第3编 Python机器学习

中国人民大学
RENMIN UNIVERSITY OF CHINA

第1讲 机器学习概述

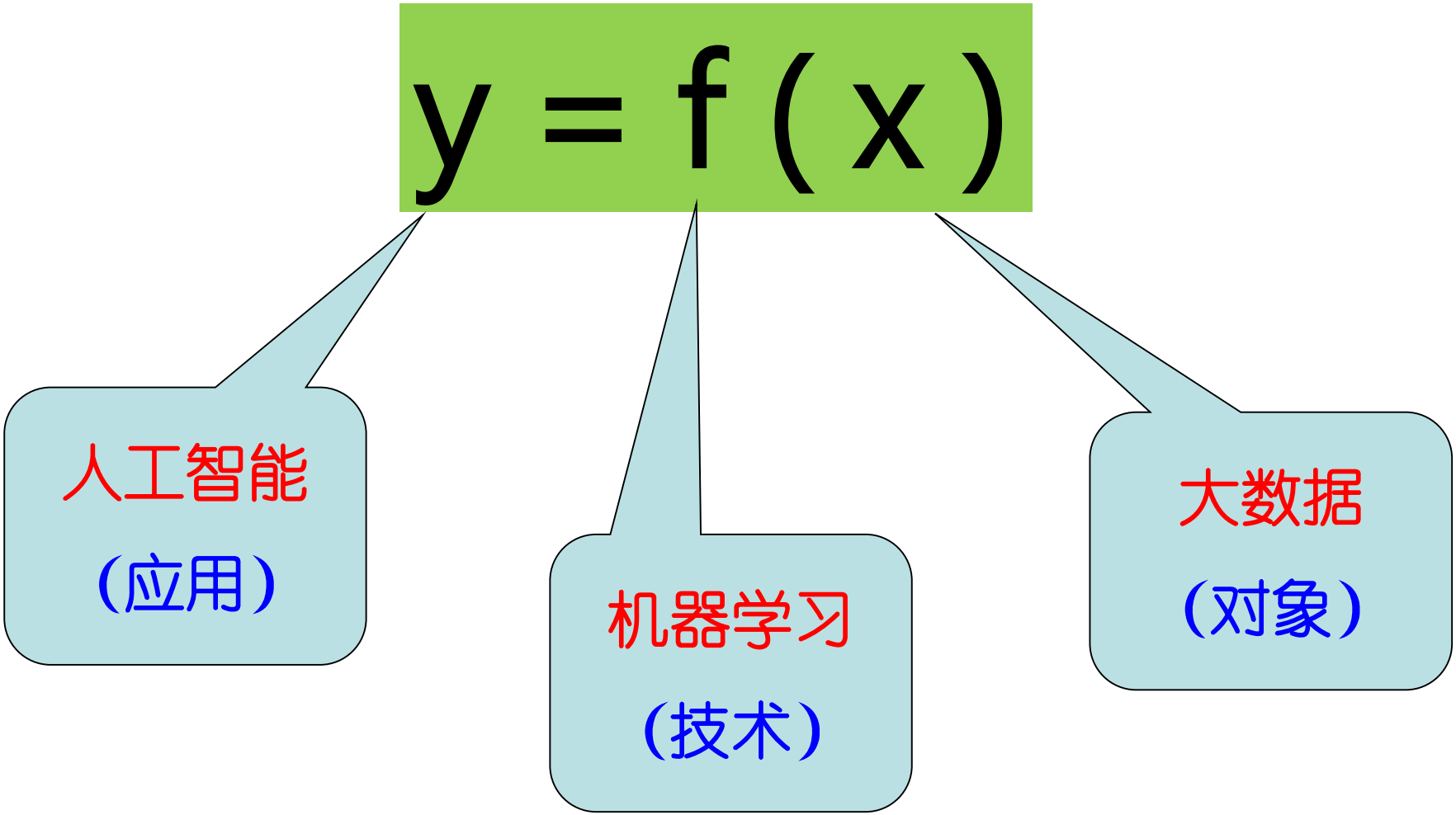
余力

buaayuli@ruc.edu.cn

中国人民大学
RENMIN UNIVERSITY OF CHINA

1. 机器学习相关概念

对“大数据分析与应用”的理解



大数据分析的内涵

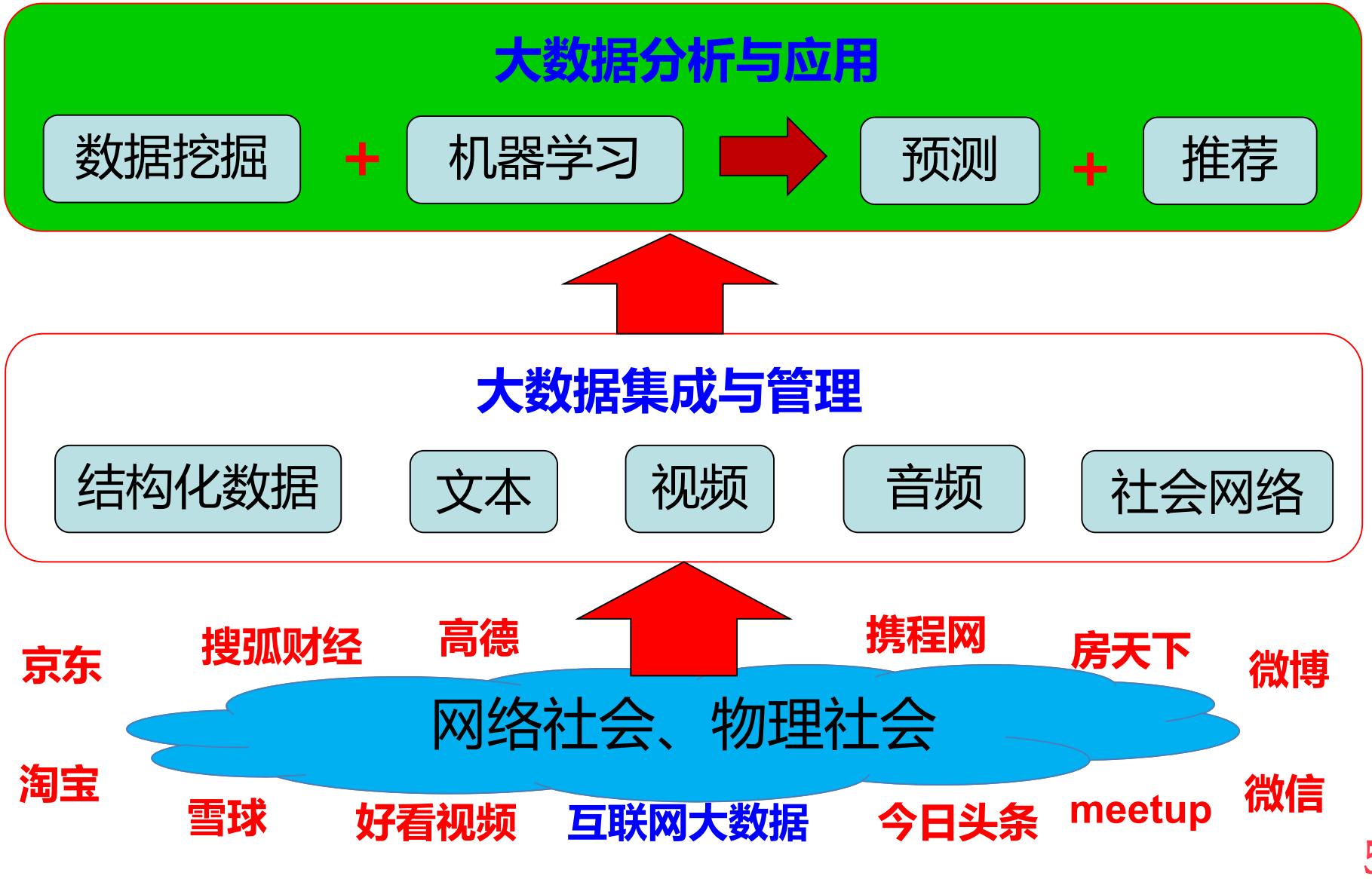
预测性分析

（机器学习）

描述性分析

（统计图、云图）

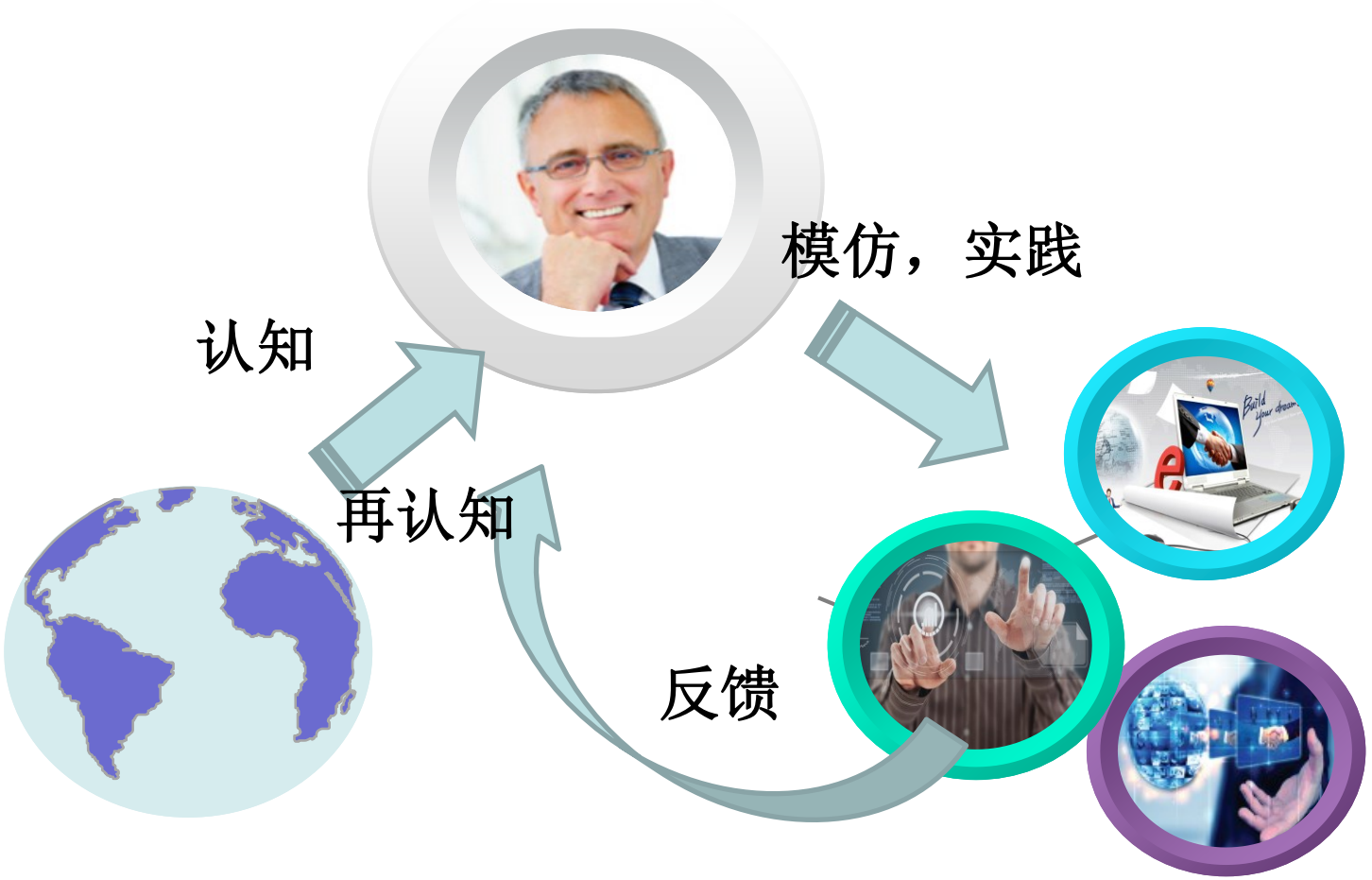
大数据分析与推荐


大数据分析与应用
数据挖掘
+
机器学习
预测
+
推荐
大数据集成与管理
结构化数据
文本
视频
音频
社会网络
京东
搜弧财经
高德
携程网
房天下
微博
网络社会、物理社会
淘宝
雪球
好看视频
互联网大数据
今日头条
meetup
微信

中国人民大学
RENMIN UNIVERSITY OF CHINA

2. 机器学习过程

人类学习过程



什么是机器学习?

机器学习就是让计算机从大量的数据中学习到相关的规律和逻辑，然后利用学习来的规律来进行决策，推理和识别等。

发现规律

规则

1. 发件人邮件地址异常；
2. 标题含有“低价促销”
3. …

测试数据

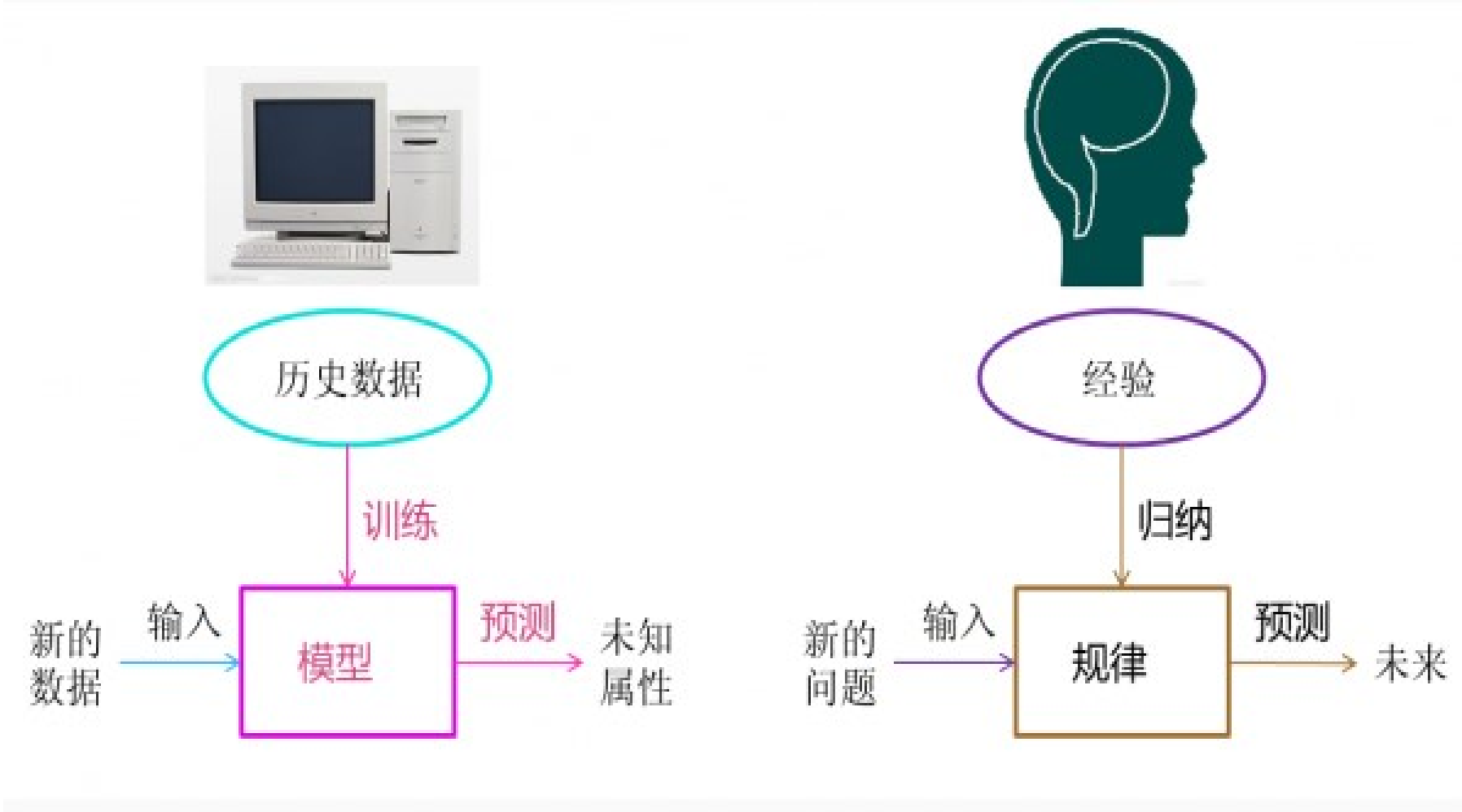
邮件X_i

测试

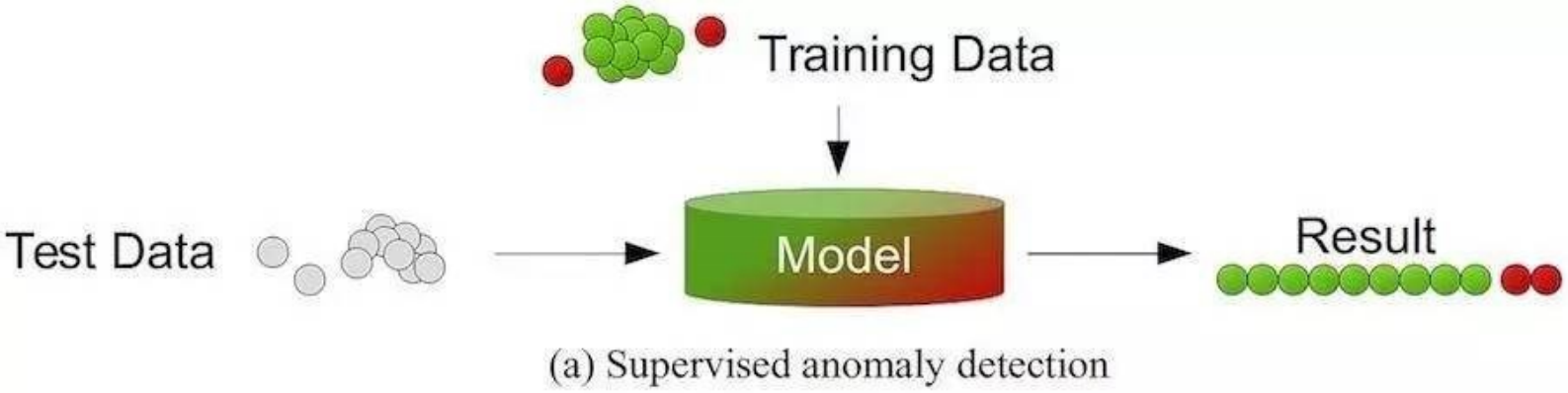
结果评估

Y_i：垃圾or正常

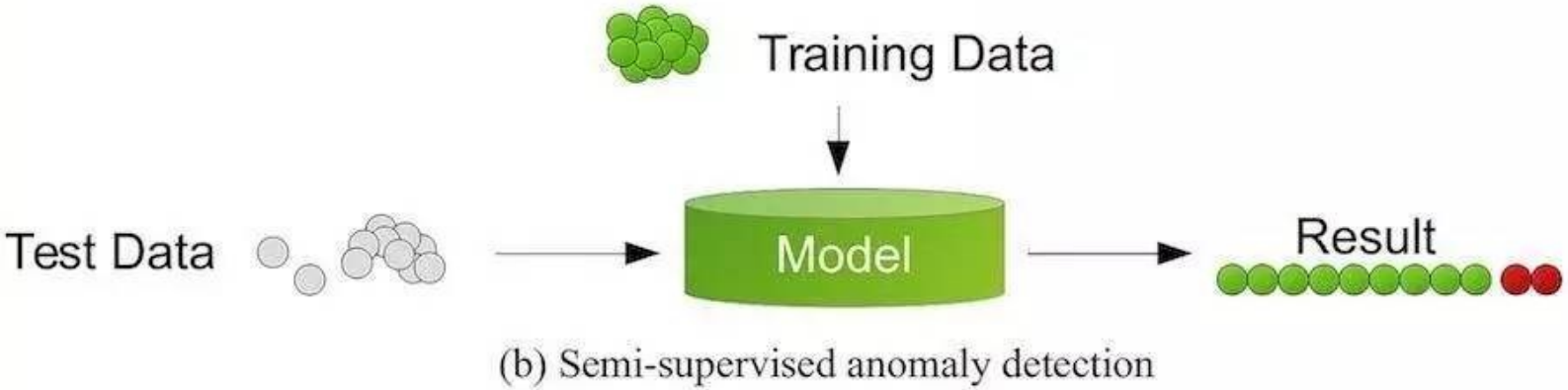
机器学习 vs. 人类学习



机器学习分类



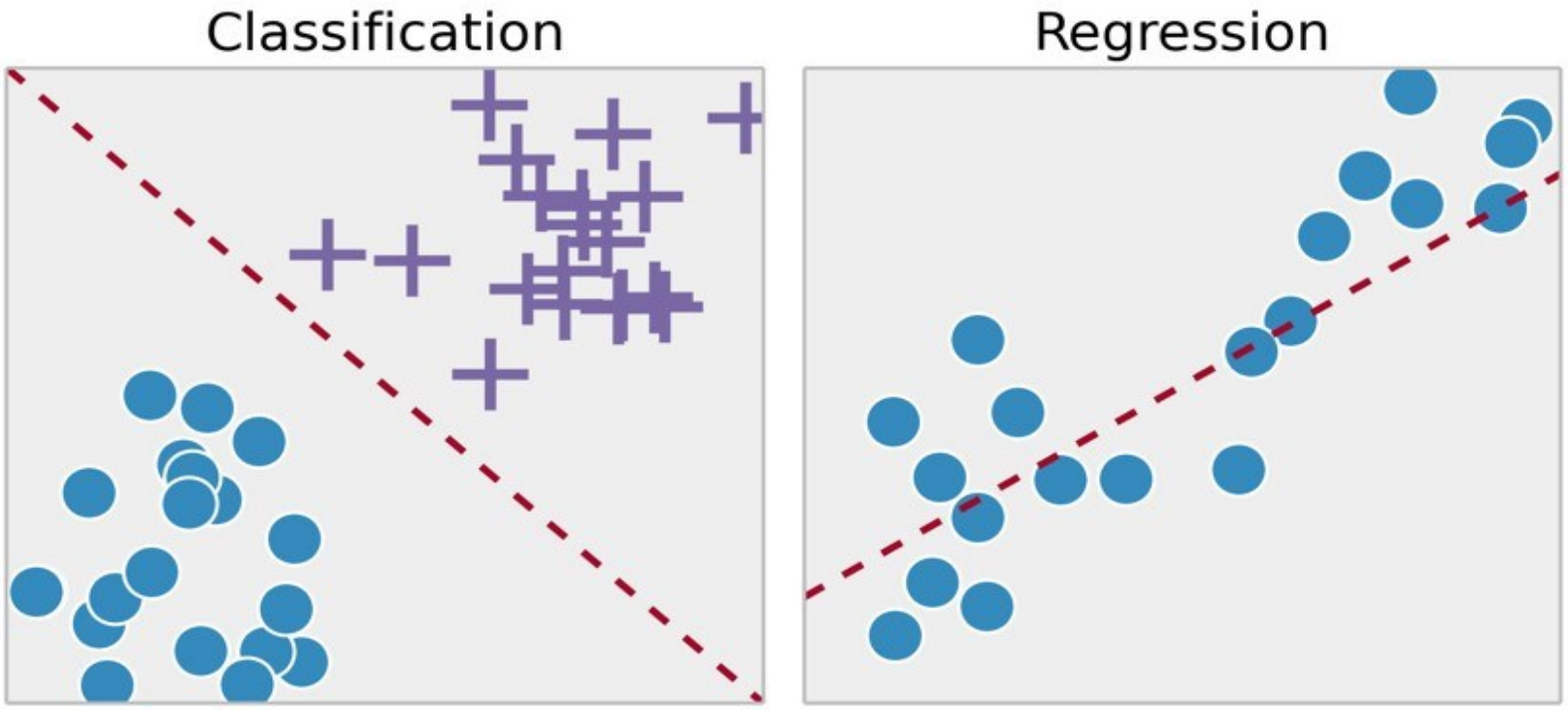
(a) Supervised anomaly detection



(b) Semi-supervised anomaly detection

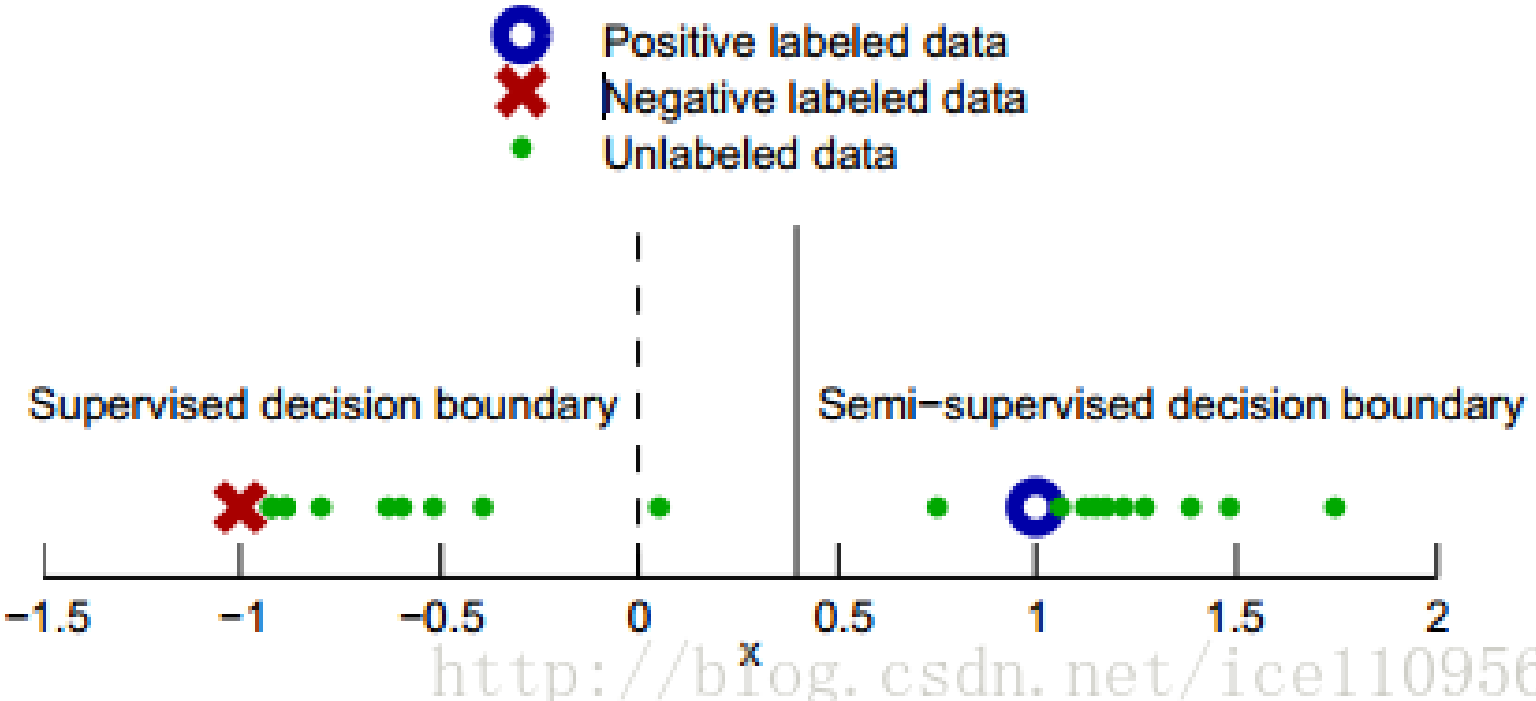
(c) Unsupervised anomaly detection

监督学习

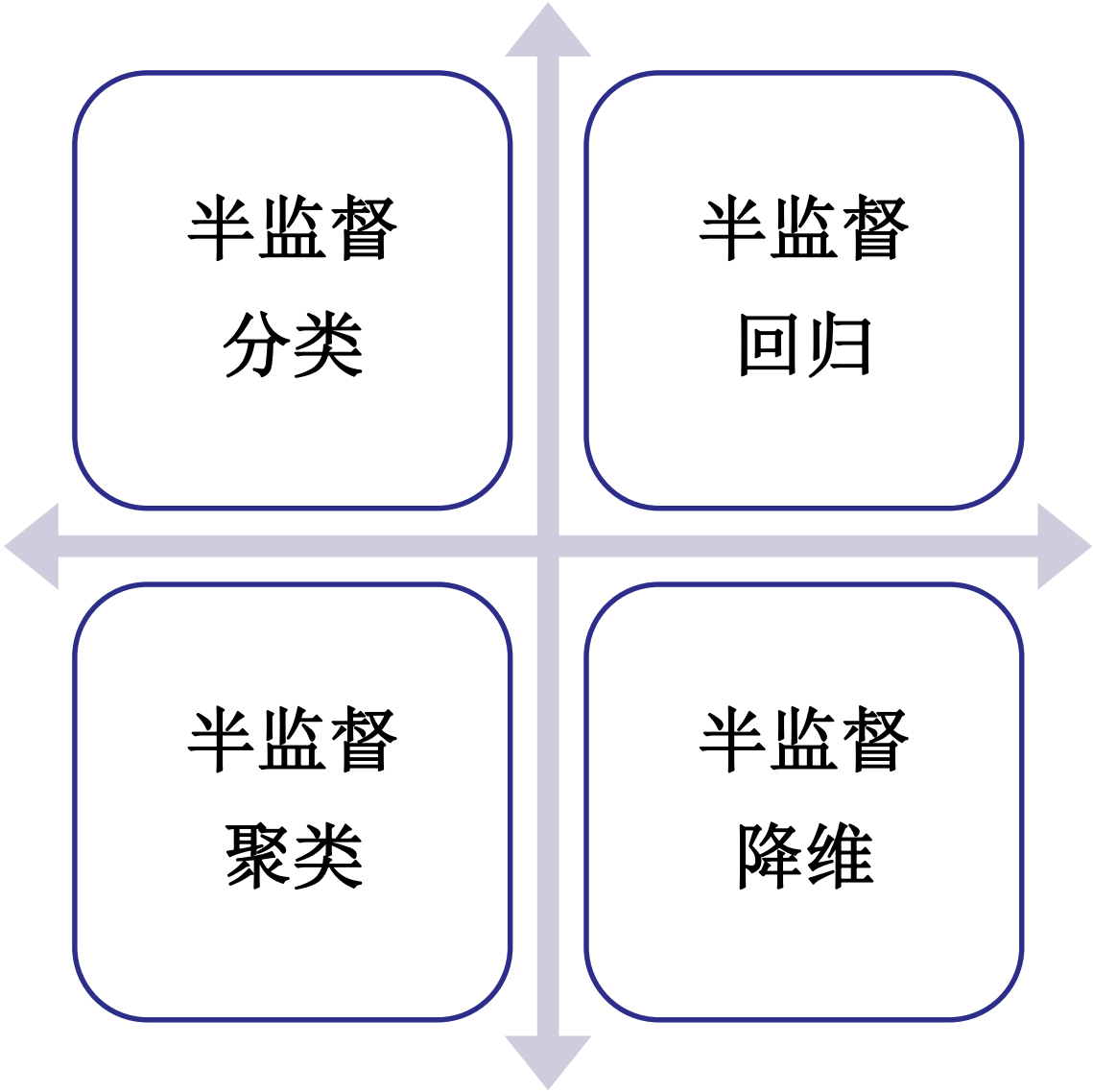


半监督学习

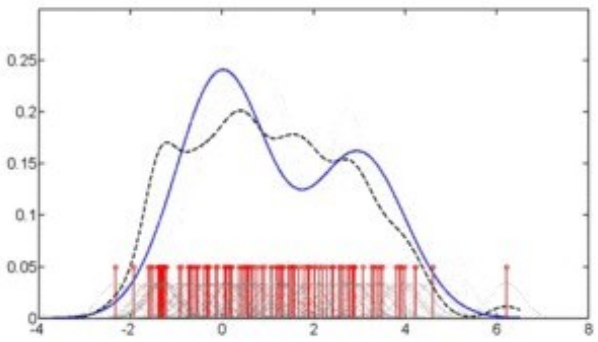
- 1. 有标记样本难以获取.
  - 需要专门的人员,特别的设备,额外的开销等等.
- 2. 无标记的样本相对而言是很廉价的.



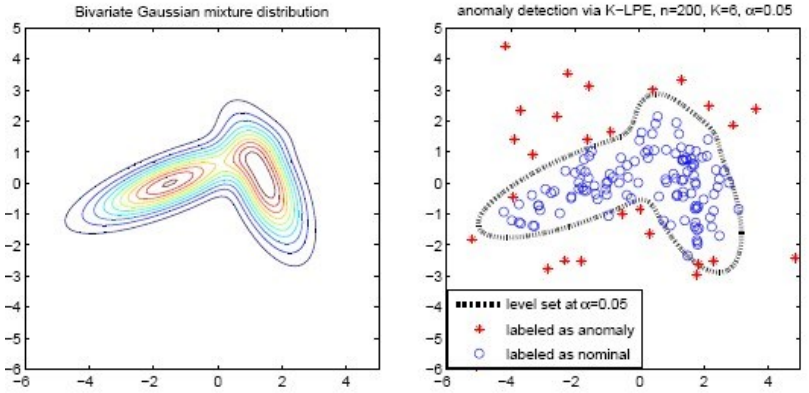
半监督学习的应用场景



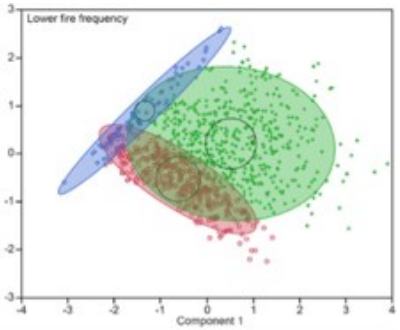
无监督学习



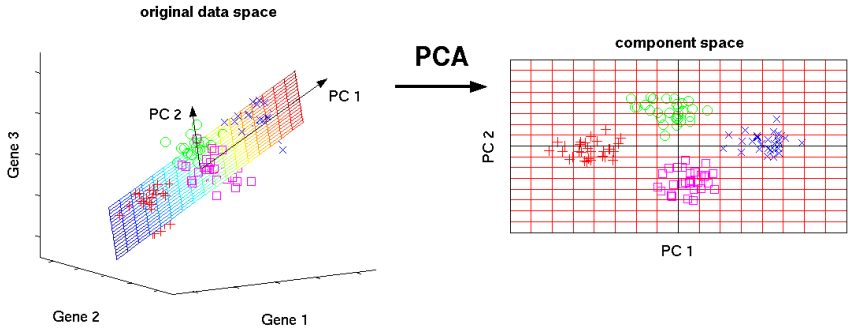
密度估计



异常检测



聚类



降维

机器学习应用实例

1、对语言、文字的认知与识别

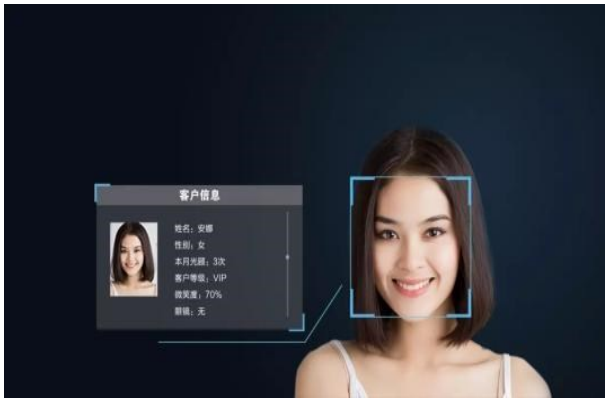
2、对图像、场景、自然物体的认知与识别

3、对规则的学习与掌握

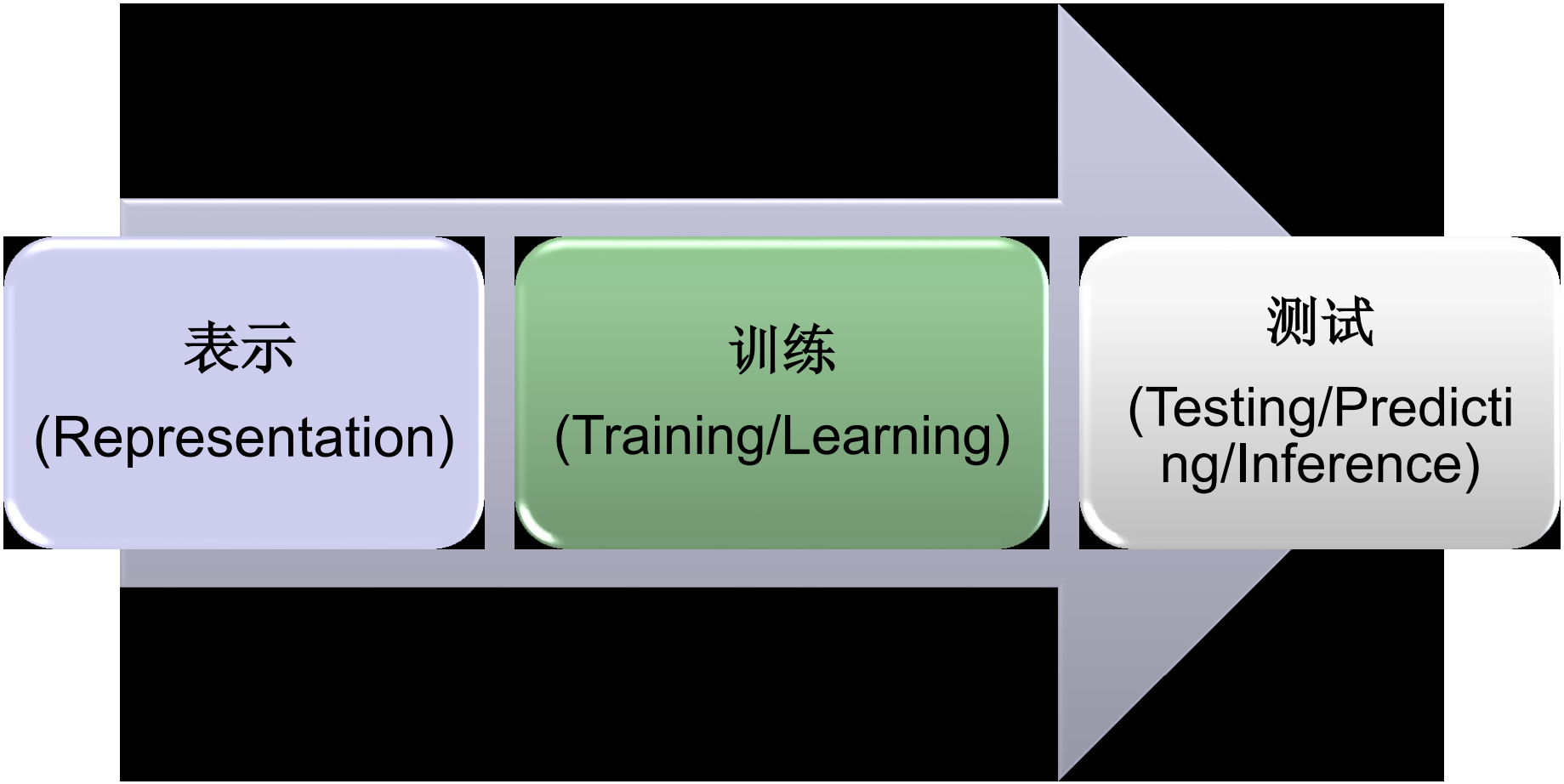
例如：下雨天要带伞，生病要吃药，天冷了要穿厚衣服等

4、对复杂事物的推理与判断能力

例如：好人与坏人的辨别能力，事物的正误的判断能力



机器学习基本过程


表示
(Representation)
训练
(Training/Learning)
测试
(Testing/Predicting/Inference)

例子：天气预报

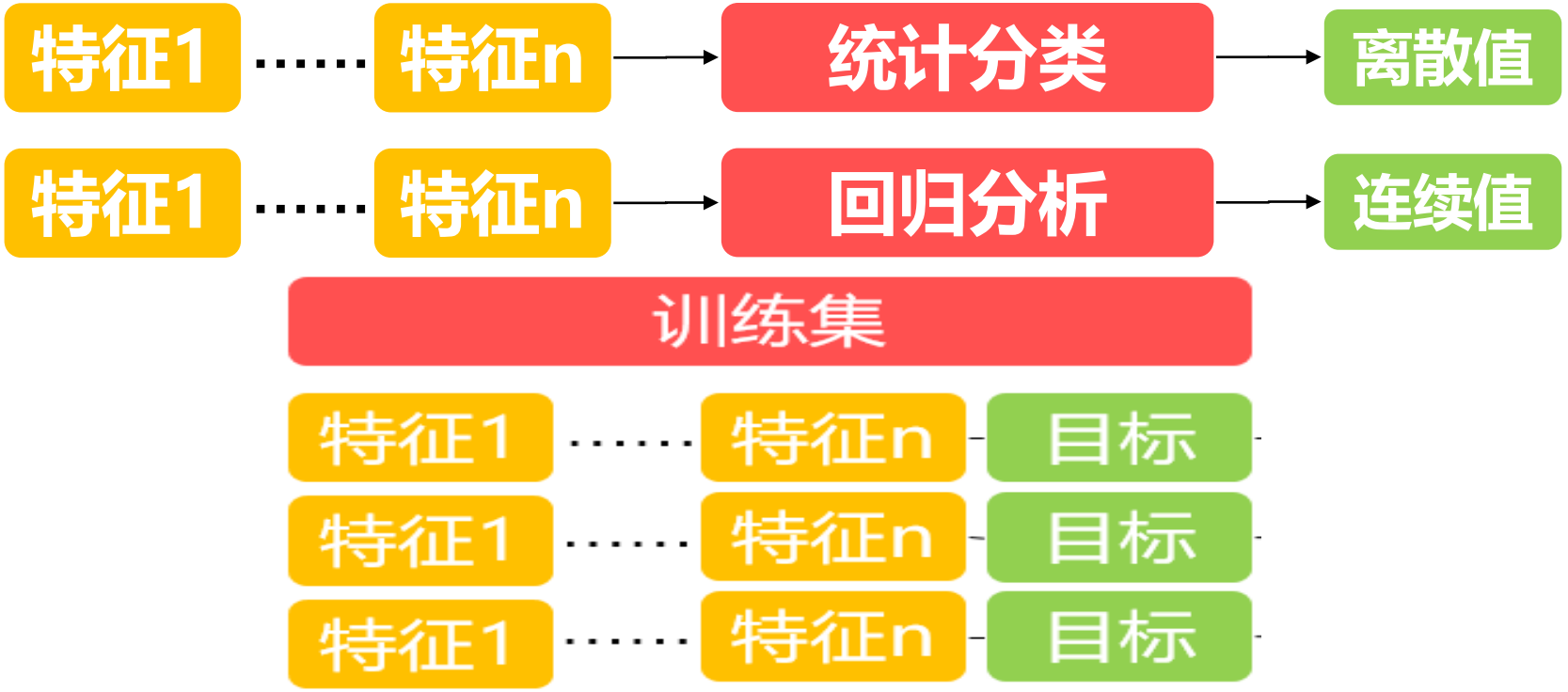
- 目标：预测明天北京会不会下雨
- 数据：过去10年北京每一天的天气数据
  - 那天是否下雨：是/否
  - 那天的前一天傍晚18点的气温、相对湿度、风向、风速、气压等(特征)
  - 某条数据：<18, 20, 东, 15, 80, 是>
- 训练：学习得到规律（模型）
- 预测：给定今天傍晚18点的气温、相对湿度、风向、风速、气压等、根据模型预测明天是否下雨

机器学习的关键问题

- 【表示】如何表示数据样本?
  - 通常用一个向量来表示一个样本，向量中选用哪些特征是关键
- 【训练】如何找出规律【模型+策略+算法】*
  - 通常变成一个选择题，给你n个候选的模型让你选。【模型】
  - 确定选择的标准（什么样的模型才叫好模型）【策略】
  - 如何快速地从n个模型中选出最好的【算法】
- 【测试】如何根据找到的规律进行预测

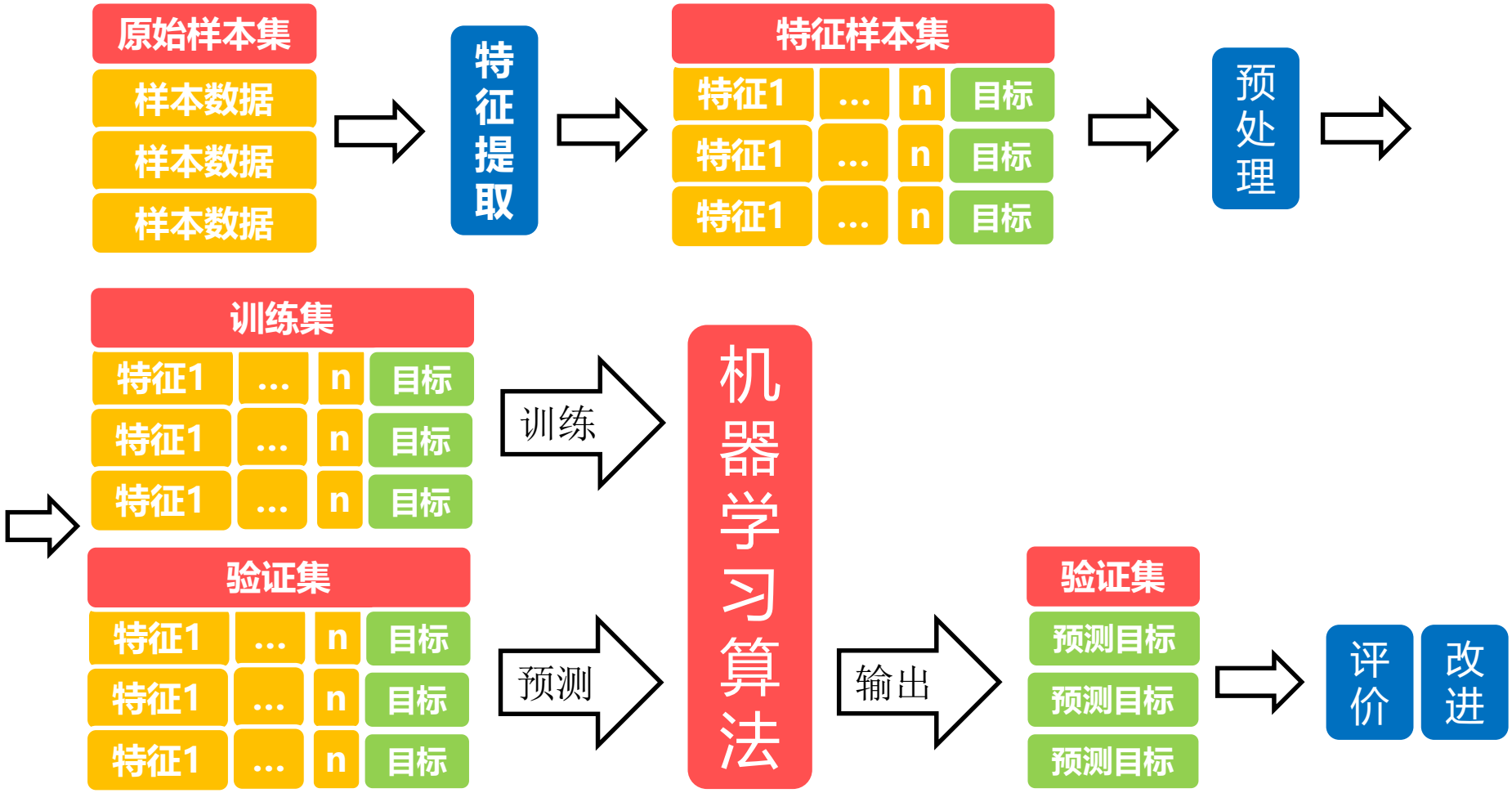
机器学习任务

$$y=f(x)$$

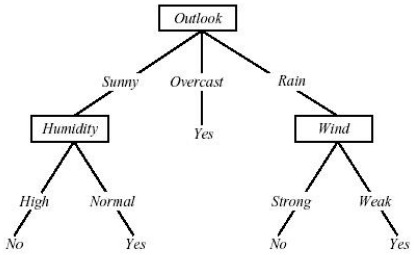


身高	发长	抽烟	性别
1.88	1.4cm	是	男
1.66	15.3cm	否	女
1.78	22.6cm	否	女

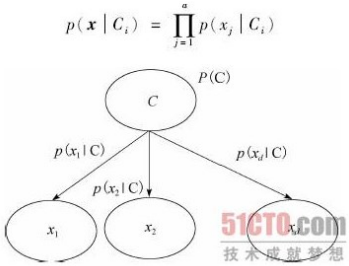
机器学习-实施过程


原始样本集
样本数据
样本数据
样本数据
特征提取
特征样本集
特征1 … n 目标
特征1 … n 目标
特征1 … n 目标
预处理
训练集
特征1 … n 目标
特征1 … n 目标
特征1 … n 目标
验证集
特征1 … n 目标
特征1 … n 目标
特征1 … n 目标
训练
预测
机器学习算法
输出
验证集
预测目标
预测目标
预测目标
评价
改进

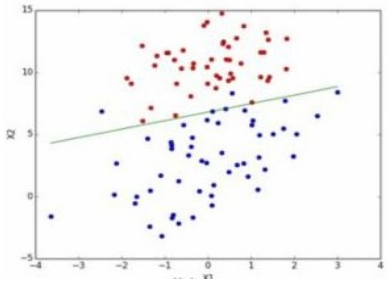
传统机器学习算法



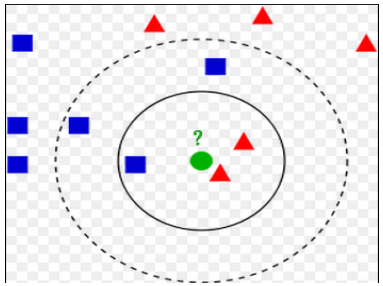
决策树



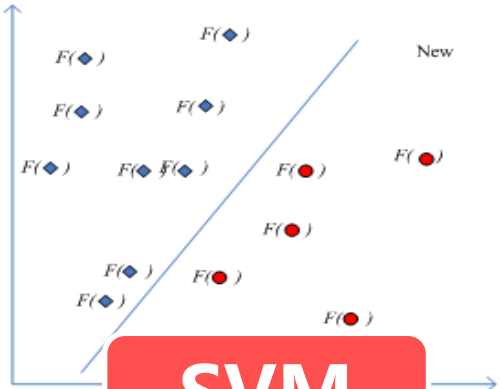
朴素贝叶斯



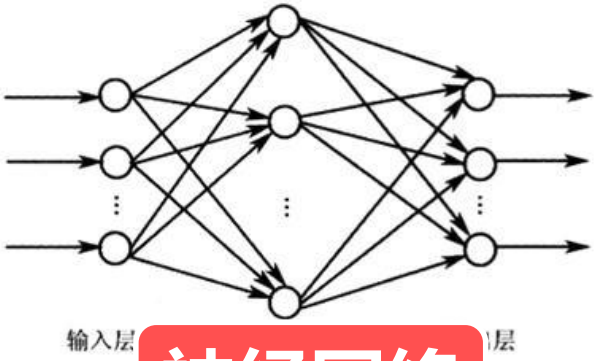
逻辑回归



KNN



SVM



神经网络

随机森林

中国人民大学
RENMIN UNIVERSITY OF CHINA

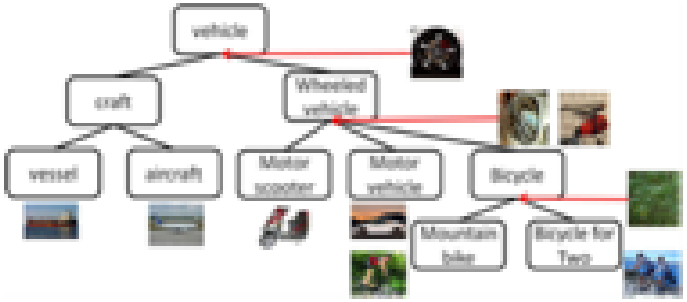
3. 机器学习分类评估

什么是分类?

- 简单地说，分类(Categorization or Classification)就是按照某种标准给对象贴标签(label)

分类体系

- 分类体系一般由人工构造
  - 政治、体育、军事、娱乐……
- 分类体系可能是层次结构



PRINCETON UNIVERSITY
WordNet
A lexical database for English

分类模式

二类问题：属于或不属于

多类问题：多个类别，可拆分成2类问题

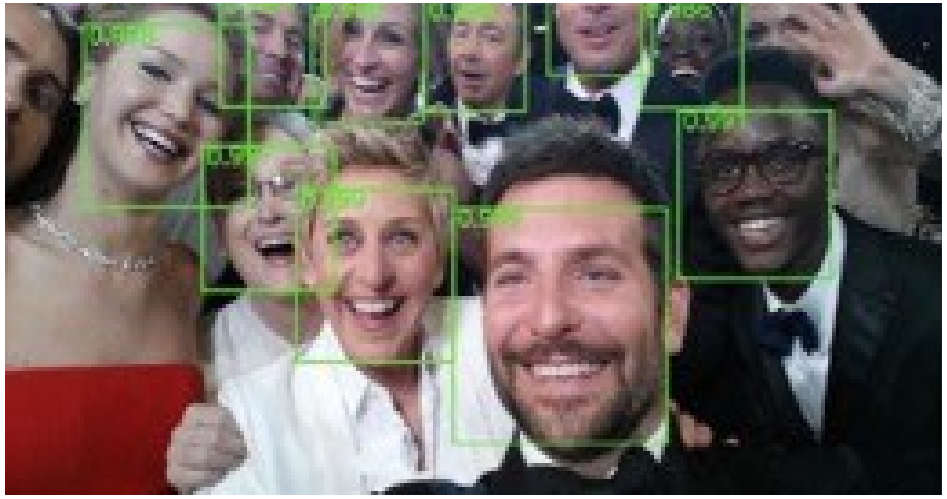
互斥

多标记问题：一个对象可以属于多类

可相互关联

应用举例：二类分类

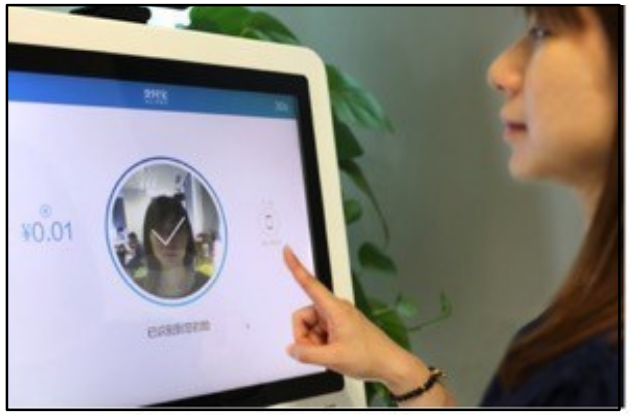
性别识别



人脸检测

应用举例：多类分类（1）

人脸识别监控

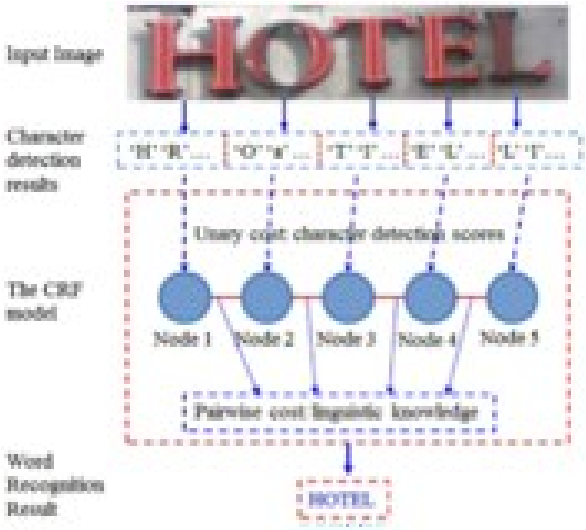


“刷脸支付”

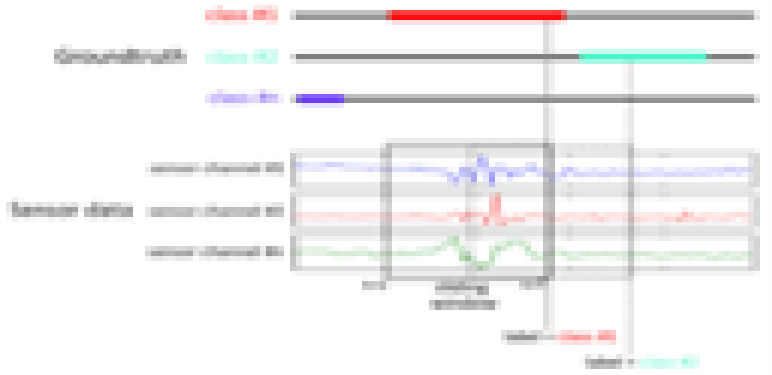
火车站刷脸进站

场馆刷脸进站

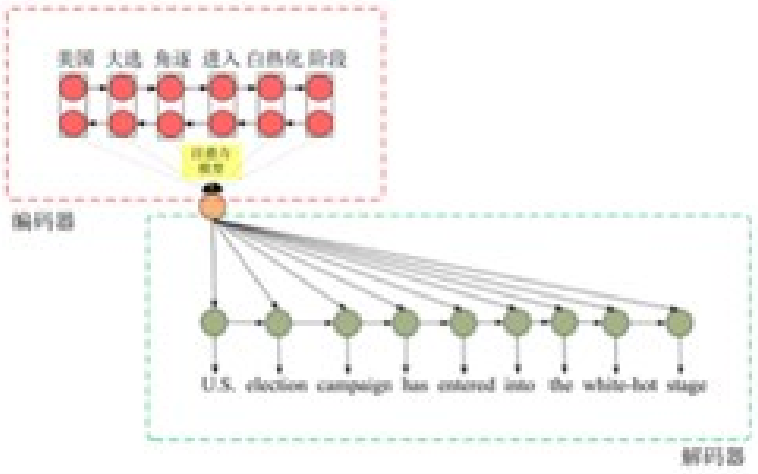
应用举例：多类分类（2）



字符识别

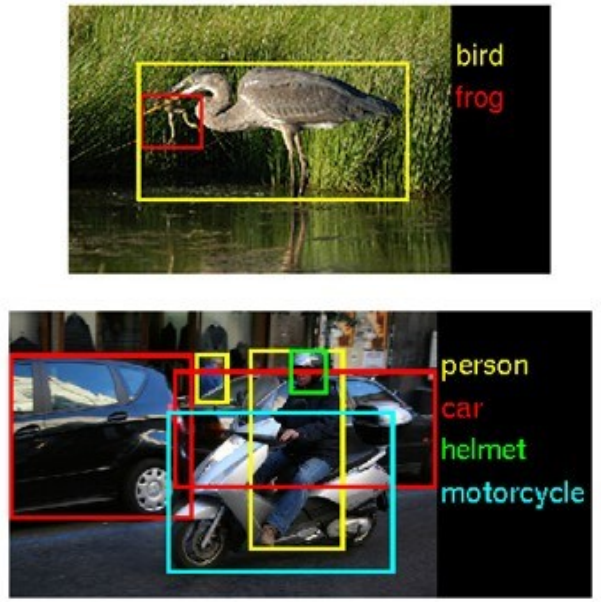


语音识别

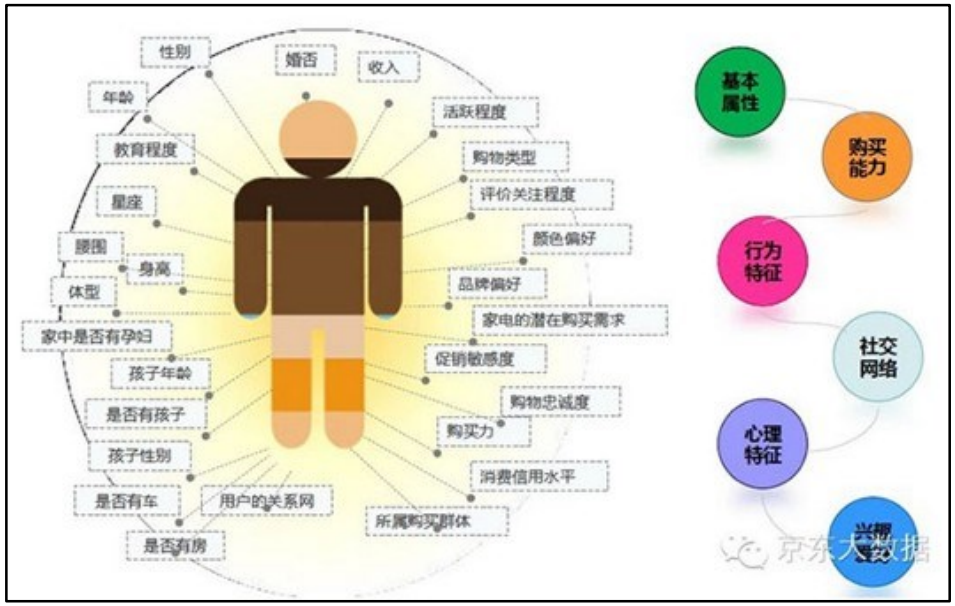


机器翻译

应用举例：多标记问题 (1)

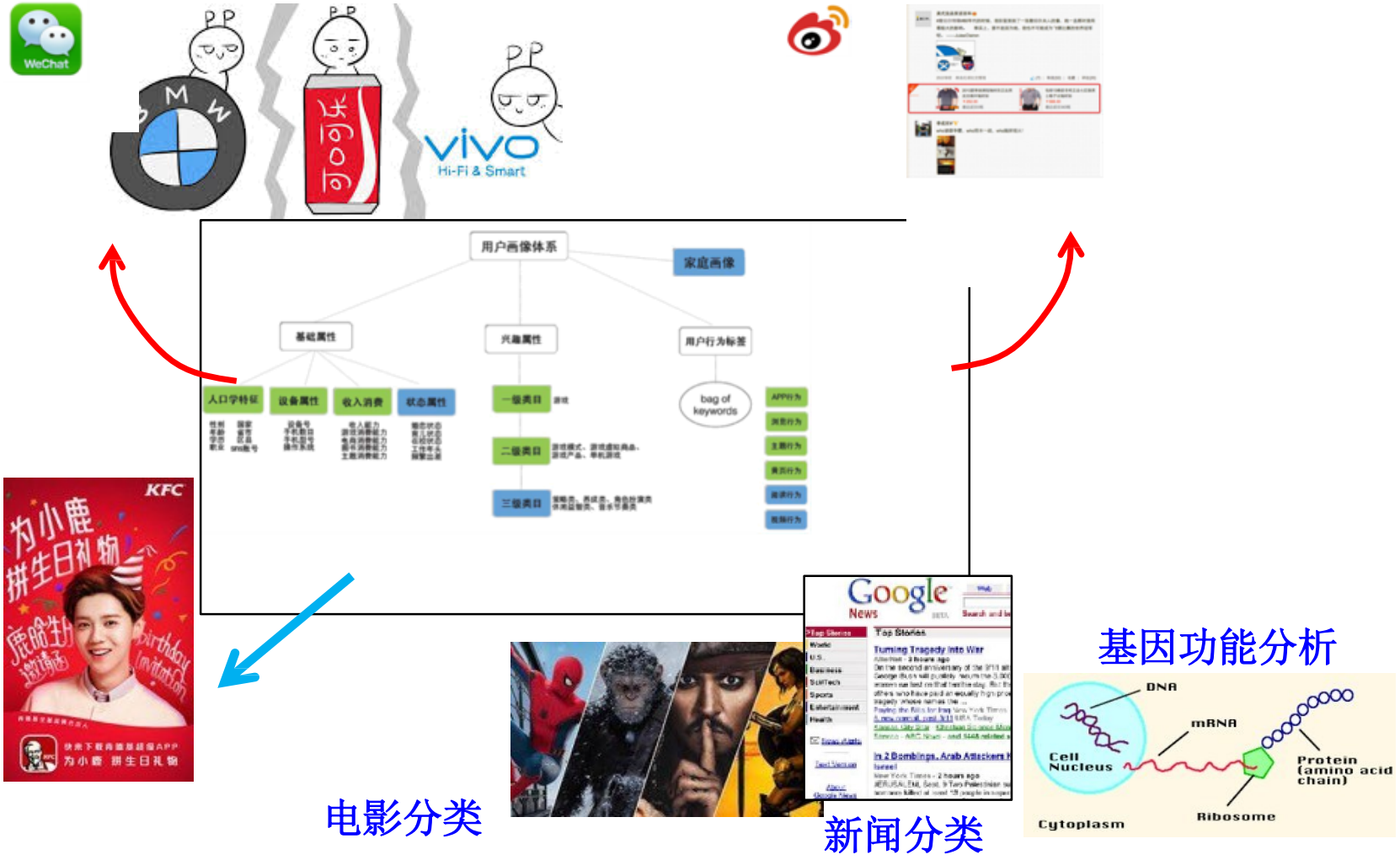


图像标注

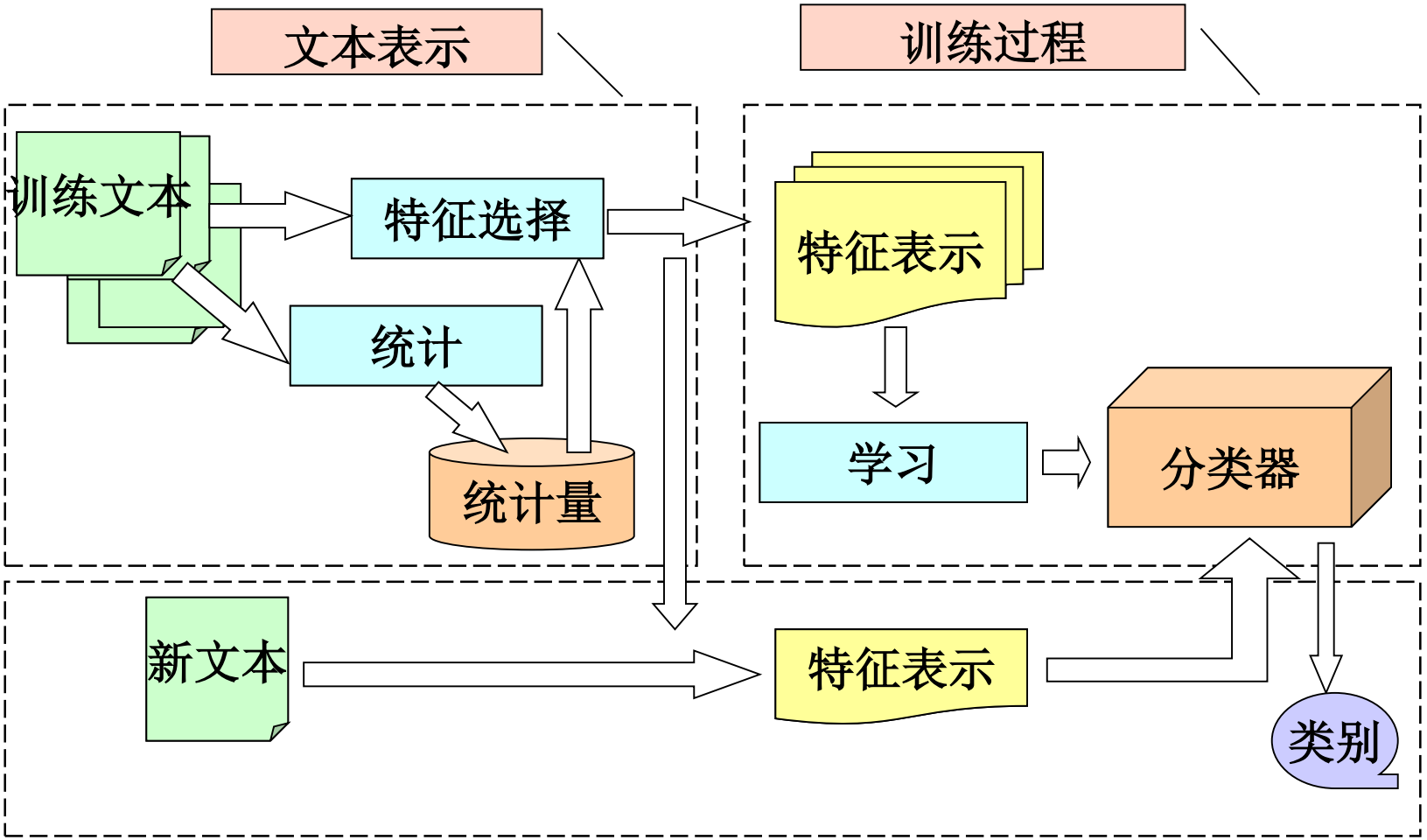


用户画像

应用举例：多标记问题 (2)


WeChat
vivo
Hi-Fi & Smart
用户画像体系
家庭画像
基础属性
兴趣属性
用户行为标签
人口学特征
设备属性
收入消费
状态属性
一级类目
二级类目
三级类目
bag of keywords
KFC
为小鹿
拼生日礼物
Google
News
Top Stories
电影分类
新闻分类
基因功能分析
DNA
mRNA
Cell Nucleus
Protein (amino acid chain)
Cytoplasm
Ribosome

分类过程



分类算法

- 线性回归、逻辑回归
- 决策树、
- SVM
- 贝叶斯分类
- 集成学习
- 神经网络
- 深度学习

分类评价

- 训练集(training set)与测试集(test set)
- k折交叉验证(k-cross validation)
- 留一法(leave-one-out)
- 指标：正确率、召回率、 F1值

正确率P 及召回率R

	in the class	not in the class
predicted to be in the class	true positives (TP)	false positives (FP)
predicted to not be in the class	false negatives (FN)	true negatives (TN)

正确率 Precision = TP / (TP + FP)

召回率 Recall = TP / (TP + FN)

精确率 Accuracy = (TP+TN)/(TP + FP + FN + TN)

一个计算的例子

	In the class	Not in the class
Predicted to be in the class	100	50
Predicted to not be in the class	40	80

正确率Precision = 100 / (100 + 50)

召回率Recall = 100 / (100 + 40)

精确率 Accuracy = (100+80)/(100 + 50 + 40 + 80)

F1值(F Measure)

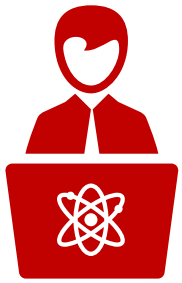
- F_1 允许在正确率和召回率之间达到某种均衡

$$F_1 = \frac{1}{\frac{1}{2}\frac{1}{P} + \frac{1}{2}\frac{1}{R}} = \frac{2PR}{P+R}$$

- 也就是P和R的调和平均值：

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)$$

中國人民大學
RENMIN UNIVERSITY OF CHINA



谢谢大家！