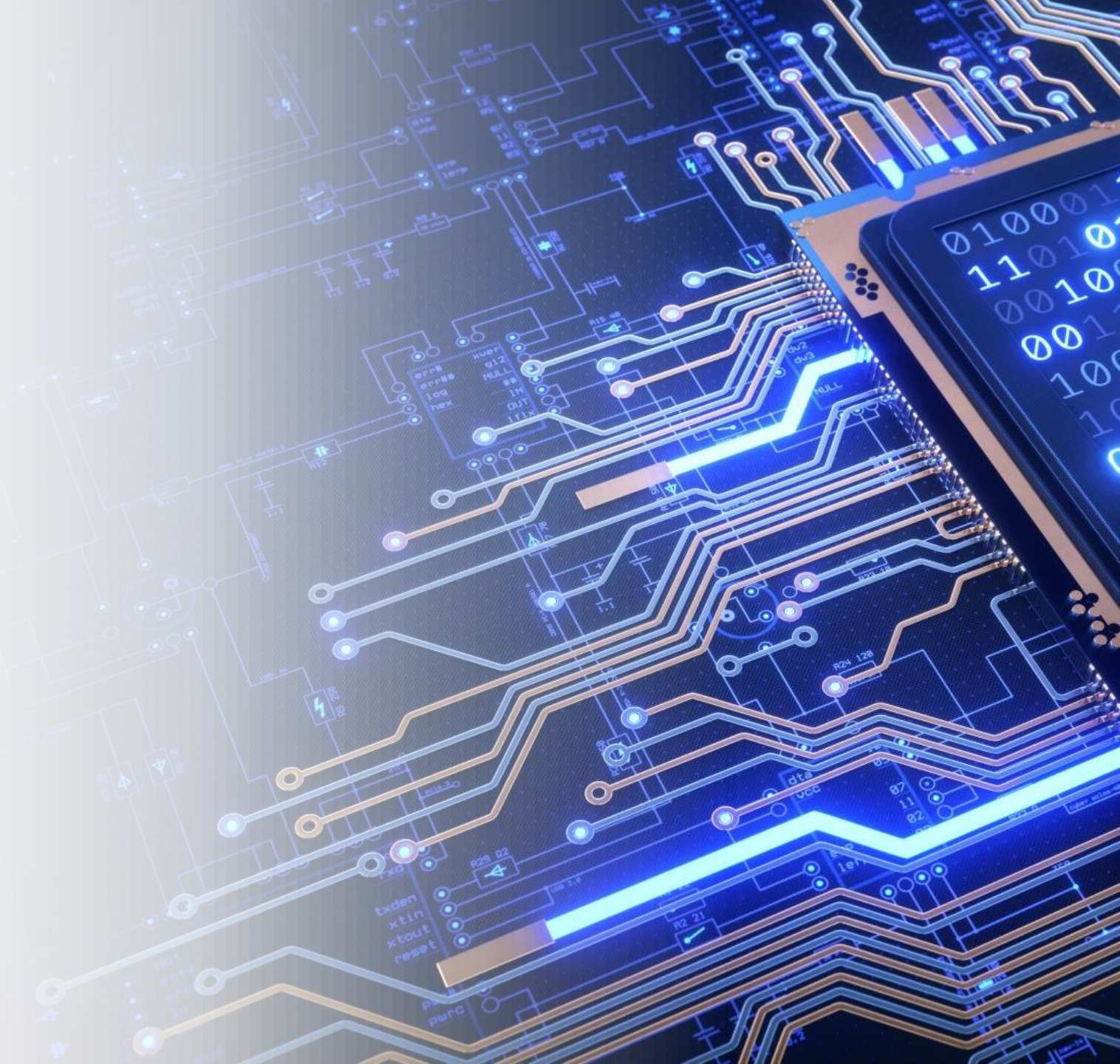
機器學習入門

講師：陳仁政 Ph.D
clement1972@gmail.com



講師經歷

- 中原大學電子所博士
- 中央研究院資訊科學所博士後研究員
- 吉鴻電子資深工程師
- 冠捷科技正工程師
- 104人力銀行人資學院資料科學家
- 長庚大學工商管理系兼任實務教師
- 104人力銀行人資學院顧問
- 台灣人工智能產業協會講師
- 實踐大學推廣中心講師
- 緯育講師

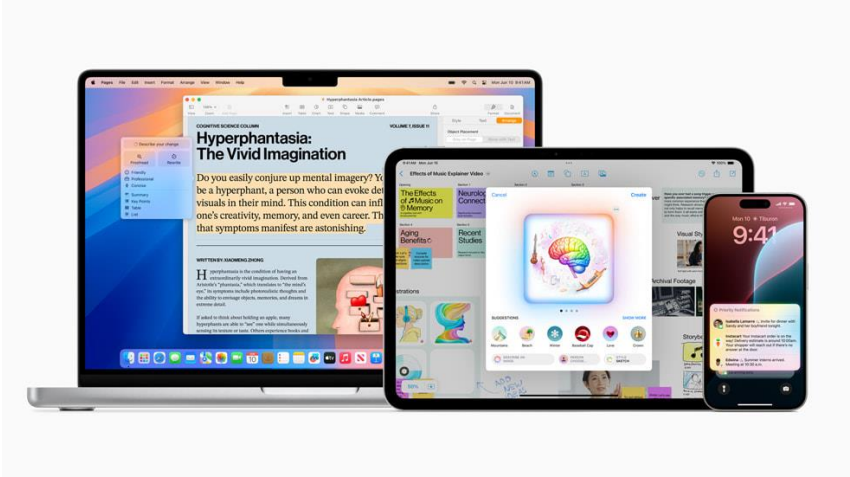
- 收鈔機韌體(偽鈔偵測)
- 電視韌體
- 交通執法系統
- 人才適任與久任度評估系統

人工智慧簡介

自動駕駛系統

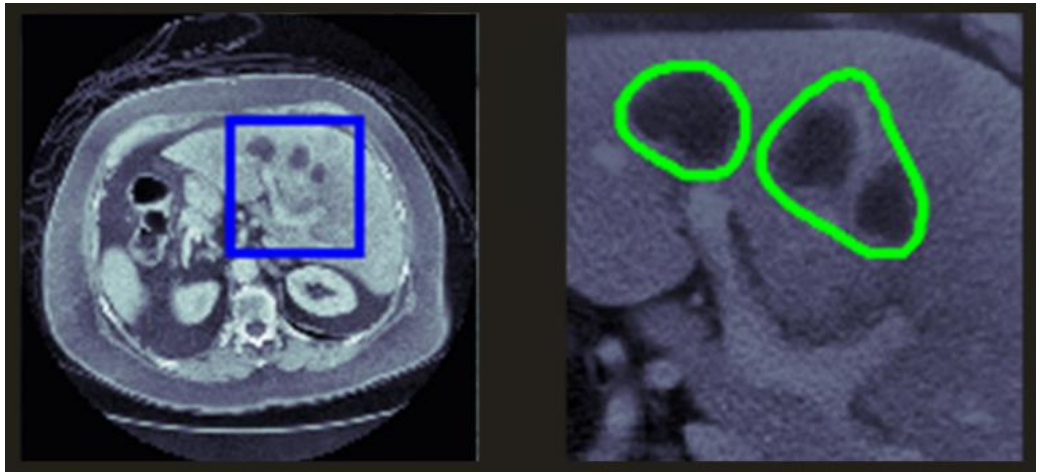
新聞聚類系統

個人智慧助理



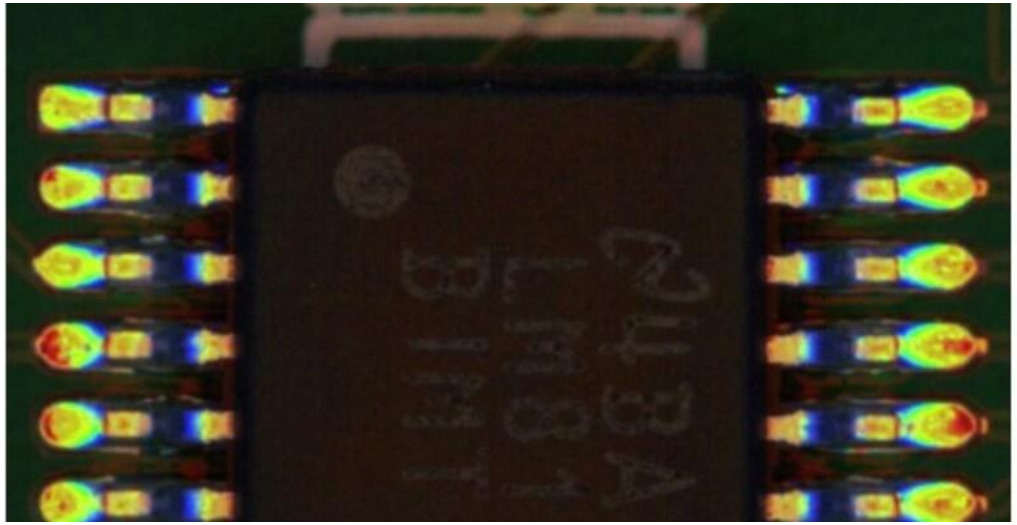
商品推薦系統

醫療診斷系統



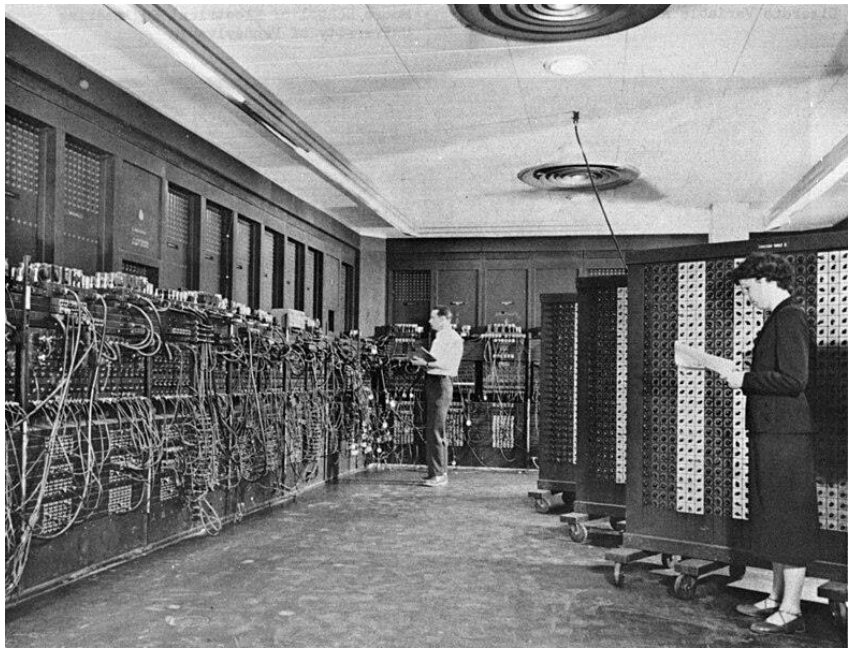
智慧交通控制系統

工業光學瑕疵診測系統

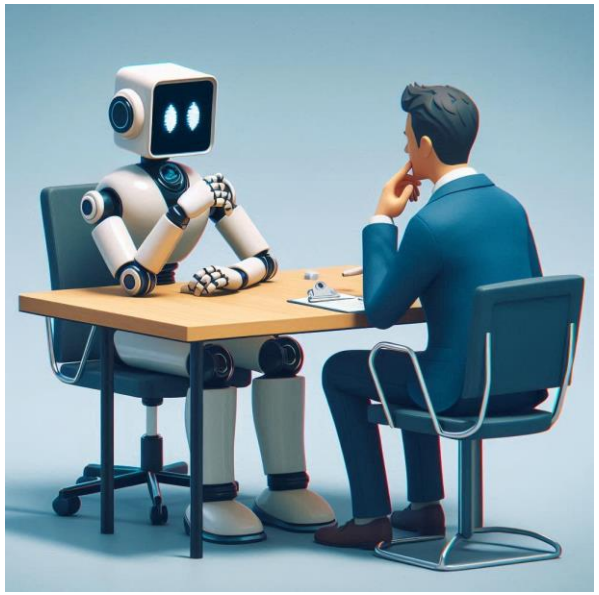



人工智慧
符號人工智慧
1950s~1980s
機器學習
1980~迄今
深度學習
2012~迄今

1946 ENIAC



1950 圖靈測試



1956 達特茅斯會議

image source: wiki

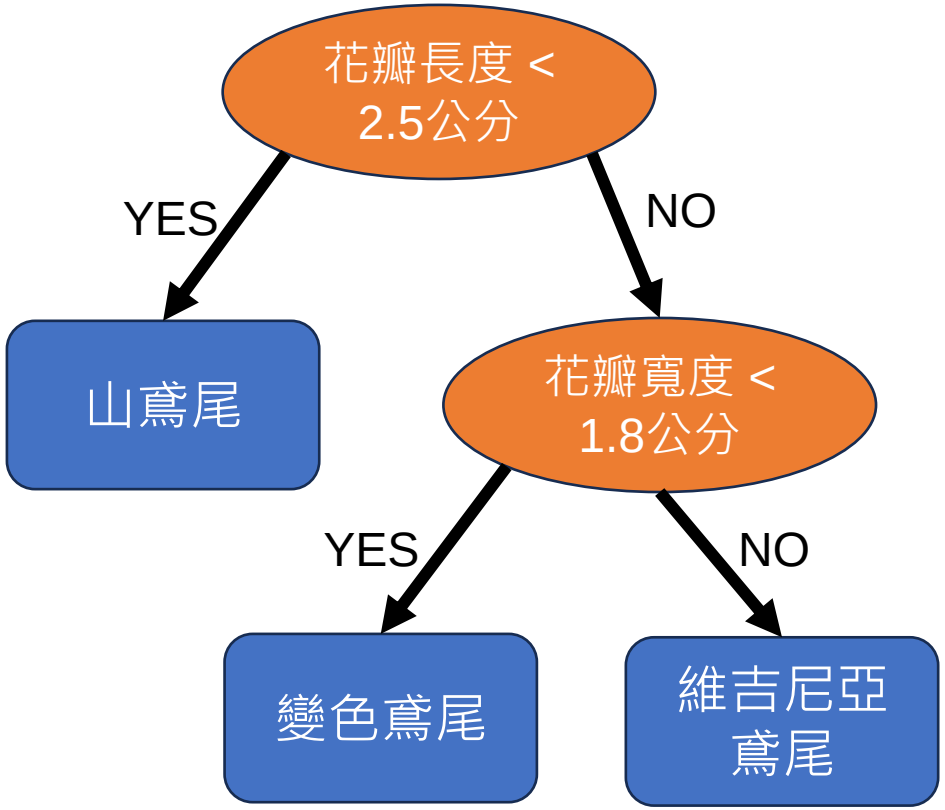
符號人工智慧 Symbolic artificial intelligence

- 1950年代中期到1980年代後期

- 專家系統：基於規則和知識庫進行推理
  - MYCIN : 醫療診斷 (1970年代)
  - Exsys : 第一個成功商業化的專家系統 (1983)

符號人工智慧 Symbolic AI

- 使用規則當成知識庫
  - 以鳶尾花分類為例
    - 如果花瓣長度小於2.5公分，則是山鳶尾
    - 如果花瓣長度大於2.5公分 且 花瓣寬度小於1.8公分，則是變色鳶尾
    - 如果花瓣長度大於2.5公分 且 花瓣寬度大於1.8公分，則是維吉尼亞鳶尾
- 規則由領域專家與AI工程師合作建立
- 常見專家系統語言：Prolog、Lisp



符號人工智慧的缺點

- 規則知識庫過於複雜
- 專家知識難以系統化和總結
- 缺乏自主學習，無法適應新情境
- 容易受偏見影響，決策不夠客觀

機器學習 Machine Learning

- 計算型智慧
  - 數值計算 (數據驅動)
- 1980~迄今
- 常見演算法
  - K近鄰算法(KNN)
  - 決策樹、隨機森林
  - 主成分分析(PCA)、線性回歸、支持向量機(SVM)
  - 類神經網路(Neural Network)

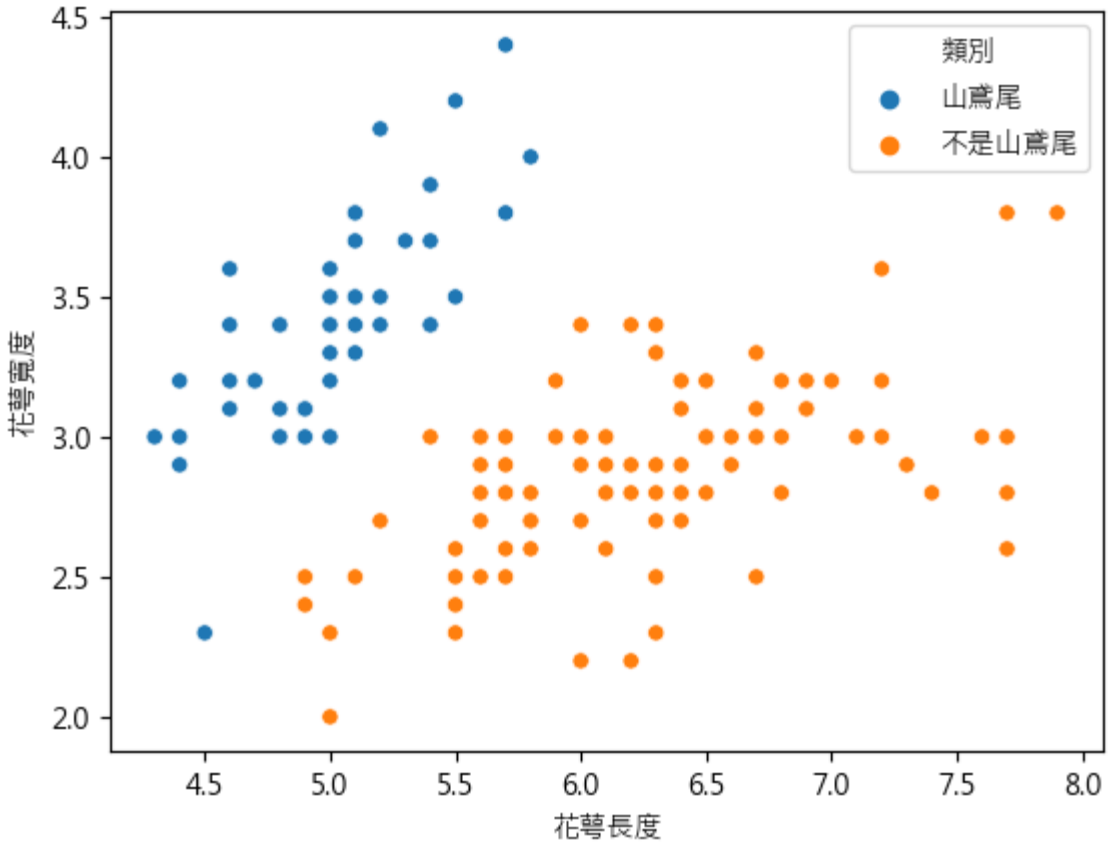
機器學習數據驅動示範

-3.4*花萼長度 + 3.15*花萼寬度 + 8.44 > 0

花萼長度	花萼寬度	類別	
5.1	3.5	山鳶尾	-3.4*5.1 + 3.15*3.5 + 8.44 = **2.125**
4.9	3	山鳶尾	-3.4*4.9 + 3.15*3 + 8.44 = **1.23**
4.7	3.2	山鳶尾	
7	3.2	不是山鳶尾	-3.4*7 + 3.15*3.2 + 8.44 = **-5.28**
6.4	3.2	不是山鳶尾	
6.9	3.1	不是山鳶尾	
6.3	3.3	不是山鳶尾	
5.8	2.7	不是山鳶尾	-3.4*5.8 + 3.15*2.7 + 8.44 = **-2.77**
7.1	3	不是山鳶尾	

訓練資料範例

花萼長度	花萼寬度	類別
5.1	3.5	山鳶尾
4.9	3	山鳶尾
4.7	3.2	山鳶尾
7	3.2	不是山鳶尾
6.4	3.2	不是山鳶尾
6.9	3.1	不是山鳶尾
6.3	3.3	不是山鳶尾
5.8	2.7	不是山鳶尾
7.1	3	不是山鳶尾

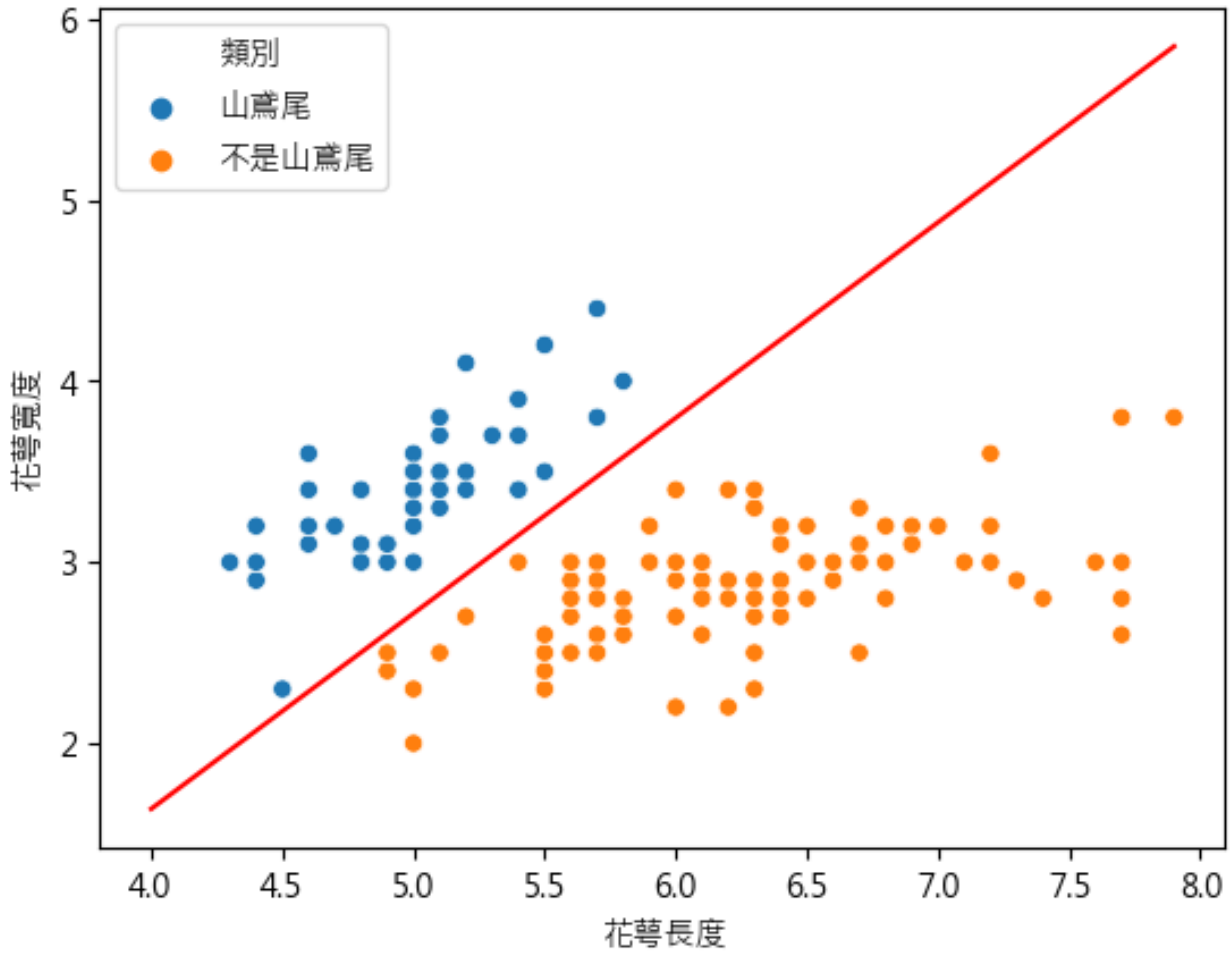


$$y = f(\text{花萼長度}, \text{花萼寬度})$$

$$w1*\text{花萼長度} + w2*\text{花萼寬度} > b$$

$$-3.4*\text{花萼長度} + 3.15*\text{花萼寬度} > -8.44$$

f(x) = -3.4*花萼長度 + 3.15*花萼寬度 + 8.44 > 0



機器學習

機器學習
Machine Learning

- 1980年代開始成為研究主流
- 訓練階段
  - 透過資料**學習**並建立模型
- 預測階段
  - 使用訓練好的模型進行預測
- 能反覆的從新資料修改預測模型

機器學習（Machine Learning）是一種人工智慧（Artificial Intelligence）的分支，指的是通過使用統計和數學模型，使計算機系統能夠自動從數據中學習和改進，而不需要明確的編程指令。

機器學習的目標是通過對大量數據進行學習和模式識別，使計算機系統能夠自主地進行預測、分類、集群、優化和決策等任務。它通常涉及以下步驟：

1. 數據收集：收集和整理相關數據，包括特徵和標籤（如果有）。
2. 數據預處理：對數據進行清洗、轉換和縮放，以消除噪音和不必要的變異性。
3. 模型選擇：選擇適合問題的機器學習模型，如監督學習、非監督學習、強化學習等。
4. 模型訓練：使用訓練數據對選定的模型進行學習和參數優化。
5. 模型評估：使用測試數據評估模型的性能和準確度，進行評估指標的計算和比較。
6. 模型應用：將訓練好的模型應用於新的、未知的數據進行預測、分類或其他相關任務。

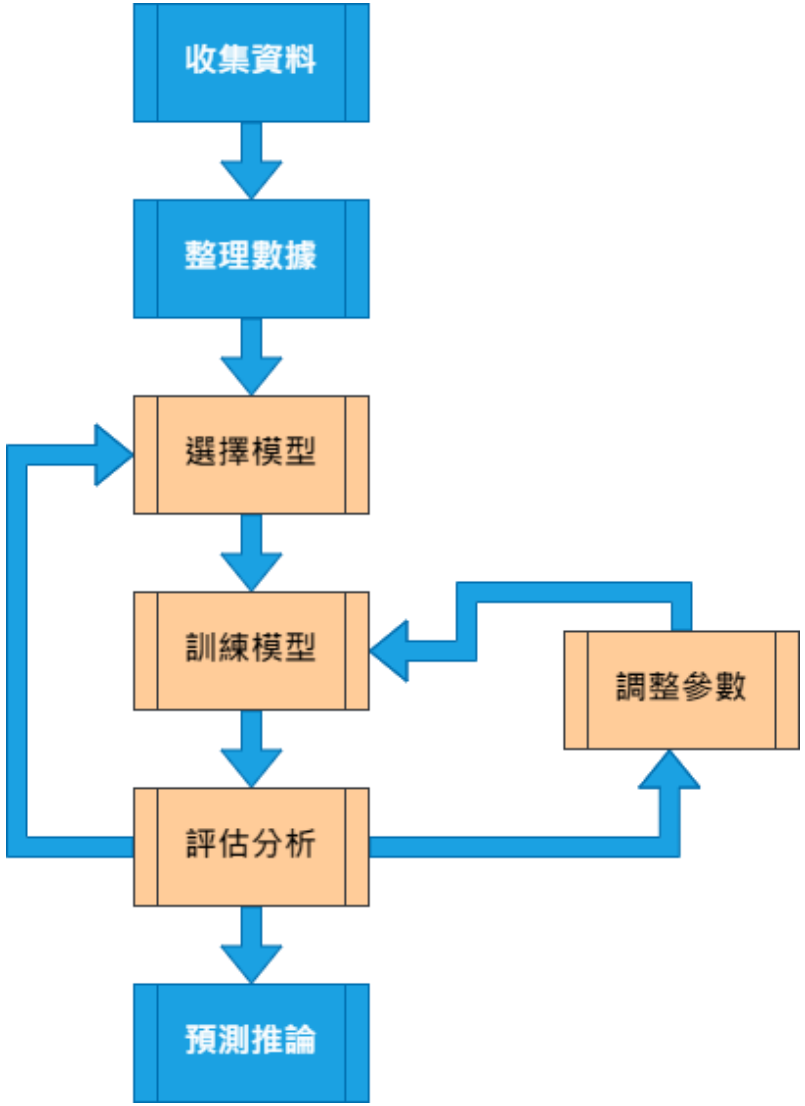
機器學習使用各種統計和數學技術，如回歸分析、分類算法、支持向量機、決策樹、深度神經網絡等。這些技術允許計算機系統從數據中提取特徵和模式，並使用這些知識做出預測或做出相應的行動。

機器學習在各個領域有廣泛的應用，包括圖像和語音辨識、自然語言處理、推薦系統、金融預測、醫學診斷等。它的發展和應用對於解決複雜問題、優化效能和改進決策過程具有重要意義。

訓練資料範例

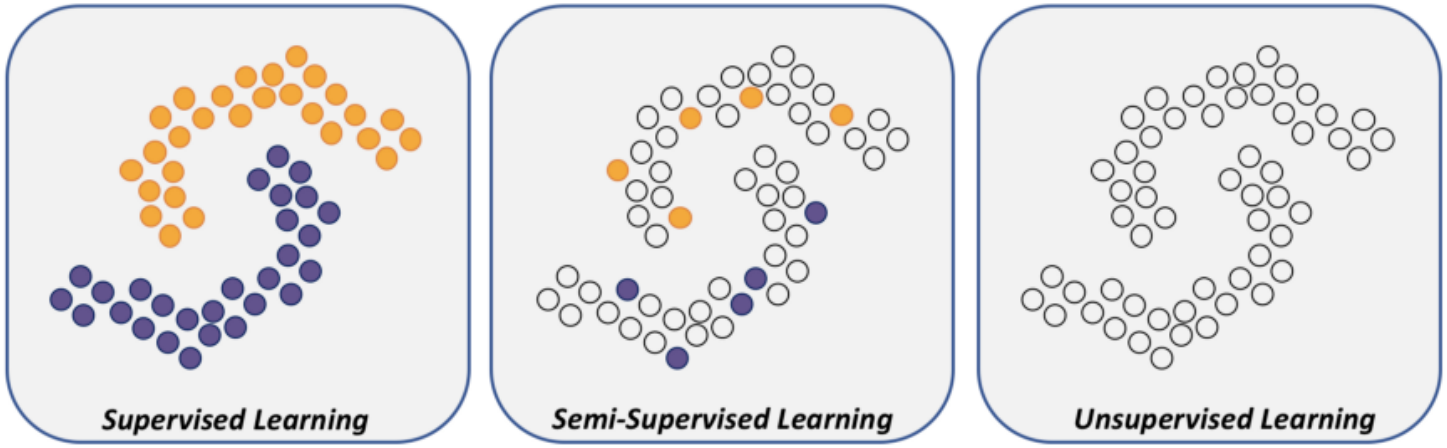
	花萼長度	花萼寬度	花瓣長度	花瓣寬度	類別
Train Data	5.1	3.5	1.4	0.2	山鳶尾
	4.9	3	1.4	0.2	山鳶尾
	4.7	3.2	1.3	0.2	山鳶尾
	7	3.2	4.7	1.4	變色鳶尾
	6.4	3.2	4.5	1.5	變色鳶尾
	6.9	3.1	4.9	1.5	變色鳶尾
	6.3	3.3	6	2.5	維吉尼亞鳶尾
	5.8	2.7	5.1	1.9	維吉尼亞鳶尾
	7.1	3	5.9	2.1	維吉尼亞鳶尾
Test Data	5	2.3	3.3	1	?

機器學習的步驟



機器學習
方法種類
(資料面來區分)

- 監督式學習（Supervised learning）
- 非監督式學習（Un-supervised learning）
- 半監督式學習（Semi-supervised learning）
- 強化學習（Reinforcement learning）



監督式學習

- 每筆資料樣本必須包含以下兩個部分
  - 輸入特徵（Features）：模型用來做出預測的觀察或描述性信息。
  - 標籤（Labels）：目標變量，即模型需要預測的結果。

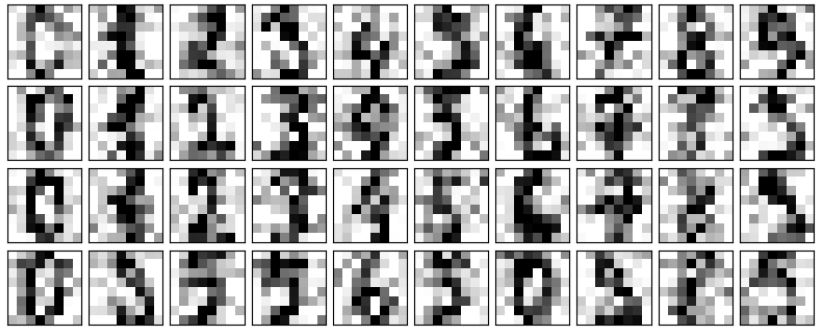
應用	特徵	標籤
信用風險評估	借貸記錄、收入水平、還款歷史等數據	是否有違約記錄
人才適任系統	性向測驗數據	考績
垃圾郵件檢測	電子郵件的內容	是否垃圾郵件
房價評估系統	房屋資訊	房屋價格
語言翻譯系統	被翻譯的語言文字	翻譯後的語言文字
醫療診斷	X光片	是否有肺炎

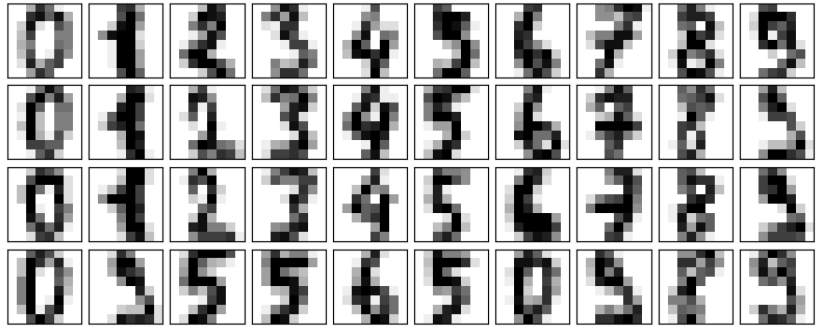
非監督式學習

- 模型在沒有任何標註的情況下進行訓練
- 特點：
  - 無需標註數據
  - 自動模式發現
  - 數據探索與特徵學習
- 主要方法
  - 聚類(Clustering)
    - K均值(K-Means)、層次聚類(Hierarchical Clustering)、DBSCAN
  - 降維(Dimensionality Reduction)
    - 主成分分析(PCA)、線性判別分析(LDA)、自編碼器(Autoencoders)
  - 關聯規則學習(Association Rule Learning)
    - Apriori、FP-Growth

非監督式學習應用

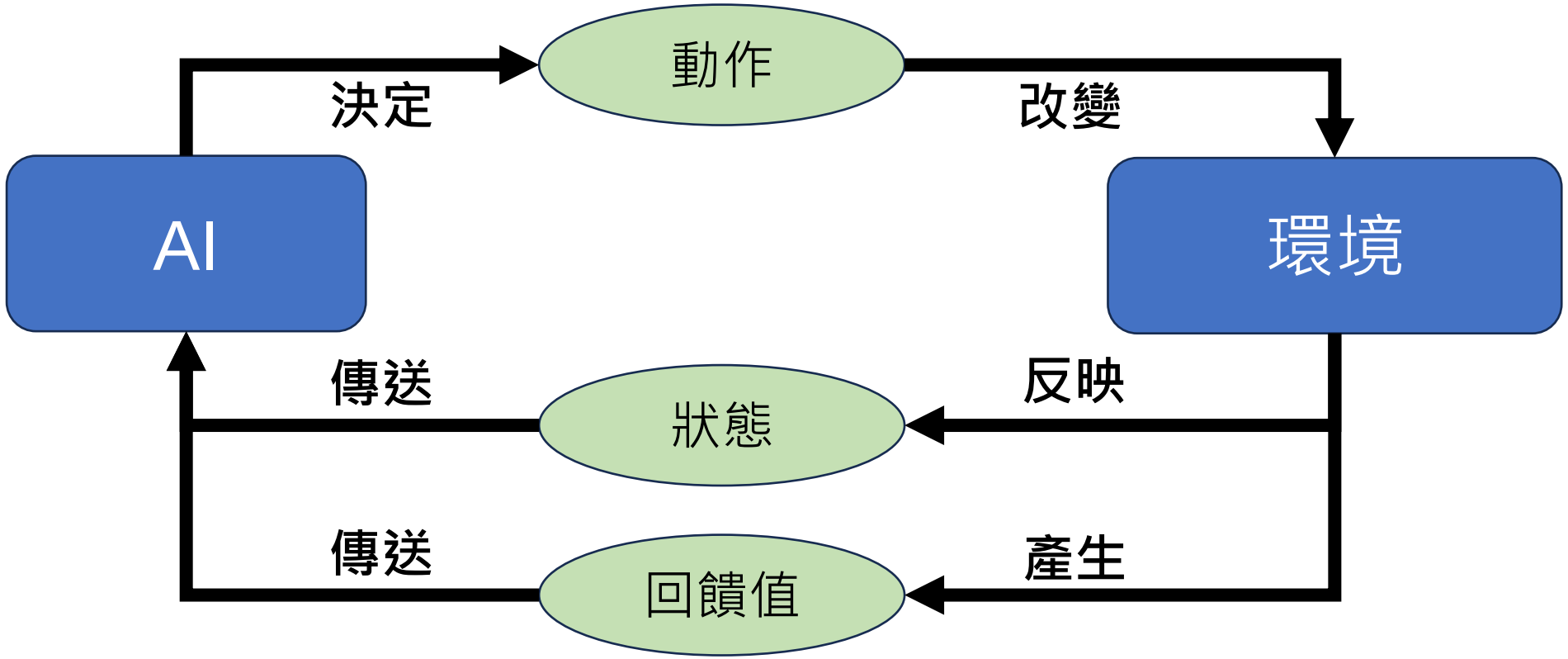
- 客戶細分
- 異常檢測
  - 信用卡欺詐檢測
  - 網絡入侵檢測
  - 偽鈔偵測
- 新聞分類
- 圖像壓縮



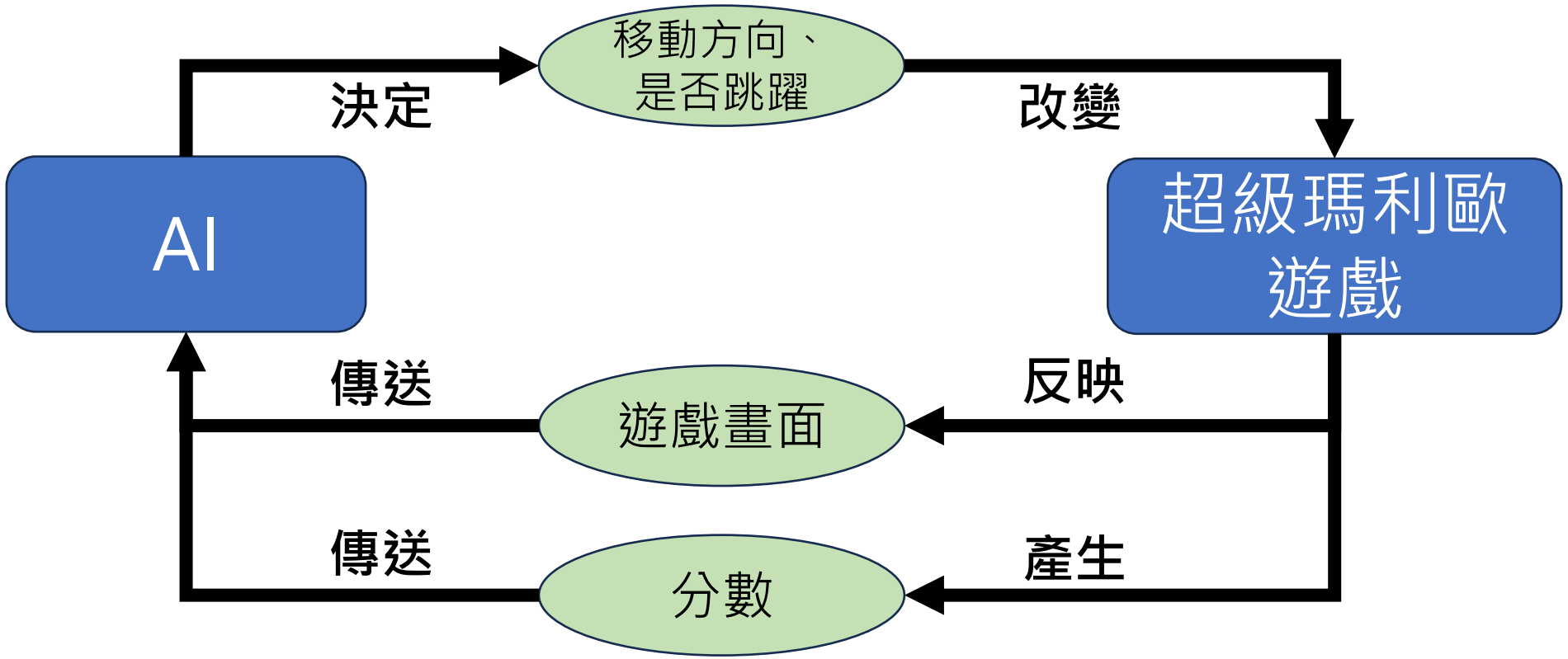


強化學習

- 模型通過與環境互動並學習如何在不同情況下做出決策，以最大化其累積獎勵



強化學習_以遊戲為範例



強化學習

優點

- 自主學習
- 適應性強
- 無需大量標註數據

缺點

- 訓練時間長
- 高計算成本
- 不穩定性和收斂問題
- 缺乏解釋性
- 環境設計的挑戰

RLHF(Reinforcement Learning from Human Feedback,帶有人類反饋的強化學習)

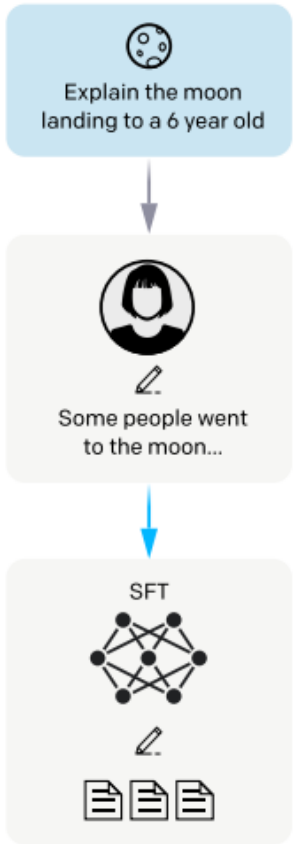
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



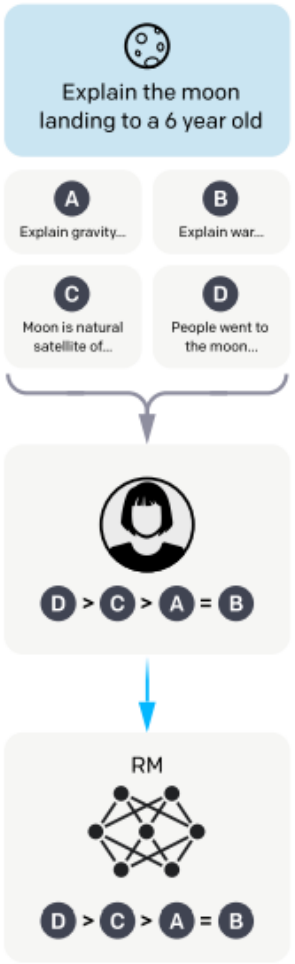
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

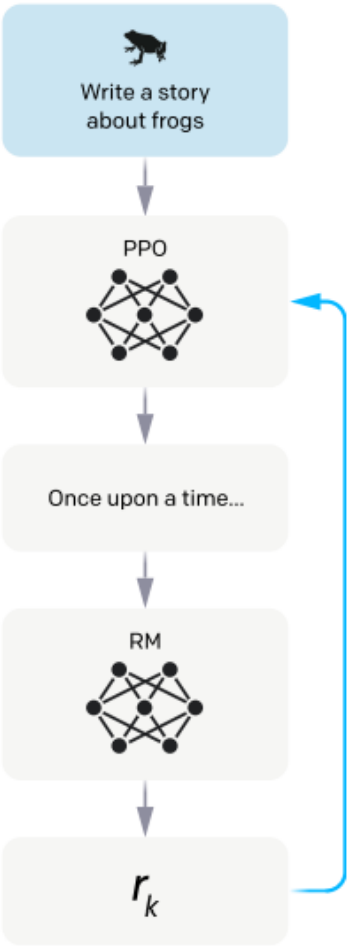
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



強化學習的應用

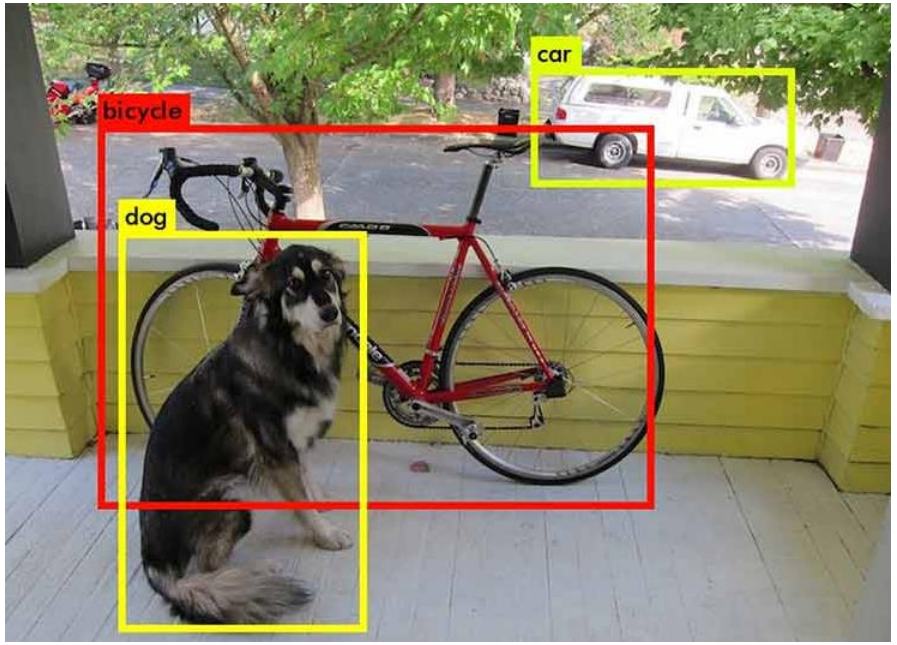
- 遊戲
- 機器人控制
- 自動駕駛
- 金融投
- 推薦系統
- 自然語言處理

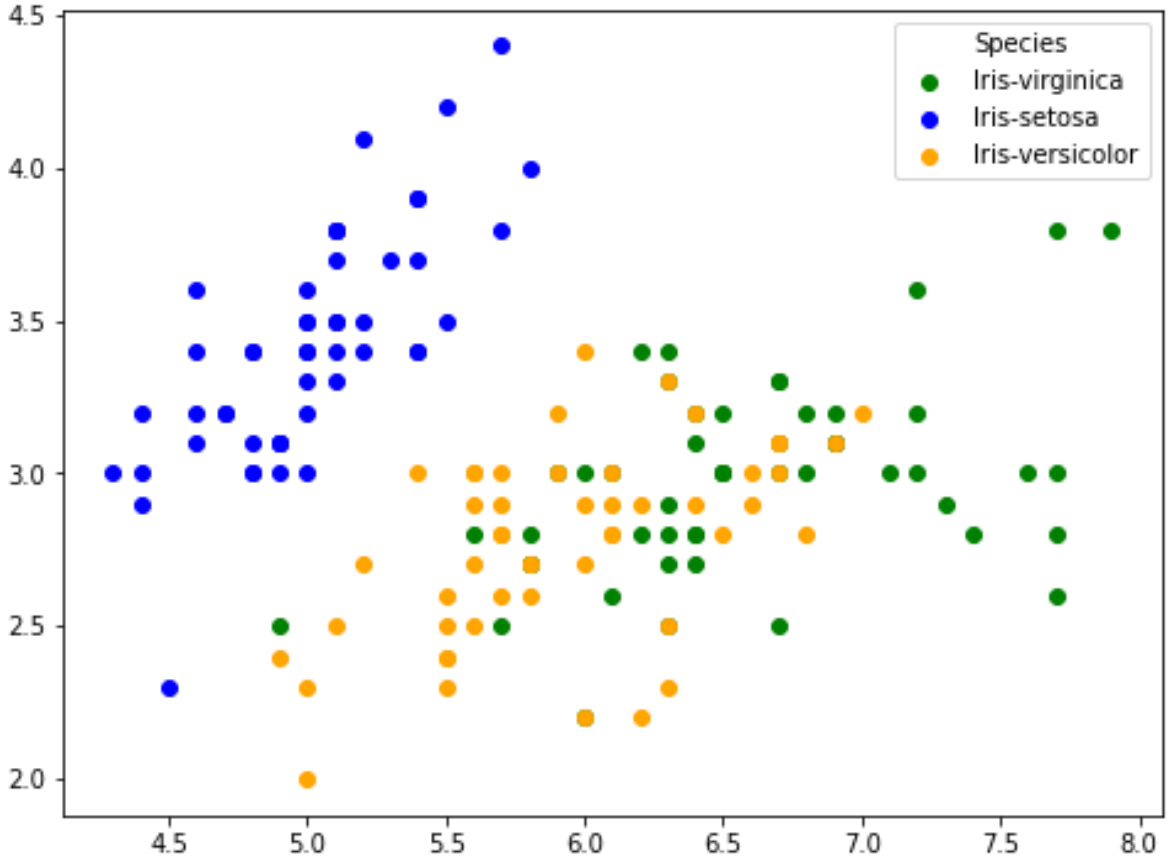
機器學習
問題種類
(預測面來區分)

- 分類 (classification)
- 聚類 (clustering)
- 回歸 (regression)
- 特徵處理
  (feature processing)
- 異常檢測
  (anomaly detection)
- 關聯分析
  (associative snalysis)

分類

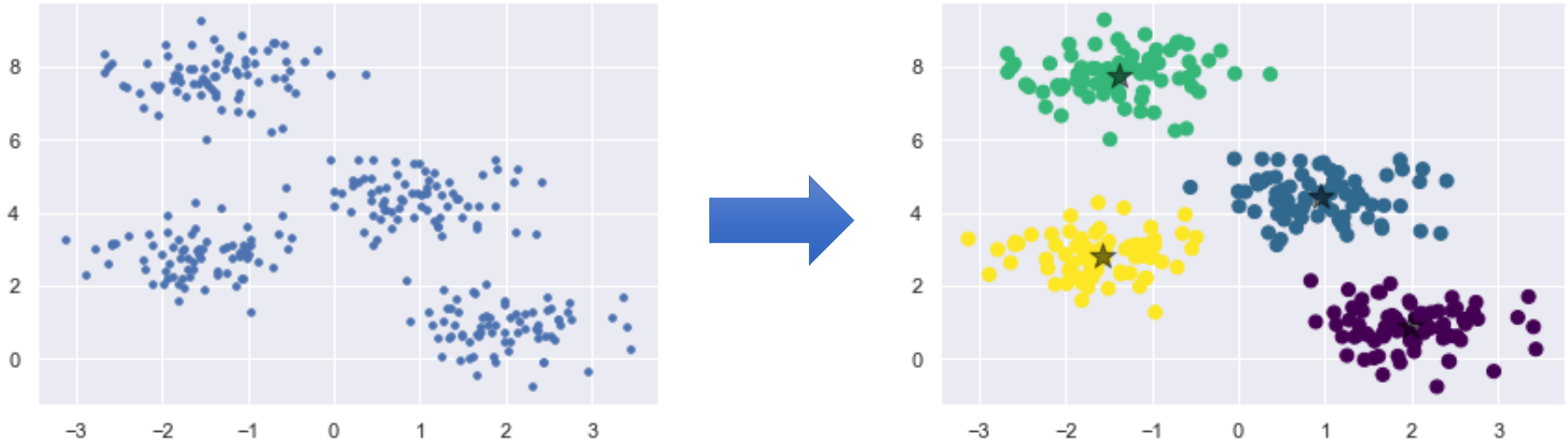
- 預測兩種以上的類別資料
  - 人臉辨識
  - 鈔票面額





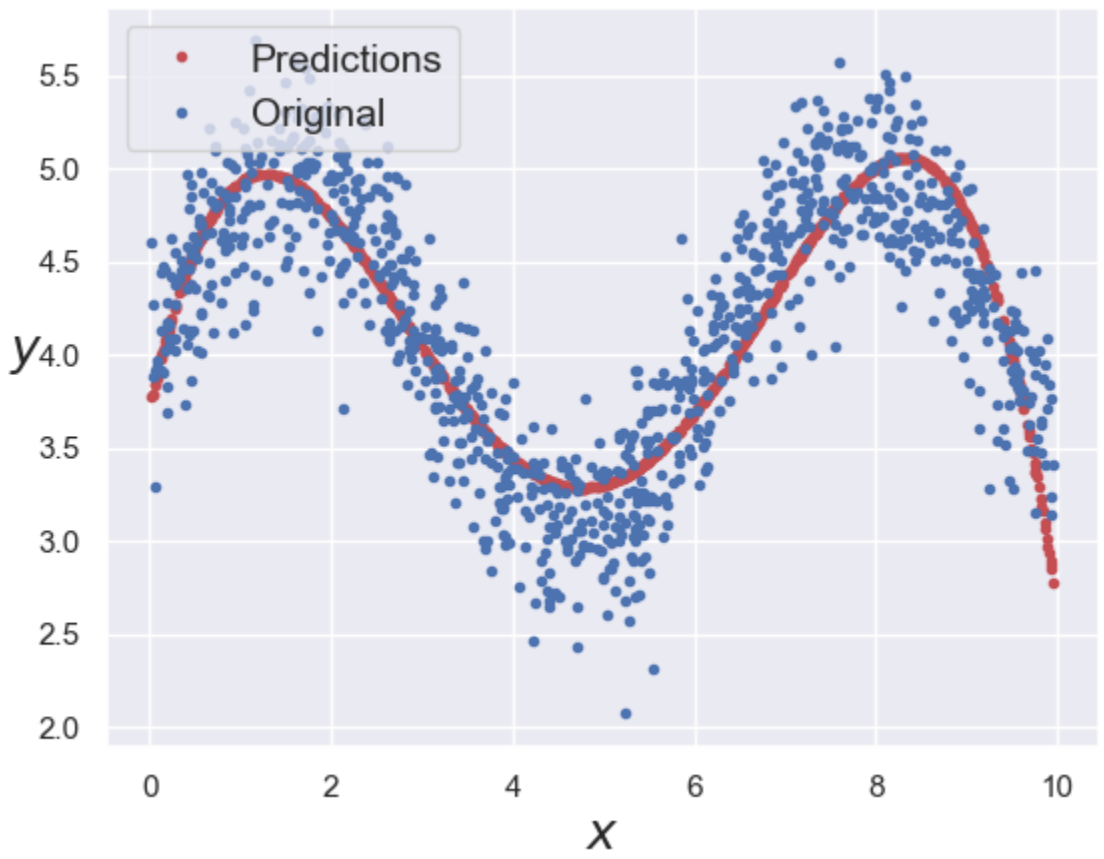
聚類

- 分類的非監督式學習版本
  - 將沒有類別標記的資料進行分類
  - 文件分類



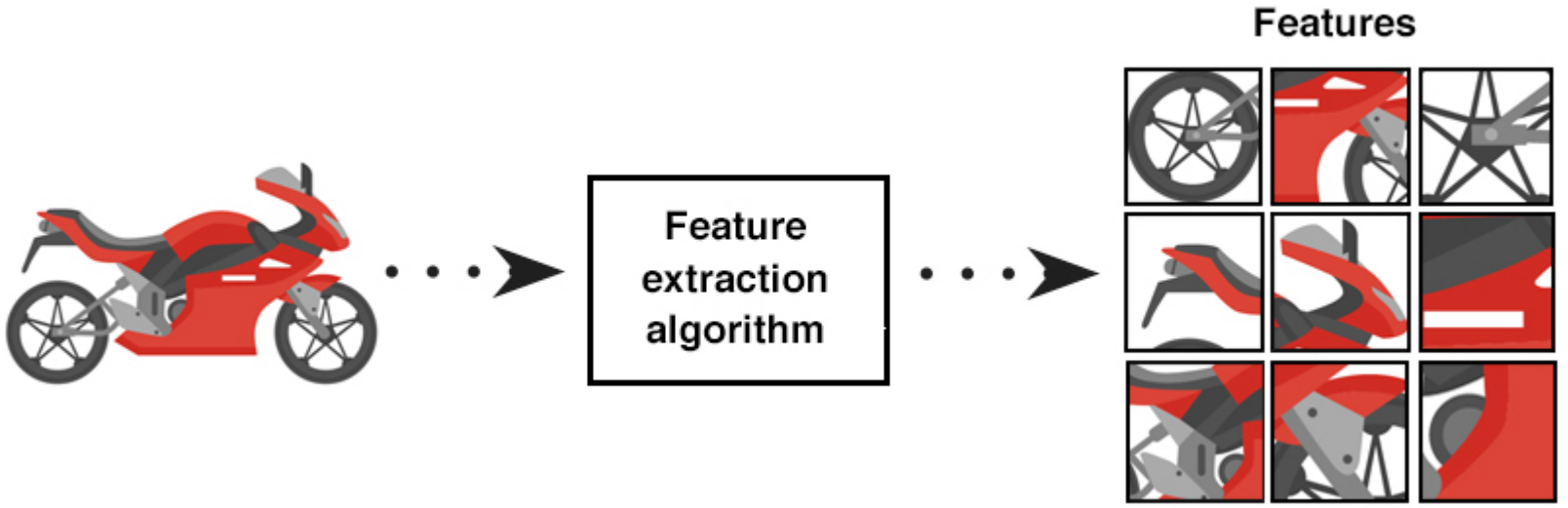
回歸

- 預測的目標為連續性數值
  - 房價、人才適任度、氣象



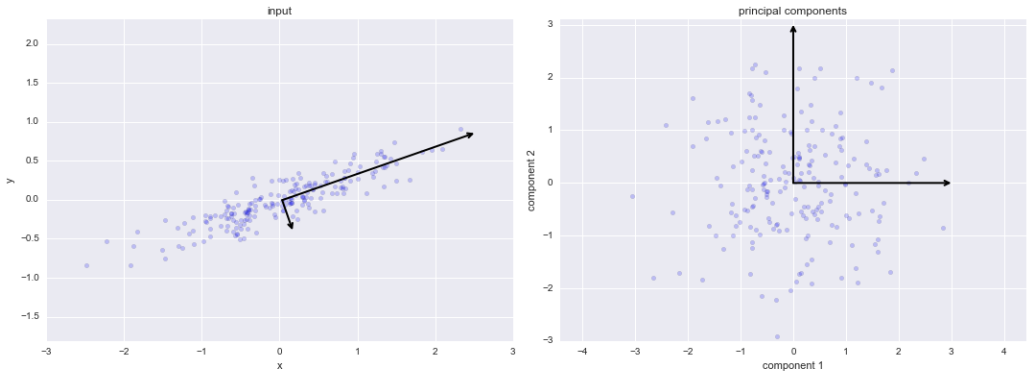
特徵處理

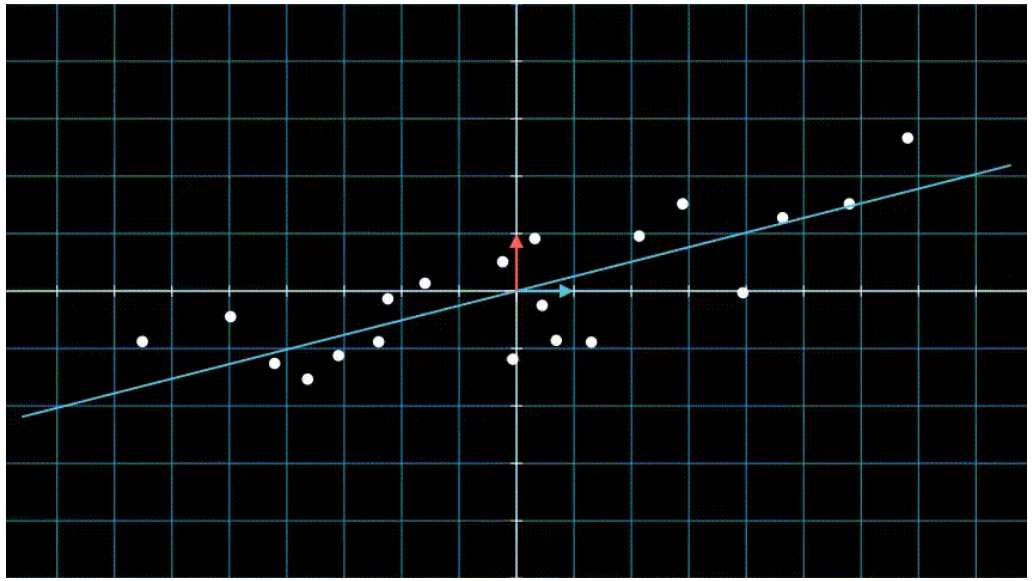
- 將原始資料的欄位(維度)轉化成更具有意義或鑑別力的形式
- 縮減資料的維度(欄位數目)



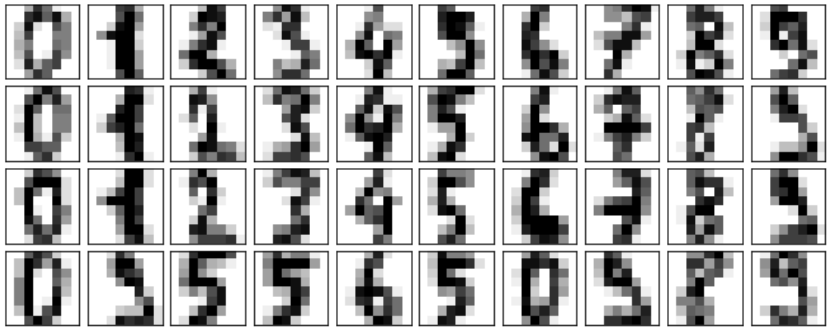
圖片來源: https://www.educative.io/answers/what-is-feature-extraction

PCA

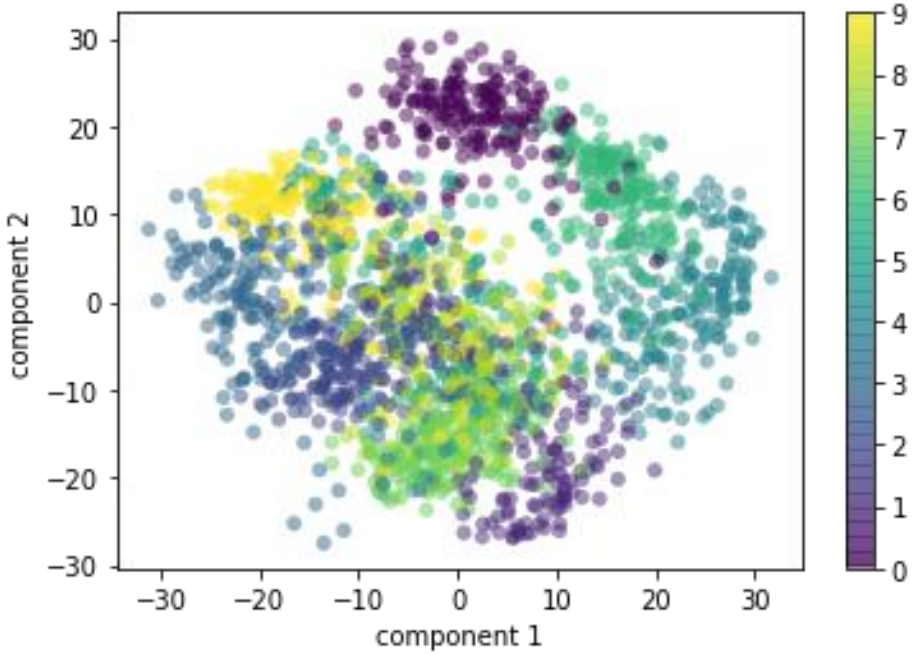




64(8*8) 維的手寫文字圖形



PCA轉化後只取前面兩維

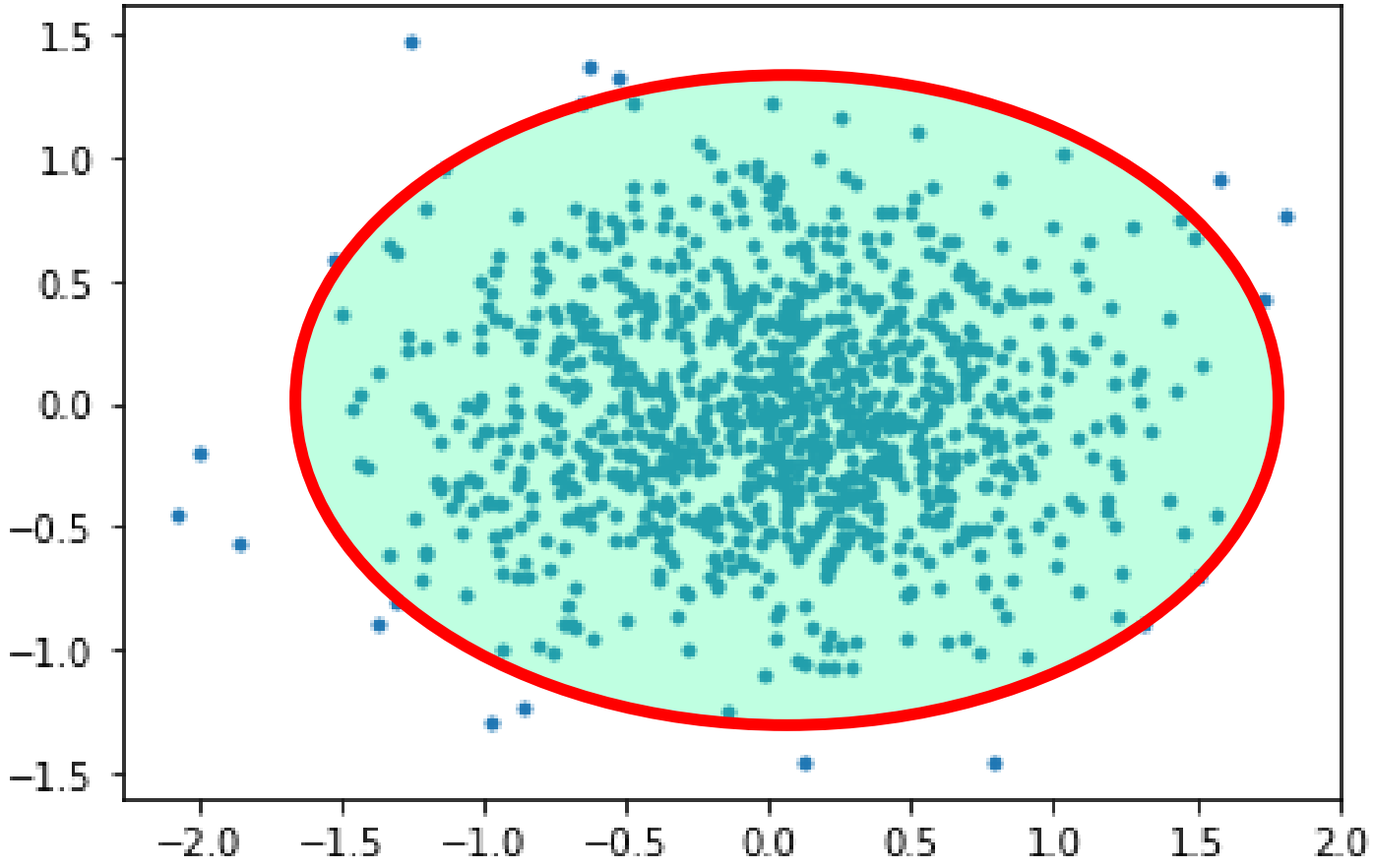


特徵選擇 Feature Selection

- 去除不重要的欄位
  - 降低系統複雜度
  - 可解釋性
  - 避免維度災難
- 演算法
  - StepWise
  - AMFES

異常檢測 (anomaly detection)

- 檢測規律性系統中的異常
- 收集正常值的資料
- 應用
  - 機台數據監控
  - 偽鈔辨識
  - 網路攻擊
  - 員工出勤



關聯分析 (associative snalysis)

- 計算項目間的關聯性
- 應用
  - 購物推薦
  - 職缺推薦

機器學習

優點

- 數據驅動
- 自主學習
  - 持續改進
  - 自動化和效率提升
  - 個性化服務
  - 實時反應
- 降低錯誤率
- 眾多演算法

缺點

- 需要大量資料
- 黑盒子

機器學習應用

表格式資料 預測模型

特徵欄位(Features)　　標籤欄位(Label)

花萼長度	花萼寬度	花瓣長度	花瓣寬度	類別
5.1	3.5	1.4	0.2	山鳶尾
4.9	3	1.4	0.2	山鳶尾
4.7	3.2	1.3	0.2	山鳶尾
7	3.2	4.7	1.4	變色鳶尾
6.4	3.2	4.5	1.5	變色鳶尾
6.9	3.1	4.9	1.5	變色鳶尾
6.3	3.3	6	2.5	維吉尼亞鳶尾
5.8	2.7	5.1	1.9	維吉尼亞鳶尾
7.1	3	5.9	2.1	維吉尼亞鳶尾

系統名稱	特徵欄位(Features)	標籤欄位(Label)
人才適任系統	性向測驗向度：如領導性、情緒調適、謹慎性、社交性...	員工過去績效
房價評估系統	房屋坪數、樓層、地區、樓層、房間數目、公設、是否有公園、學校、電梯....	過去成交價格
信用評估	年齡、性別、工作類型、收入水平、債務水平、是否有違約記錄	過去核准的貸款額度
急診流感檢測 https://doi.org/10.1016/j.bj.2022.09.002	年齡、性別、體重、BMI、體溫、脈搏率、呼吸頻率、氧飽和度、病程天數、流感季節期間訪問醫院、頭痛、咳嗽、喉嚨痛、身體疼痛、呼吸急促、噁心、腹瀉...	是否感染流感
客戶流失預測	購買記錄、服務使用頻率、客戶服務互動	客戶是否會停止使用服務或產品

異常偵測 (Anomaly Detection)

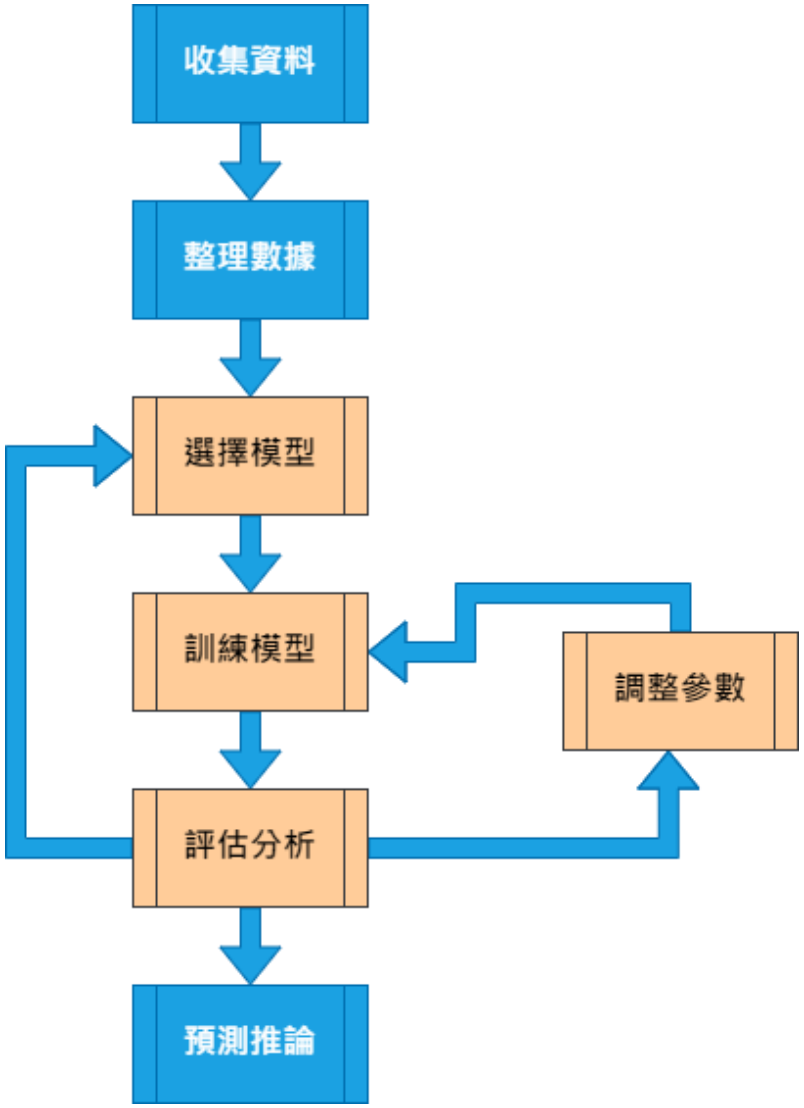
系統名稱	特徵欄位(Features)
信用卡欺詐檢測	過去交易紀錄：如交易頻率、交易金額、交易地點、商店類型...
網絡入侵檢測	過去網路紀錄：IP位址、協議類型、流量大小、持續時間、封包數量
機器故障預測和維護	設備的感測器數據：溫度、壓力、震動...
偽鈔辨識	真鈔光學感應訊號
零售業損失預防	交易日期和時間、交易金額、付款方式、交易類型、優惠券使用、庫存變化、存貨周轉率、商品損耗報告、員工折扣使用、註銷和退貨操作...

表格式資料預測模型導入需求

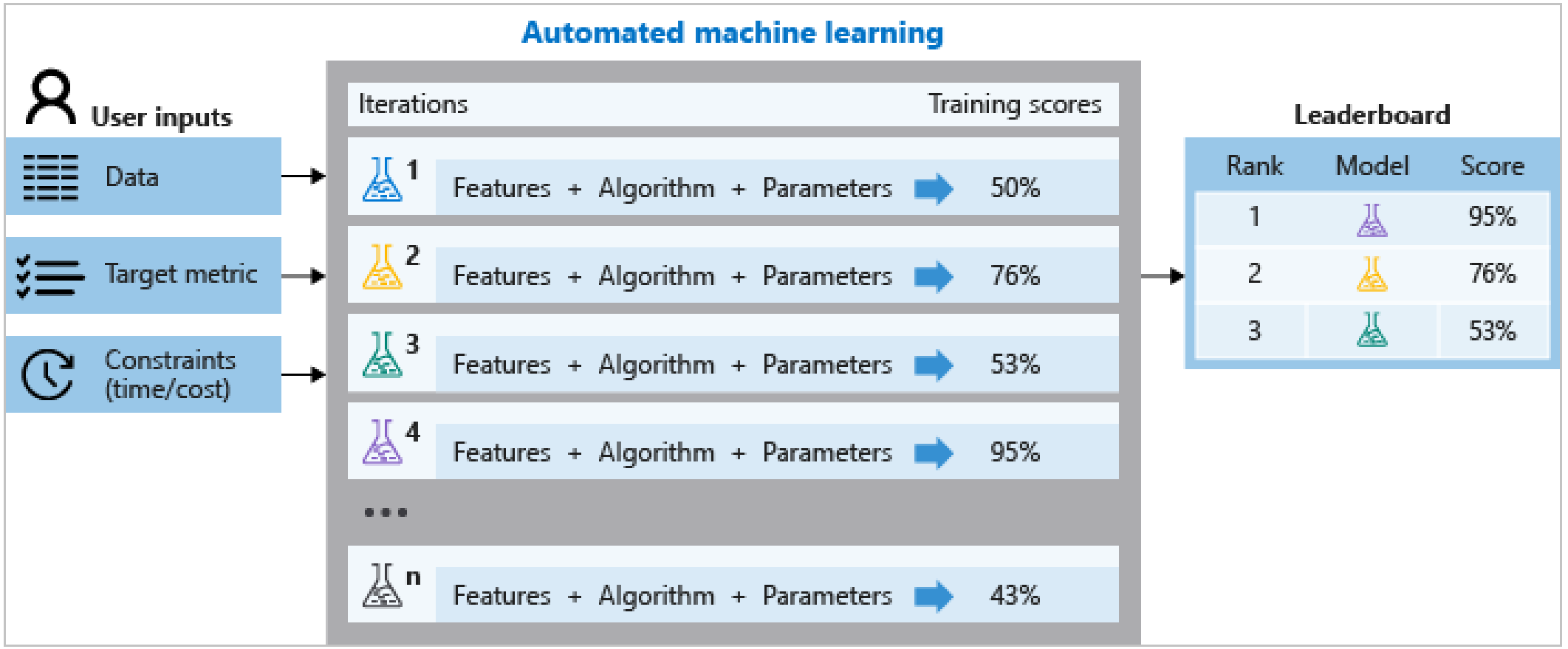
- 設備要求相對低，只需要CPU
- 訓練資料
  - 至少要有數百筆以上
  - 特徵欄位(Features)必須跟標籤欄位(Label)有關連性
  - 資料需要多樣化
- 開發工具
  - 程式：Python, scikit-learn
  - 無程式碼(No Code)：AutoML

AutoML

- 簡化和加速機器學習模型的構建、優化和部署過程
- 自動建立預測模型
  - 數據前處理
  - 模型選擇與參數設定
  - 模型效能評估
  - 特徵重要性評估
  - 模型部署



AutoML



source:https://learn.microsoft.com/zh-tw/azure/machine-learning/concept-automated-ml

AutoML產品

- Azure Machine Learning
  - https://azure.microsoft.com/zh-tw/solutions/automated-machine-learning
- AWS AutoML
  - https://aws.amazon.com/tw/machine-learning/automl/
- Google AutoML
  - https://cloud.google.com/automl
- MoBagel Decanter AI
  - https://mobagel.com/tw/decanter-ai/
- DataRobot AI Platform
  - https://www.datarobot.com/platform/

資料收集與前處理需要注意的事情

資料收集

- 增加數據的欄位
- 增加數據量
- 增加數據的廣度(多樣性)
- 減少資料的缺失值
- 減少離異資料

資料收集常見的問題

- 重要的欄位未收入變量中
- 欄位的值沒有變化
- 人才募集中薪資應該是重要因素
- 但是公司內的員工薪資變化不大，所以收集的資料中薪資欄位會顯示不出重要性
- 欄位值需要進行有意義的轉化
- 原始欄位 進入公司的日期
- 轉化為 進入公司年數

資料前處理

- 轉換成有意義的數值
- 缺值補齊
- 排除異常資料
- 特徵選擇

轉換成有意義的數值

將日期轉換成數值

- 出生日期 --> 年齡
- 到職日期 --> 年資

將類別型欄位量化

- 學歷 --> 數值
- 職稱 --> 職等

缺值處理

移除

- 移除缺值的整筆資料
- 移除缺值的欄位

手動填值

- 平均值
- 中位數
- 出現頻率最高的值

差值法

- 找出兩筆資料最接近的紀錄
- 填入兩筆資料的中間值

找出異常資料

Q1 = Quartile 25%

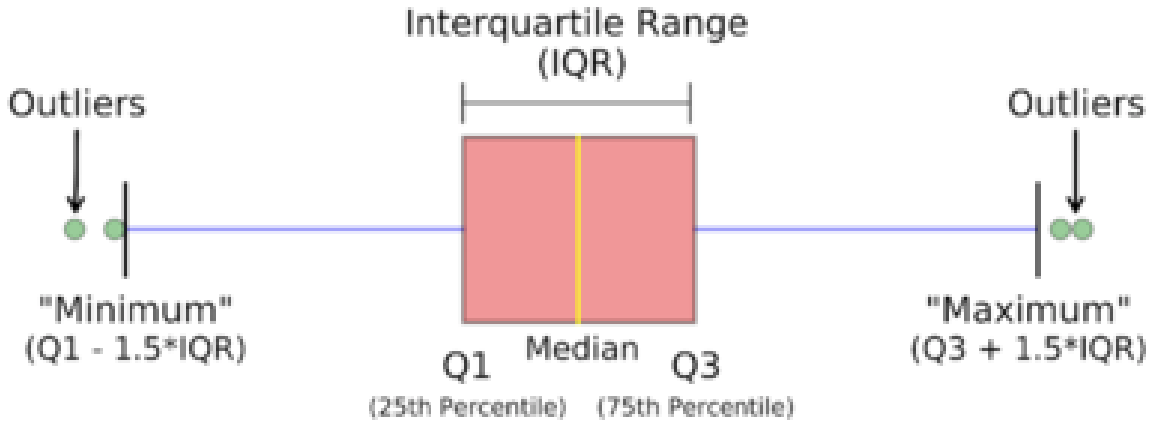
Q2 = Quartile 50% (中位數)

Q3 = Quartile 75%

IQR = Q3 - Q1

Outlier data =

(Q1 - 1.5 IQR) U (Q3 + 1.5 IQR)



特徵選擇

人工

- 根據經驗篩選有價值的欄位

特徵選取演算法

- PCA
- Feature Selection
  - Step Wise
  - AMFES
- Convolution (影像)
- AutoEncoder

END