HUAWEI

1.1 机器学习概述

目录

Contents

什么是机器学习

机器学习这门学科所关注的问题是：计算机程序如何随着经验积累自动提高性能。

对于某类任务T和性能度量P，如果一个计算机程序在T上以P衡量的性能随着经验E而自我完善，那么我们称这个计算机程序在从经验E学习。

-- Tom Mitchell, Machine Learning

模式识别起源于工程学，而机器学习产生于计算机科学。然而这些领域可以看做成是同一领域的两个方面。

--Bishop, Pattern Recognition and Machine Learning

机器学习是一类从数据中自动发现模式，并基于发现的模式预测未来数据或者在不确定条件下执行某类决策的方法。

-- Murphy, Machine Learning: A Probabilistic Perspective

简单解释：

机器学习是一种从数据当中发现复杂规律，并且利用规律对未来时刻、未知状况进行预测和判定的方法。

机器学习与人工智能

人工智能是指使用某种方法使得计算机能够自行对外部世界的变化进行判断并做出反应，并且这类反应并非是由人类工程师预先定义好。

人工智能的目标是通过图灵测试，图灵测试是由艾伦·麦席森·图灵发明，指测试者与被测试者（一个人和一台机器）隔开的情况下，通过一些装置向被测试者随意提问。进行多次测试后，如果有超过30%的测试者不能确定出被测试者是人还是机器，那么这台机器就通过了测试，并被认为具有人类智能。

机器学习是当下被认为最有可能实现人工智能的方法，随着大数据+机器学习的组合，使得机器学习算法从数据中发现的规律越来越普适。

人工智能是一种无招胜有招的境界，机器学习是达成这种境界的一个门派武功。

机器学习类型

按学习方式分为三大类

	说明	解决问题
监督学习 Supervised learning	从给定的训练数据集（历史数据）中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集需要包括输入和输出，也可以说是特征和目标/Label。训练集中的目标是由人标注的。	分类（类别预测） 回归（数值预测）
无监督学习 Unsupervised learning	与监督学习相比，输入的数据没有人为标注的结果，模型需要对数据的结构和数值进行归纳。比如根据用户的基本信息把所有用户划分为不同的用户群，再对不同人群采取不同的销售策略	聚类（簇分群）
强化学习 Reinforcement learning	输入数据可以刺激模型并且使模型做出反应。反馈不仅从监督学习的学习过程中得到，还从环境中的奖励或惩罚中得到。	机器人 Alpha GO

机器学习类型

按算法分为三大类

	特点	来处
传统机器学习 Machine Learning	适用于结构化的数据（就是非常规整的表格型数据），适用于需要进行预测的场景（预测类别型结果、数值型结果）：信用风险检测（预测类别）、销售预测（预测金额）、用户画像（预测类别）、商品推荐（预测类别、预测评分）等等。 模型可解释性强。	基于早期数据挖掘学
深度学习 Deep learning	适用于非结构化数据，比如图像、语音等，适用于识别类场景：图像识别、语音识别、语音合成、语义识别。 模型可解释性弱。	基于神经网络
强化学习 Reinforcement learning	适用于需要探索和优化的场景，不一定需要结构化的数据，对于模拟环境的准确度有强要求，能够根据环境中参数的变化自动给出最优选择：制造业某种设备运行时参数自动调控、智能温控、智能污水处理、智能交通信号灯、Alpha Go围棋。	基于机器人学

传统机器学习

本文主要讲授传统机器学习的方法，基于早期数据挖掘学的传统机器学习算法具有模型可解释性强、计算资源占用低等优点，在一些实时场景（设备检测）、需要给出数据规律原因的场景（如商业决策）能否发挥巨大的作用

武功秘籍：

行走AI江湖，出山之前肯定要学好功夫。学习传统机器学习知识的过程中，有以下的一些工具需要熟练掌握。
编程语言：python，现在有众多机器学习的库支持python，也因为python脚本语言的特性、python语法简洁优雅，非常适合做数据处理。
库：数据处理的库（pandas、numpy）、算法库（scikit-learn、xgboost）。
参考书：《数据挖掘-概念与技术》

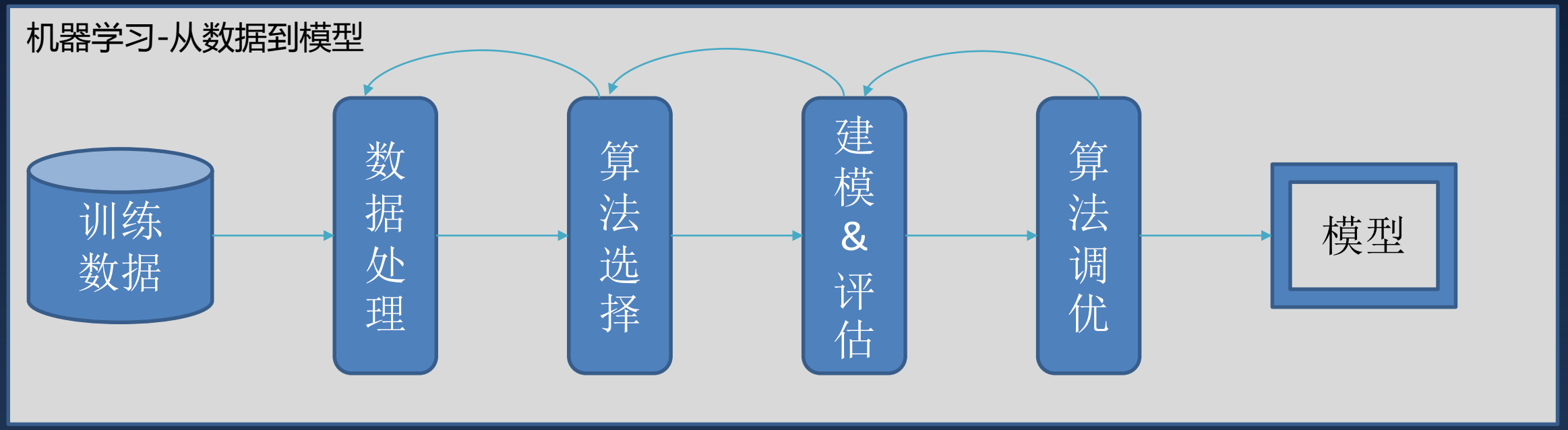
神兵利器：

实践机器学习算法时，一个强大的、不必关心细节的平台或者工具是非常重要的，尤其是涉及大量数据计算的时候个人电脑根本扛不住，还得自己费尽功夫搭集群，幸好现在有云计算了，推荐使用华为云机器学习服务。

目录

Contents

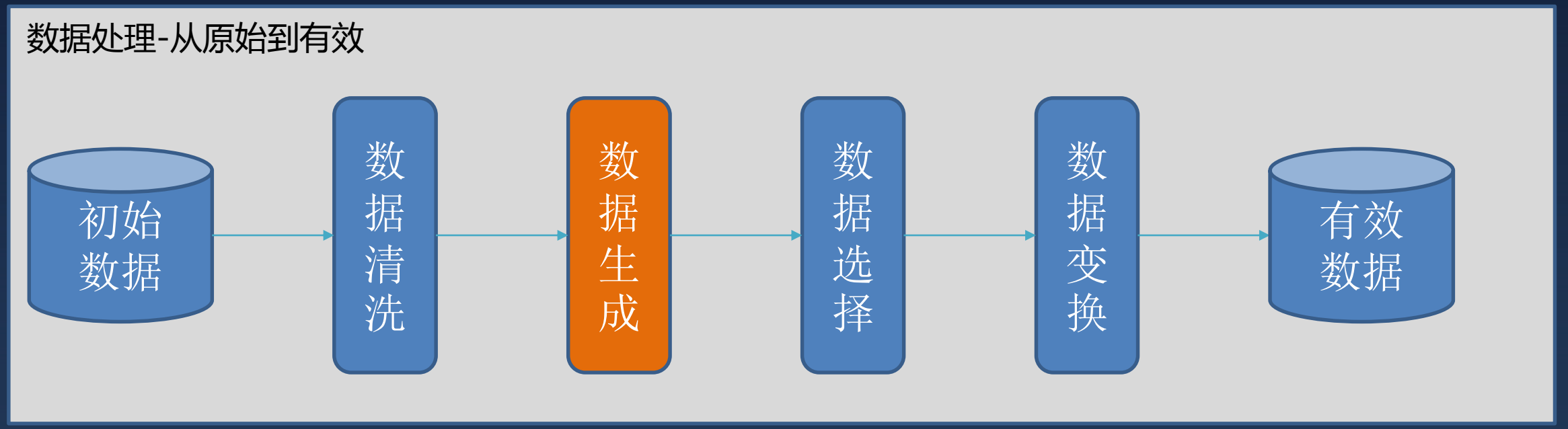
机器学习的一般过程



数据处理到算法调优这个过程是一个不断完善、循环往复的过程，这个过程相当于做实验，直到得出一个在接受度范围内的模型，但是这个过程是可以被一些先验经验指导的，需要识别问题、识别场景、算法原理掌握等等。

数据处理

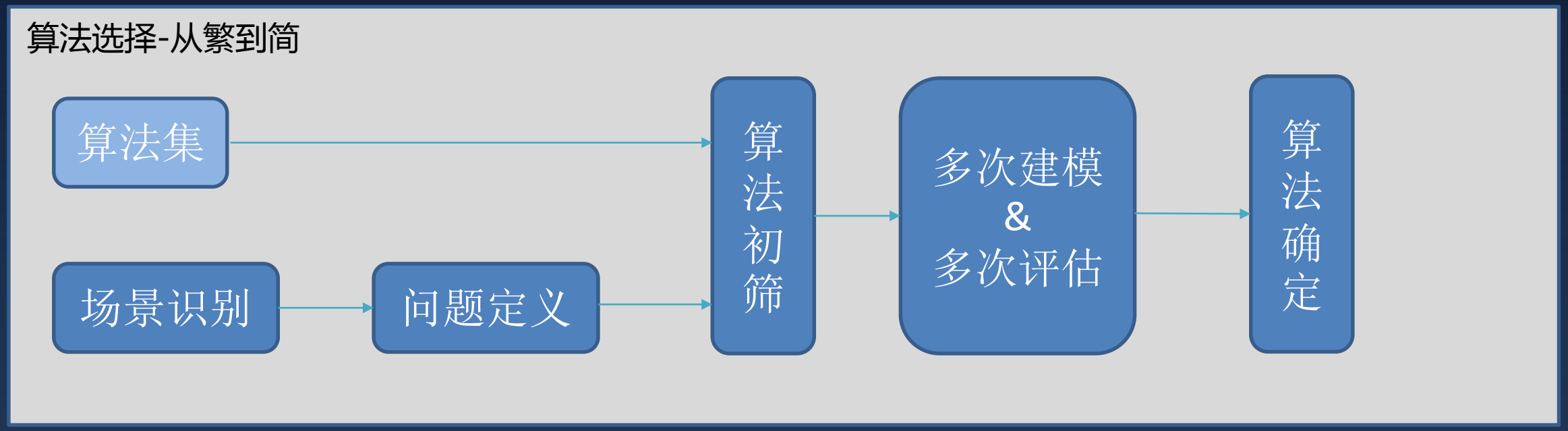
没有完美的数据，实际场景当中的数据往往存在着各种问题，不能直接用于建模。即便是数据从格式、内容上是完整的，一些适当的数据变换、数据选择可以直接提升模型的精度。



数据清洗：异常值处理、缺失值填充、样本选择
数据生成：生成一部分的样本，解决样本不均衡（可选）
数据选择：特征相关性分析，训练集、验证集、测试集的有效划分
数据变换：离散化、标准化、特征降维、特征膨胀

算法选择

算法选择与其说是一个技术活，不如说是一个经验活



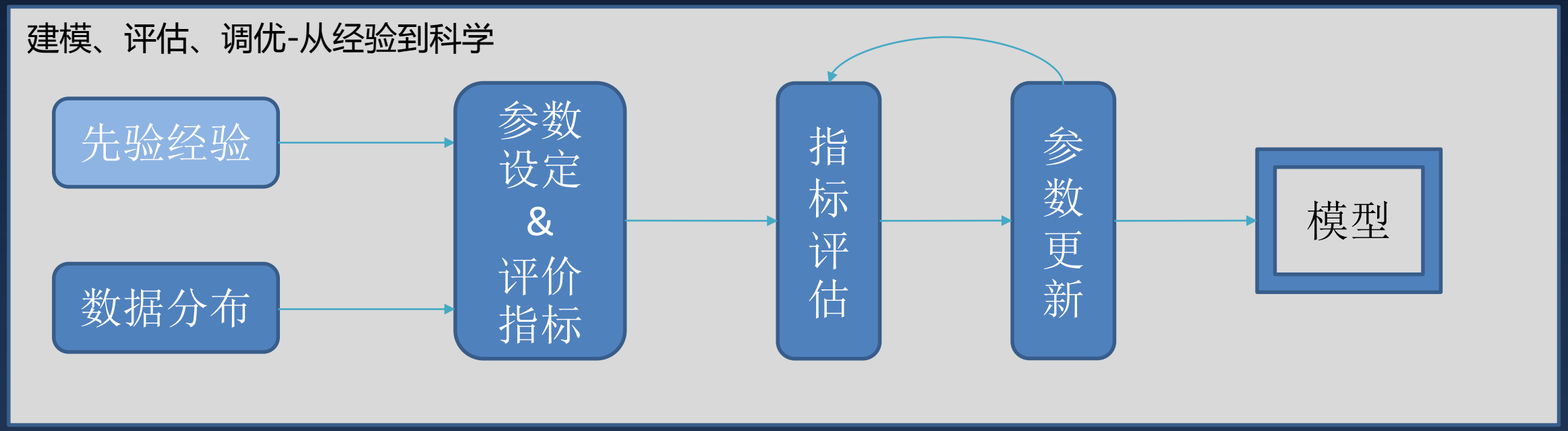
场景识别：在实际的应用当中，首先要进行业务场景的识别，将复杂的业务分解为多个场景问题
问题定义：单个的场景问题要进行问题定义，转化为机器学习定义，（比如：销售预测→回归）
算法初筛：利用历史经验、数据分布规律在算法集中进行算法初步筛选
算法确定：经过对多个算法的多次建模与评估，最终确立可用算法

行业通用的机器学习算法选择

分类	回归	推荐
朴素贝叶斯：模型参数较少，对缺失值不敏感，要求特征之间相互独立	线性回归：适用于预测目标与特征之间线性关系强的数据集	域分解机：特征选择不敏感，推荐精确度高，计算量大
随机森林分类：适用范围广精度高，广泛应用于分类场景	随机森林回归：多分类器计算量小，回归误差小	交替最小二乘：计算量小，适用于特征相关性较高的数据集
支持向量机：适用于小样本、非线性的数据集	K最近邻回归：计算量大无需建模过程，解释性好	

建模、评估、调优

先验经验、多次尝试、指标先行



参数设定：根据先验经验和数据分布规律对于算法的默认参数进行初次更改

评价指标：不同的机器学习问题采用不同的评价指标、不同的场景问题采用不同的评价指标

指标评估：确立评价指标的接受度范围（比如：可接受准确率为70%）

参数更新：通过查看模型，找到评价指标的影响因子，定向进行参数更新

常见评价指标

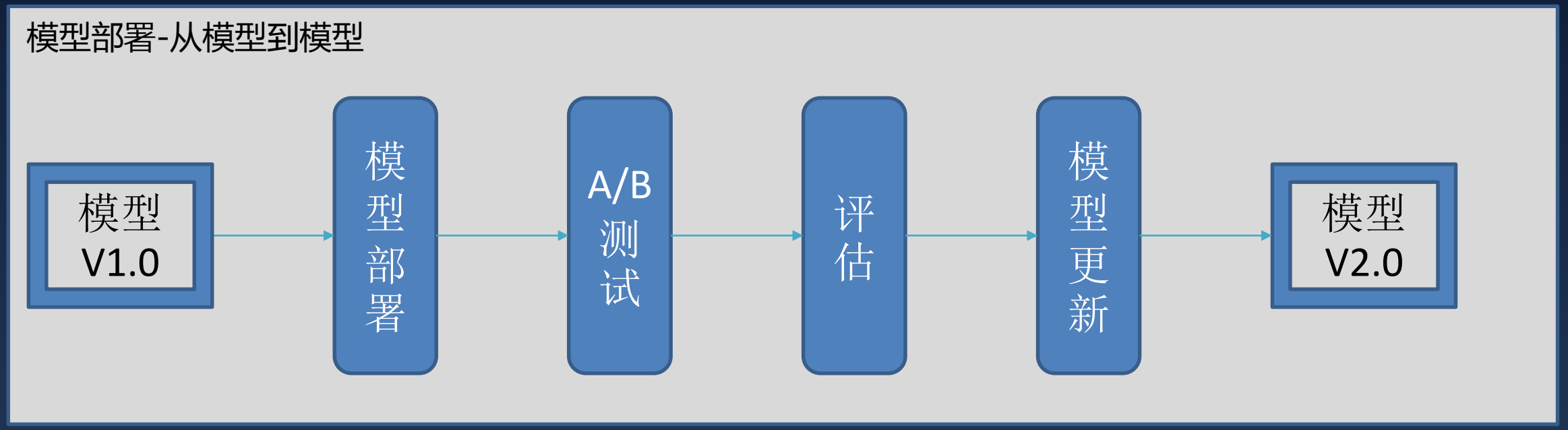
分类模型的评估

- 准确率（accuracy）
- 精确率（查准率，precision）
- 召回率（查全率，recall）
- F1 score
- ROC曲线
- AUC曲线
- 混淆矩阵（Confusion matrix）

回归问题的评估

- 平均绝对误差（ mean absolute error，MAE ）
- 均方根误差（root mean squared error，RMSE）
- 确定性系数（coefficient of determination，R2）

机器学习的后续过程



1、机器学习模型建立好之后，不仅仅要在训练数据上表现良好，更好在未来时间、未知数据上表现达到一定的程度。

2、模型部署到业务系统，经过A/B Test，将结果进行评估，然后进行模型的更新。

3、验证有效的模型也需要更新，按照定时or条件的方式进行更新。

Thank You.

华为云机器学习服务MLS
www.huaweicloud.com/product/mls.html