机器学习

初识

涂文婷
tu.wenting@mail.shufe.edu.cn

什么是机器学习

- 人类 PK 机器

微湿路面

感到和风

看到晚霞

什么是机器学习

- 人类 PK 机器

色泽青绿

根蒂蜷缩

敲声浊响

什么是机器学习

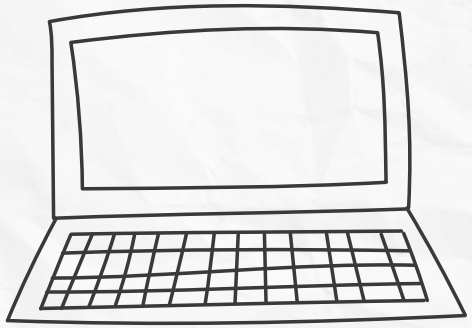
- 人类 PK 机器

人类专家积累了许多经验，而通过对经验的利用，就能对新情况做出有效的决策

什么是机器学习

- 人类 PK 机器

VS



对经验的利用是靠我们人类自身完成的. 计算机能帮忙吗?

什么是机器学习

- 定义

“经验”通常以“数据”形式存在，因此，机器学习所研究的主要内容，是关于在计算机上从数据中产生“模型”的算法，即“学习算法”．可以说机器学习是研究关于“学习算法”的学问。

基本术语

- “数据集”

数据 

西瓜记录1：（色泽=青绿；根蒂=蜷缩；敲声=浊响）
西瓜记录2：（色泽=乌黑；根蒂= 稍蜷；敲声=沉闷）
西瓜记录3：（色泽=浅白；根蒂=硬挺；敲声=清脆）

记录的集合称为一个“数据集”（data set）

基本术语

- “示例”/“样本”

数据

西瓜记录1：(色泽=青绿；根蒂=蜷缩；敲声=浊响)
西瓜记录2：(色泽=乌黑；根蒂= 稍蜷；敲声=沉闷)
西瓜记录3：(色泽=浅白；根蒂=硬挺；敲声=清脆)

每条记录是关于一个事件或对象(这里是一个西瓜)的描述，称为一个“示例” 或“样本”

基本术语

- “属性”/“特征”

数据

西瓜记录1：（色泽=青绿；根蒂=蜷缩；敲声=浊响）
西瓜记录2：（色泽=乌黑；根蒂= 稍蜷；敲声=沉闷）
西瓜记录3：（色泽=浅白；根蒂=硬挺；敲声=清脆）

反映事件或对象在某方面的表现或性质的事项，例如“色泽” “根蒂”“敲声”，称为“属性”或“特征”

基本术语

- “属性值”/“特征值”

数据

西瓜记录1：(色泽=青绿；根蒂=蜷缩；敲声=浊响)
西瓜记录2：(色泽=乌黑；根蒂= 稍蜷；敲声=沉闷)
西瓜记录3：(色泽=浅白；根蒂=硬挺；敲声=清脆)

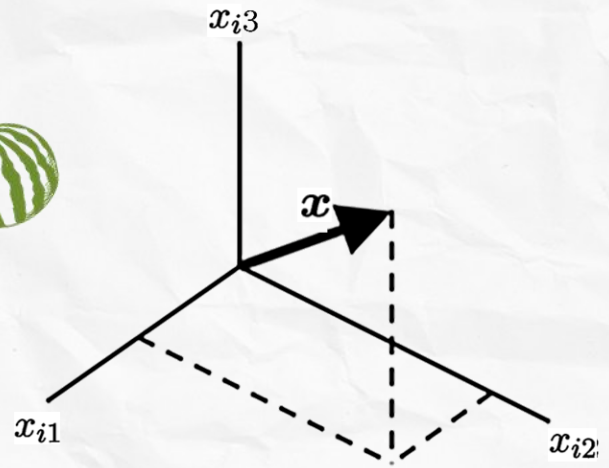
属性上的取值，例如“青绿”“乌黑”，称为“属性值”或“特征值”

基本术语

- “属性空间”/“样本空间”/“输入空间”

数据

西瓜记录1：（色泽=青绿；根蒂=蜷缩；敲声=浊响）
西瓜记录2：（色泽=乌黑；根蒂= 稍蜷；敲声=沉闷）
西瓜记录3：（色泽=浅白；根蒂=硬挺；敲声=清脆）



把“色泽”“根蒂”“敲声”作为三个坐标轴，则它们张成一个用于描述西瓜的三维空间，每个西瓜都可在这个空间中找到自己的坐标位置.

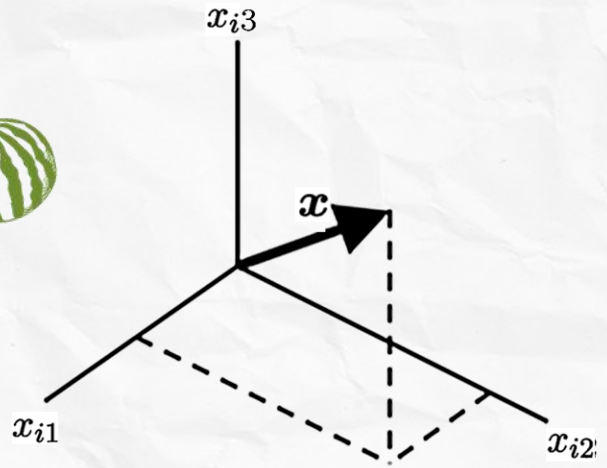
属性张成的空间称为“属性空间” 、“样本空间” 或“输入空间”

基本术语

- “特征向量”

数据

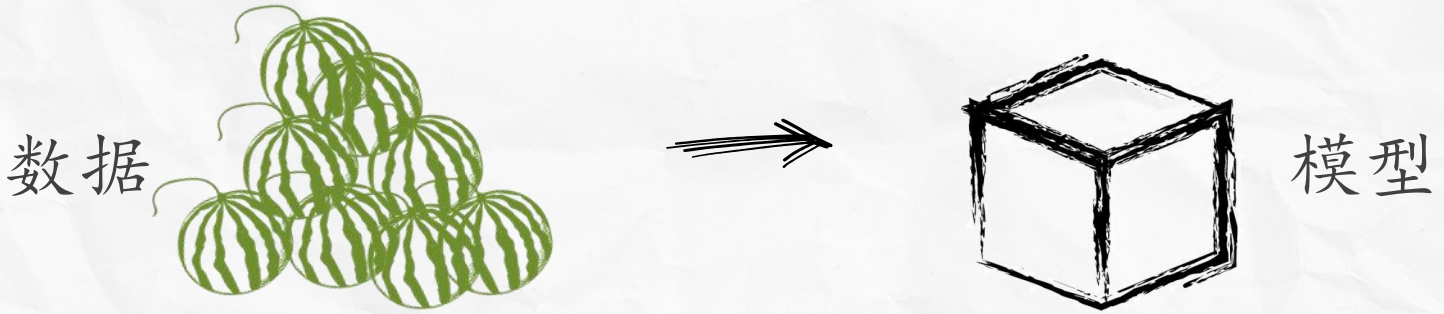
西瓜记录1：（色泽=青绿；根蒂=蜷缩；敲声=浊响）
西瓜记录2：（色泽=乌黑；根蒂= 稍蜷；敲声=沉闷）
西瓜记录3：（色泽=浅白；根蒂=硬挺；敲声=清脆）



由于空间中的每个点对应一个坐标向量，
因此我们也把一个示例称为一个
“特征向量”

基本术语

- “训练”/“学习”

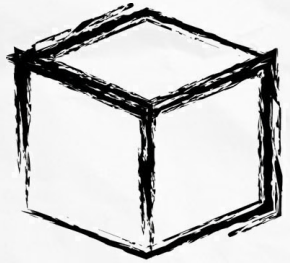


从数据中学得模型的过程称为“学习” 或“训练”，这个过程通过执行某个学习算法来完成。

训练过程中使用的数据称为“训练数据”，其中每个样本称为一个“训练样本”，训练样本组成的集合称为“训练集”。

基本术语

- “标记”/“标签”

数据    模型

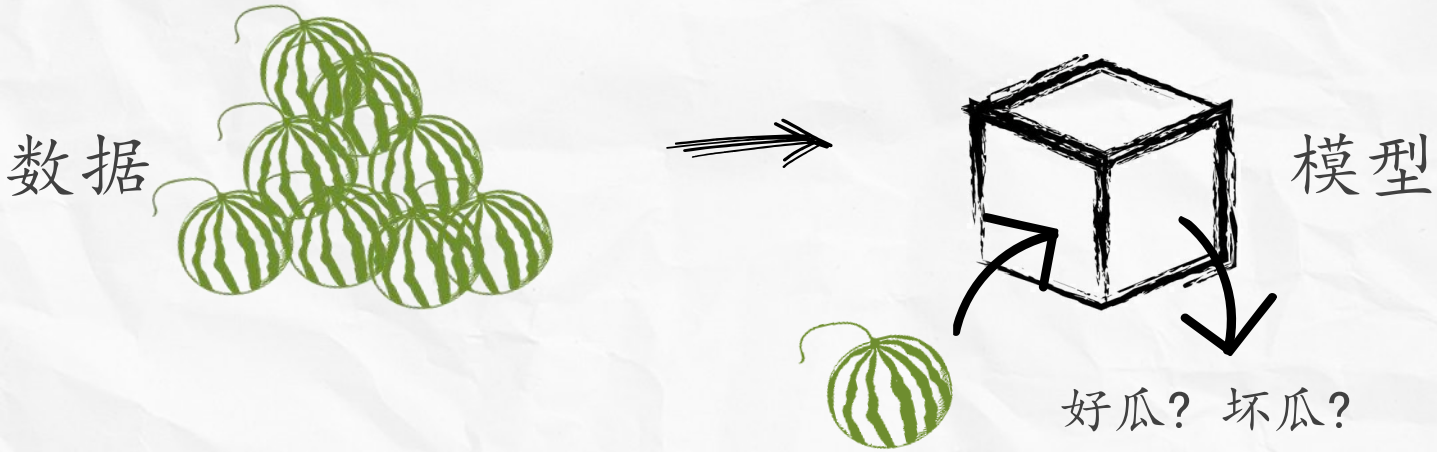
训练出的模型，要能够给出是不是好瓜的预测，通常需要训练集里是包含结果信息的：

西瓜记录1：((色泽=青绿；根蒂=蜷缩；敲声=浊响)，好瓜)
西瓜记录2：(色泽=乌黑；根蒂= 稍蜷；敲声=沉闷)，坏瓜)
西瓜记录3：(色泽=浅白；根蒂=硬挺；敲声=清脆)，好瓜)

这里关于示例结果的信息，例如“好瓜”，称为“标记”或“标签”

基本术语

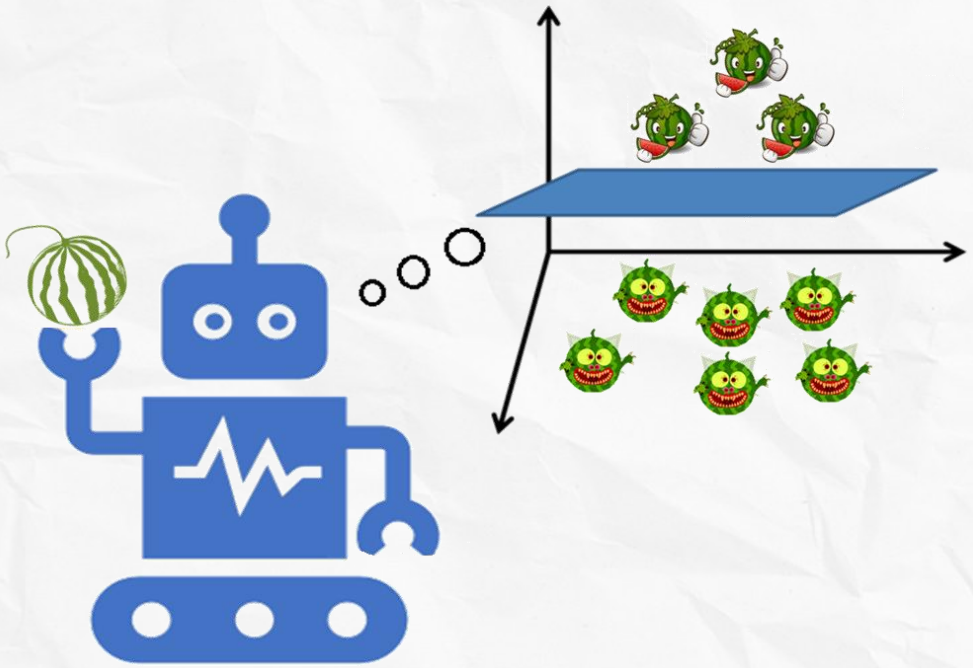
- “测试”



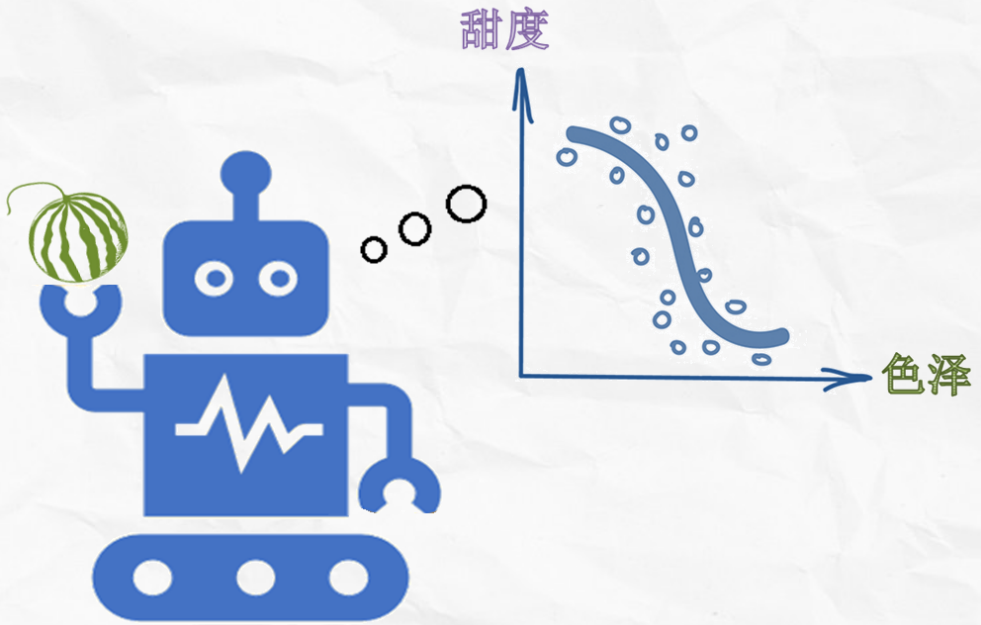
学得模型后，使用其进行预测的过程称为“测试”，被预测的样本称为“测试样本”。

基本术语

- “分类”&“回归”



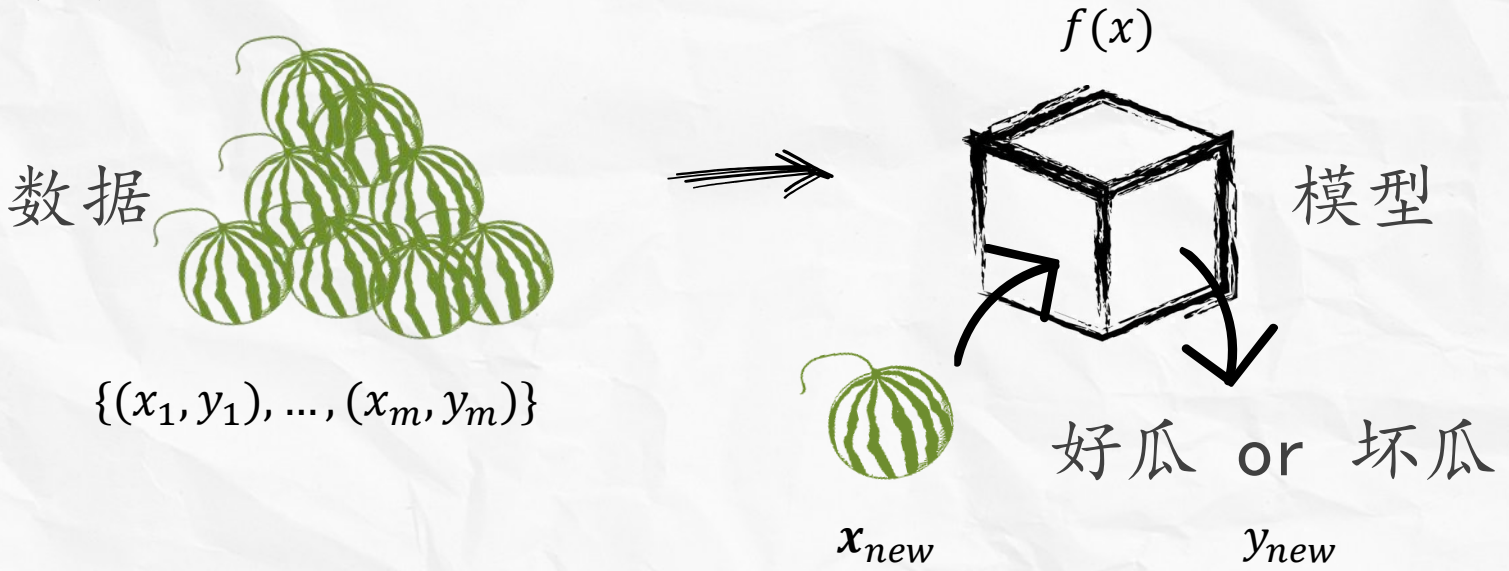
若我们欲预测的是离散值，例如“好瓜”“坏瓜”，此类学习任务称为“分类”



若欲预测的是连续值，例如西瓜甜度 0.95、0.37，此类学习任务称为“回归”

基本术语

- “真相”&“假设”

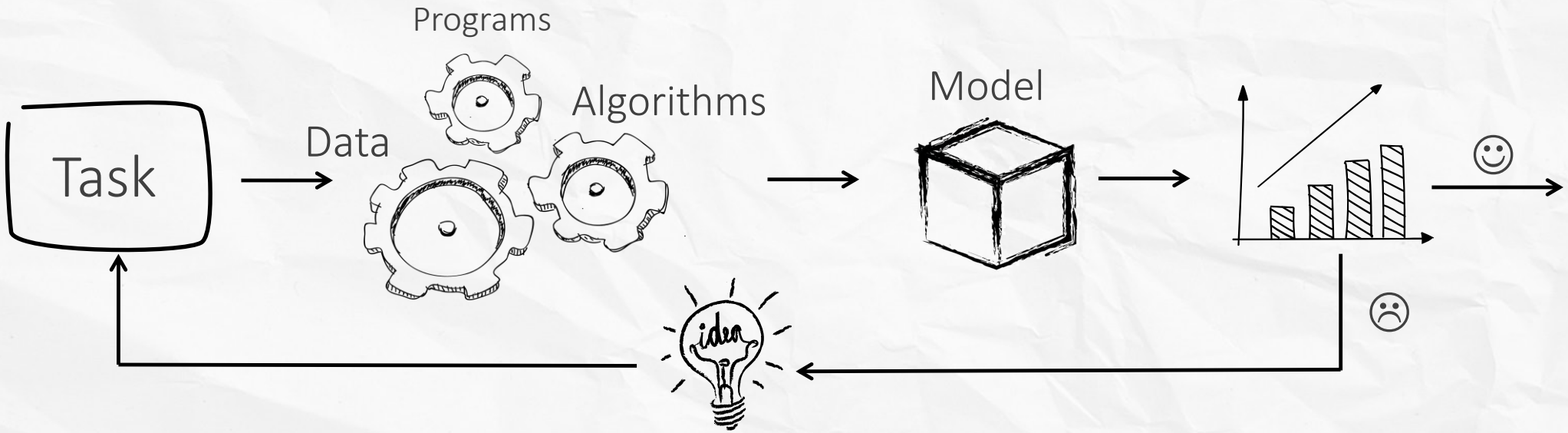


$f(x)$ 学得模型对应了关于数据的某种潜在的规律，因此亦称“假设”(hypothesis)

$g(x)$ 潜在规律自身，则称为“真相”或“真实”(ground-truth)

总结

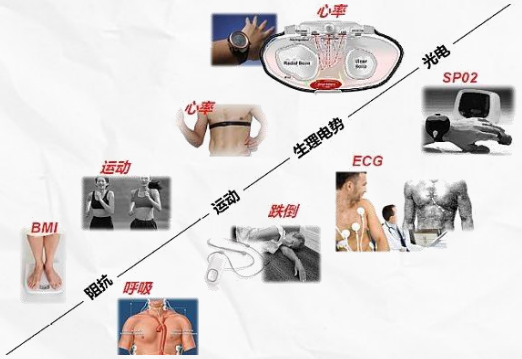
机器学习研究了一套算法：能够令到计算机通过机器学习算法利用和某个任务T相关的经验数据E来学习/训练模型。模型能够在任务T上在评估准则P上获得性能改善。

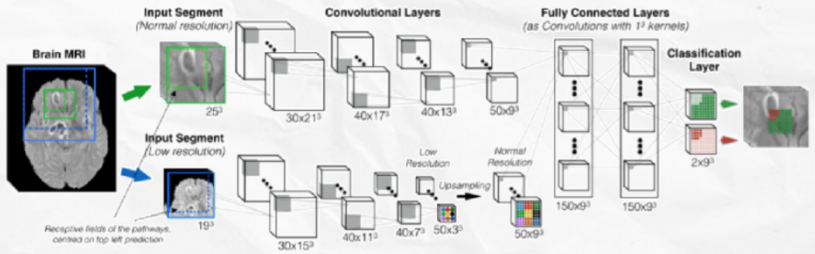


现实应用

- **金融领域**
  - 信贷风控
  - 精准营销
  - 智能投顾
  - 产品定价
  - 电子支付
  - …

- **医疗领域**
  - 医学影像识别
  - 个性化治疗
  - 智能医疗咨询
  - 药物效果评估
  - 基因编辑
  - …

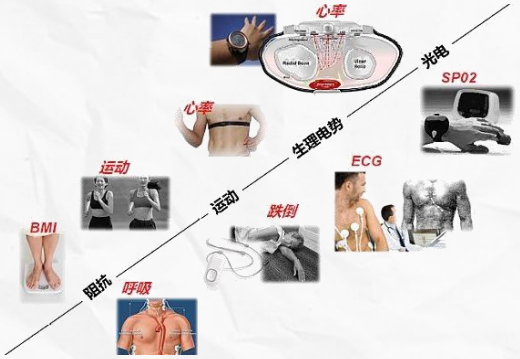


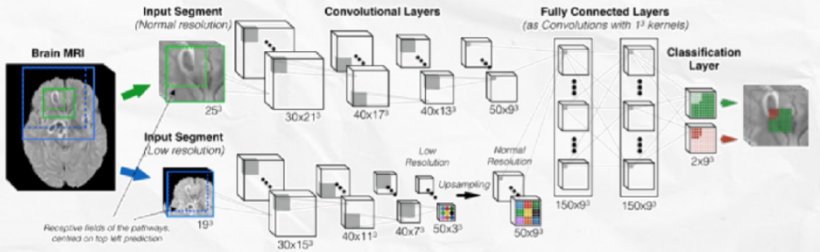


现实应用

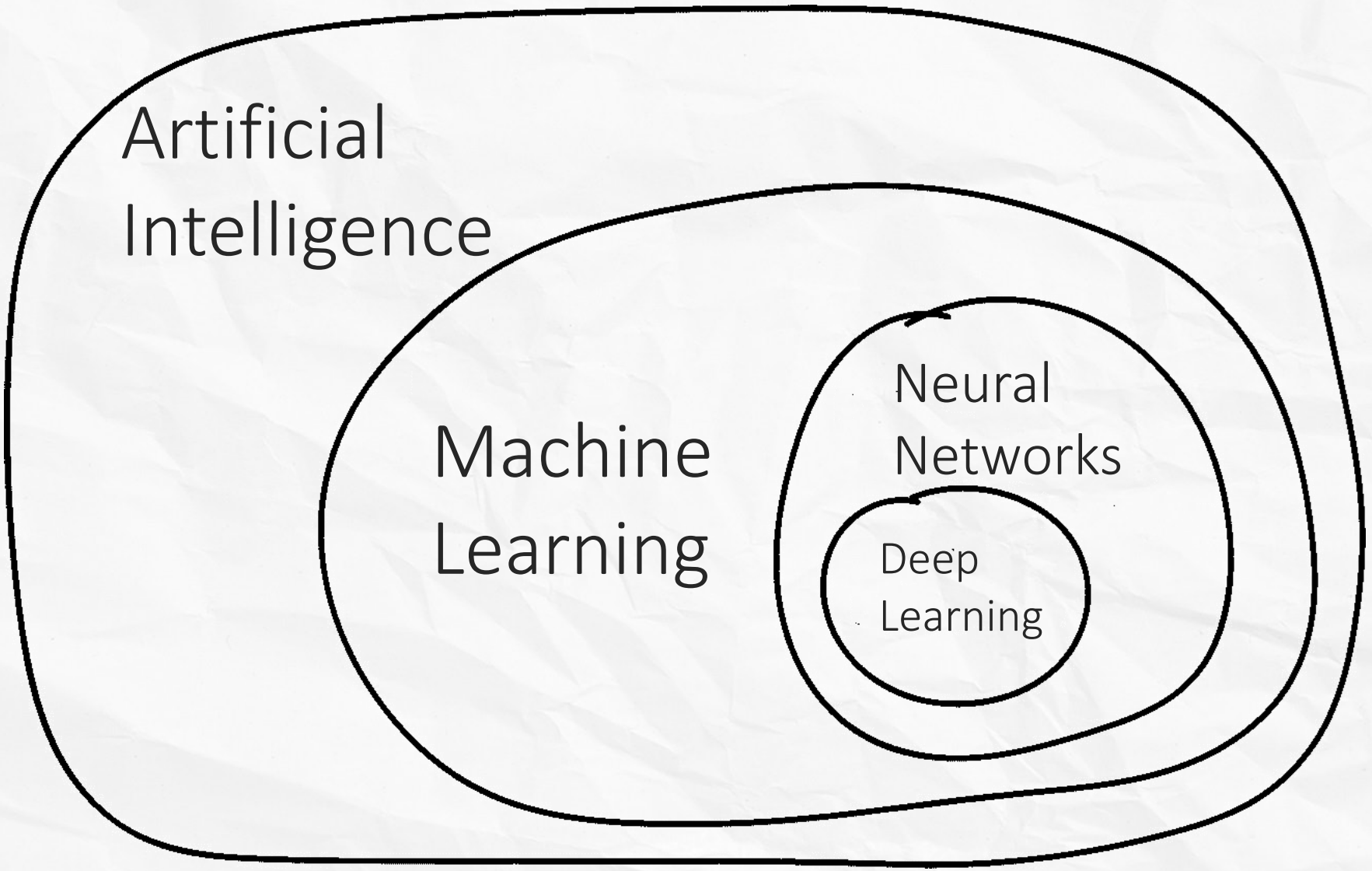
- **交通领域**
  - 车牌号码识别
  - 路径规划
  - 航班晚点率预测
  - 智能交通疏导
  - 自动驾驶
  - …

- **教育领域**
  - 教学水平监测
  - 学习诊断与预警
  - 教育资源配置
  - 学生发展预测
  - 机器人辅导
  - …

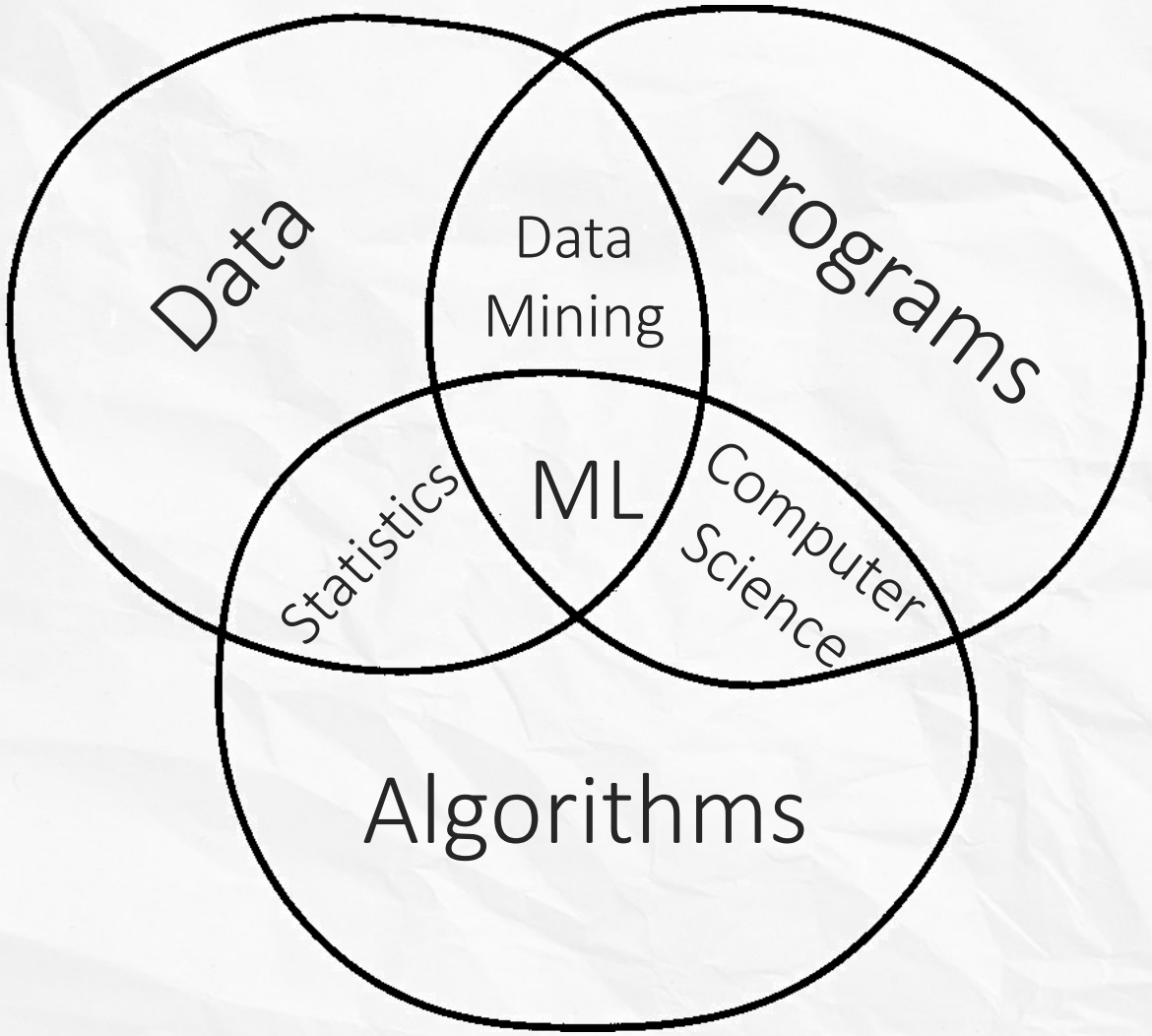




学科对比


Artificial
Intelligence
Machine
Learning
Neural
Networks
Deep
Learning

学科对比


Data
Programs
Data
Mining
Statistics
ML
Computer
Science
Algorithms

附录

- 机器学习相关工具包

附录

- 编程环境配置

  - 安装Python + 依赖包（e.g., pandas, jupyter, scikit-learn…）
    https://scikit-learn.org/stable/install.html

  - 安装Anaconda
    https://www.anaconda.com/products/individual