# 机器学习的定义：

机器学习是让机器具备找一个函数的能力，该函数可用于解决语音识别、影像识别、下围棋等复杂任务，这些任务的函数复杂到无法人工编写，需机器自动找出。

# 机器学习的任务分类：

## 回归（Regression）：

输出是数值，例如预测未来某一时间的PM2.5数值，函数输入是与预测PM2.5有关的指数，输出是明天中午的PM2.5数值。

## 分类（Classification）：

输出是从设定好的选项中选择一个，例如垃圾邮件检测，选项是“是垃圾邮件”或“不是垃圾邮件”；AlphaGo下围棋也是分类问题，选项是棋盘上19乘19个落子位置。

## 结构化学习：

输出是有结构的物件，例如机器画图、写文章，可拟人化地理解为让机器学会创造。

# 机器学习找函数的步骤：

第一步：写出带未知参数的函数（模型）：以预测YouTube频道隔天点阅率为例，假设函数为y = b + w×xₗ，其中y是预测的点阅人数，xₗ是前一天的点阅人数，b和w是未知参数，该函数被称为Model，xₗ是Feature，w是weight，b是Bias。

第二步：定义损失函数（Loss）：Loss是关于模型参数的函数，用于衡量参数的好坏。以预测YouTube点阅率为例，用训练资料（2017 - 2020年的点阅次数）计算预测值与真实值的差距，差距的平均值即为Loss，常用的计算差距的方法有平均绝对误差（MAE）、均方误差（MSE）等。

第三步：解优化问题：通过梯度下降法找到使Loss最小的参数w和b。随机选取初始参数，计算参数对Loss的微分（切线斜率），根据斜率调整参数，重复此过程直到Loss不再减小或达到设定的迭代次数。梯度下降法可能会陷入局部最小值（local minima），但在深度学习中，局部最小值问题其实影响不大，真正的难点之后会讲。

# 模型优化与评估：

仅考虑前一天点阅次数的模型，在训练资料上的Loss是0.48k，在2021年新资料上的误差是0.58k。

观察到数据有每周五、六点阅率低的周期性规律后，改进模型为考虑前七天的点阅次数，训练资料上的Loss降至0.38k，新资料上的误差为0.49k。

进一步考虑前28天、56天的点阅次数，训练资料上的Loss继续降低，但新资料上的误差在0.46k左右不再明显改善，说明模型优化存在极限。

# 线性模型（Linear model）：

上述考虑不同天数点阅次数并乘以权重再加上偏差的模型，都属于线性模型，其共同特点是将输入特征乘以权重再加上偏差得到预测结果。