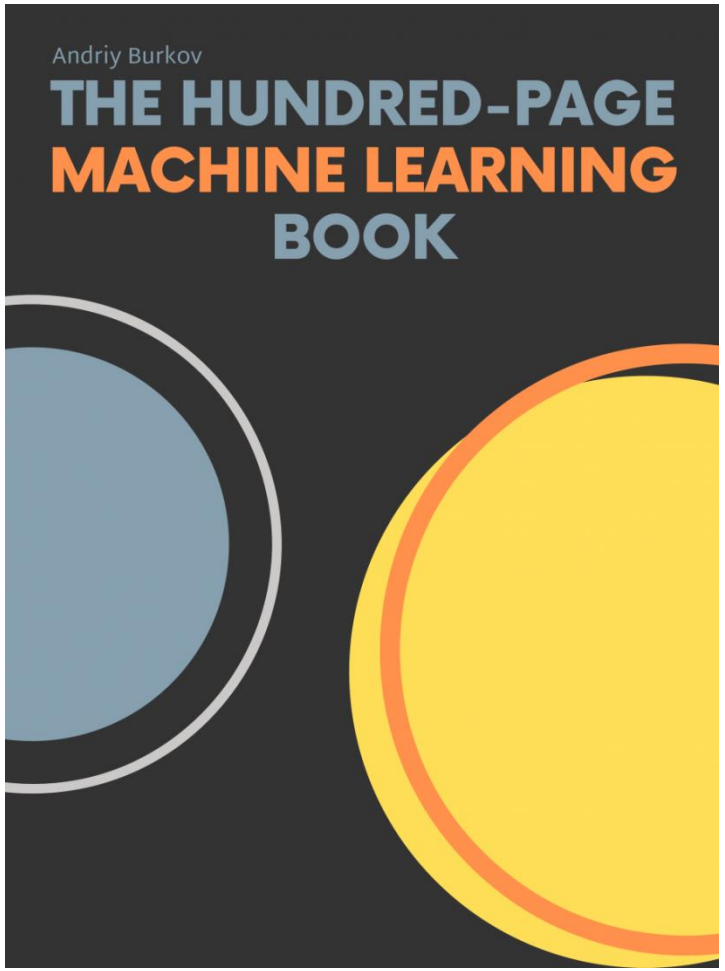
应用数据科学导论

Ht_song@163.com

2020.02

参考书


O'REILLY
TURING 图灵程序设计丛书
数据科学入门
Data Science from Scratch
[美] Joel Grus 著
高蓉 韩波 译
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS


Andriy Burkov
THE HUNDRED-PAGE
MACHINE LEARNING
BOOK

主要内容

- 数据科学简介
- Python语言
- 数据统计可视化
- 数据分析方法——机器学习
- 分析项目实践

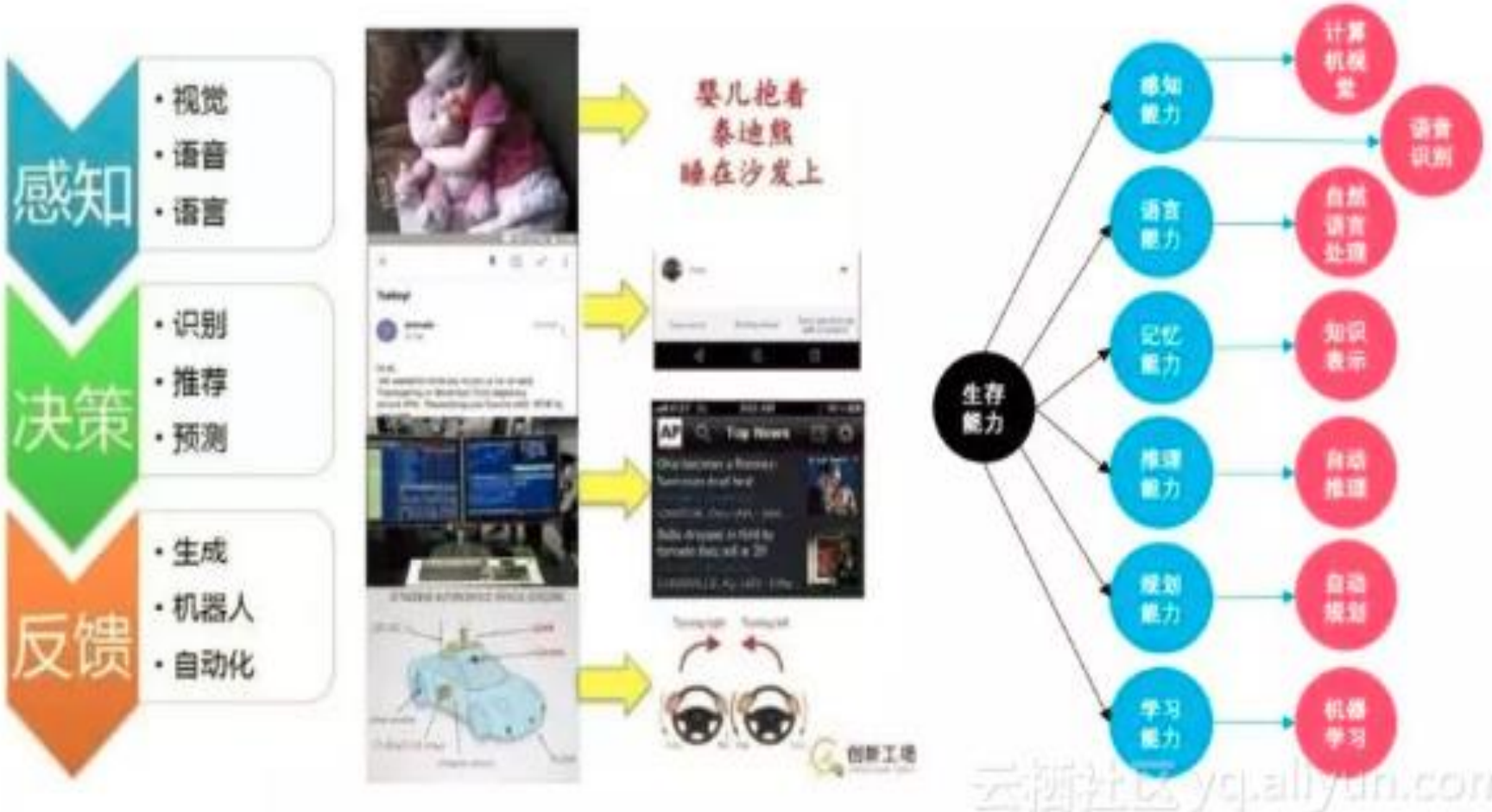
0. 以ML实现数据智能

- 人工智能大体上可以分为“推理期”，“知识期”和“学习期”。
  - 推理期主要注重逻辑推理但是感知器过于简单；
  - 知识期虽然建立了各种各样的专家系统，但是自主学习能力和神经网络资源能力都不足。
  - 学习期机器能够自己学习知识，而直到1980年后，机器学习因其在很多领域的出色表现，才逐渐成为热门学科。

智能的要素

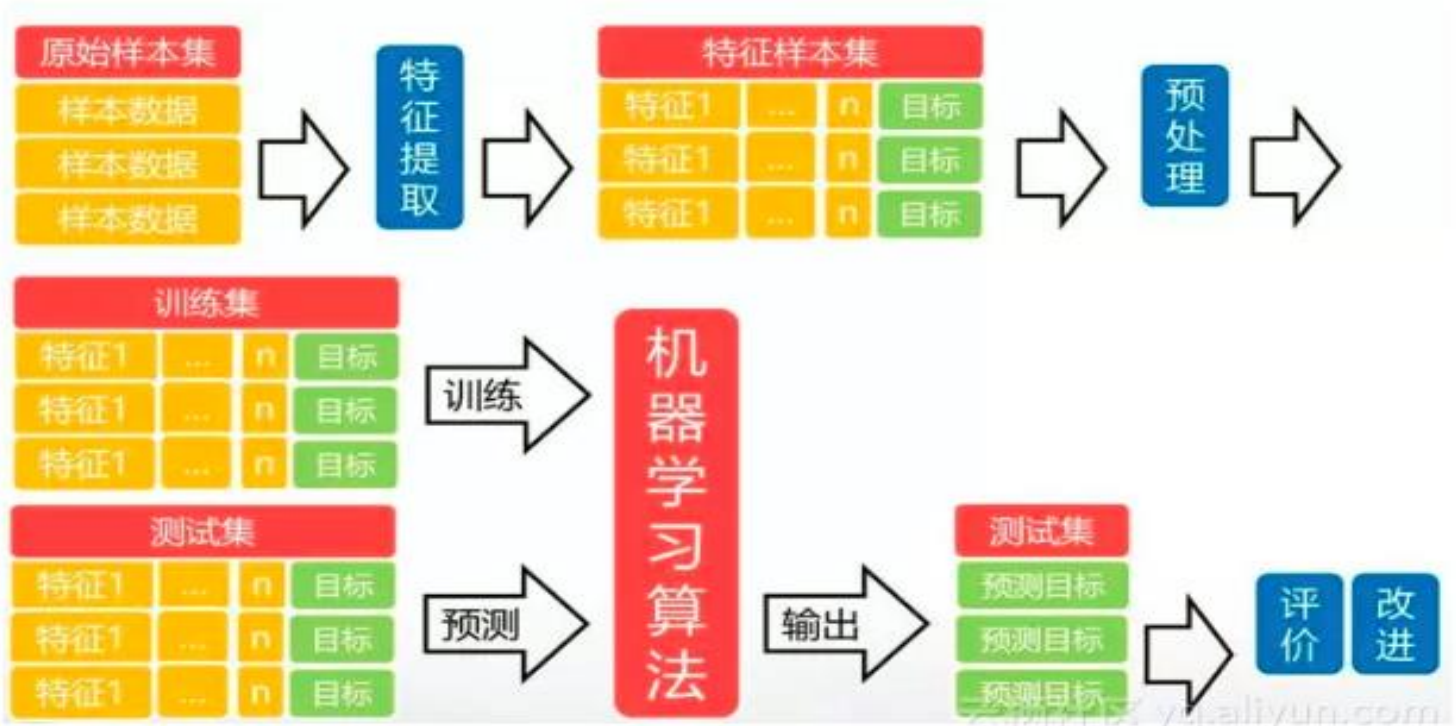
- 人工智能的主要包含：
  - 首先是感知，包括视觉、语音、语言;
  - 然后是决策，例如做出预测和判断;
  - 最后是反馈，如果想做一套完整的系统，就像机器人或是自动驾驶，则需要一个反馈。
- 人工智能众多要素中，学习能力特别重要
  - 机器学习，它是人工智能的核心，是使计算机具有智能的关键。
  - 不能自我学习，人工智能也只是徒有其表。

智能的要素


感知
• 视觉
• 语音
• 语言
决策
• 识别
• 推荐
• 预测
反馈
• 生成
• 机器人
• 自动化
婴儿抱着
泰迪熊
睡在沙发上
生存
能力
感知
能力
语言
能力
记忆
能力
推理
能力
规划
能力
学习
能力
计算
机视
觉
语音
识别
自然
语言
处理
知识
表示
自动
推理
自动
规划
机器
学习

机器学习

- 机器学习最大的特点是利用数据而不是指令来进行各种工作，其学习过程主要包括：数据的特征提取、数据预处理、训练模型、测试模型、模型评估改进等几部分。



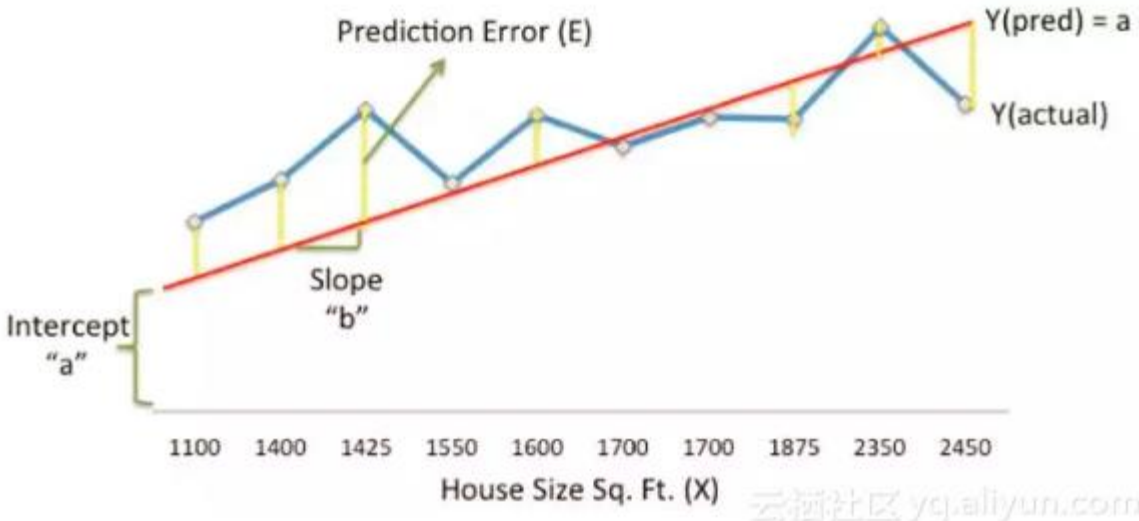
1. ML算法类别

- 算法是通过使用已知的输入和输出以某种方式“训练”以对特定输入进行响应。代表着用系统的方法描述解决问题的策略机制。人工智能的发展离不开机器学习算法的不断进步。 传统机器学习算法主要包括以下五类：
- **回归**
  - 建立一个回归方程来预测目标值，用于连续型分布预测
- **分类**
  - 给定大量带标签的数据，计算出未知标签样本的标签取值
- **聚类**
  - 将不带标签的数据根据距离聚集成不同的簇，每一簇数据有共同的特征
- **关联分析**
  - 计算出数据之间的频繁项集合
- **降维**
  - 原高维空间中的数据点映射到低维度的空间中

ML算法类别

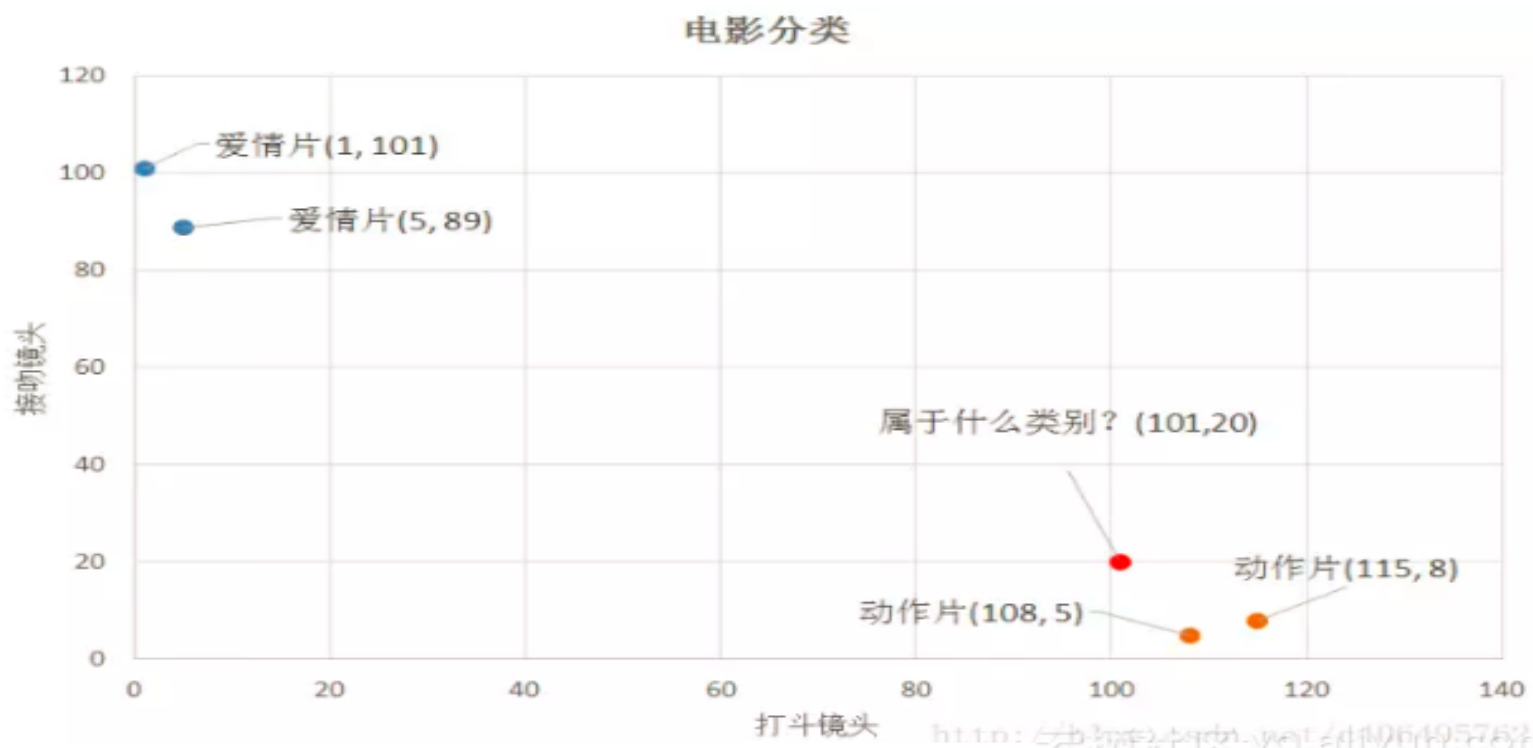
2. 回归

- **线性回归：找到一条直线来预测目标值**
- 一个简单的场景：已知房屋价格与尺寸的历史数据，问面积为2000时，售价为多少？



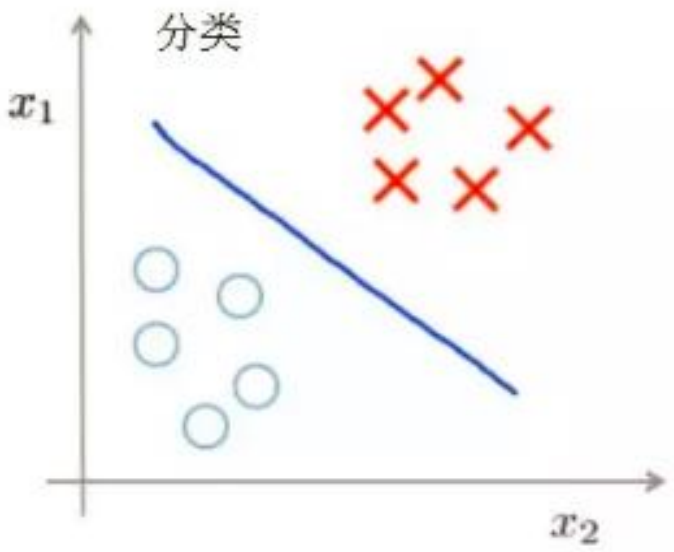
3. 分类

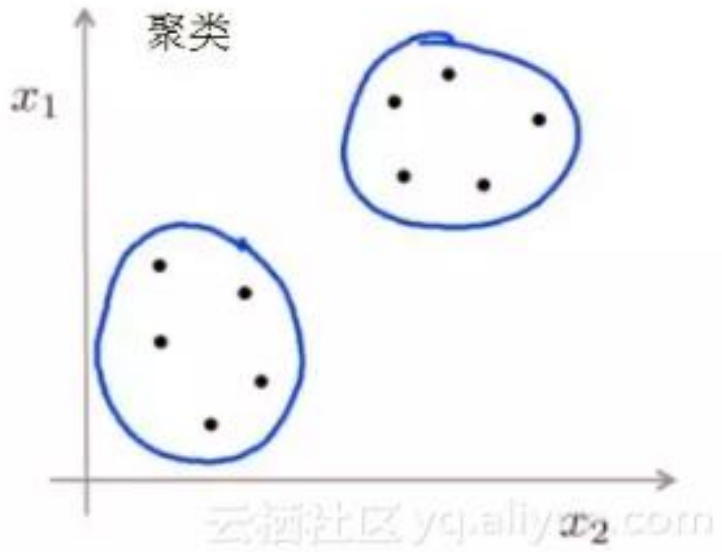
- **K-近邻：用距离度量最相邻的分类标签**
- 一个简单的场景：已知一个电影中的打斗和接吻镜头数，判断它是属于爱情片还是动作片。当接吻镜头数较多时，根据经验我们判断它为爱情片。那么计算机如何进行判别呢？
-



4. 聚类

- **K-means：计算质心，聚类无标签数据**
- 在上面介绍的分类算法中，需要被分类的数据集已经有标记，而对于没有标记的数据集，希望能有一种算法能够自动的将相同元素分为紧密关系的子集或簇，这就是聚类算法。





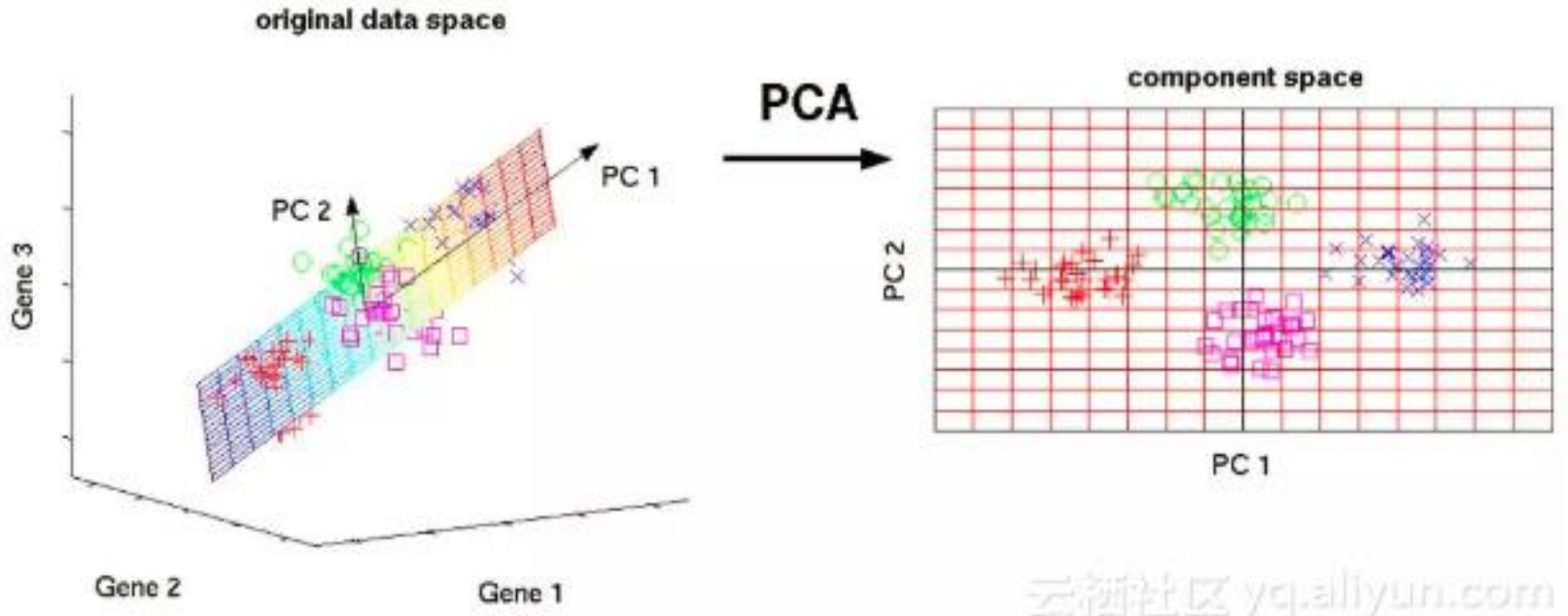
5. 关联分析

- **关联分析：挖掘啤酒与尿布（频繁项集）的关联规则**
- 20世纪90年代美国沃尔玛超市中，超市管理人员分析销售数据时发现“啤酒”与“尿布”两件看上去毫无关系的商品会经常出现在同一个购物篮中。
- 调查发现，这种现象出现在年轻的父亲身上，年轻的父亲去超市买尿布时，往往会顺便为自己购买啤酒。如果在卖场只能买到两件商品之一，他很有可能会放弃购物而去另一家可以同时买到啤酒与尿布的商店。

- 由此，沃尔玛发现了这一独特的现象，开始在卖场尝试将啤酒与尿布摆放在相同区域，让年轻的父亲可以同时找到这两件商品，从而获得了很好的商品销售收入。

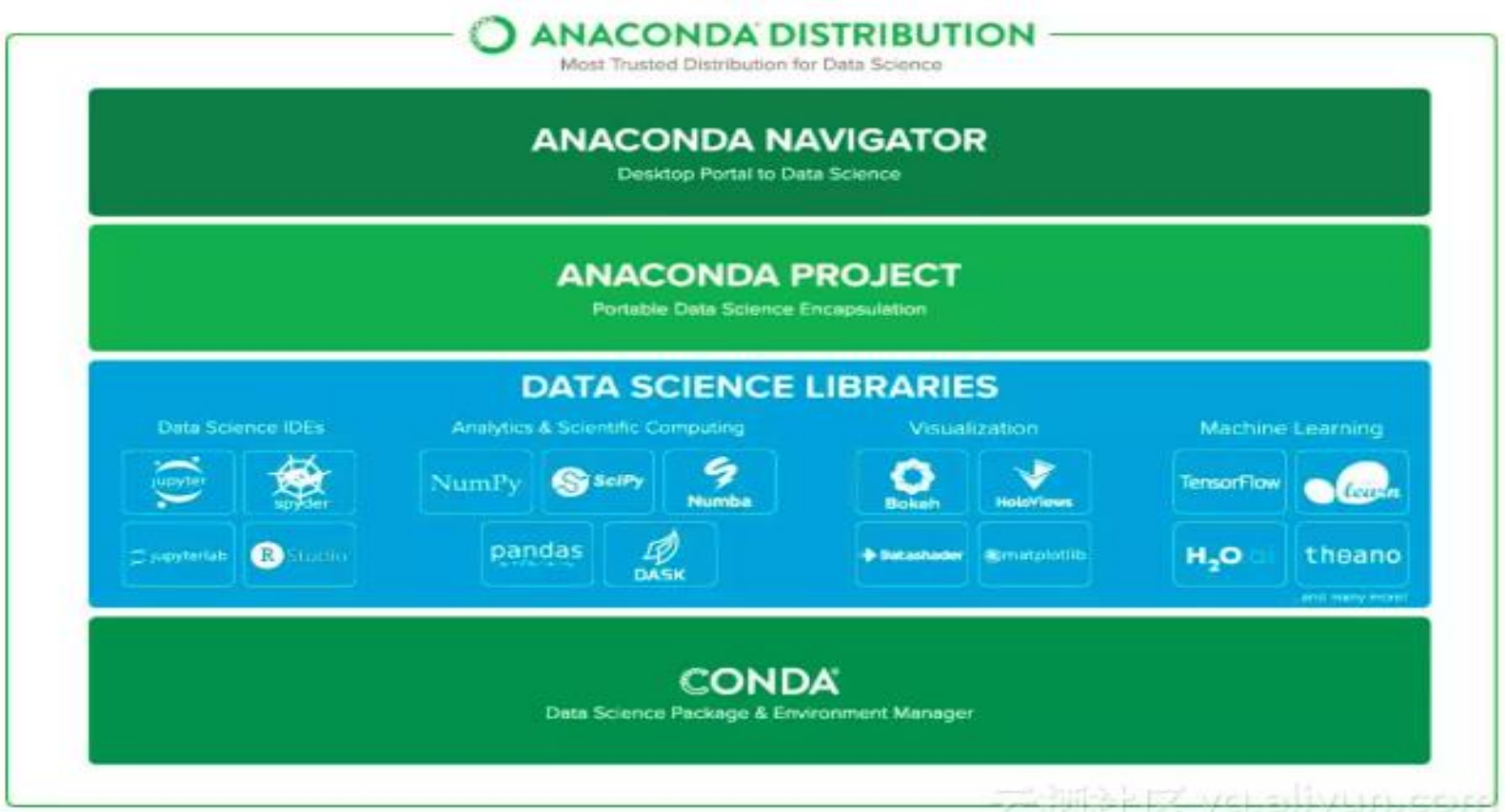
6. 降维

- **PCA降维：减少数据维度，降低数据复杂度**
- 降维是指将原高维空间中的数据点映射到低维度的空间中。因为高维特征的数目巨大，距离计算困难，分类器的性能会随着特征数的增加而下降；减少高维的冗余信息所造成的误差,可以提高识别的精度。



7. 算法之外

- **Anaconda：初学Python、入门ML首选平台**
- 如何动手实践呢？ Anaconda是一个用于科学计算的Python发行版，提供了包管理与环境管理的功能，可以很方便地解决多版本python并存、切换以及各种第三方包安装问题。



数据，是信息时代的真相！