中国科学技术大学
University of Science and Technology of China

《人工智能数学原理与算法》
第2章：机器学习基础

2.1 机器学习介绍

冯福利
fengfl@ustc.edu.cn

目录

目录

什么是机器学习?

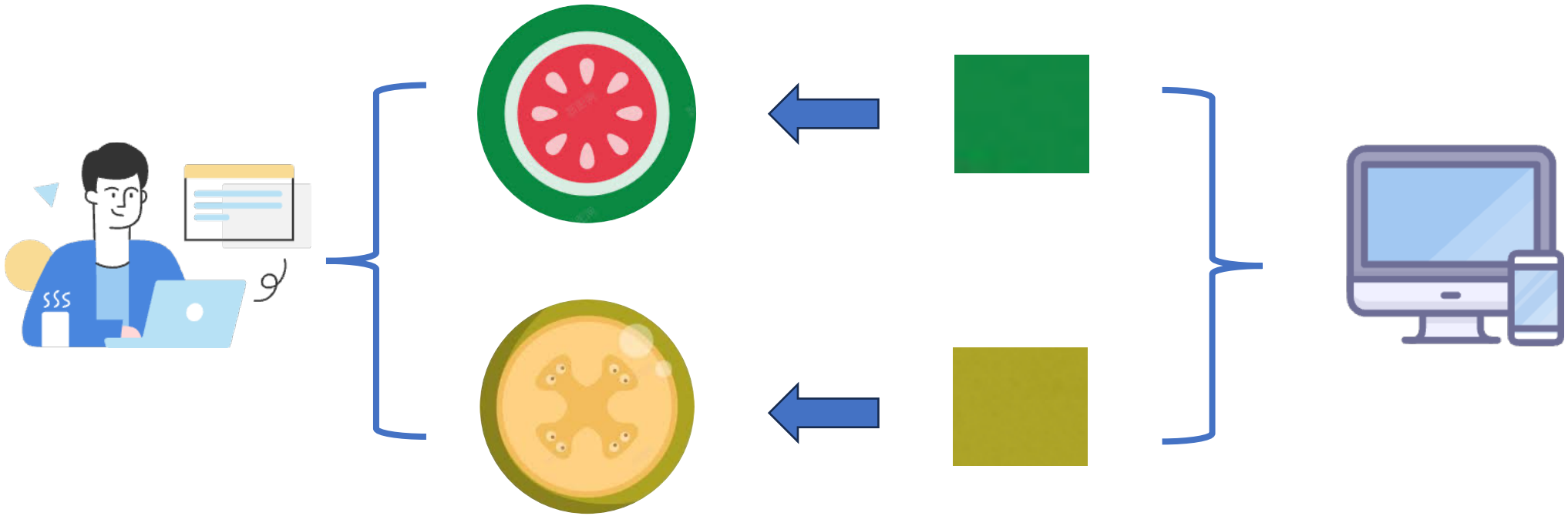
□ Arthur Samuel (1959):

Fields of study that gives computers the ability to learn without being explicitly programmed.

什么是机器学习?

□ 显示编程:

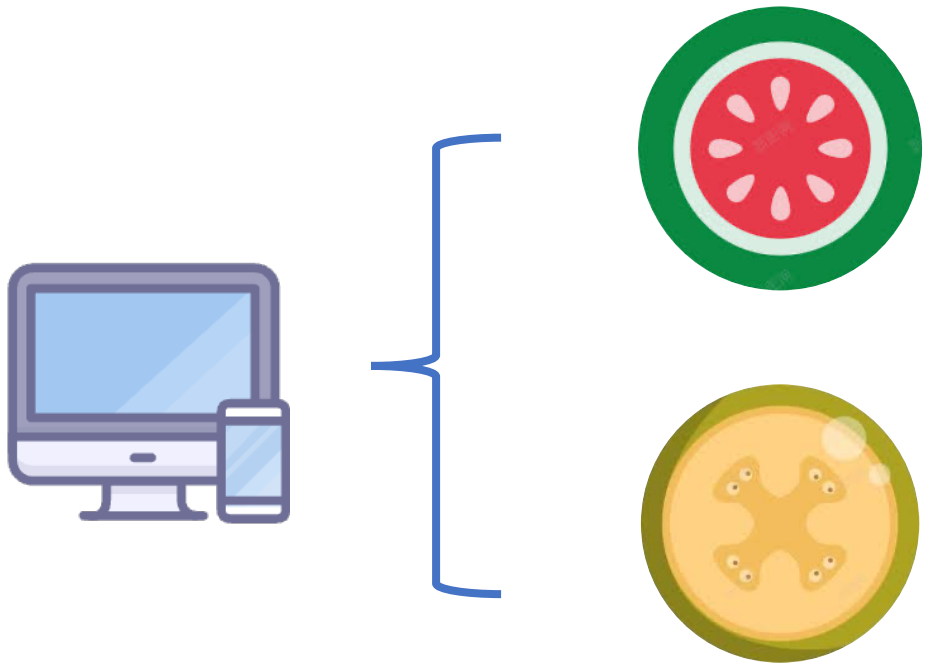
人为告知计算机什么样的输入应该得到什么样的输出



什么是机器学习?

□ 非显示编程:

让计算机自己总结输入与输出之间的规律



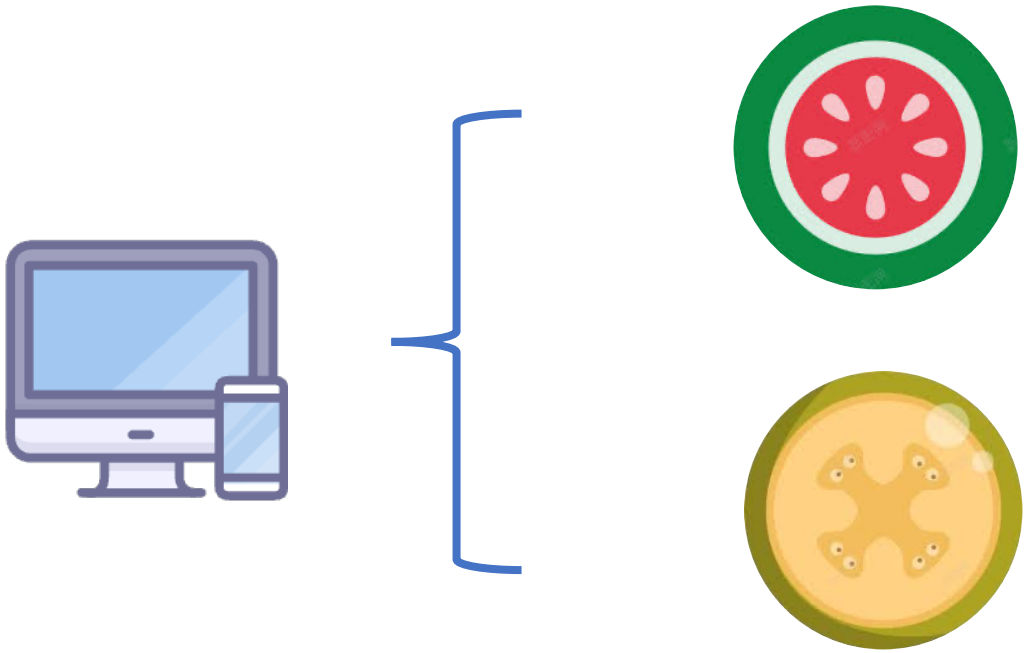
什么是机器学习?

□ Tom Mitshell (1997):

假设用P来评估计算机程序在某类任务T上的性能，若一个程序通过利用经验E在任务T上获得了性能改善，则我们就说关于T和P，该程序对E进行了学习。

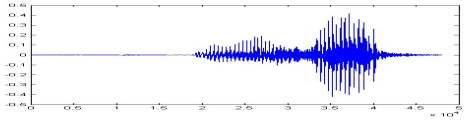
总结：利用经验改善任务性能

什么是机器学习?



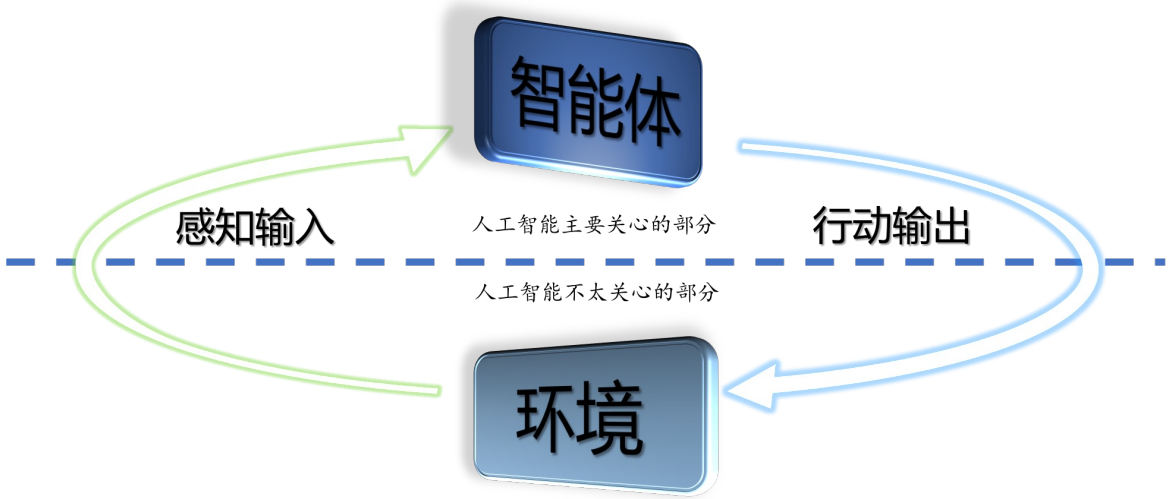
- **任务T**：识别好瓜和坏瓜
- **性能P**：识别正确率
- **经验E**：标注好的好瓜和坏瓜的图片

什么是机器学习?

任务T	经验E			性能P
• 图像识别	$f_\theta($		$)=$ “猫”	识别正确率
• 语音识别	$f_\theta($		$)=$ “你好”	识别正确率
• 围棋系统	$f_\theta($		$)=$ “5-5”(落子位置)	胜率
• 对话系统	$f_\theta($	“你好”	$)=$ “今天天气真不错”	回答质量

人工智能：从智能的外延到智能体（回顾）

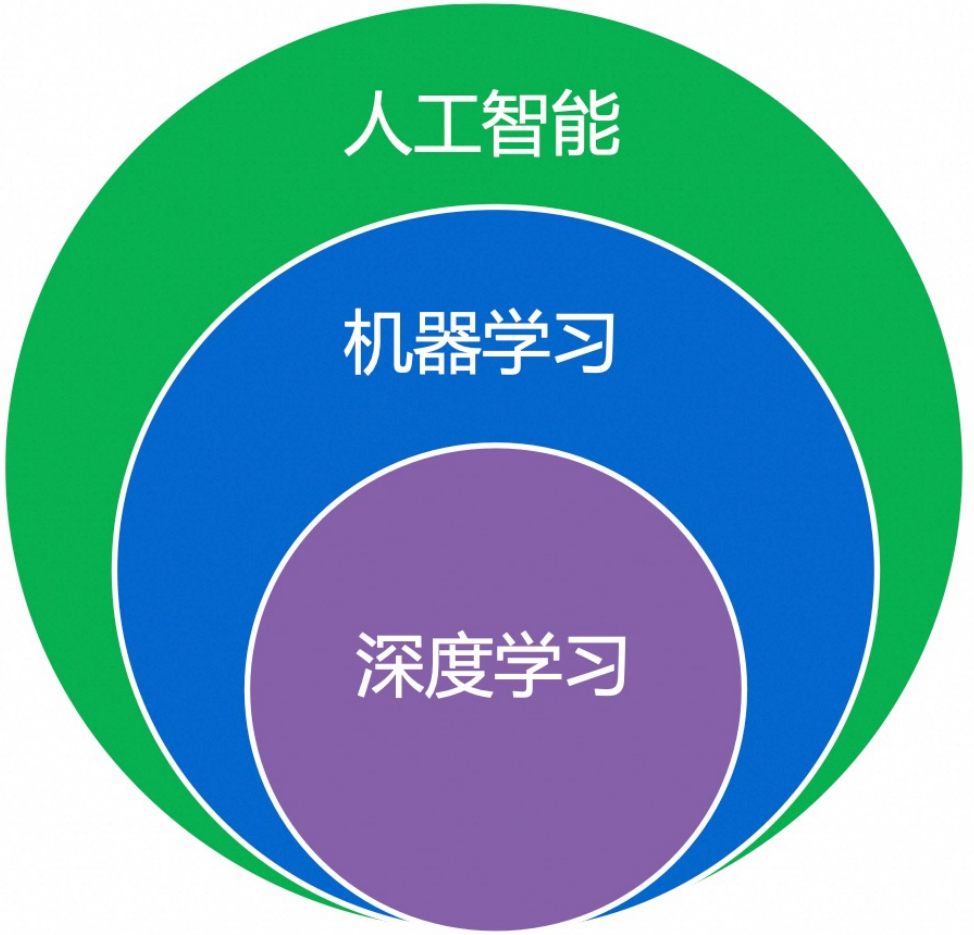
- 每一种智能行为X都对应着一种人工X智能，行为X与环境需要进行交互



	人脸识别	对话问答	围棋象棋	机器翻译	数学证明	……
input	人脸	问题	棋盘状态	语言1句子	题目	……
output	ID	回答	下一步落子	语言2句子	答案	……
feedback	正确与否	正确与否	输赢（多步）	正确与否	正确与否（单/多步）	……

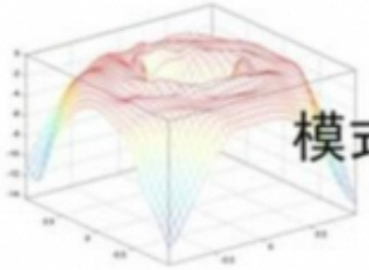
机器学习与人工智能、深度学习的关系

- **人工智能**：机器展现人类智能
- **机器学习**：计算机利用已有的数据（经验），得到某种模型，并利用此模型预测新数据的方法
- **深度学习**：机器学习的一种技术



目录

机器学习能做什么?



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

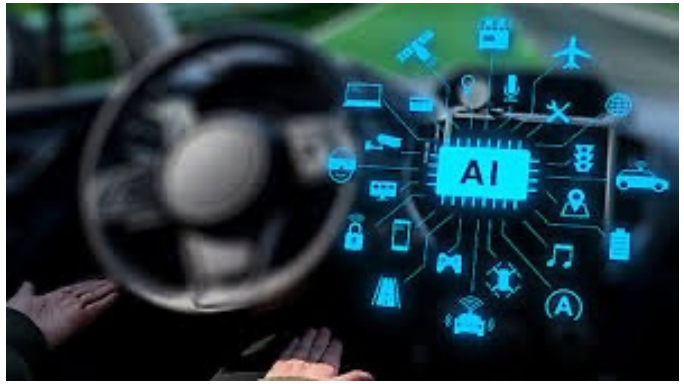
自然语言处理

机器学习能做什么?

□ 机器学习已经与人们的生活密切相关

搜索引擎

自动驾驶



超速检测

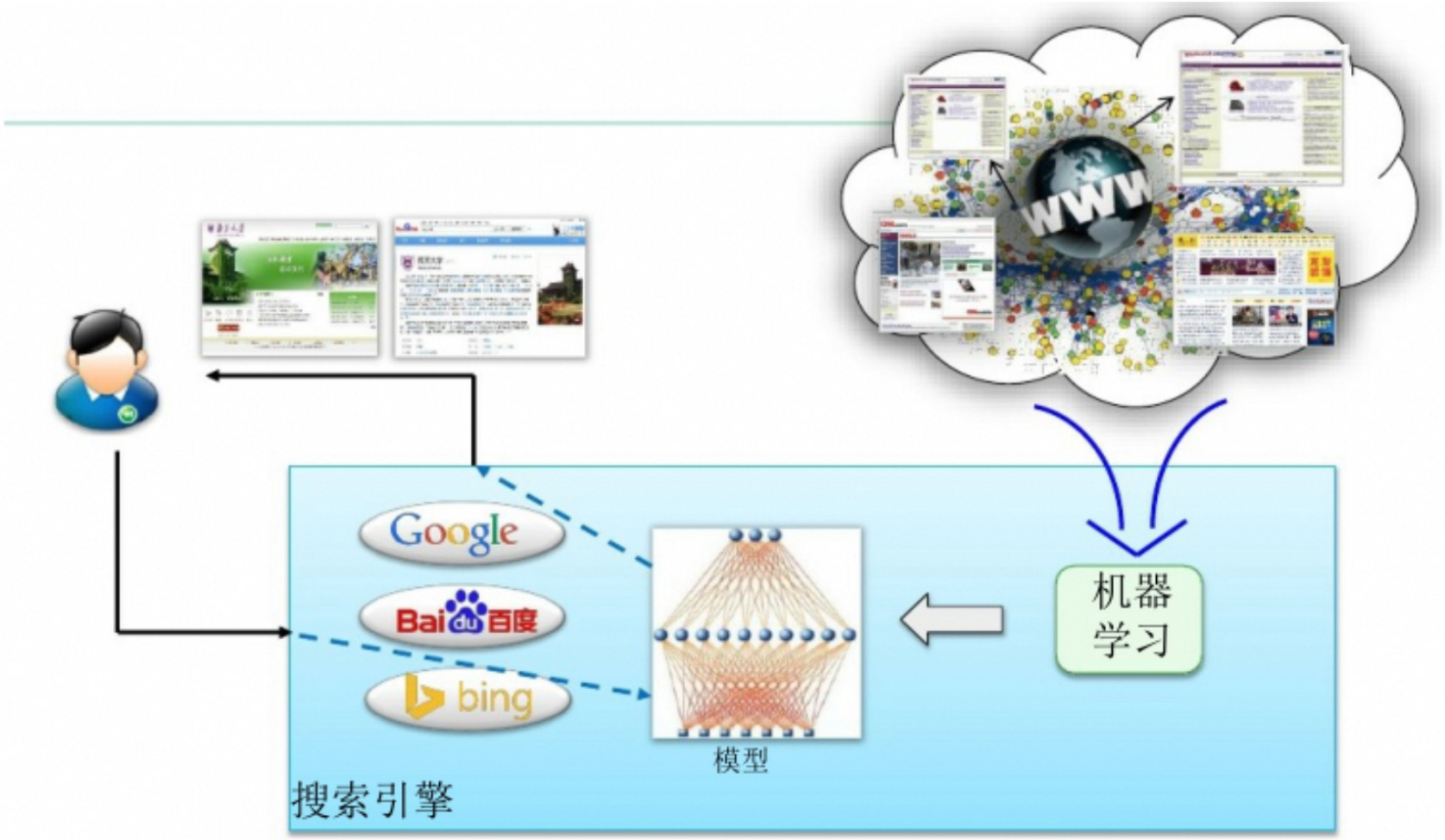
人脸识别

推荐系统

机器翻译

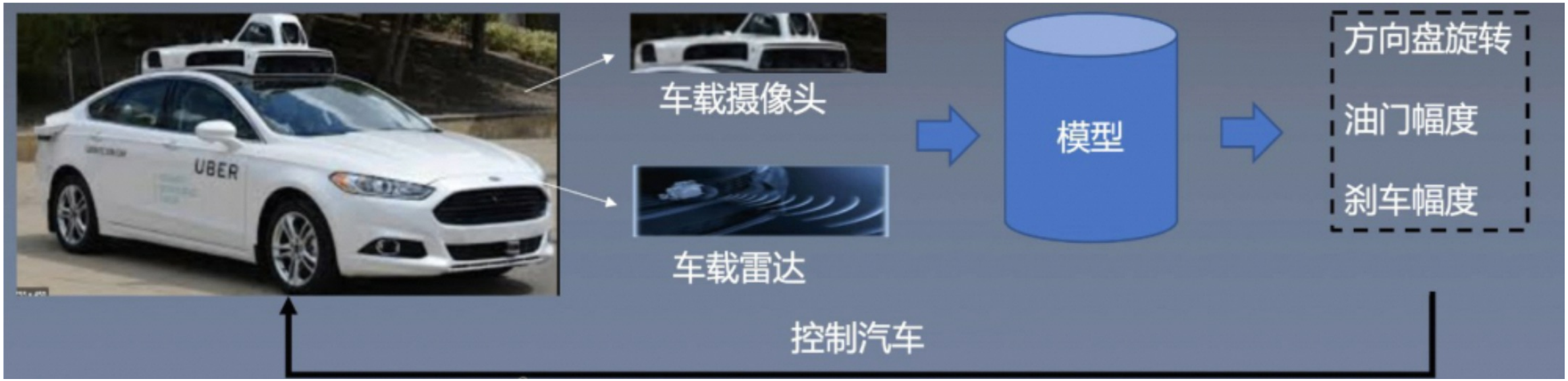
机器学习能做什么？——搜索引擎

□ 机器学习技术支撑各种搜索引擎



机器学习能做什么？——自动驾驶

□ 机器学习技术支撑无人驾驶



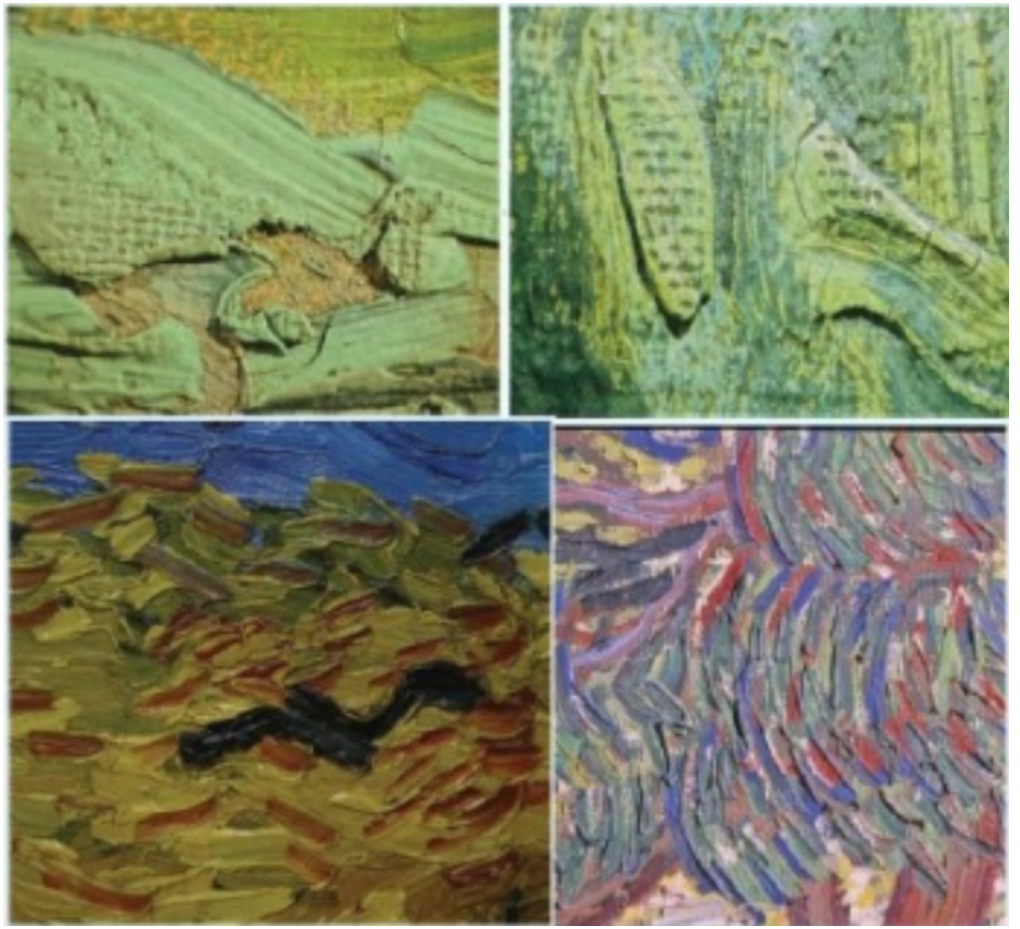
机器学习能做什么？——画作鉴别

□ 机器学习技术应用于艺术领域

笔触分析是画作鉴定的重要工具，旨在从视觉上判断画作中是否具有艺术家的特有“笔迹”。

该工作对专业知识要求极高，需要

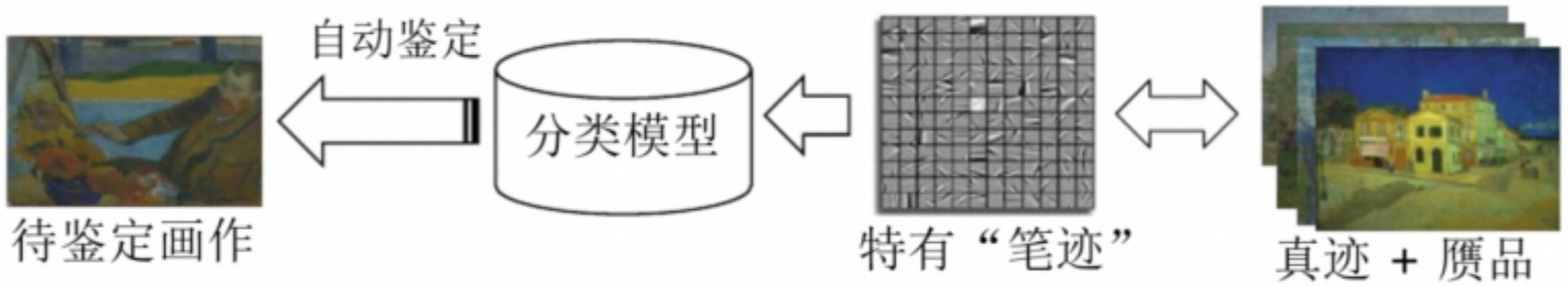
- 具有较高的绘画艺术修养
- 掌握画家的特定绘画习惯



很难同时掌握不同时期、不同流派多位画家的绘画风格！

机器学习能做什么？——画作鉴别

□ 为了降低分析成本，机器学习技术被引入



Kraller Miller美术馆与Cornell等大学的学者对82幅梵高真迹和6幅赝品进行分析，自动鉴别精度达 95%。 [C. Johnson et al., IEEE-SP, 2008]

Dartmouth学院、巴黎高师的学者对8幅勃鲁盖尔真迹和5幅赝品进行分析，自动鉴别精度达100%。 [J. Hughes et al., PNAS 2009][J. Mairal et al., PAMI' 12]

机器学习技术对用户要求低、准确高效、适用范围广

机器学习能做什么？——古文献修复

□ 机器学习技术应用于文化领域

古文献是进行历史研究的重要素材，一个重要问题是原书籍已经变成分散且混杂的多个书页，如何拼接相邻的书页？

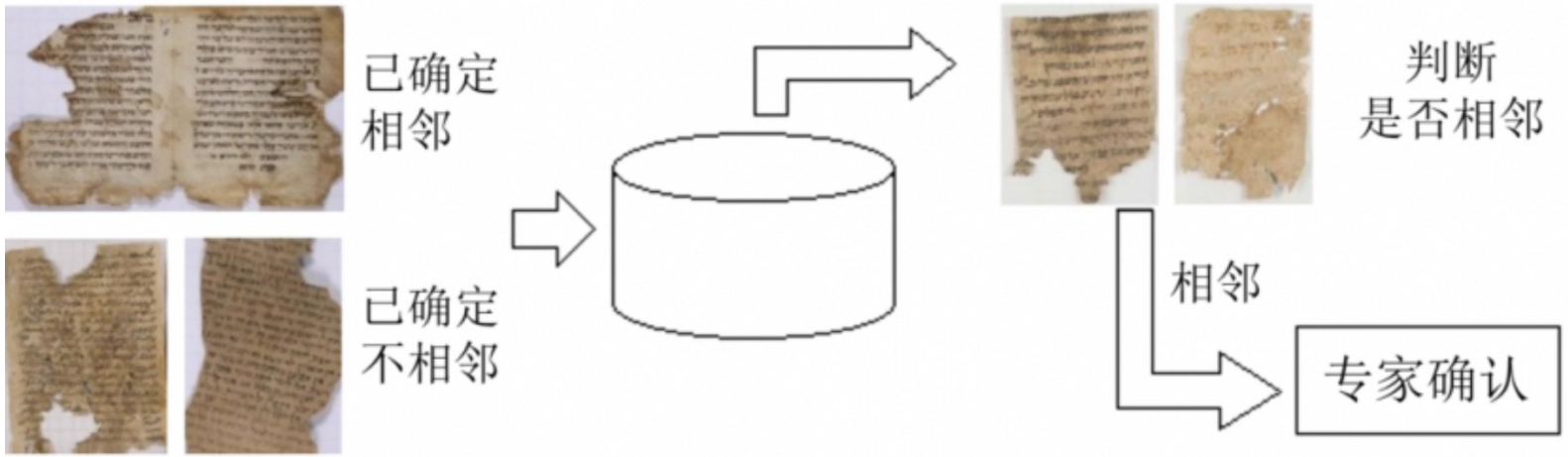
人工完成书页拼接困难在于

- 书页数量大，且分布在多处
- 部分损毁较严重，字迹模糊
- 需要大量掌握古文字的专业人才

高水平专家的大量精力被用于古文献修复

机器学习能做什么？——古文献修复

□ 古文献的数字化浪潮给自动文学修复提供了机会



在Cairo Genizah测试数据上，系统的自动判断精度超过93%，新完成约 1,000篇Cairo Genizah文章的拼接。 [R. Laidani et al., 2011]

对比：过去整个世纪，数百人类专家只完成了几千篇文章拼接

目录

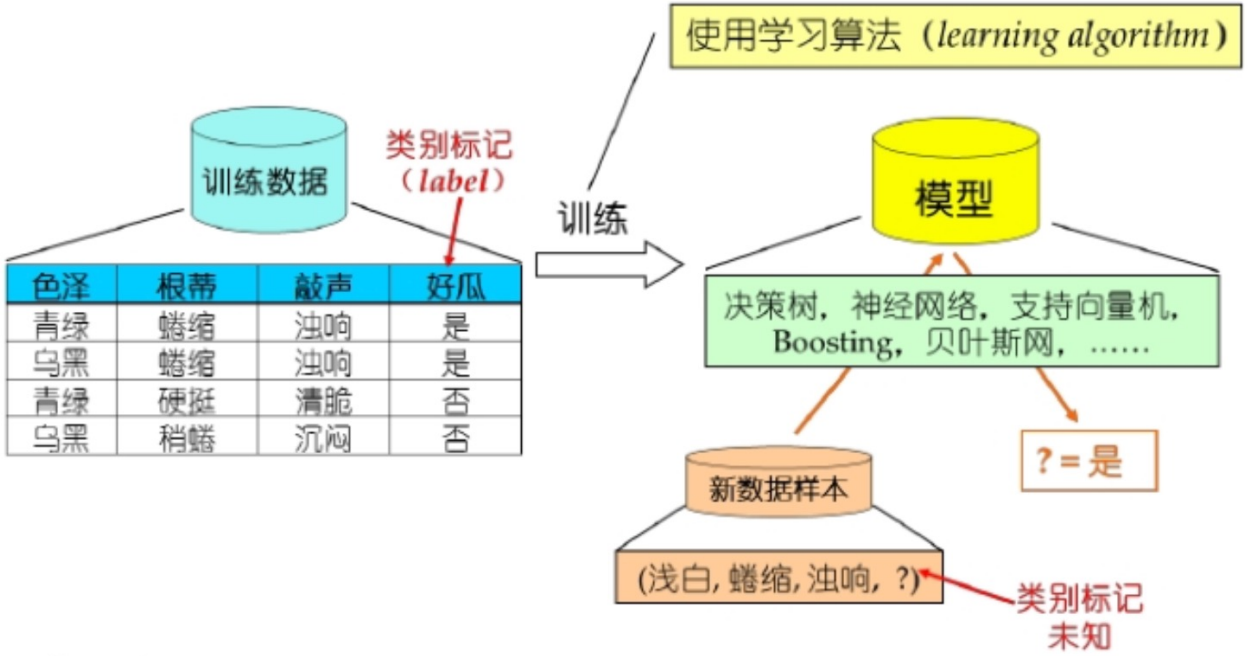
机器学习基本概念

□ 数据集

色泽（x_1）	根蒂（x_2）	敲声（x_3）	好瓜（y）
0（青绿）	0（蜷缩）	0（浊响）	0（是）
1（乌黑）	0（蜷缩）	0（浊响）	0（是）
0（青绿）	1（硬挺）	1（清脆）	1（否）
1（乌黑）	2（稍蜷）	2（沉闷）	1（否）

样本

特征　　标签



□ 模型

$y = ax_1 + bx_2 + cx_3 + d$　　假设$y = x_1 + x_2 + x_3 - 1$，代入x_1，x_2，x_3

$\hat{y}$ 是预测值

若≤0，$\hat{y} = 0$

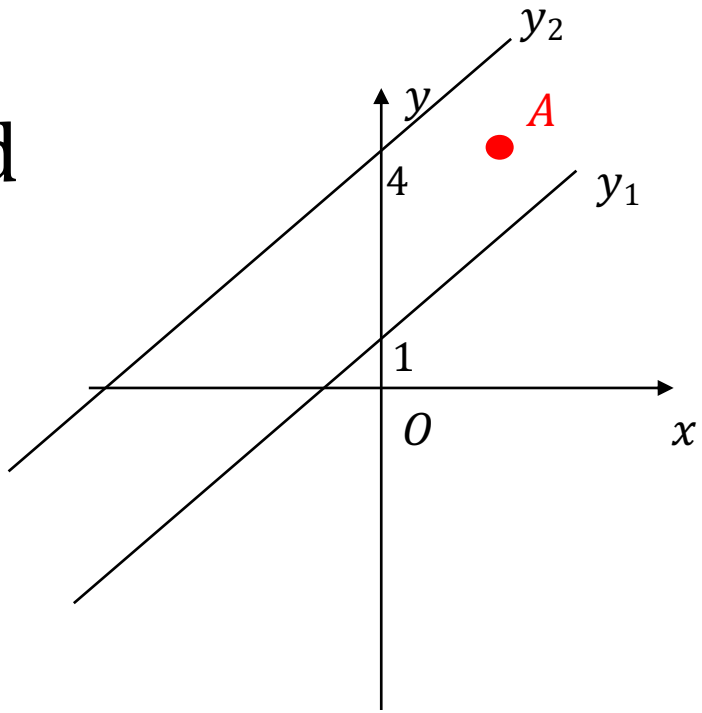
若>0，$\hat{y} = 1$

机器学习基本概念

□ 不同的θ有什么影响呢?

$$y = ax_1 + bx_2 + cx_3 + d$$

$$\theta = \{a, b, c, d\}$$



$(1) y_1$: $a = 1, b = 1, c = 1, d = -1$

$(2) y_2$: $a = 1, b = 1, c = 1, d = -4$

y_1 设置下A是坏瓜， y_2 设置下A是好瓜 。

不同的θ会导致预测的不同。

机器学习基本概念

□ 机器学习实际就是从训练数据中对θ进行学习。

样本	色泽（x_1）	根蒂（x_2）	敲声（x_3）	好瓜（y）
1	0（青绿）	0（蜷缩）	0（浊响）	0（是）
2	1（乌黑）	0（蜷缩）	0（浊响）	0（是）
3	0（青绿）	1（硬挺）	1（清脆）	1（否）
4	1（乌黑）	2（稍蜷）	2（沉闷）	1（否）

$$y = ax_1 + bx_2 + cx_3 + d$$

$(1) y_1$: $a = 1, b = 1, c = 1, d = -1$

$(2) y_2$: $a = 1, b = 1, c = 1, d = -4$

y_1 设置下样本1，2是好瓜，样本3，4是坏瓜

y_2 设置下样本1，2，3是好瓜，样本4是坏瓜

哪一组参数更好呢?

机器学习基本概念

□ 损失函数可以判断哪组参数更好。

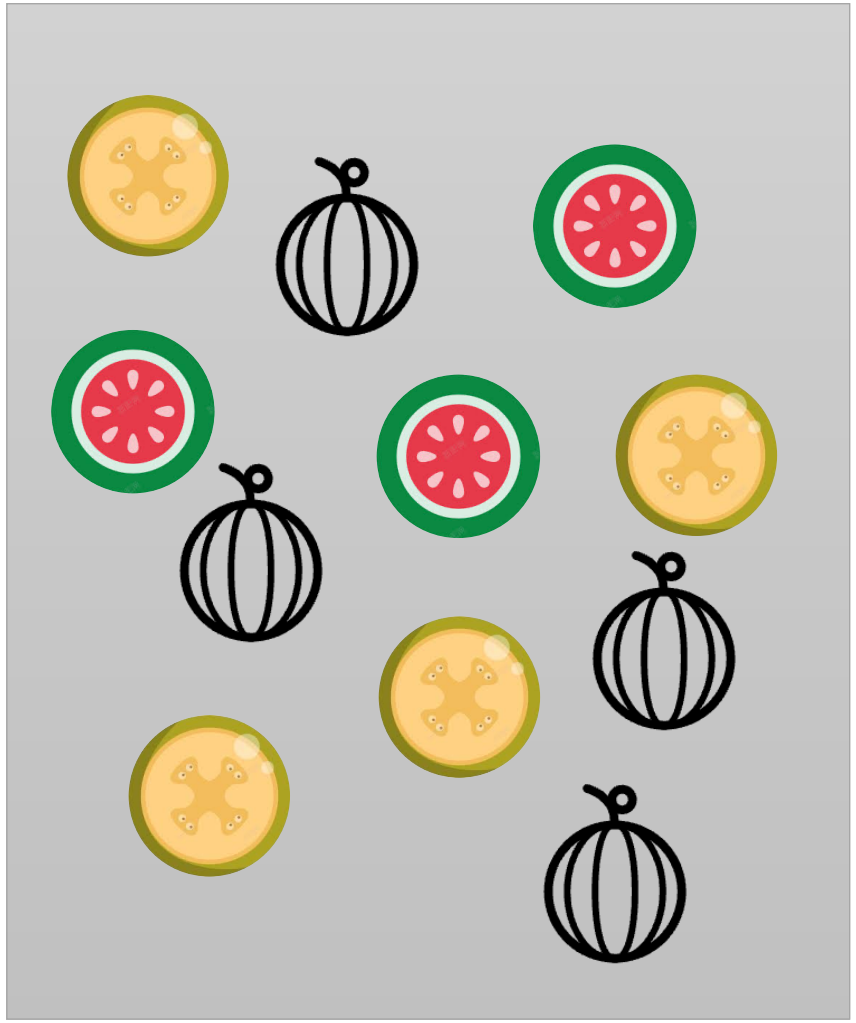
- **0-1损失函数：** $L(y,\hat{y}) = \sum_{i=1}^{N} I(y \neq \hat{y}), I = \begin{cases} 1, y \neq \hat{y} \\ 0, y = \hat{y} \end{cases}$ $\quad y_1: L(y_1,\hat{y_1}) = 0 \quad < \quad y_2: L(y_2,\hat{y_2}) = 1$
- **绝对值损失函数：** $L(y,\hat{y}) = \sum_{i=1}^{N} |y - \hat{y}|$ $\quad L(y_1,\hat{y_1}) = 0 \quad < \quad L(y_2,\hat{y_2}) = 1$
- **平方损失函数：** $L(y,\hat{y}) = \sum_{i=1}^{N} (y - \hat{y})^2$ $\quad L(y_1,\hat{y_1}) = 0 \quad < \quad L(y_2,\hat{y_2}) = 1$

y_1 设置的损失更小，参数更好。

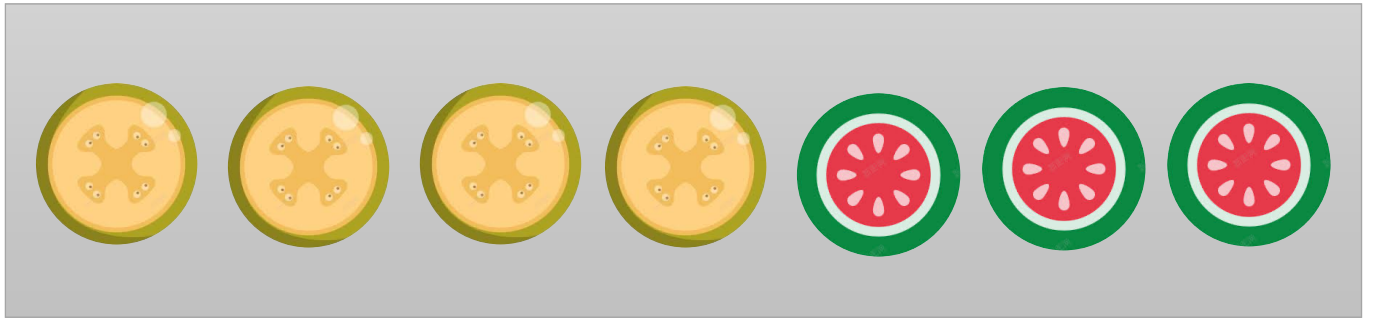
训练的过程就是在损失函数的指引下找到最好的参数设置的过程，实际上是一个优化问题。

机器学习基本概念

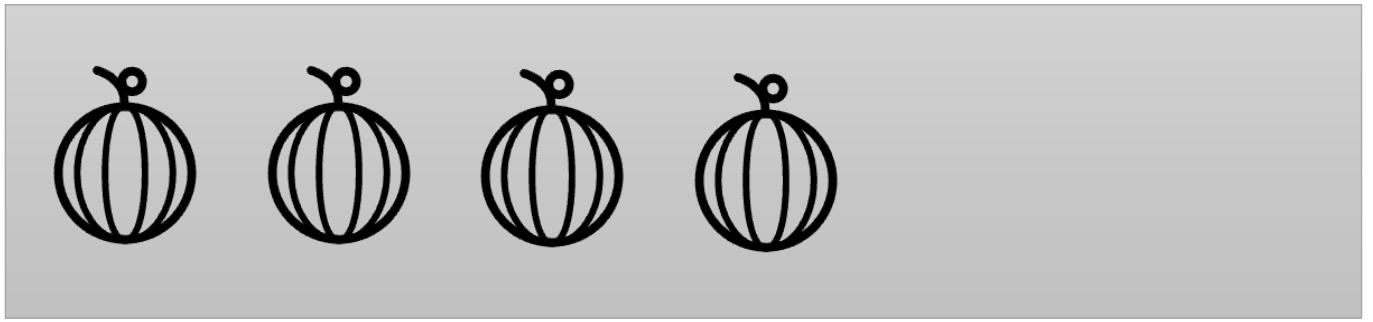
□ 数据集



□ 训练集：帮助训练模型，简单说就是通过训练集的数据选出明确的参数。



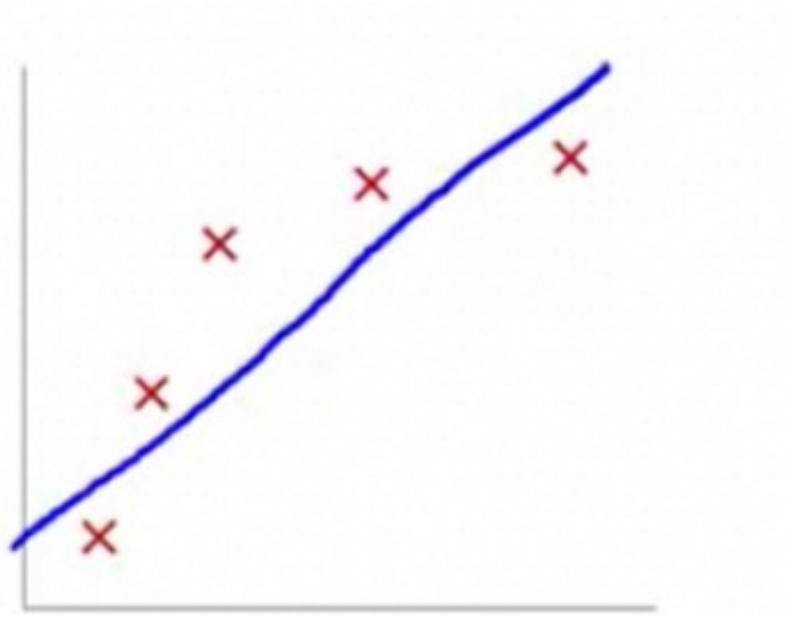
□ 测试集：为了测试已经训练好的模型的性能。



机器学习基本概念

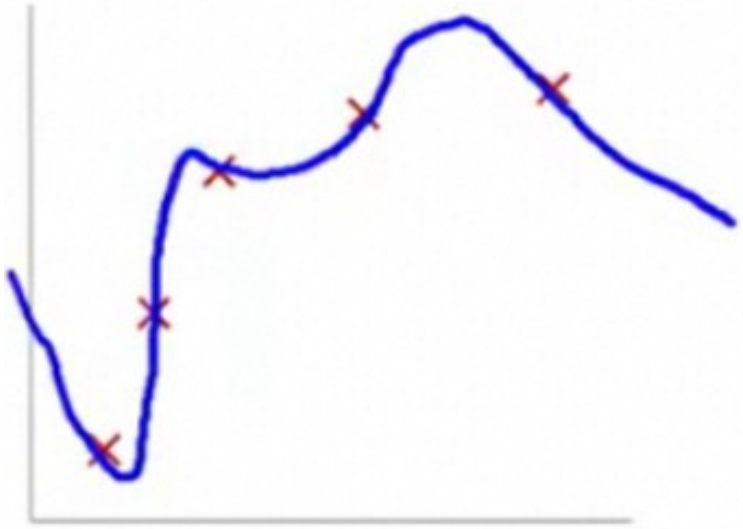
□ 高考备考中的"测试集"

- 平时的模拟考试题 = 训练集
- 真正的高考试题 = 测试集
- 目标：在真正高考时取得好成绩



□什么是过拟合?

- 学生小明死记硬背往年题库
- 模拟考试成绩优秀(训练集表现好)
- 高考遇到新题型完全懵了(测试集表现差)
- 这就是典型的"过拟合"



目录

人工智能：从智能的内涵到人工智能四要素与数据形态（回顾）

- 表示：（知识/模型）长什么样？
  机器编码f_θ、input、output、feedback。
- 推理：（知识/模型）怎么用来解决问题？
  给定input，机器实现f_θ计算output。
- 学习：（知识/模型）怎么来的？
  基于数据<input，output，feedback>集，给定f，更新计算θ。

人工智能四要素（“知识”有待商榷）

1. 算法/模型：f（及部分θ）
2. 计算：f_θ/input/output/feedback转换
3. 数据：<input，output，feedback>
4. 知识：θ（及部分f）

数据：<input，output，feedback>

- 有监督：<input，output，feedback>
- 无监督：<input，output，空缺>
- 强化：<input，output，多步>
- 自监督：<input，input*，正/1>
- ……

机器学习类型

- **监督学习（supervised learning）**

  有标签，学习从输入数据到标签的映射关系

- **无监督学习 （unsupervised learning）**

  无标签，学习数据内在的规律

- **半监督学习（semi-supervised learning）**

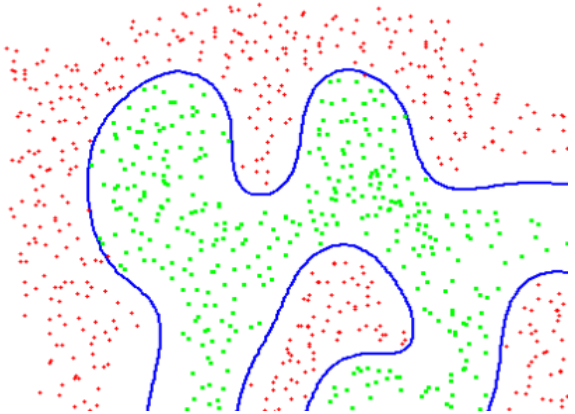
  利用大量未标注数据信息辅助少量标注数据的学习

- **强化学习（reinforcement learning）**

  计算机通过与环境互动逐渐强化自己的学习方式

机器学习类型——监督学习（重点）

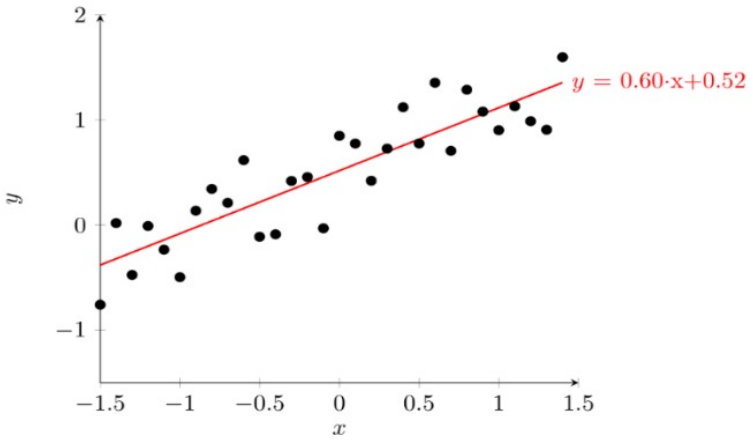
口分类（classification）：



预测的是离散值

- 比如：垃圾邮件识别、图片识别等

口回归（regression）：



预测的是连续值

- 比如：预测合肥的房价、未来股票的走势等

机器学习类型——无监督学习

- **聚类（clustering）：**

  将数据集划分为若干个子集

  - K-means, GMM

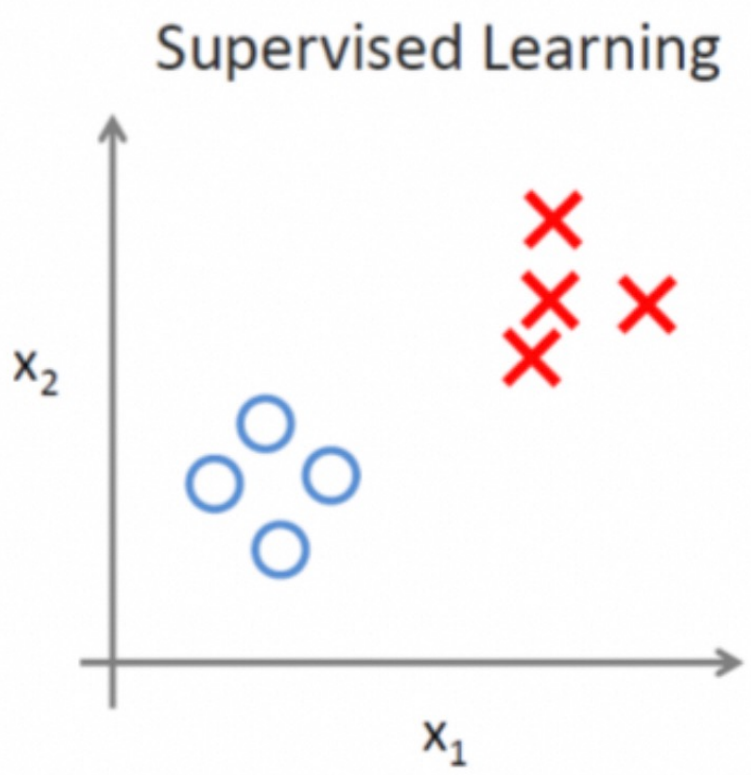
- **降维（dimensionality reduction）：**

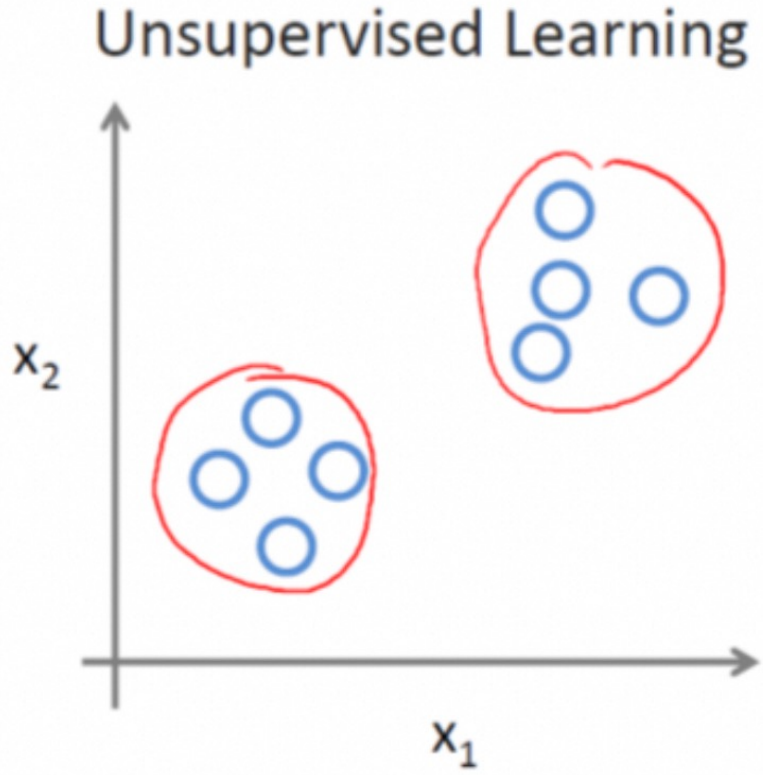
  将高维空间转变为与一个低维子空间

  - PCA

机器学习类型

□监督学习与无监督学习的区别：

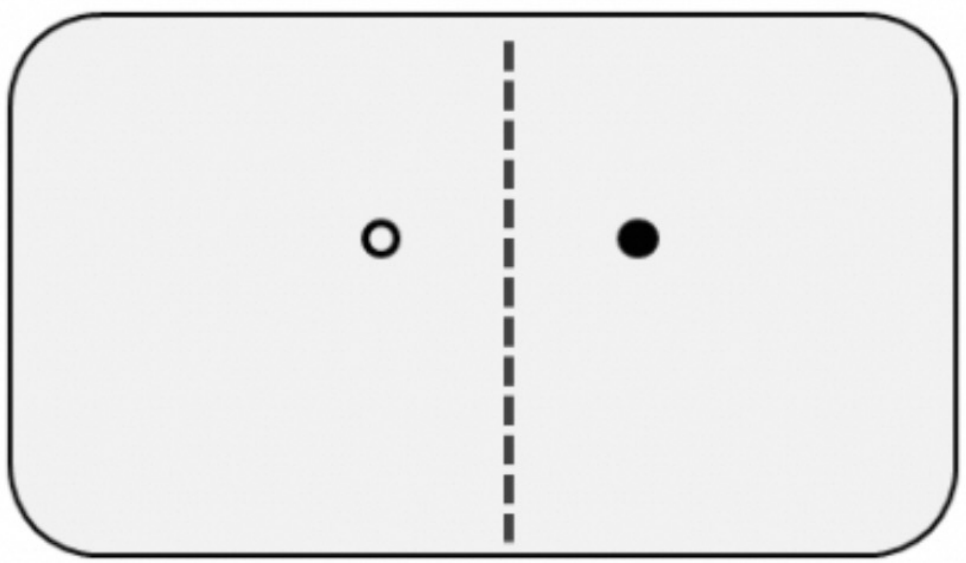


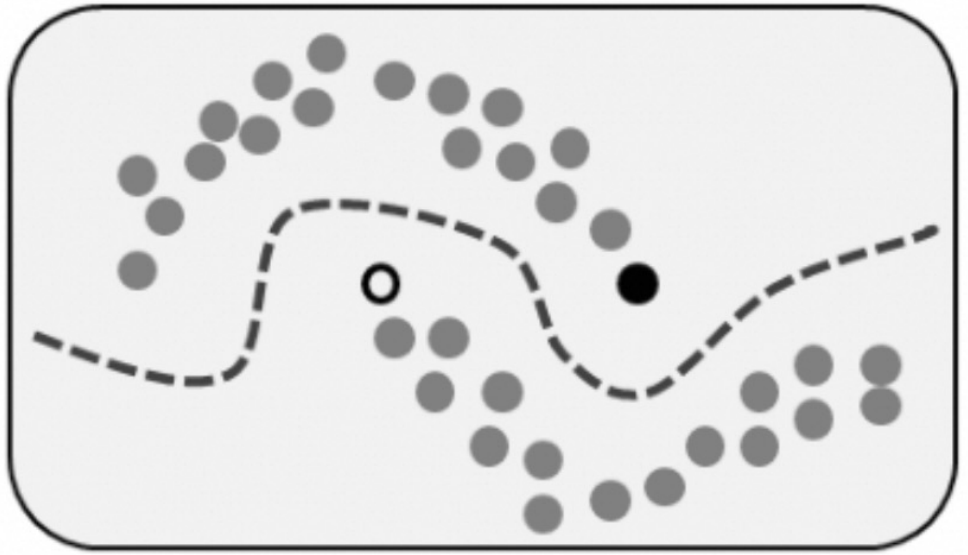


机器学习类型——半监督学习

口半监督学习（semi-supervised learning）：

少量标注数据，大量未标注数据





机器学习类型——强化学习

□ 强化学习（reinforcement learning）：

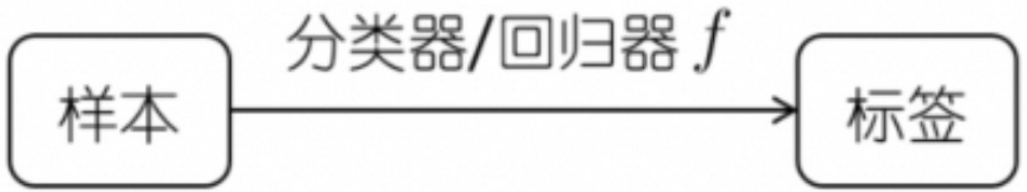
经验E是由计算机与环境互动获得的，计算机产生行为，同时获得这些行为的结果我们的程序只需要定义这些行为的收益函数，对行为进行奖励或惩罚。通过设计算法让计算机改变自己的行为模式去最大化收益函数，完成机器学习的过程

- AlphaGo

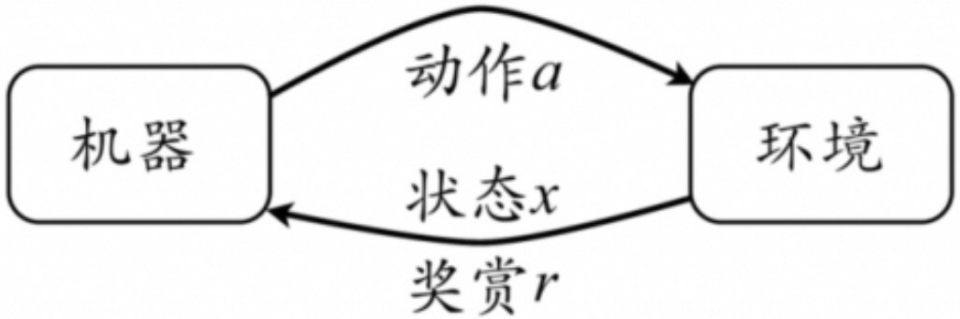
机器学习类型

强化学习与监督学习的区别：

- 监督学习：提供有标签样本



- 强化学习：没有有标签的样本，通过执行动作之后反馈的奖励来学习



强化学习在某种意义上可以认为是具有“延迟标签信息”的监督学习

目录

机器学习发展历程

诞生
1943-1956
控制论
符号推理

高潮
1956-1974
搜索式推理
自然语言处理
反向传播算法

1970
在三到八年的时间里我们将得到一台具有人类平均智能的机器。
——马文·闵斯基

1967
一代之内……创造 ‘人工智能’ 的问题将获得实质上的解决。
——马文·闵斯基

1965
二十年内，机器将能完成人能做到的一切工作。
——艾伦·纽厄尔和赫伯特·西蒙

1958
十年之内，计算机将成为国际象棋世界冠军。
十年之内，计算机将发现并证明一个重要的数学定理。
——艾伦·纽厄尔和赫伯特·西蒙

低谷/冬天
1974-1980
运算能力不够
复杂性指数级爆炸
没有常识
“中文房间”问题

繁荣
1980-1987
专家系统五代机
神经网络

低谷/冬天
1987-1993
PC性能超过专用的AI硬件
五代机没有实用的成果

发展
1993-2006
1997深蓝
贝叶斯网络
隐马科夫模型
信息论的发展

突破
2007-
深度学习
视觉/翻译在比赛中超过人类

突破
2016 –
AlphaGO在围棋比赛中击败李世石

突破
2022 –
ChatGPT问世！

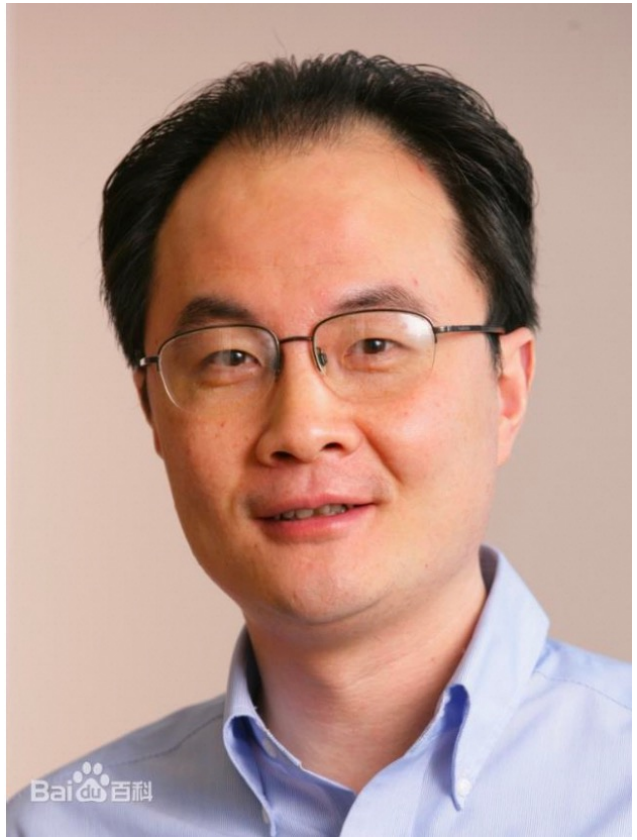
具有非常惊艳的语言理解、生成、知识推理能力

机器学习届的执牛耳者

杨立昆（Yann LeCun）
杰弗里•欣顿（Geoffrey Hinton）
本吉奥（Bengio）
共同获得了2018年计算机科学的最高奖项**ACM图灵奖**。

Andrew Ng
中文名**吴恩达**，斯坦福大学副教授，前“百度大脑”的负责人与百度首席科学家。

机器学习届的国内泰斗



李航，现任字节跳动科技有限公司人工智能实验室总监，北京大学、南京大学客座教授，IEEE 会士，ACM杰出科学家，CCF 高级会员。
代表作：《统计学习方法》

周志华，南京大学计算机科学与技术系主任、人工智能学院院长。
代表作：《机器学习》（西瓜书）

机器学习届的青年才俊

陈天奇，陈天奇是机器学习领域著名的青年华人学者之一，本科毕业于上海交通大学ACM班，博士毕业于华盛顿大学计算机系。
主要贡献：设计了XGBoost算法。

何恺明，本科就读于清华大学，博士毕业于香港中文大学多媒体实验室。2016年，加入Facebook AI Research 担任研究科学家。
主要贡献：设计了ResNets

国内外知名人工智能企业榜单

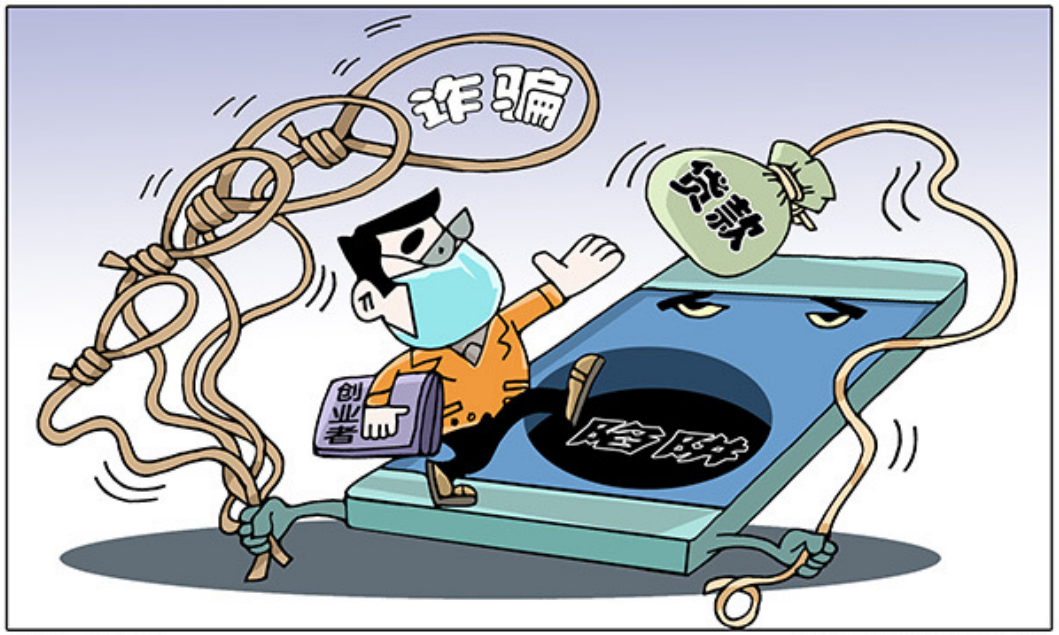
编码	企业名称	人工智能技术	应用领域	所属国家	成立时间	资本市场状态	市值/估值/融资额
1	Microsoft（微软）	计算机视觉技术、自然语言处理技术 等	办公	美国	1975年	上市	市值1.21万亿美元
2	Google（谷歌）	计算机视觉技术、自然语言处理技术 等	综合	美国	1998年	上市	市值9324亿美元
3	Facebook（脸书）	人脸识别、深度学习等	社交	美国	2004年	上市	市值5934亿美元
4	百度	计算机视觉技术、自然语言处理技 术、知识图谱等	综合	中国	2001年	上市	市值438亿美元
5	大疆创新	图像识别技术、智能引擎技术等	无人机	中国	2006年	战略融资	估值210亿美元
6	商汤科技	计算机视觉技术、深度学习	安防	中国	2014年	D轮融资	估值70亿美元
7	旷视科技	计算机视觉技术等	安防	中国	2011年	D轮融资	估值40亿美元
8	科大讯飞	智能语音技术	综合	中国	1999年	上市	市值108亿美元
9	Automation Anywhere	自然语言处理技术、非结构化数据认知	企业管理	美国	2003年	B轮融资	估值68亿美元
10	IBM Watson（IBM沃森）	深度学习、智适应学习技术	计算机	美国	1911年	上市	市值1198亿美元
11	松鼠AI 1对1	智适应学习技术、机器学习	教育	中国	2015年	A轮融资	估值11亿美元
12	字节跳动	跨媒体分析推理技术、深度学习、自 然语言处理、图像识别	资讯	中国	2012年	Pre-IPO轮融资	估值750亿美元
13	Netflix（网飞）	视频图像优化、剧集封面图片个性 化、视频个性化推荐	媒体及内容	美国	1997年	上市	市值1418亿美元
14	Graphcore	智能芯片技术、机器学习	芯片	英国	2016年	D轮融资	估值17亿美元
15	NVIDIA（英伟达）	智能芯片技术	芯片	美国	1993年	上市	市值1450亿美元
16	Brainco	脑机接口	教育、医疗、智能硬件	美国	2015年	天使轮融资	融资额600万美元
17	Waymo	自动驾驶	交通	美国	2016年	C轮融资	估值1050亿美元
18	ABB Robotics	机器人及自动化技术	机器人	瑞士	1988年	上市	市值514亿美元
19	Fanuc（发那科）	机器人技术	制造	日本	1956年	上市	市值362亿美元
20	Preferred Networks	深度学习、机器学习技术	物联网	日本	2016年	C轮融资	估值20亿美元

目录

机器学习示例——贷款诈骗

口什么是贷款诈骗?

- 贷款诈骗就是骗子假装提供贷款服务，实际上是为了骗你的钱或个人信息。
- 比如，他们可能会说“先交手续费才能拿到贷款”，但你交了钱后，贷款却根本不存在。
- 或者假装是银行工作人员，要求你先交“手续费”。
- 又或者发送假短信或邮件，诱导你点击链接并输入银行卡信息。



警惕“黑手”　　新华社发 司海英 作

机器学习示例——贷款诈骗

口如何将机器学习应用于贷款诈骗?

第一步：收集数据

- 收集过去的贷款申请记录，包括哪些是正常的，哪些是诈骗的。
- 数据可能包括：申请人的年龄、收入、职业、申请时间等。
- 划分训练集和测试集。

第二步：筛选数据

- 筛选数据是为了去掉不靠谱的数据，让模型学得更好。
- 异常值是指那些特别奇怪的数据点，可能是录入错误或故意伪造的。
  - 比如：
    1. 申请人填写的收入是“1亿元”，但职业却是“学生”。
    2. 贷款申请时间是凌晨3点（大多数人不会在这个时间申请贷款）。
- 处理方法：
  - 删除异常值：直接去掉这些奇怪的数据。
  - 设定范围：比如规定“收入不能超过100万”，超出范围的数据就标记为可疑。

机器学习示例——贷款诈骗

口如何将机器学习应用于贷款诈骗?

第三步：处理特征

- 特征就是用来描述贷款申请的信息，比如年龄、收入、职业等。
- 但是如果某些特征的数值特别大（比如收入是几百万），会影响模型的学习效果。
- 解决方法：
  - 归一化：把所有特征的数值缩放到一个固定的范围，比如0到1之间。
    比如：把收入从“1000元到100万元”变成“0到1”。
  - 标准化：把特征的数值调整为平均值为0、标准差为1的形式。
    比如：把年龄从“18到80岁”变成“-2到2”。
- 比如：
  - 如果收入是“100万元”，年龄是“30岁”，直接用这两个数字会让模型觉得收入更重要。但归一化后，收入和年龄的数值变得差不多，模型就能公平对待它们。

机器学习示例——贷款诈骗

口如何将机器学习应用于贷款诈骗?

第四步：模型选择与训练

- 在贷款诈骗问题中，我们用的是**分类模型** 。

  分类模型的目标是把数据分成两类：正常 vs. 诈骗。

- 我们可以使用**平方损失函数**。

  我们的目标是在训练过程找到让损失最小的参数设置，这样模型就能更准确地识别诈骗。

- 举个例子：

  如果模型说“这笔贷款有90%的概率是诈骗”，但实际是正常的，损失就会很高，我们就更倾向于不选择这组参数。

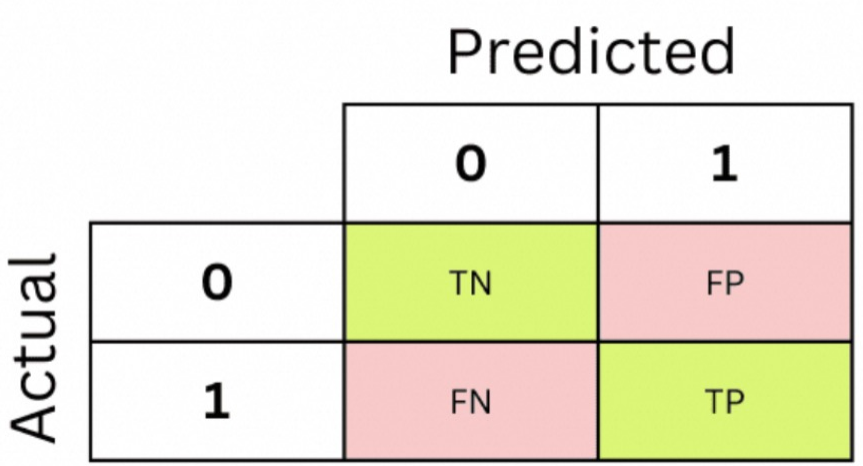
  如果模型说“这笔贷款有90%的概率是诈骗”，而实际确实是诈骗，损失就会很低，我们就更倾向于选择这组参数。

第五步：模型测试

在测试集上测试模型的性能，如果性能优秀，就可以上线模型。当有新的贷款申请时，模型可以判断这是不是诈骗。

机器学习示例——贷款诈骗

- **常用指标：**
  - **准确率（Accuracy）：**
    - 正确预测的比例。
    - 公式：(TP + TN) / (TP + TN + FP + FN)
  - **精确率/查准率（Precision）：**
    - 预测为诈骗的贷款中有多少是真正的诈骗。
    - 公式：TP / (TP + FP)
  - **召回率/查全率（Recall）：**
    - 所有诈骗贷款中有多少被正确识别。
    - 公式：TP / (TP + FN)

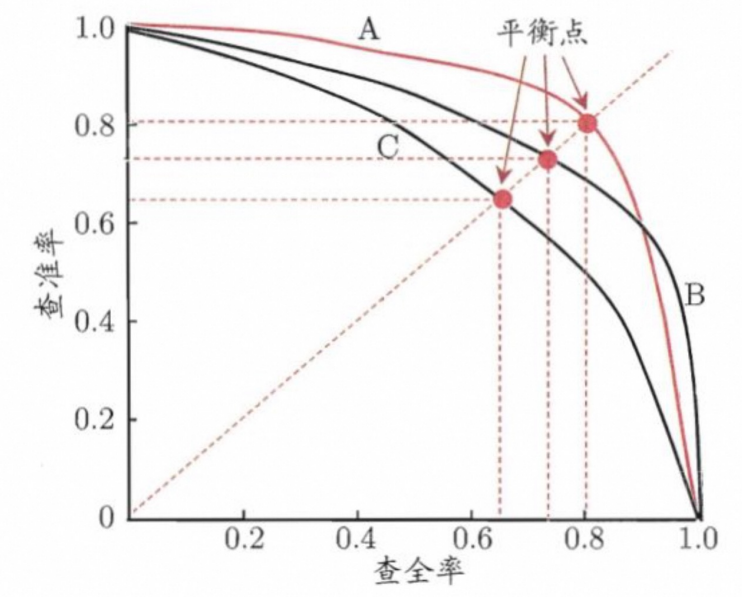


这些指标之间有什么关系吗?

机器学习示例——贷款诈骗

□ F1 分数（F1 Score）：

- 平衡Precision和Recall的综合指标，即下图的平衡点。
- 公式：2 * (Precision * Recall) / (Precision + Recall)



如果这里是判断好瓜和坏瓜任务，y_1相当于模型A，y_2相当于模型C， y_1的查准率和查全率一直高于y_2，所以y_1参数更好。

□贷款诈骗实际意义：

- **高召回率：**尽量减少漏报，确保尽可能多地识别贷款诈骗。
- **高精确率：**避免误报，防止正常贷款被误判为诈骗。

课后作业

1. 贷款诈骗还包含哪些特征可以用于分析?
2. 电子邮件垃圾分类包含哪些特征可以用于分析?

中国科学技术大学
University of Science and Technology of China

谢谢！