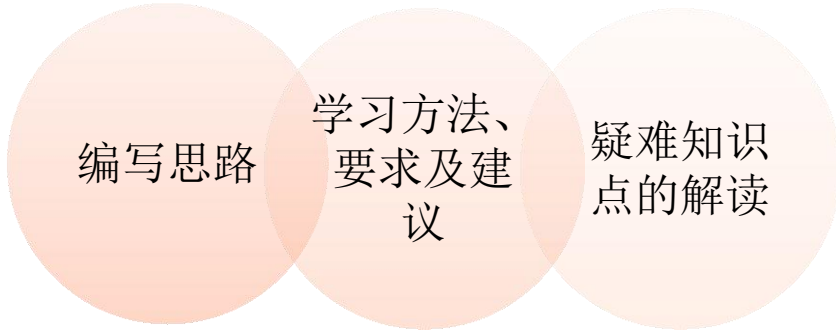
第3章 机器学习与算法

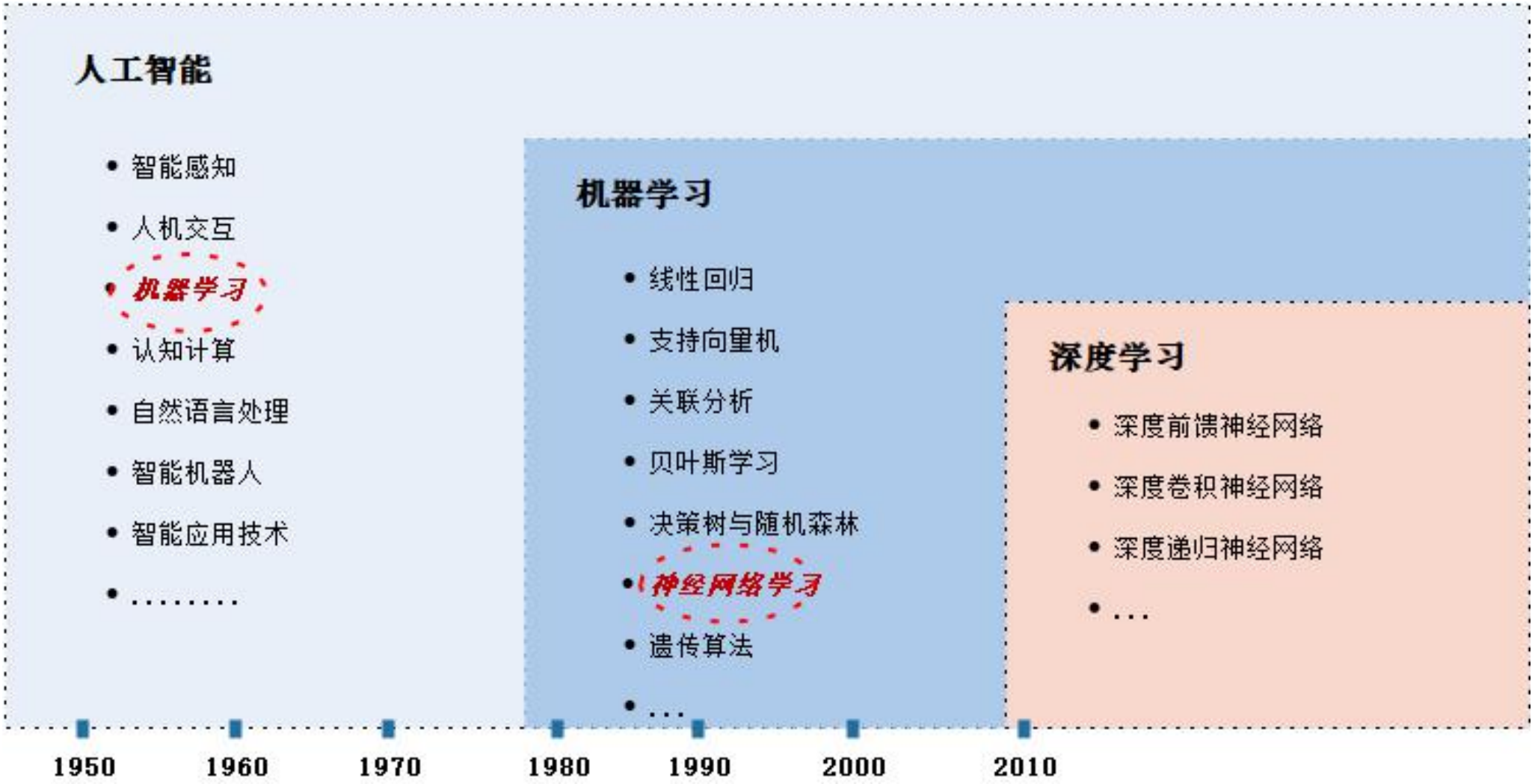


1.本章定位与内容简介

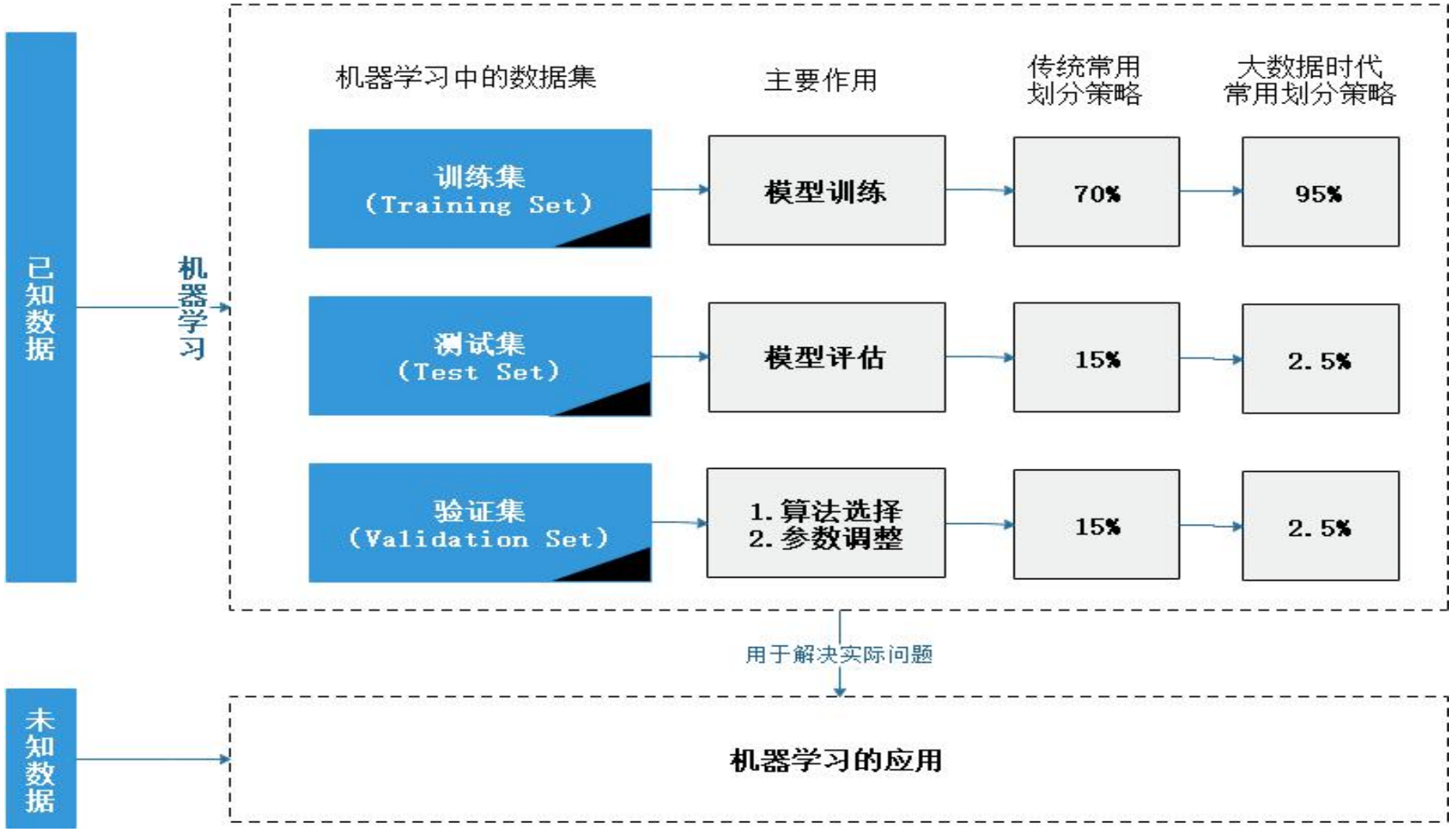
2.本章学习提示及要求

了解	理解	掌握	熟练掌握
• 机器学习与数据科学的区别与联系 • 大数据环境下机器学习面临的主要挑战	• 数据科学中应用机器学习的基本步骤 • 算法的类型及选择方法	• 面向机器学习的数据划分及准备方法 • 机器学习中对模型的评估方法	• 基于Python的机器学习编程实践

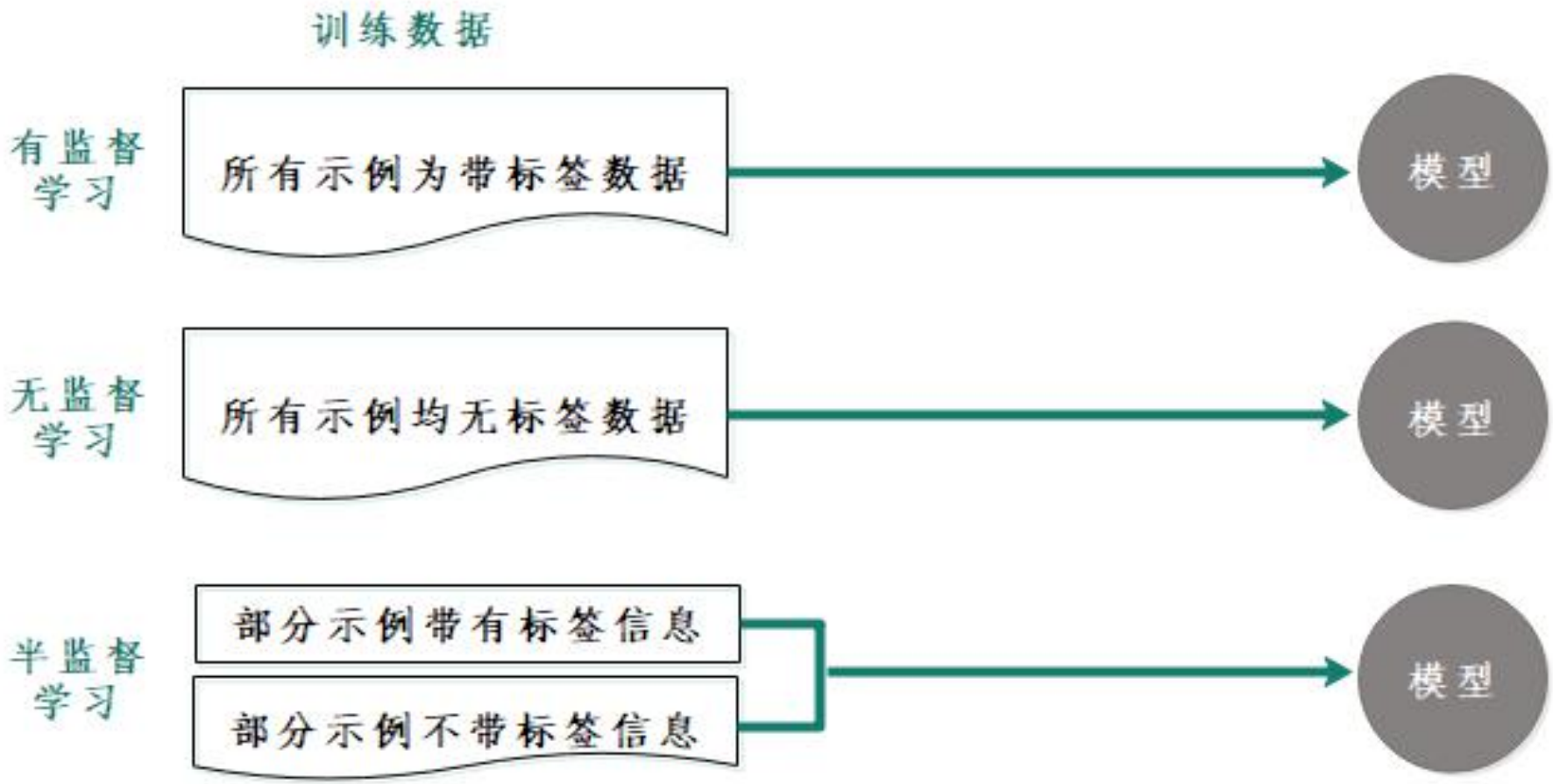
3.数据智能及其实现


人工智能
智能感知
人机交互
机器学习
认知计算
自然语言处理
智能机器人
智能应用技术
.....
机器学习
线性回归
支持向量机
关联分析
贝叶斯学习
决策树与随机森林
神经网络学习
遗传算法
...
深度学习
深度前馈神经网络
深度卷积神经网络
深度递归神经网络
...
1950
1960
1970
1980
1990
2000
2010

4.机器学习中的数据加工



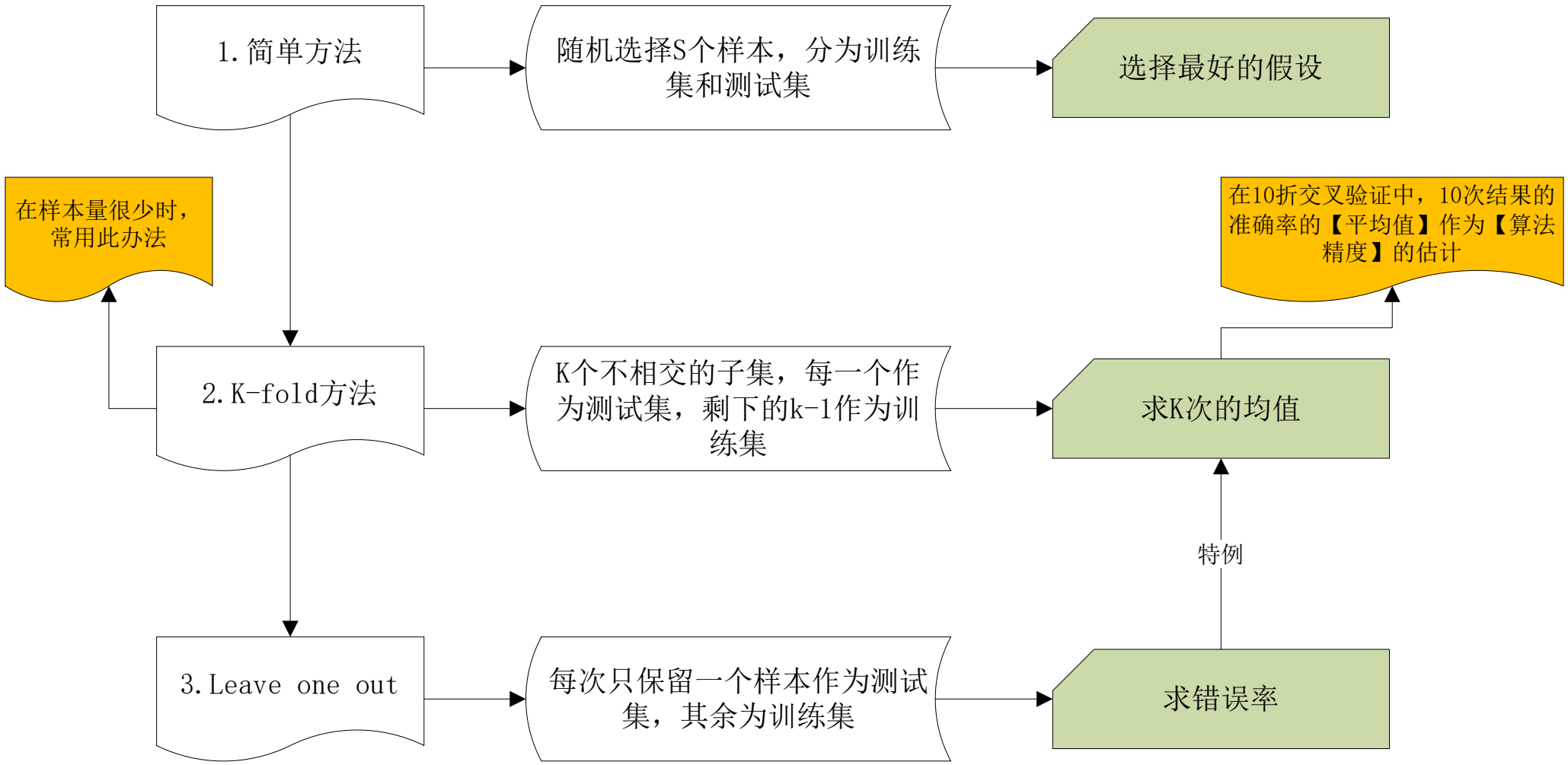
5.有监督学习与无监督学习

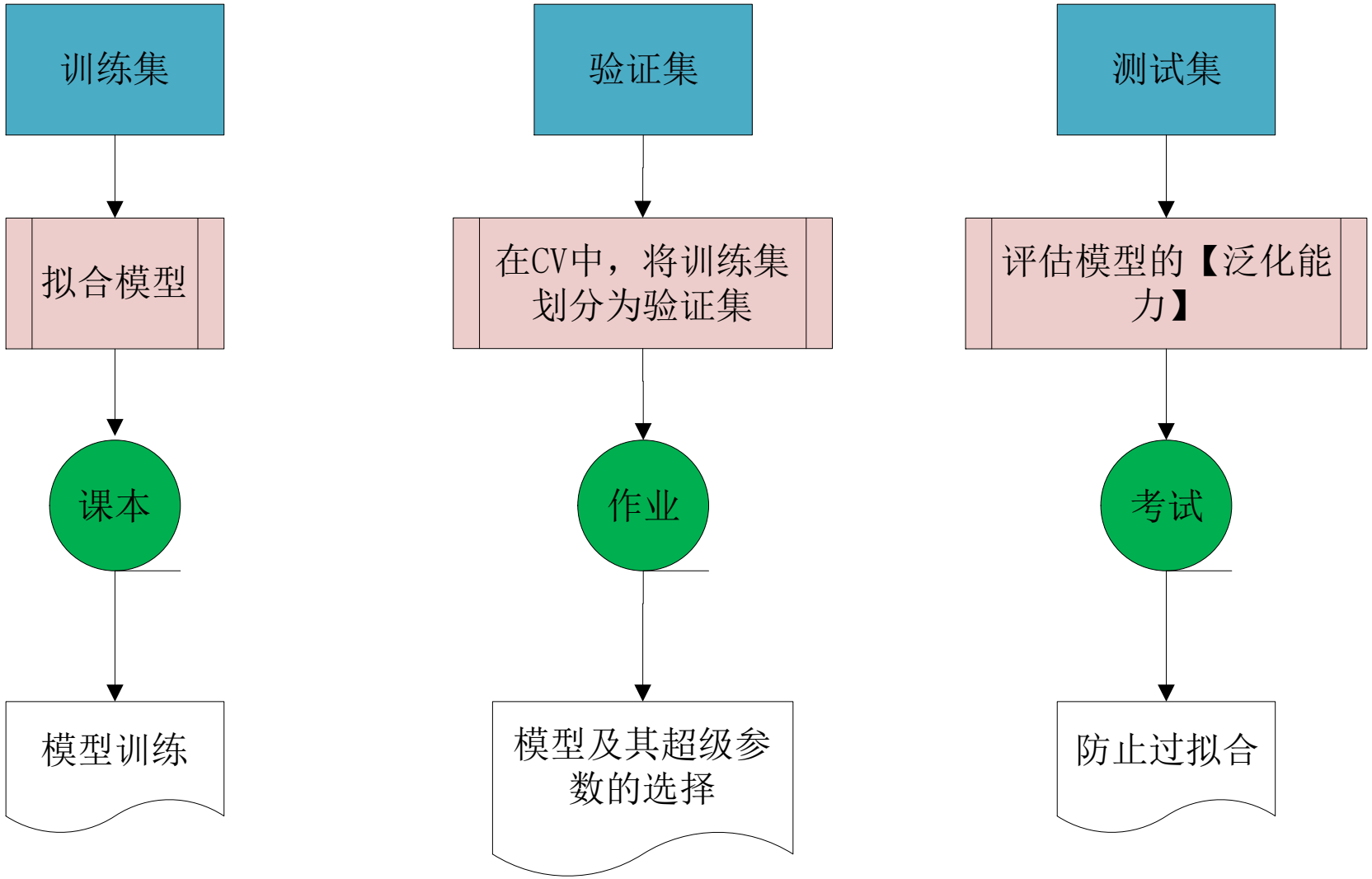


6.机器学习算法的类型

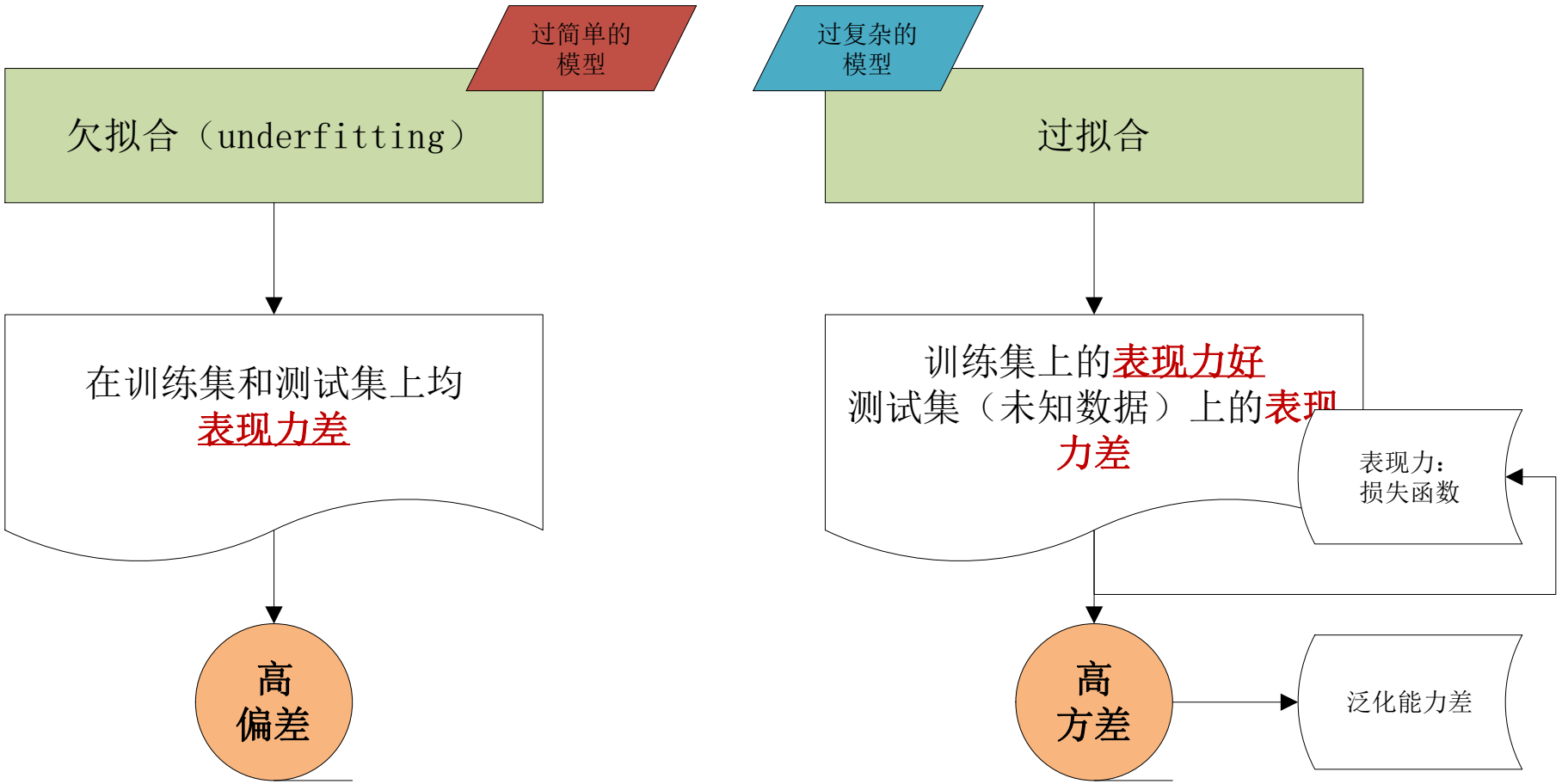
	无监督	有监督
连续型	聚类与维度下降 SVD PCA K-Means	回归 线性回归 多项式回归 决策数 随机森林
分类型	关联分析 Apriori FP-Growth 隐马尔可夫模型	分类 KNN 逻辑回归 朴素贝叶斯 SVM

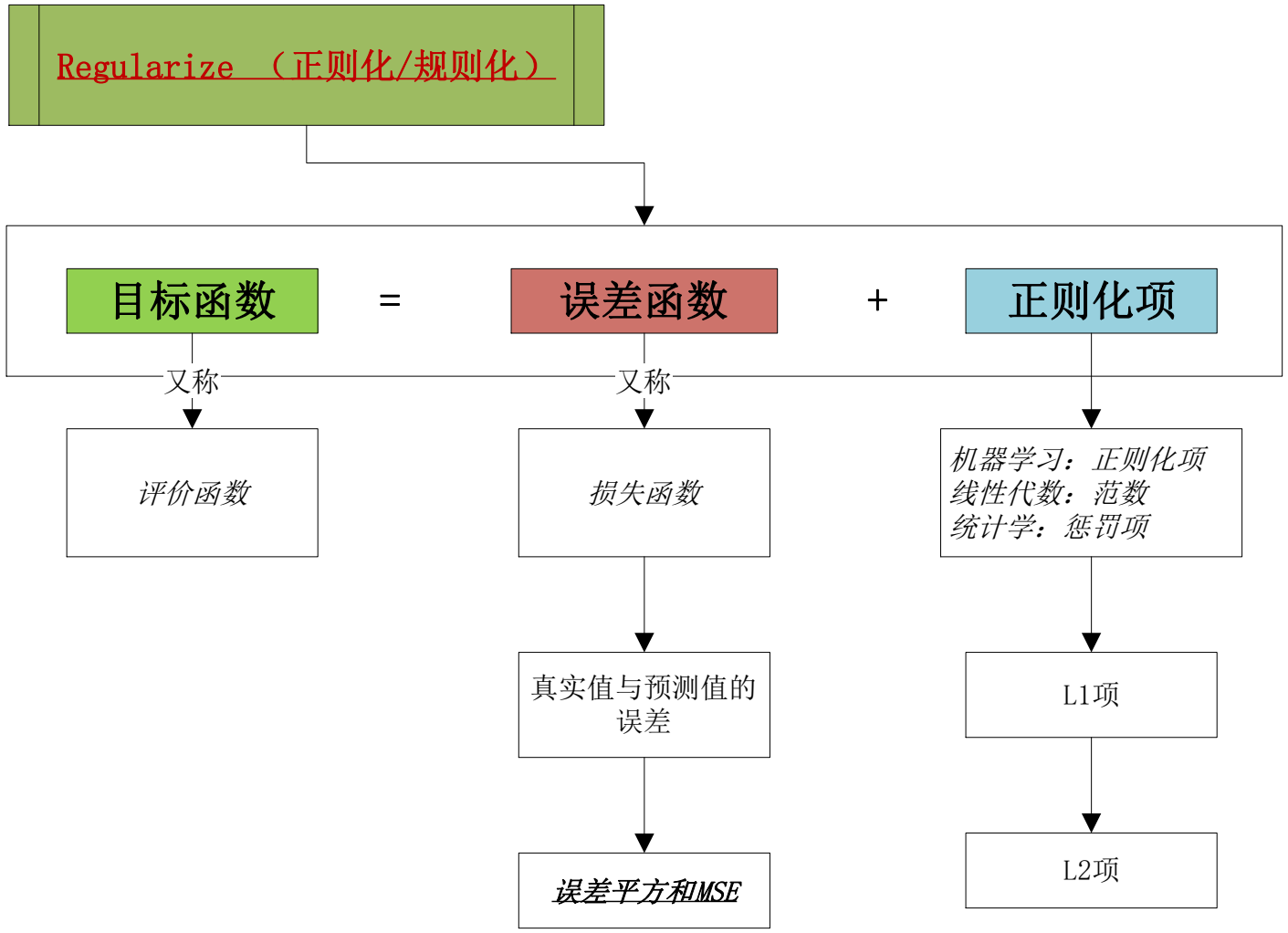
7.机器学习中的交叉校验




训练集
拟合模型
课本
模型训练
验证集
在CV中，将训练集划分为验证集
作业
模型及其超级参数的选择
测试集
评估模型的【泛化能力】
考试
防止过拟合

8.机器学习中的过拟合问题





损失函数的优化方法： ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag

9.机器学习面临的主要挑战

过拟合（Overfiting）

- 训练集→测试集

维度灾难（Curse of Dimensionality）

- 低维度→高维度

特征工程（Feature Engineering）

- 训练集的特征+领域知识

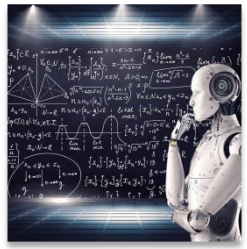
算法的可扩展性（Scalability）

- 训练集的规模、目标函数的复杂度、算法运行效率之间的平衡

模型集成

- Bagging、Boosting、Stacking

10.如何继续学习本章知识



统计学与机器学习的区别与联系

数据科学中常用的统计学算法（P78）

表 3-2　　统计学与机器学习的术语对照表

	机器学习	统计学
1	训练（Train）	拟合（Fit）
2	算法（Algorithm）	模型（Model）
3	分类器（Classifier）	假设（Hypothesis）
4	无监督学习（Unsupervised Learning）	聚类（Clustering）
5	有监督学习（Supervised Learning）	分类（Classification）
6	网络（Network）/图（Graph）	模型（Model）
7	权重（Weights）	参数（Parameters）
8	变量（Variable）	特征（Feature）

小结


1.本章定位与内容简介
2.本章学习提示及要求
3.数据智能及其实现
4.有监督学习与无监督学习
5.机器学习算法的类型
4.机器学习中的数据加工
7.机器学习中的交叉校验
8.机器学习中的过拟合问题
9.机器学习面临的主要挑战
10.如何继续学习本章知识