

Day 17

探索式數據分析

把連續型變數離散化



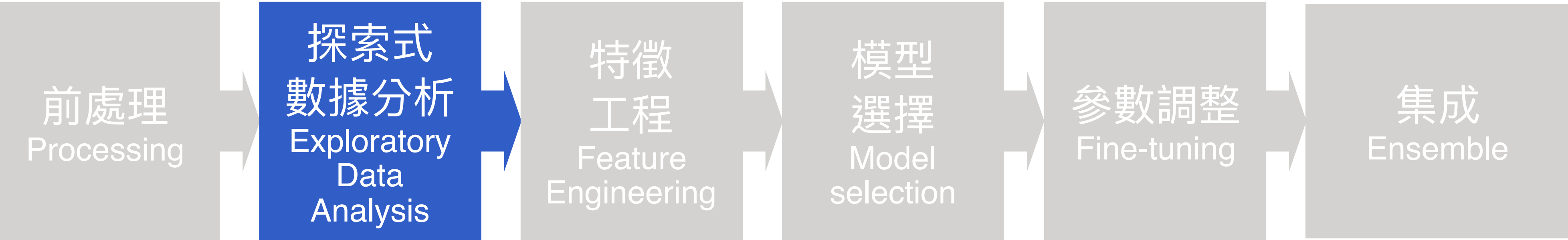
出題教練

游為翔 / 杜靖愷

# 知識地圖 探索式數據分析 離散化與EDA

## 機器學習概論 Introduction of Machine Learning

### 監督式學習 Supervised Learning

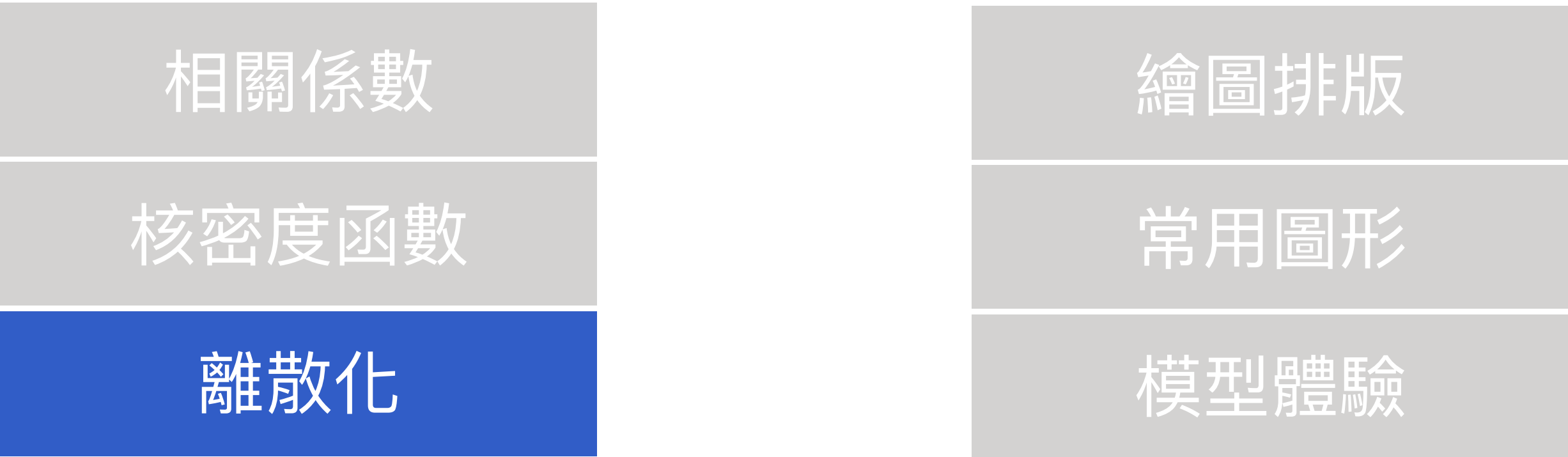


### 非監督式學習 Unsupervised Learning



## 探索式數據分析 Exploratory Data Analysis (EDA)

### 統計值的視覺化





# 本日知識點目標

了解離散化連續數值的意義以及方法

# 連續型變數離散化

## Goal

- 變得更簡單 (可能性變少了)
  - 假設年齡 0-99 (100 種可能性) >> 每 10 歲一組 (10 種可能性)
- 離散化的變數較穩定，假設年齡 > 30 是 1，否則 0。  
如果沒有離散化，outlier 「年齡 300 歲」 會給模型帶來很大的干擾。

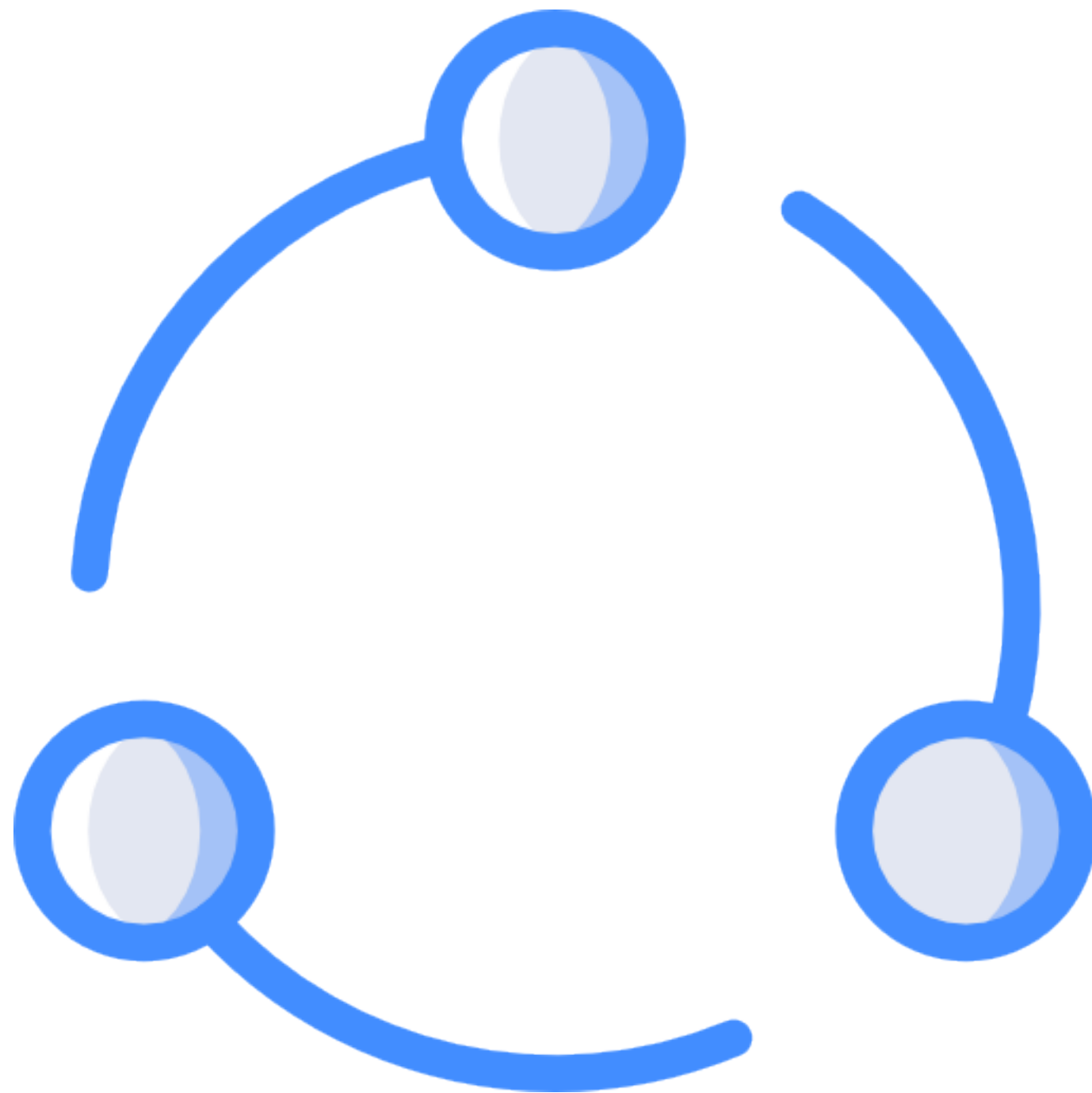
## 關鍵點

- 組的數量
  - 一樣以年齡為例子，每 10 歲一組就會有 10 組
- 組的寬度
  - 一組的寬度是 10 歲

## 主要的方法

- **等寬劃分**：按照相同寬度將資料分成幾等份。缺點是受到異常值的影響比較大。
- **等頻劃分**：將資料分成幾等份，每等份資料裡面的個數是一樣的。
- **聚類劃分**：使用聚類演算法將資料聚成幾類，每一個類為一個劃分。

除了以上的主要方法，也會因需求而需要自己定義離散化的方式，如何離散化是一門學問！



- 離散化的目的是讓事情變簡單、減少 outlier 對分析以及訓練模型的影響
- 主要的方法是等寬劃分 (對應 pandas 中的 cut) 以及等頻劃分 (對應 pandas 中的 qcut)
- 可以依實際需求來自己定義離散化的方式

連續特徵的離散化：在什麼情況下可以獲得更好的效果(知乎)

## 網頁連結

- 這個網頁是個討論串，經由幾個網友的討論與補充，很好地說明了離散化的理由：儲存空間小，計算快，降低異常干擾與過擬合(overfitting)的風險，主要想請同學參考第1位的回答，至於其他的討論則請同學參考即可。

# 解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

