【硬塞科技字典】什麼是資料探勘（Data Mining）？

[Hsinlan Chen](https://www.inside.com.tw/author/hsinlanchen)

資料探勘（Data Mining），意指利用一個龐大數據庫建立模型（Model），並從中找出隱藏的特殊關聯性及特徵。例如：某公司握有自身客戶的資訊（包含：年齡、資產、交易頻率、交易量等），利用此資料庫找出其客戶消費的模式、習慣，並據此將客戶群分類，藉此針對不同客群做出精準行銷，就是所謂的資料探勘。

資料探勘所建立的模型可分為六種：

1. **分類分析（Classification）**：透過研究數據庫中的特徵，將已知資料做出分類，並根據已知的特徵預測未經分類的新進數據。如前段所述的客戶群分類。
2. **群集分析（Clustering）**：和分類分析的概念相似，亦是將一數據庫的資料做出分類，並歸納出組間的差異性及組中的相似性。其不同點在於，分類分析在劃分後又有明確對應的類別／函數（換句話說，數據間有「已知」的特徵），但群集分析在演算法運算時無法得知分類的依據及數據的特徵，也就是分類後並沒有明確的類別／函數，因此必須在分類後另行解讀各個分類的意義。
3. **迴歸分析（Regression）**：透過一系列的現有數據去預測未知數據的可能值。例如：我們可以透過分析某一地區多筆房地產交易的數據庫（包含：坪數、地點、房型、交易金額等），對另一個未售出的房產做出成交金額預測。
4. **時間序列分析（Time Series Forecasting）**：和迴歸分析的概念相同，也是藉由已知的數據來預測未來數據的可能值。其不同點在於：時間序列分析模型中的數據中必須含有時間關聯性。透過時間序列分析，可得知事件沿著時間軸（如季節性、節日、過去與未來的相關性）所產生的變化情形，進一步使用歷史資料來預測未來趨勢。過去，時間序列分析通常用於一國家／地區的經濟發展政策，目前已進一步地延伸到醫療資料分析。如：糖尿病遠端醫療照護。
5. **關聯分析（Association）**： 是分析數據庫中各資料彼此相依的機率，通常被用來分析公司各產品被同時購買的關係與頻率。例如：某顧客在已經購買該品牌洗髮精的情況下，同時購買該品牌潤髮乳的機率。
6. **順序型態分析（Sequential Pattern Analysis）**： 與關聯分析相似，只是順序型態分析中的數據中具有次序及時間的關係。例如：某顧客在已經購買該品牌洗髮乳後，一周後再度購買潤髮乳所發生的機率。順序型態分析很常用在股市分析上，例如：因應英國脫歐公投，某股票一天內下跌了 10％，另一檔股票在兩天內跟著下跌 10％ 的機率。