

Day 6

資料清理數據前處理

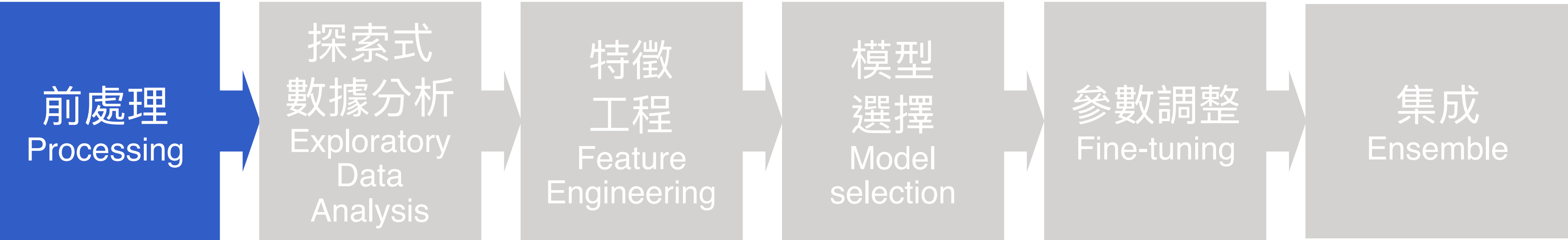
EDA與Outlier檢查



# 知識地圖 機器學習前處理 Outlier 及處理

## 機器學習概論 Introduction of Machine Learning

### 監督式學習 Supervised Learning



### 非監督式學習 Unsupervised Learning



### 前處理 Processing



# 本日知識點目標

---



目標  
知識點

了解什麼是例外值 (outlier)

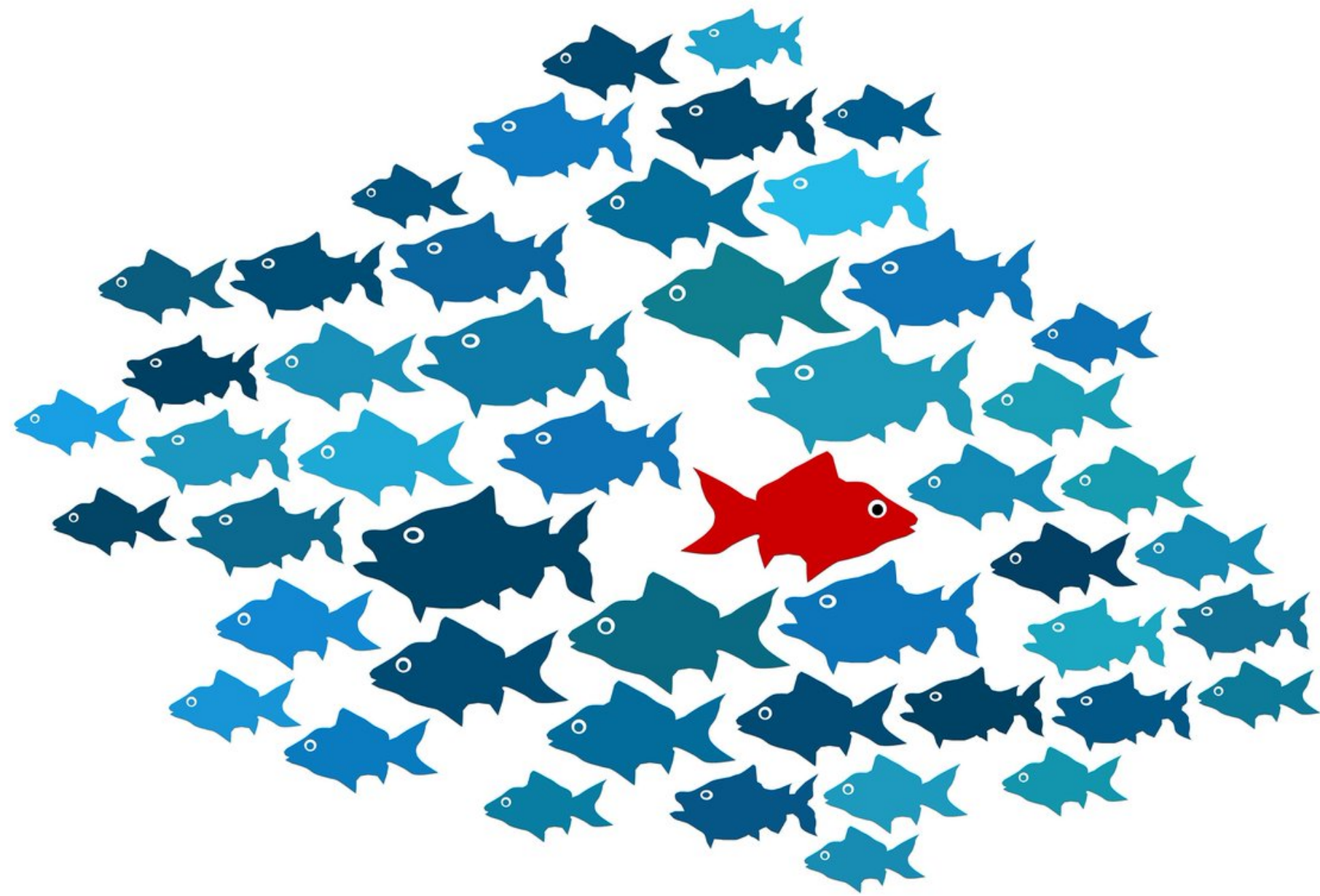


獲得  
知識點

完成今日課程後你應該可以了解

- 學會如何透過資料探勘方法找到例外值





圖片來源: [Sergio Santoyo](#)

## Dell電腦標價錯誤

	
Dell UltraSharp™ 2007FP 20" 液晶顯示器 高階平面顯示器含數位 DVI-D/類比/S-video/ Composite 輸入	Dell E2009W 20 吋寬螢幕平面顯示器
原價 NTD 13,200	原價 NTD 7,999
線上折扣 NTD 7,000	線上折扣 NTD 7,000
線上折後價 NTD 6,200	線上折後價 NTD 999
包括增值稅和運費	包括增值稅和運費
優惠	優惠
<a href="#">我要自選配備</a>	<a href="#">我要自選配備</a>

1

## 異常值 (Outliers) 出現的可能原因

1. 所以未知值，隨意填補 (約定俗成的代入)  
如年齡 = -1 或 999, 電話是 0900-123-456
2. 可能的錯誤紀錄/手誤/系統性錯誤  
如某本書在某筆訂單的銷售量 = 1000 本

2

## 檢查 Outliers 的流程與方法

- 盡可能確認每一個欄位的意義 (但有些競賽資料不會提供欄位意義)
- 透過檢查數值範圍 (五值、平均數及標準差) 或繪製散點圖 (scatter)、分布圖 (histogram) 或其他圖檢查是否有異常。

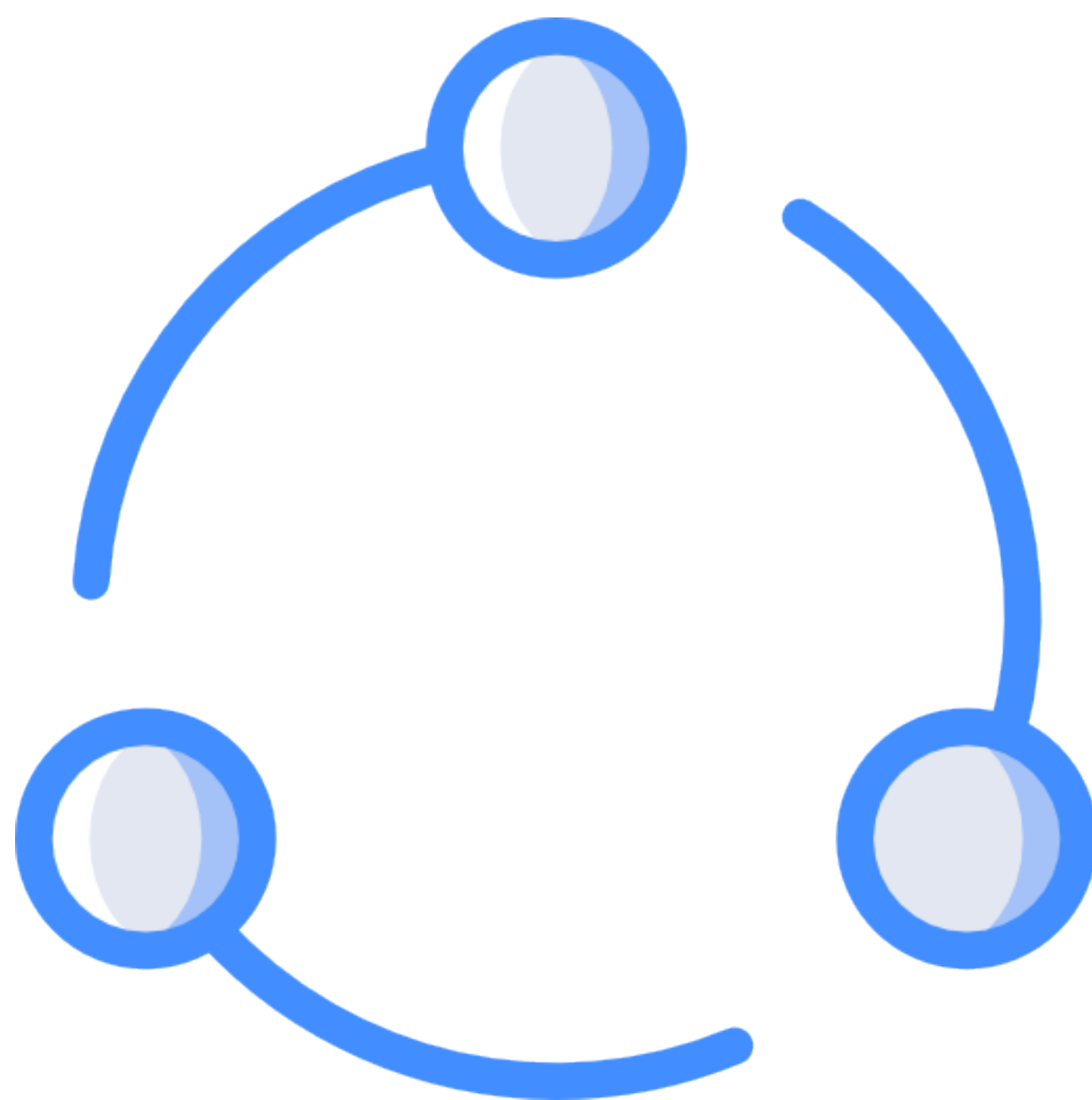
3

## 對 Outliers 的處理方法

- 新增欄位用以紀錄異常與否
- 填補 (取代)
- 視情況以中位數, Min, Max 或平均數填補(有時會用 NA)

# 重要知識點複習

---



- 檢查異常值的方法
  - 統計值：如平均數、標準差、中位數、分位數
  - 畫圖：如直方圖、盒圖、次數累積分布等
- 處理異常值
  - 取代補值：中位數、平均數等
  - 另建欄位
  - 整欄不用

# 解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

