机器学习概述与算法介绍

Machine Learning Engineer

机器学习工程师

讲师：寒小阳

稀牛学院 網易 NetEase

目 录

CONTENTS

01

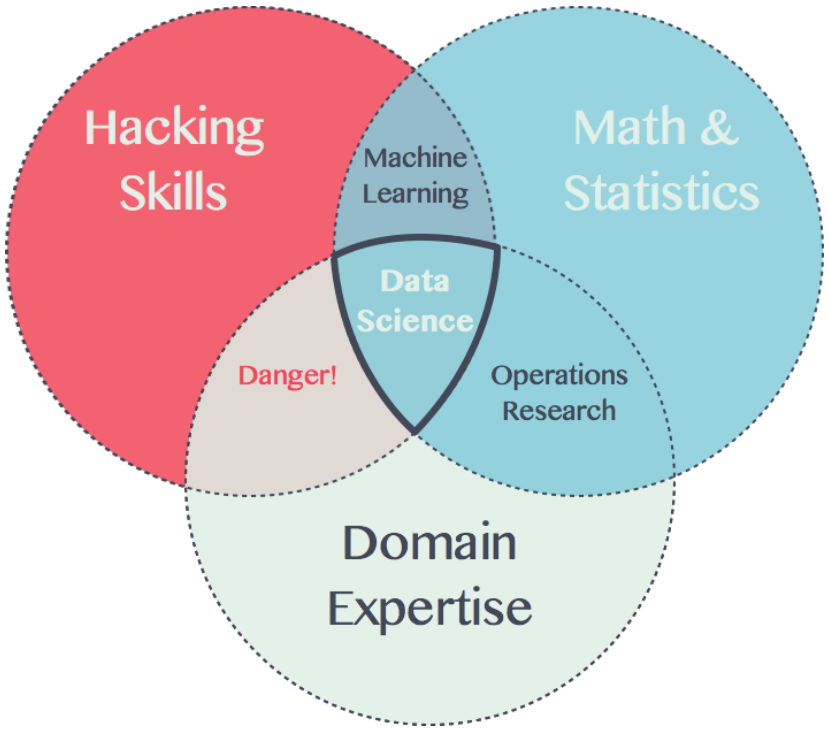
机器学习概述

1.1 机器学习是什么

1.2 机器学习几个例子

1.1 机器学习是什么

- 人工智能的一个重要学科分支多领域交叉学科。
- 数据驱动，在数据上通过算法总，结规律模式，应用在新数据上。



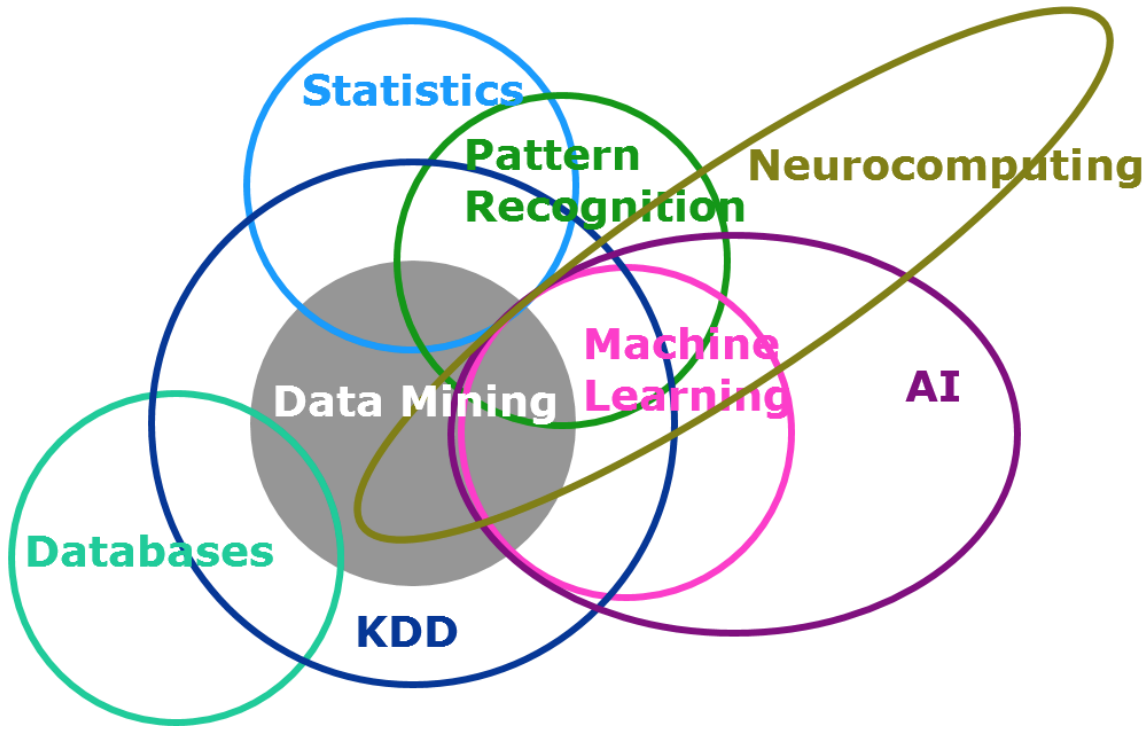
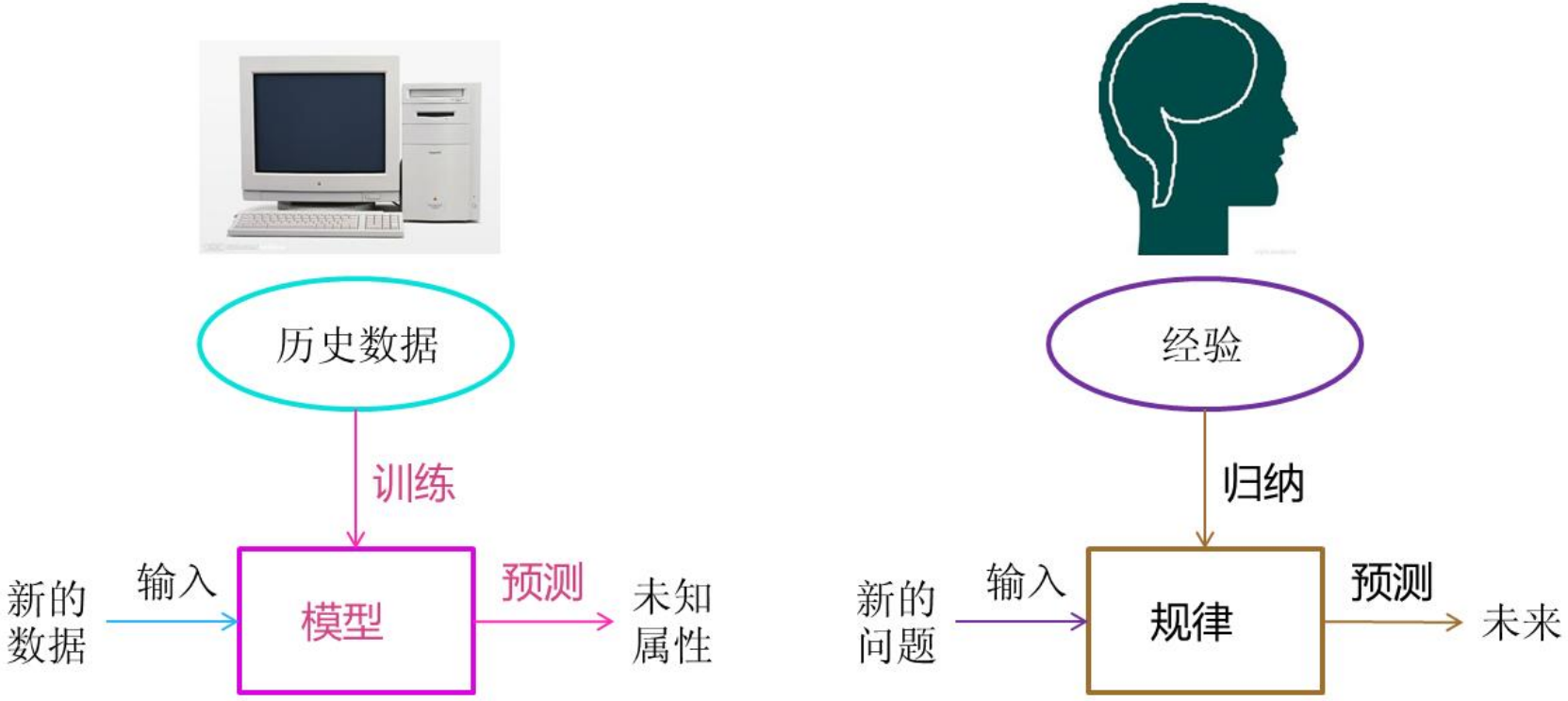


Image from http://branchandbound.net/

1.1 机器学习是什么

- 机器学习研究的是计算机怎样模拟人类的学习行为，以获取新的知识或技能，并重新组织已有的知识结构使之不断改善自身。
- 就是计算机从数据中学习出规律和模式，以应用在新数据上做预测的任务。

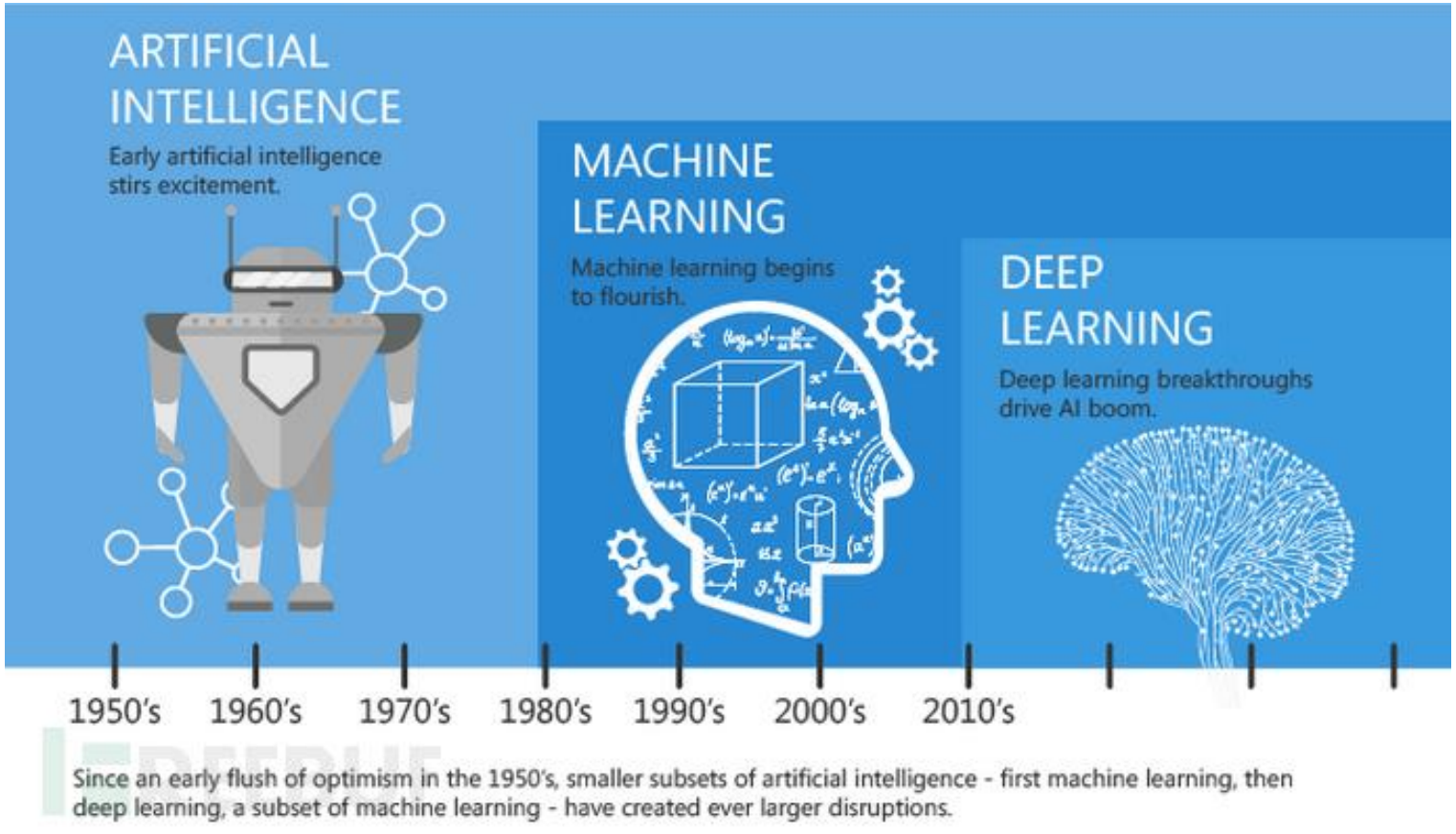


1.1 机器学习是什么

互联网时代

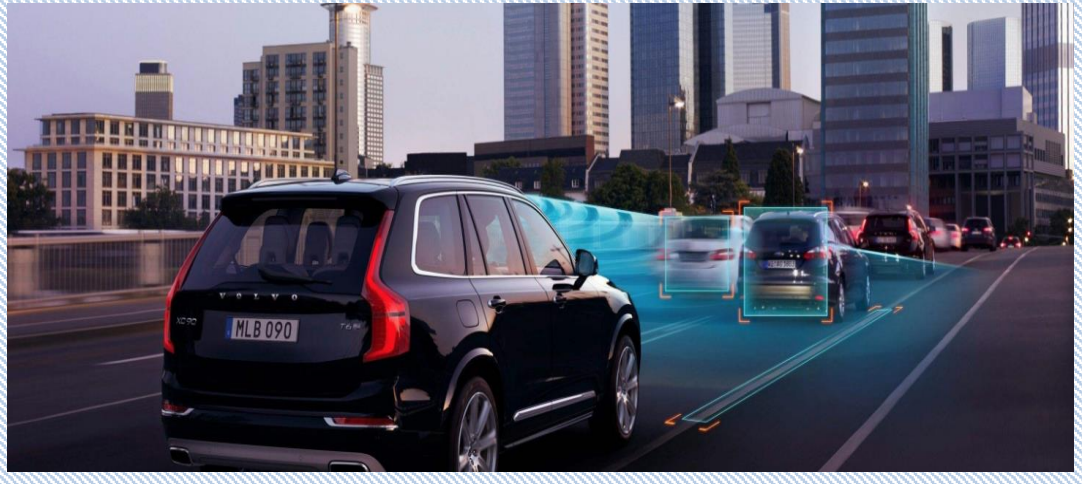
巨大的数据量

超高的数据维度



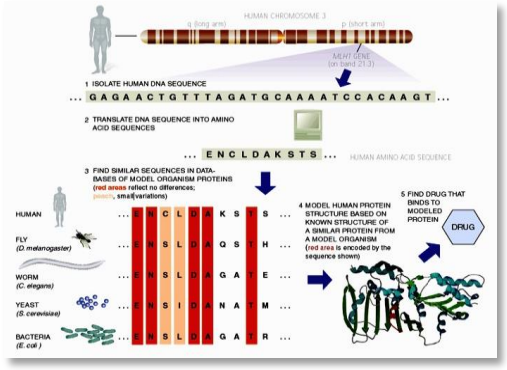
1.2 机器学习几个例子

几个例子



1.2 机器学习几个例子

- 几乎无处不在



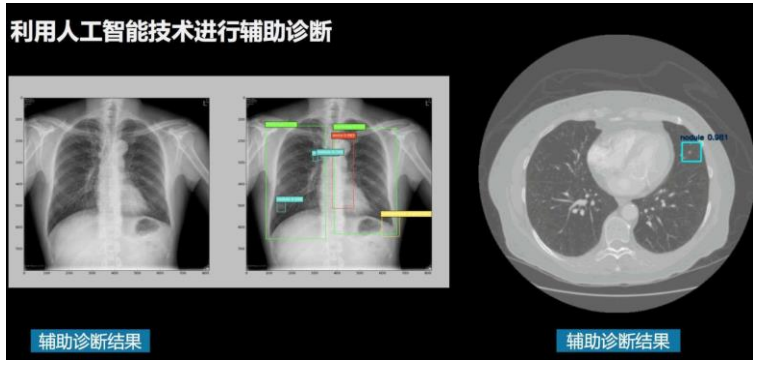
生物信息学

无人驾驶汽车

Web搜索

电商推荐

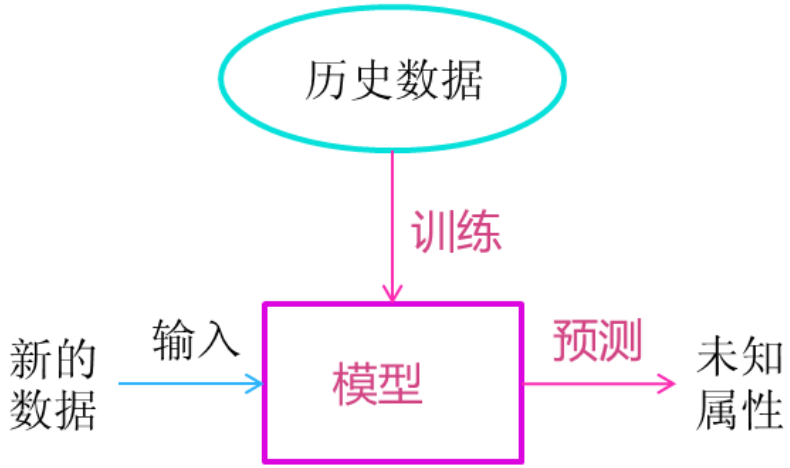
决策助手(DARPA)



智能医疗

要点总结

- 机器学习：计算机从数据中学习出规律和模式，以应用在新数据上做预测的任务。
- 作为一套数据驱动的方法，在互联网、生物、医疗、金融、能源、交通等等领域有广泛应用。



02 机器学习基本概念

2.1 不同类型的问题

2.2 基本术语与概念

2.3 工业界应用方向

2.1 不同类型的问题

01 聚类

机器学习基础
聚类算法原理及应用

02 分类

监督学习算法
分类算法原理及应用

*04 强化学习

*AlphaGo 背后的算法
*深度学习实战项目

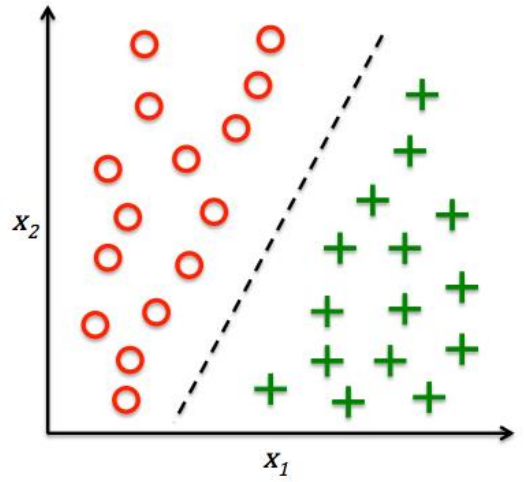
03 回归

回归分析与预测
回归算法原理及应用

2.1 不同类型的问题

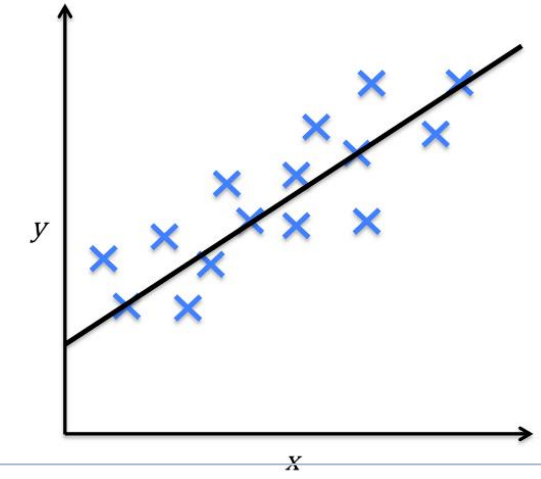
分类问题（监督学习）：

- 根据数据样本上抽取出的特征，判定其属于有限个类别中的哪一个
- 垃圾邮件识别（结果类别：1、垃圾邮件 2、正常邮件）
- 文本情感褒贬分析（结果类别：1、褒 2、贬）
- 图像内容识别识别（结果类别：1、喵星人 2、汪星人 3、人类 4、草泥马 5、都不是）



回归问题（监督学习）：

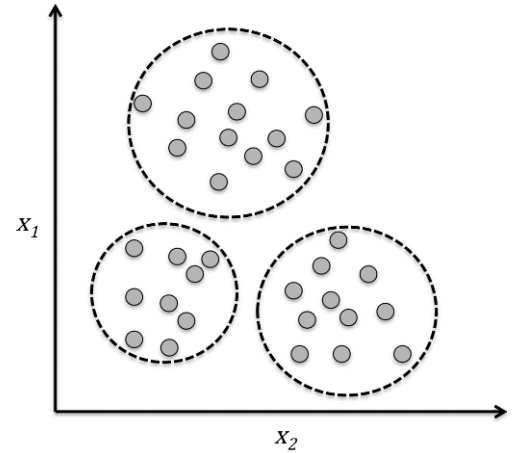
- 根据数据样本上抽取出的特征，预测连续值结果
- 《芳华》票房值
- 魔都房价具体值
- 刘德华和吴彦祖的具体颜值得分



2.1 不同类型的问题

聚类问题(无监督学习)：

- 根据数据样本上抽取出的特征，挖掘数据的关联模式
- 相似用户挖掘/社区发现
- 新闻聚类



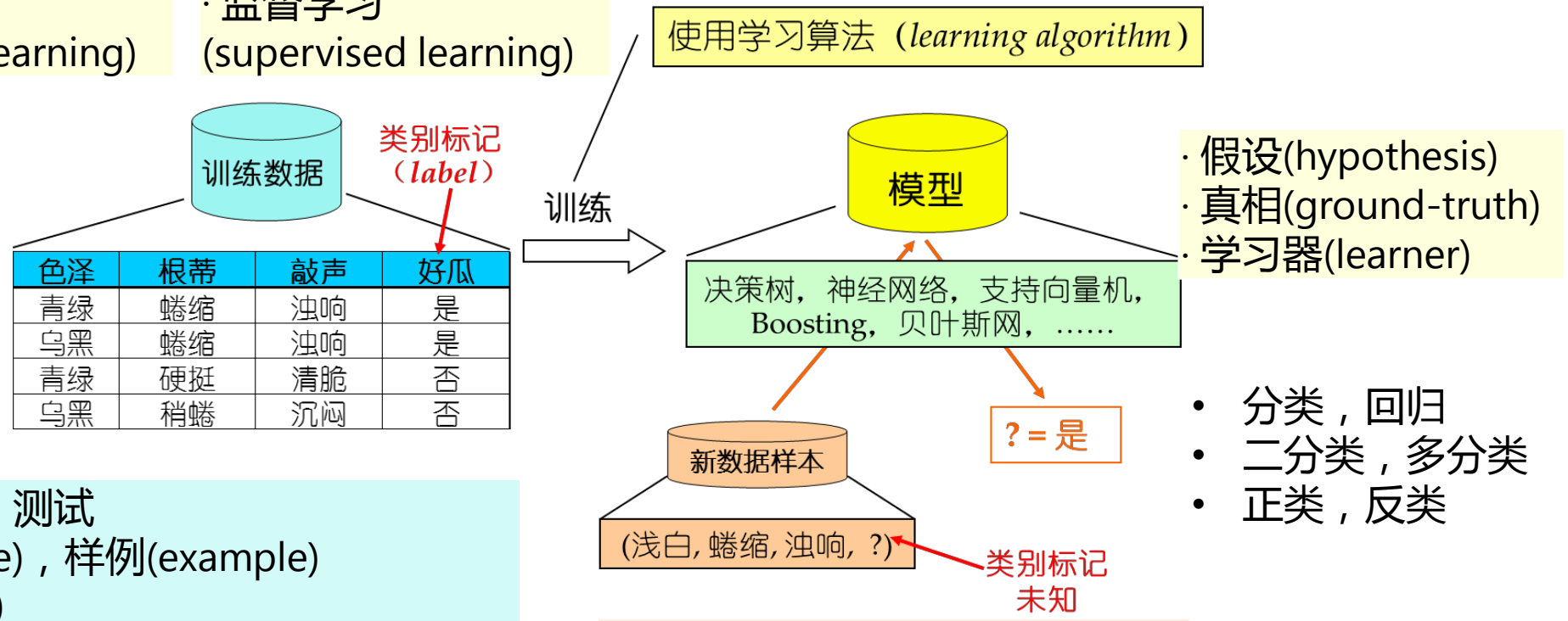
强化问题：

- 研究如何基于环境而行动，以取得最大化的预期利益
- 游戏(“吃鸡”)最高得分
- 机器人完成任务

2.2 基本术语与概念

· 无监督学习
(unsupervised learning)

· 监督学习
(supervised learning)



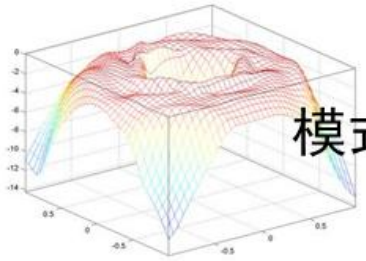
- 数据集; 训练, 测试
- 示例(instance)，样例(example)
- 样本(sample)
- 属性(attribute)，特征(feature)；属性值
- 属性空间，样本空间，输入空间
- 特征向量(feature vector)
- 标记空间，输出空间

- 分类，回归
- 二分类，多分类
- 正类，反类

- 未见样本(unseen instance)
- 未知"分布"
- 独立同分布(i.i.d.)
- **泛化(generalization)**

案例 From 周志华《机器学习》

2.3 实际应用场景



模式识别

计算机视觉

数据挖掘

机器学习

语音识别

统计学习

自然语言处理

要点总结

- 机器学习分类
  - 监督学习：特征+标签
    - 分类：离散个结果中做选择
    - 回归：输出连续值结果
  - 无监督学习：特征
    - 聚类：抱团学习
    - 关联规则
  - 强化学习：从环境到行为映射的学习

- 机器学习概念
  - 样本/示例/样例、特征/属性、训练集、测试集
- 机器学习工业应用方向
  - 自然语言处理、计算机视觉、电商推荐与预估…

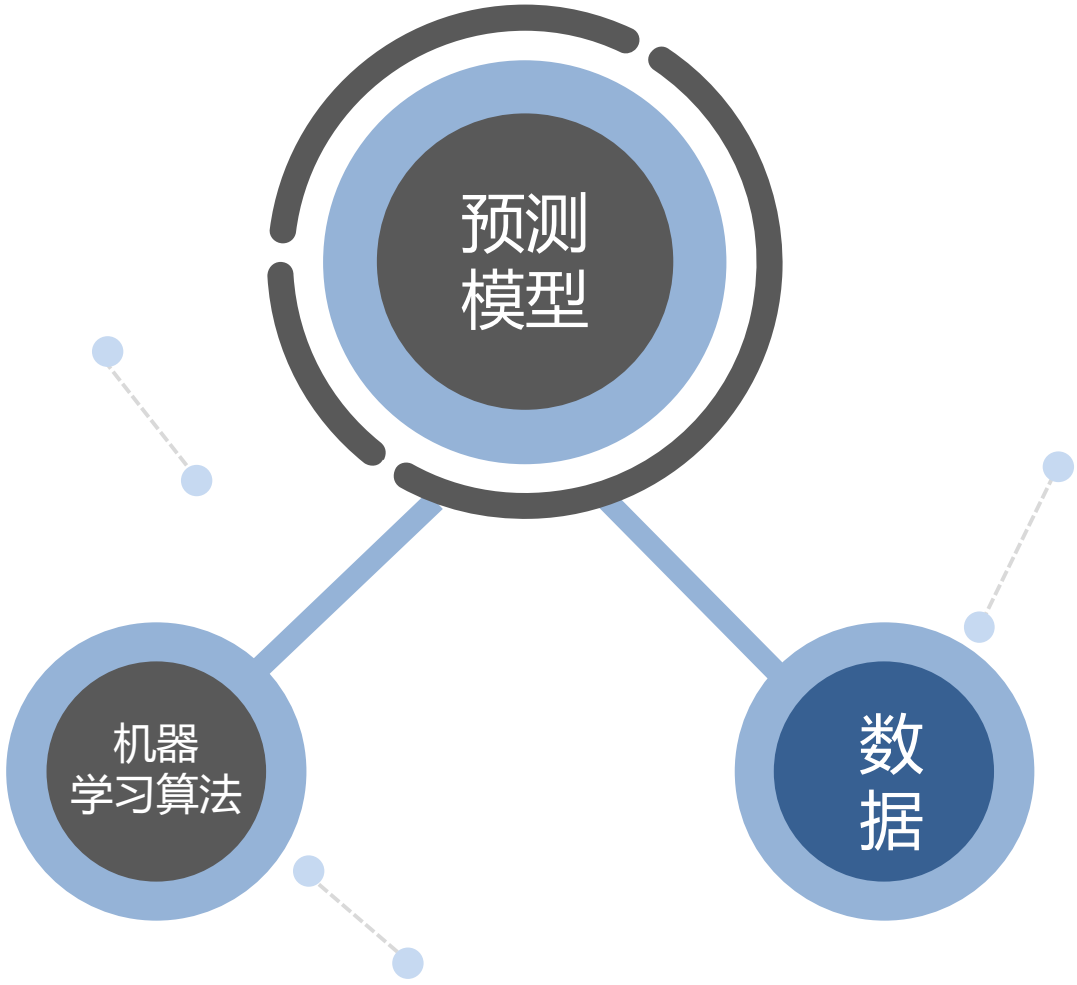
03

机器学习基本流程与工作环节

3.1 机器学习应用几大环节

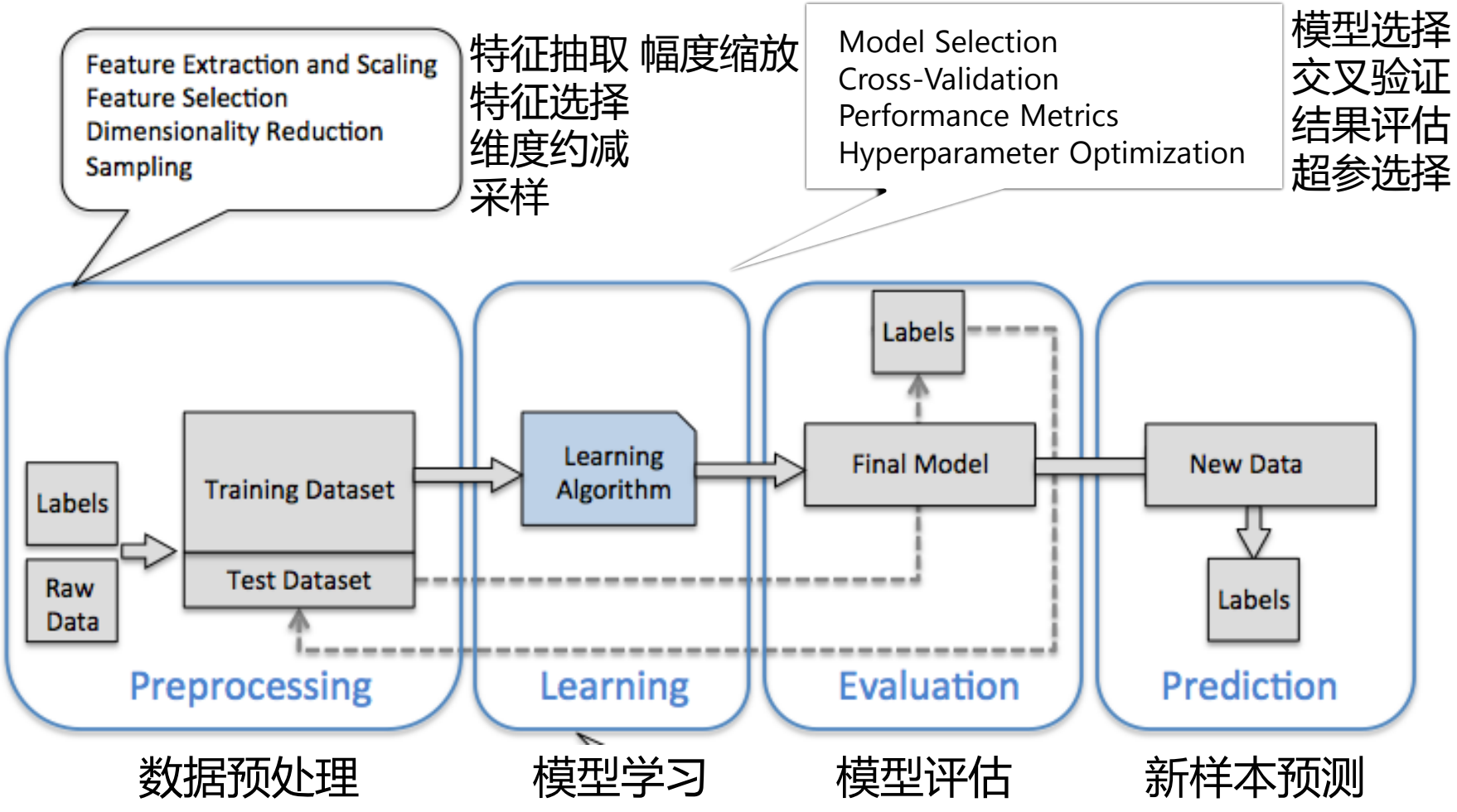
3.2 机器学习不同阶段与作用

3.1 机器学习应用几大环节

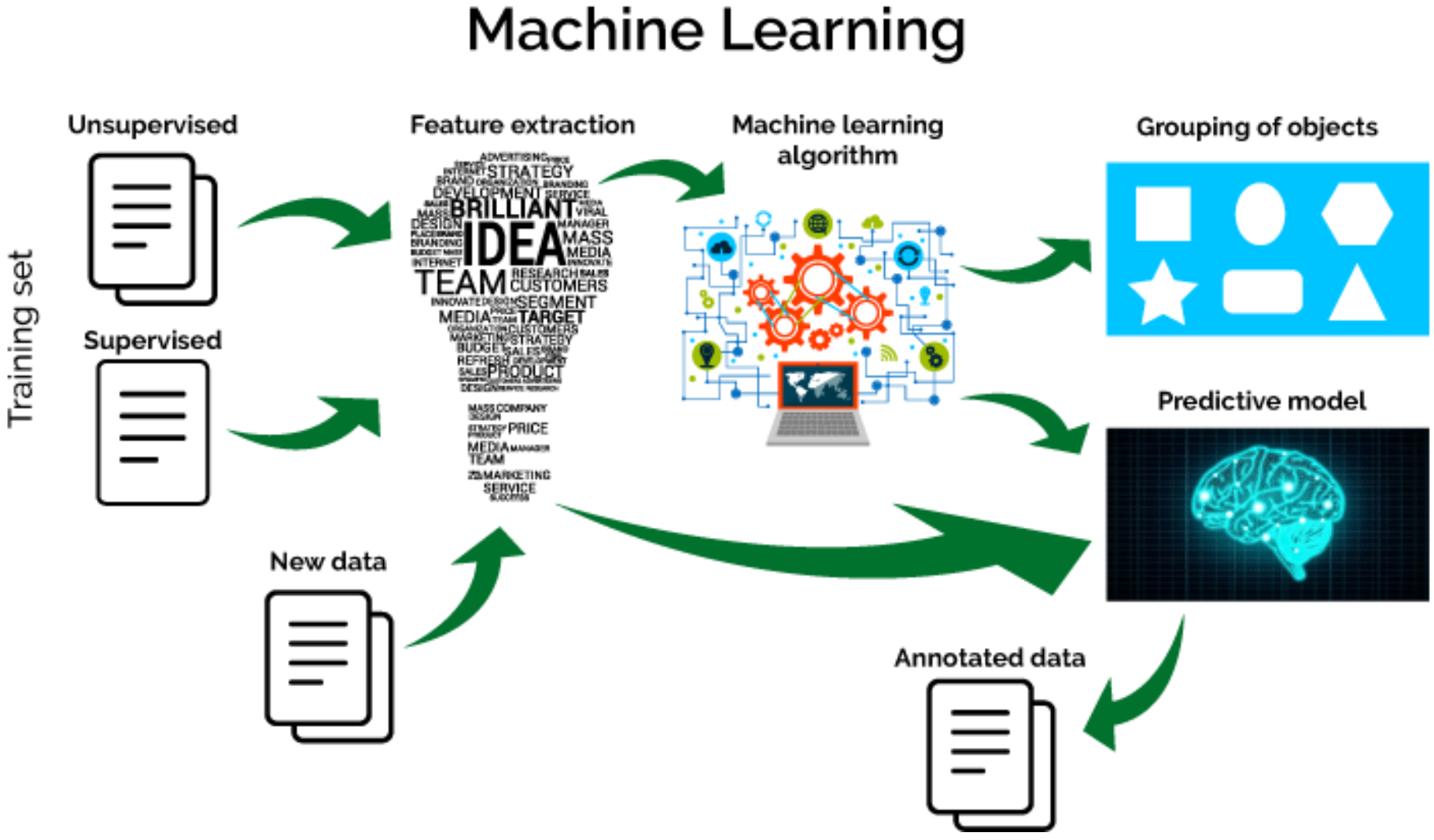


数据与算法展开的
机器学习的应用工作是围绕着

3.2 机器学习不同阶段与作用



3.2 机器学习不同阶段与作用



要点总结

- 数据驱动方法：数据+机器学习算法 = 预测模型
- 机器学习应用阶段
  ① 数据预处理
    - 数据采样、数据切分、特征抽取、特征选择、降维
  ② 模型学习
    - 超参选择、交叉验证、结果评估、模型选择、模型训练
  ③ 模型评估
    - 分类、回归、排序评估标准
  ④ 模型上线

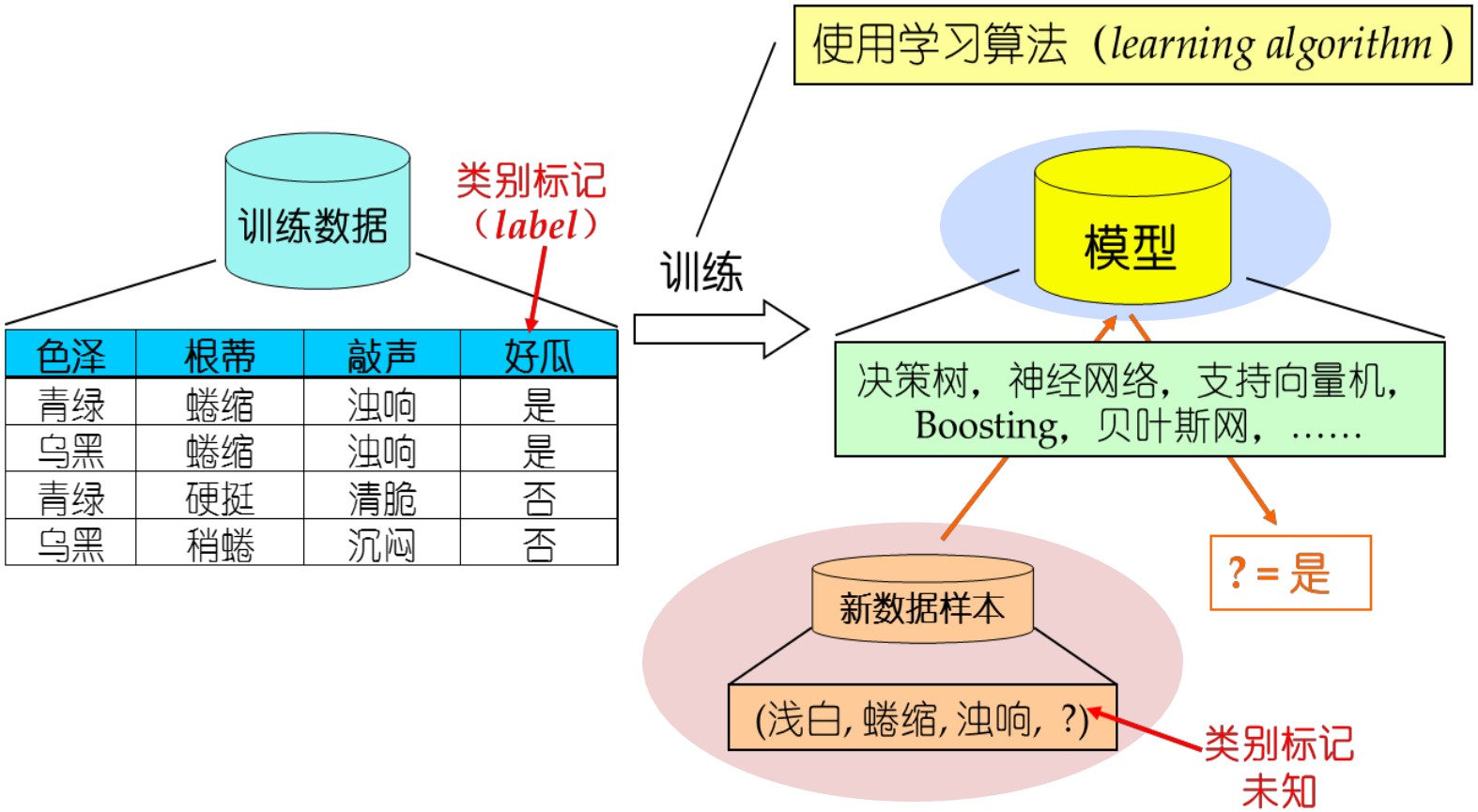
04 机器学习中的评估指标

4.1 机器学习的目标

4.2 机器学习的评估方法

4.3 机器学习的评估度量指标

4.1 机器学习的目标



色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

然而，我们手上没有未知的样本……

Q: 什么模型好？

A: 泛化能力强！

能很好地适用于没见过的样本

例如，错误率低、精度高

案例 From 周志华《机器学习》

4.2 机器学习的评估方法

我们手上没有未知的样本，如何可靠地评估？

关键：获得可靠的“测试集数据”(test set) ？

测试集(用于评估)应该与训练集(用于模型学习)“互斥”

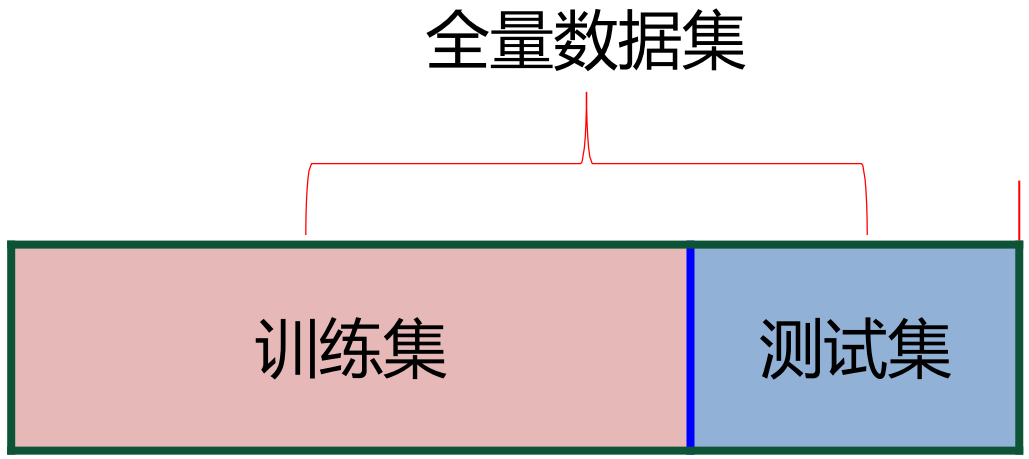
常见方法：

- 留出法(hold-out)
- 交叉验证法(cross validation)
- 自助法(bootstrap)

① 留出法

注意点：

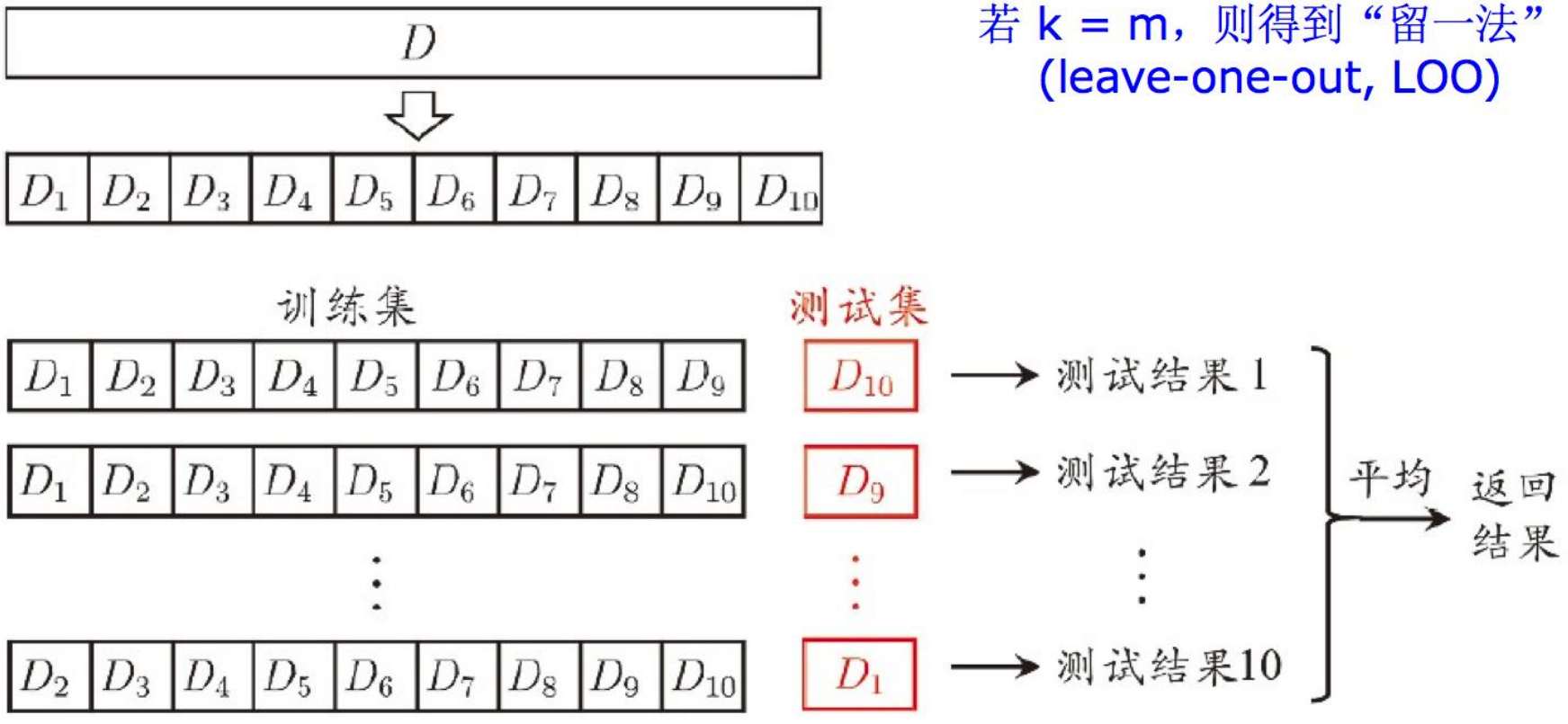
- 保持数据分布一致性
  （例如: 分层采样）
- 多次重复划分
  （例如: 100次随机划分）
- 测试集不能太大、不能太小
  （例如：1/5~1/3）



② k折交叉验证

若 k = m，则得到“留一法”
(leave-one-out, LOO)

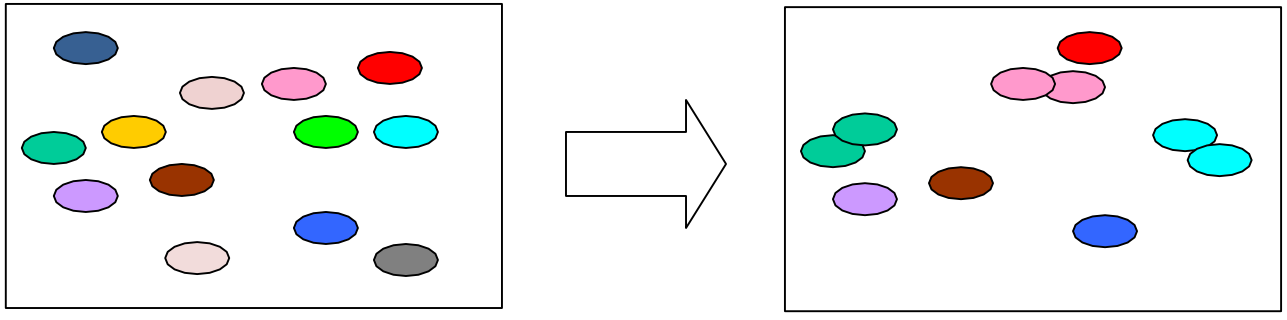
典型的
10折交叉验证



③ 自助法(bootstrap)

基于“自助采样”的方法(bootsrap sampling)

别称：“有放回采样”、“可重复采样”



约有36.8% 的样本不出现

$$\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^{m}\mapsto\frac{1}{e}\approx 0.368$$

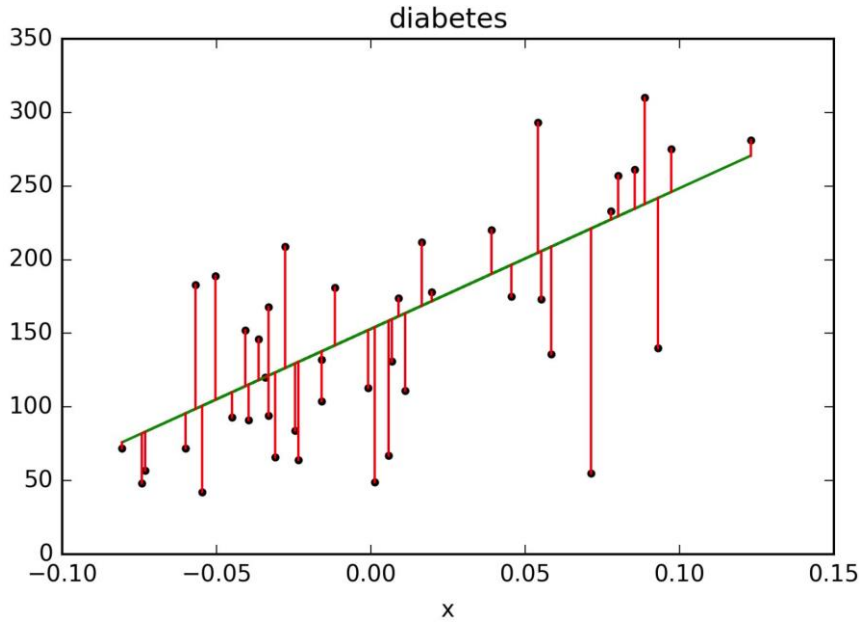
- 训练集与原样本集同规模
- 数据分布有所改变

“包外估计”(out-of-bag estimation)

4.3 机器学习的评估度量标准

- 性能度量（performance measure）是衡量模型泛化能力的数值评价标准，反映了当前问题（任务需求）
- 使用不同的性能度量可能会导致不同的评判结果

关于模型"好坏"的判断，不仅取决于算法和数据，还取决于当前任务需求。



比如：回归（regression）任务常用均方误差：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2$$

4.3 机器学习的评估度量标准

- **分类问题**的常用性能度量
  - 错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

  - 精度：

$$\begin{aligned}\operatorname{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=y_i\right)\\ &= 1-E(f;D)\ .\end{aligned}$$

4.3 机器学习的评估度量标准

- **分类问题**的常用性能度量　　二分类混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- 查准率(准确率)： $P = \frac{TP}{TP + FP}$
- 查全率(召回率)： $R = \frac{TP}{TP + FN}$

4.3 机器学习的评估度量标准

- **分类问题**的常用性能度量　　查准率 vs. 查全率

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- F1值　　$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

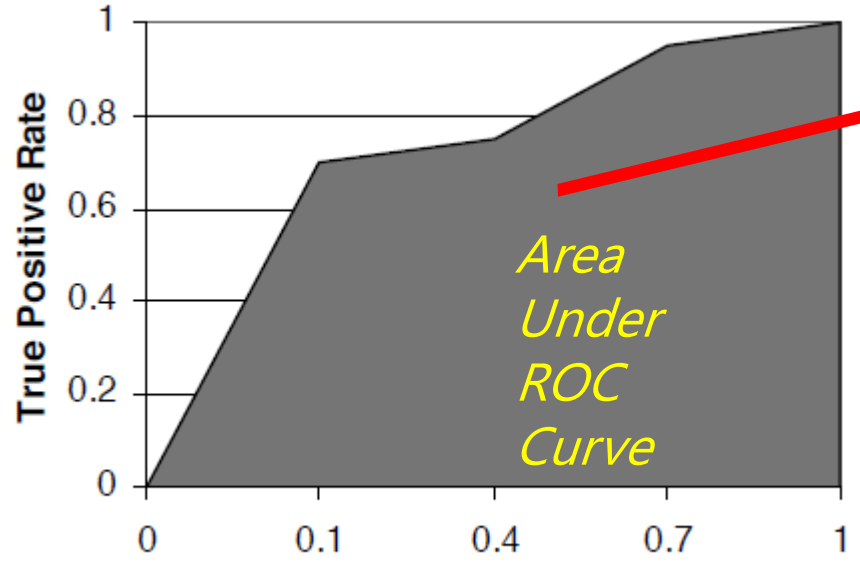
$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

$\beta > 1$ 时查全率有更大影响; $\beta < 1$ 时查准率有更大影响

机器学习的评估度量标准

- **分类问题**的常用性能度量

ROC && AUC



ROC (Receiver Operating Characteristic) Curve

[Green & Swets, Book 66; Spackman, IWML'89]

AUC (Area Under the ROC Curve)

AUC越大，结果越好

$$tpr = \frac{TP}{TP+FN} = \frac{TP}{m_+}$$

$$fpr = \frac{FP}{FP+TN} = \frac{FP}{m_-}$$

$$\mathrm{AUC} = 1 - \frac{1}{m^+ m^-} \sum_{\boldsymbol{x}^+ \in D^+} \sum_{\boldsymbol{x}^- \in D^-} \left(\mathbb{I}\left(f(\boldsymbol{x}^+) < f(\boldsymbol{x}^-)\right) + \frac{1}{2}\mathbb{I}\left(f(\boldsymbol{x}^+) = f(\boldsymbol{x}^-)\right) \right)$$

4.3 机器学习的评估度量标准

- **回归问题**的常用性能度量
  - MAE(Mean Absolute Error) 平均绝对误差

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

  - MSE(Mean Square Error) 均方误差

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2$$

  - RMSE(Root Mean Square Error) 均方根误差

$$RMSE = \sqrt{MSE}$$

  - R平方

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2}$$

要点总结

- 机器学习目标
  - 拿到有泛化能力的“好模型”
- 机器学习的评估方法
  - 留出法、交叉验证法、自助法
- 机器学习的评估度量标准
  - 分类问题
    - 错误类、精度、召回率/准确率、混淆矩阵、F1值、AUC
  - 回归问题
    - MAE、MSE、RMSE、R平方

05 机器学习算法一览

5.1 机器学习算法一览

5.2 机器学习算法可视化理解

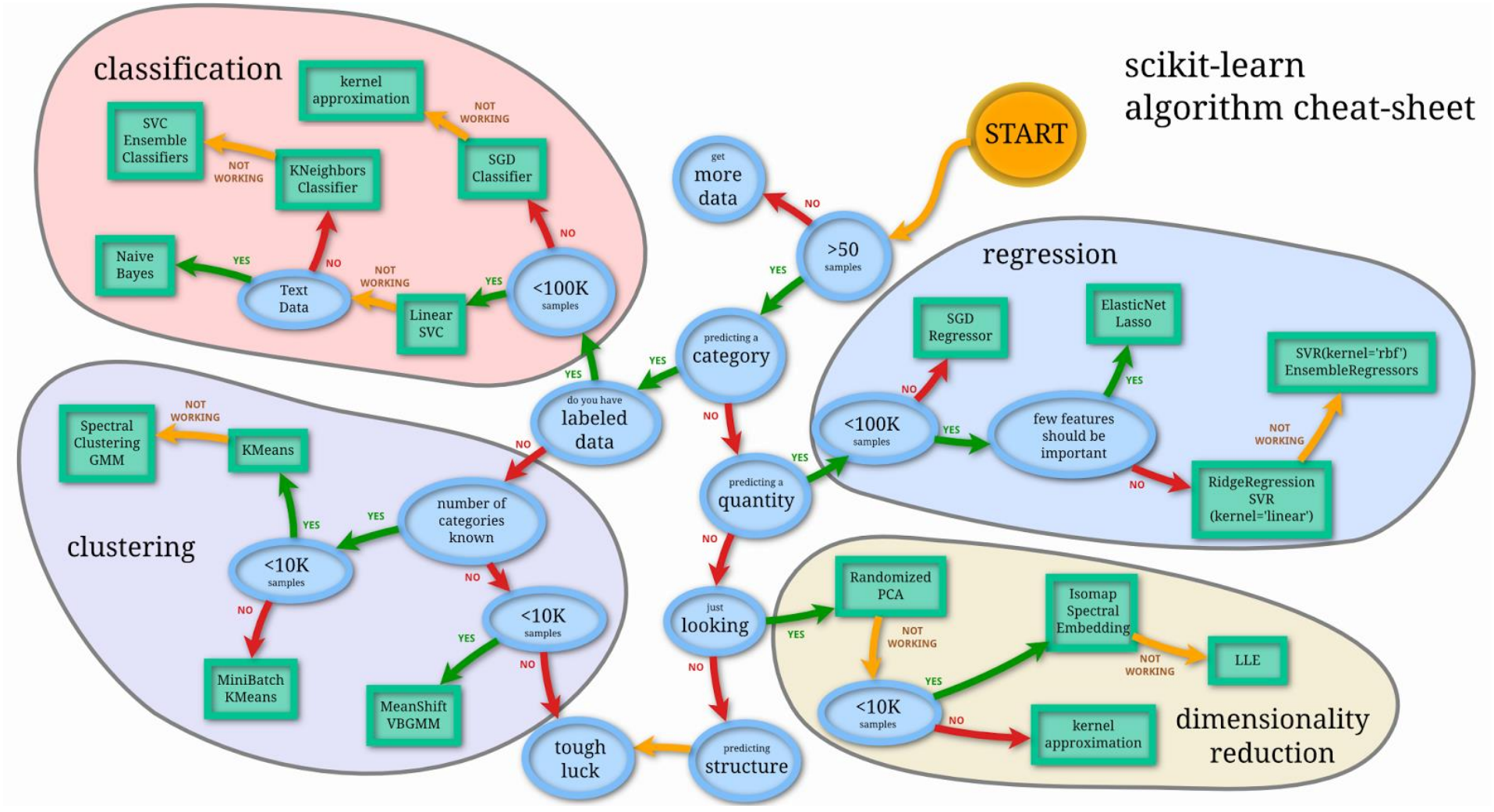
Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

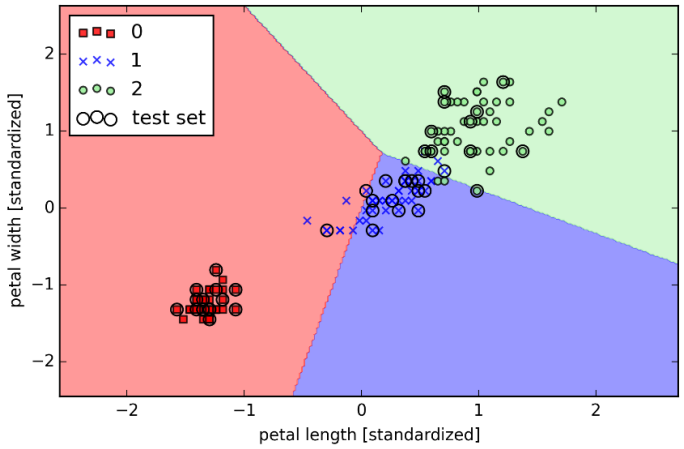
5.1 机器学习算法一览

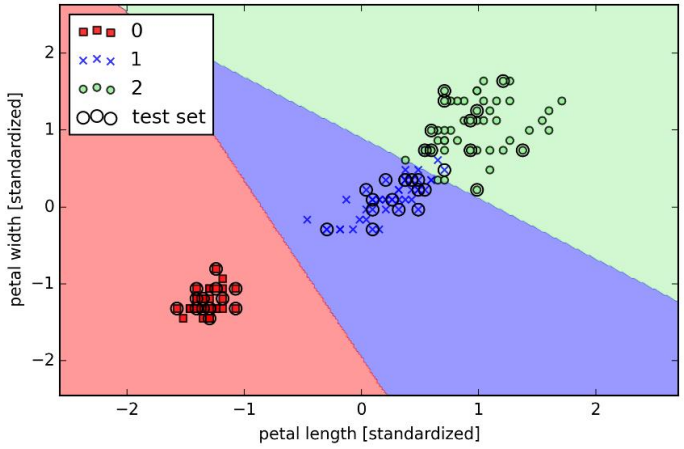

scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
NO
just looking
NO
predicting structure
tough luck
classification
kernel approximation
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
<100K samples
NO
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression SVR (kernel='linear')
clustering
Spectral Clustering GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift VBGMM
dimensionality reduction
Randomized PCA
YES
NOT WORKING
Isomap Spectral Embedding
NOT WORKING
LLE
<10K samples
YES
NO
kernel approximation

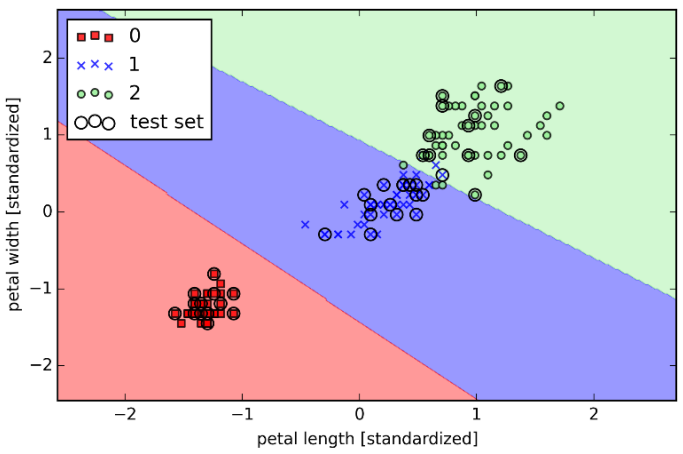
5.2 机器学习算法可视化理解

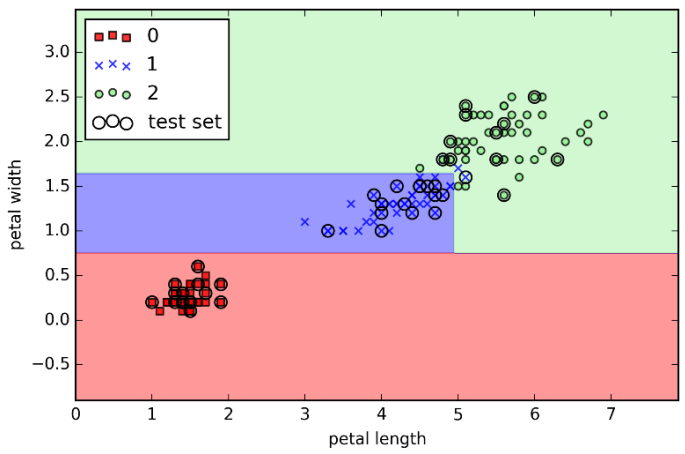
不同算法在完成分类与回归问题时候，有不同的处理方式。

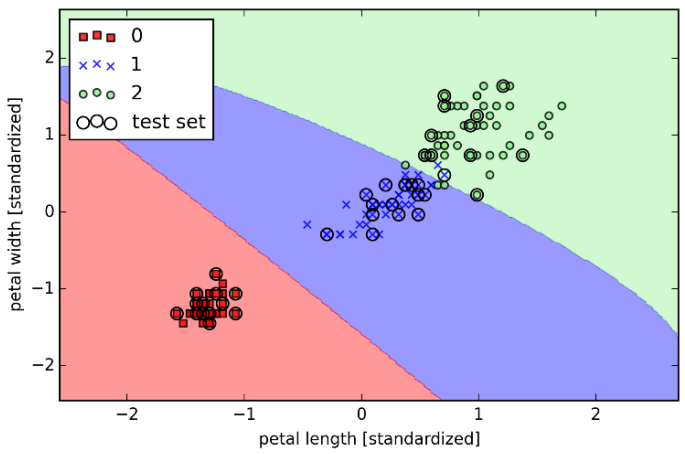
详见课程动态演示







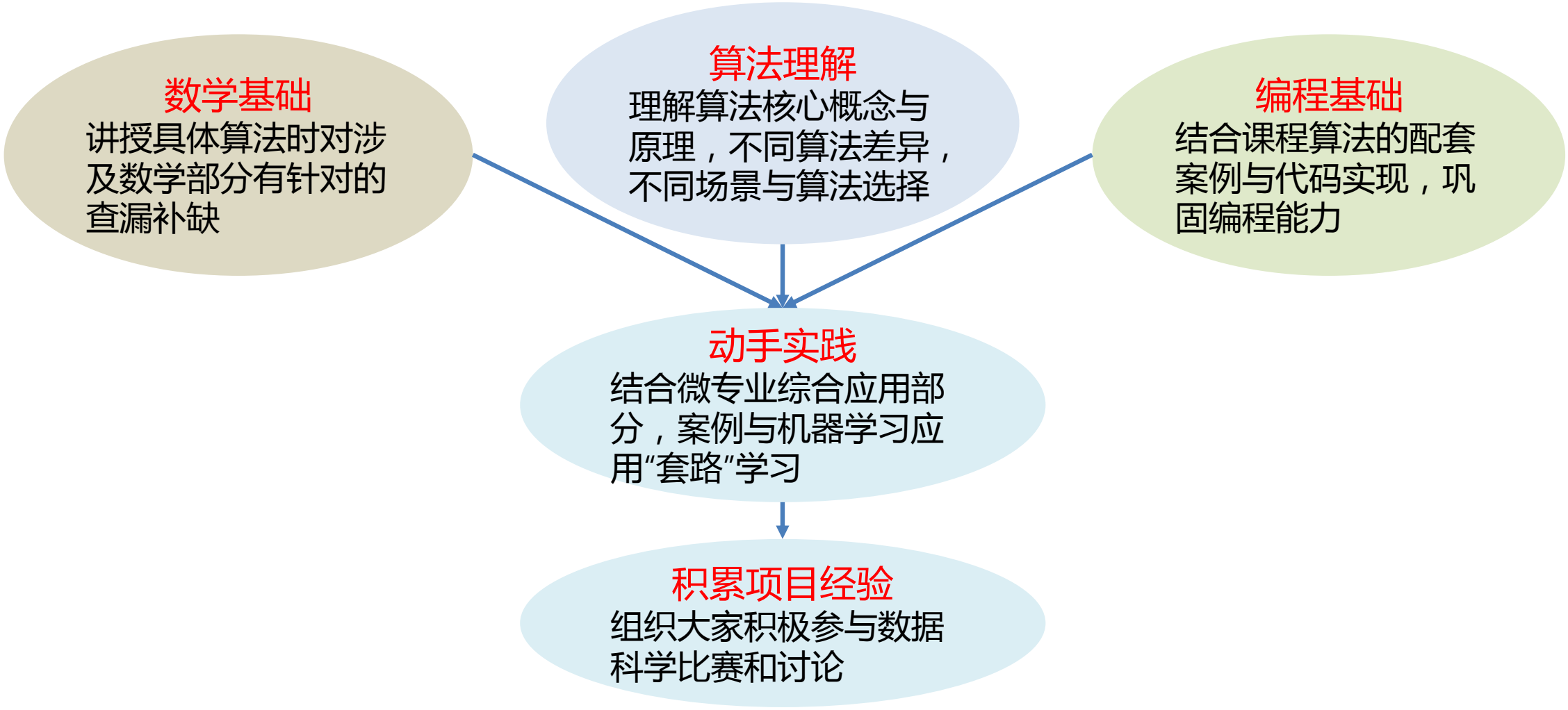




要点总结

- 机器学习算法
  - 监督学习
    - 分类：K最近邻、逻辑回归、朴素贝叶斯、支持向量机、树模型…
    - 回归：线性回归、多项式回归、岭回归、树模型回归…
  - 无监督学习
    - 聚类：K-means，层次聚类、密度聚类、GMM…
    - 关联规则：Fpgrowth
- 机器学习算法可视化理解
  - 分类问题
    - 不同的算法在尝试生成不同的决策边界，从而完成分类
    - 回归类问题有不同的拟合方式

附 结合微专业课程的学习



参考文献/Reference

- Prof. Andrew Ng. Machine Learning. Stanford University
- 李航，统计学习方法，清华大学出版社，2012
- 周志华，机器学习，清华大学出版社，2016
- Scikit-learn，http://scikit-learn.org/stable/index.html


THANK YOU !
Machine Learning Engineer
机器学习工程师微专业
稀牛学院
網易 NETEASE