# 模型训练步骤

1. 定义含未知参数 θ 的函数 fθ(x)：x 为输入特征，θ 代表模型中所有未知参数。
2. 定义 loss 函数：用于判断参数 θ 的好坏，输入为参数，输出为衡量参数优劣的 loss 值。
3. 解决优化问题：寻找使 loss 值最小的参数 θ\*，将 θ\* 代入 fθ(x)，用测试资料的 x 作为输入，得到输出结果并上传。

# 训练结果优化攻略

## （一）训练资料 loss 大的原因及解决方法

1. **model bias（模型偏差）**
   * 原因：模型过于简单，其包含的函数集合中没有能降低 loss 的函数（即 “大海中没有针”）。
   * 解决：重新设计模型以增加弹性，例如增加输入特征（如用更多历史数据）、使用更大模型或深度学习方法。
2. **优化问题**
   * 原因：模型弹性足够（函数集合中存在低 loss 函数），但优化算法（如梯度下降）未找到最优参数。
   * 判断：比较不同模型，若深层模型弹性更大但 loss 未低于浅层模型，则可能是优化问题。
   * 解决：改进优化方法。

## （二）训练资料 loss 小但测试资料 loss 大（过拟合）

1. **原因**：模型在训练资料上表现好（记住训练数据细节），但在未见过的测试资料上表现差（泛化能力弱）。
2. **解决方法**
   * 数据增强：基于对问题的理解创造新训练资料（如影像辨识中对图片左右翻转、局部截取放大等），需保证增强合理。
   * 限制模型弹性：减少参数（如减少神经元数量）、让模型共用参数（如 CNN 针对影像特性设计的限制，后续课程讲解）、使用较少特征、早停（Early stopping）、正则化（Regularization）、Dropout 等。

# 四、模型选择方法

1. **避免的误区**：仅根据公开测试集（public testing set）结果选择模型，可能导致在私有测试集（private testing set）上表现差（因可能过拟合公开测试集）。
2. **合理方法**
   * 划分训练集与验证集：将训练资料分为训练集（如 90%）和验证集（如 10%），用训练集训练模型，根据验证集分数选择最优模型。
   * N 折交叉验证：将训练资料分为 N 等份，轮流用其中 1 份作为验证集、其余作为训练集，计算每个模型在 N 次验证中的平均分数，选择最优模型后，用全量训练资料训练并用于测试。

# 五、其他注意事项

* 数据分布不匹配：当训练资料与测试资料分布不同时，增加训练资料无效，作业十一专门针对此问题设计。
* 公开测试集的局限性：公开测试集结果可能被 “随机结果” 偶然优化，需通过限制上传次数避免过度依赖。