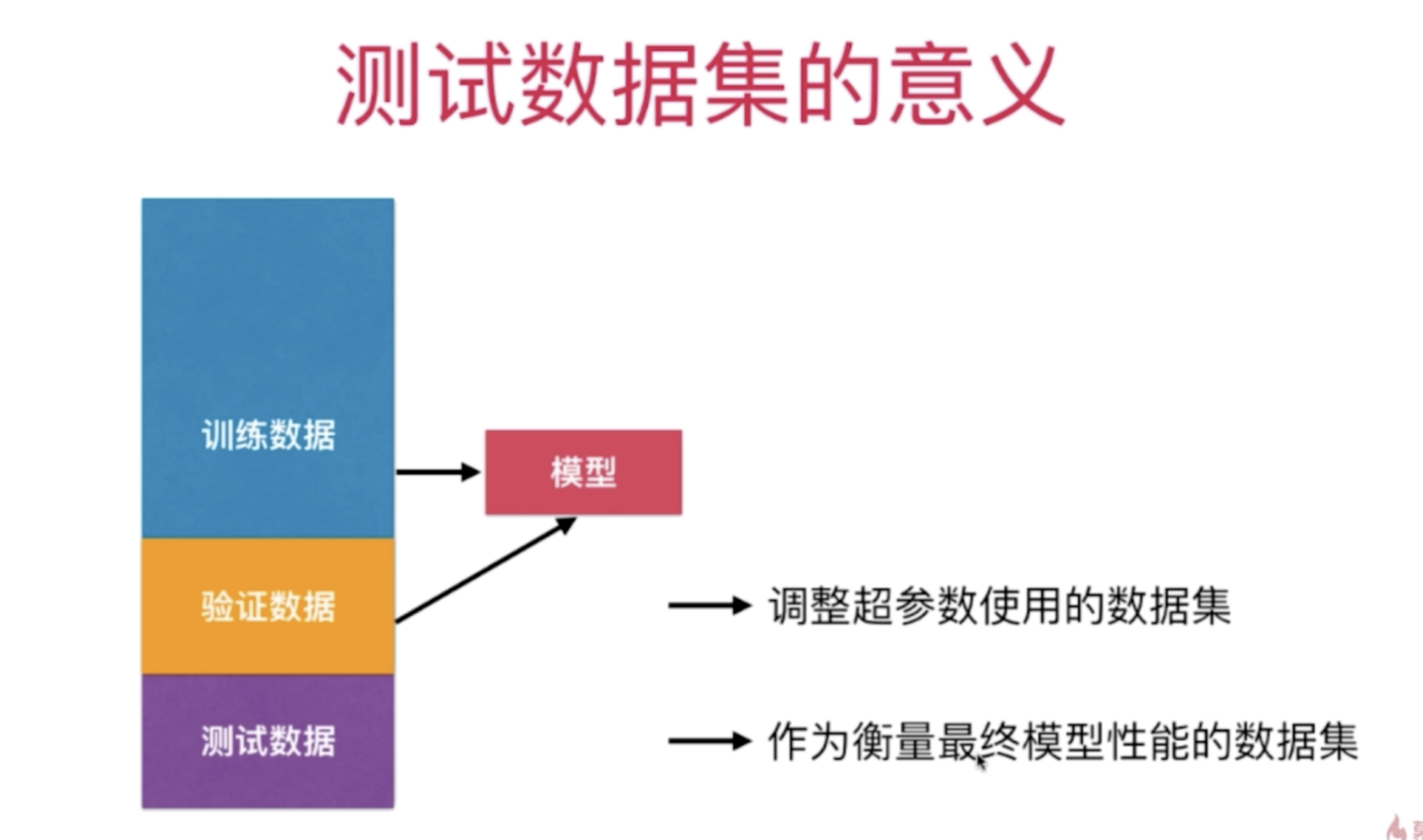
如果把所有的数据都作为训练数据集的话，有下面的问题：



那我们之前用的 train test split 也有一个问题：

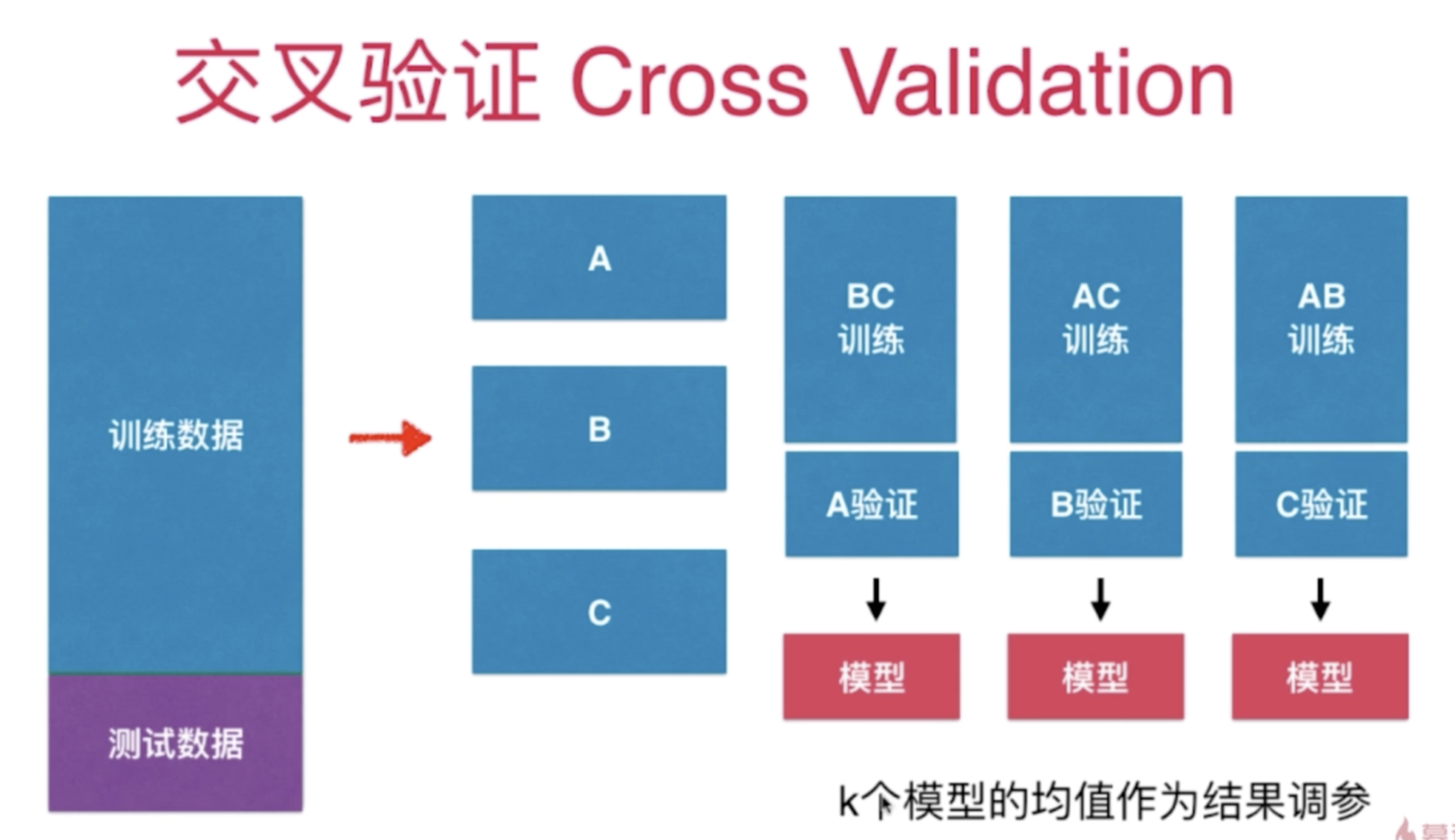


于是有了3个数据集：



但是，还要一个问题，验证数据集是随机抽取的，那有可能里面包含有比较极端的数据，就会导致模型生成的不准确。

最终方案：把训练数据集和验证数据集分做多份，分别相互验证生成模型，再取均值：



实验代码：[08-06 Cross-Validation.ipynb]



如果样本数据有m份，就训练m份模型，每个模型使用m-1数据来训练，用剩下的1个来做验证。一般更接近于模型本身的性能，但是计算量太大，只用在一个学术上要求比较严谨是才用。