

Day 24 特徵工程

# 類別型特徵 - 其他進階處理



出題教練

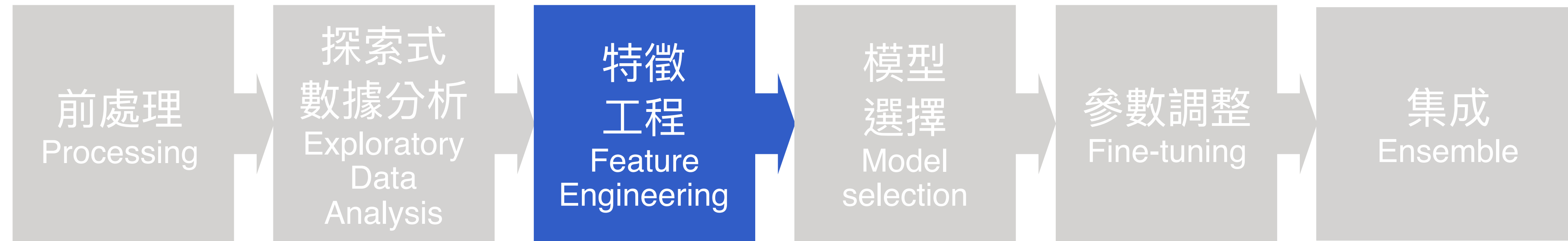
陳明佑



# 知識地圖 特徵工程 類別型特徵 - 其他進階處理

## 機器學習概論 Introduction of Machine Learning

### 監督式學習 Supervised Learning

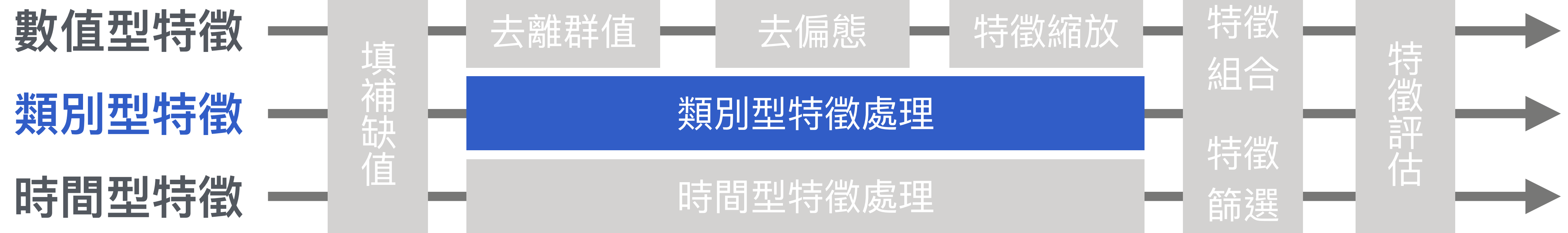


### 非監督式學習 Unsupervised Learning



## 特徵工程 Feature Engineering

### 概論





# 本日知識點目標

- 什麼是計數編碼，在什麼條件下可以考慮使用
- 雜湊編碼在什麼情況下可以考慮使用

# 計數編碼 ( Counting )

如果類別的目標均價與類別筆數呈正相關 ( 或負相關 )，也可以將筆數本身當成特徵  
例如：購物網站的消費金額預測

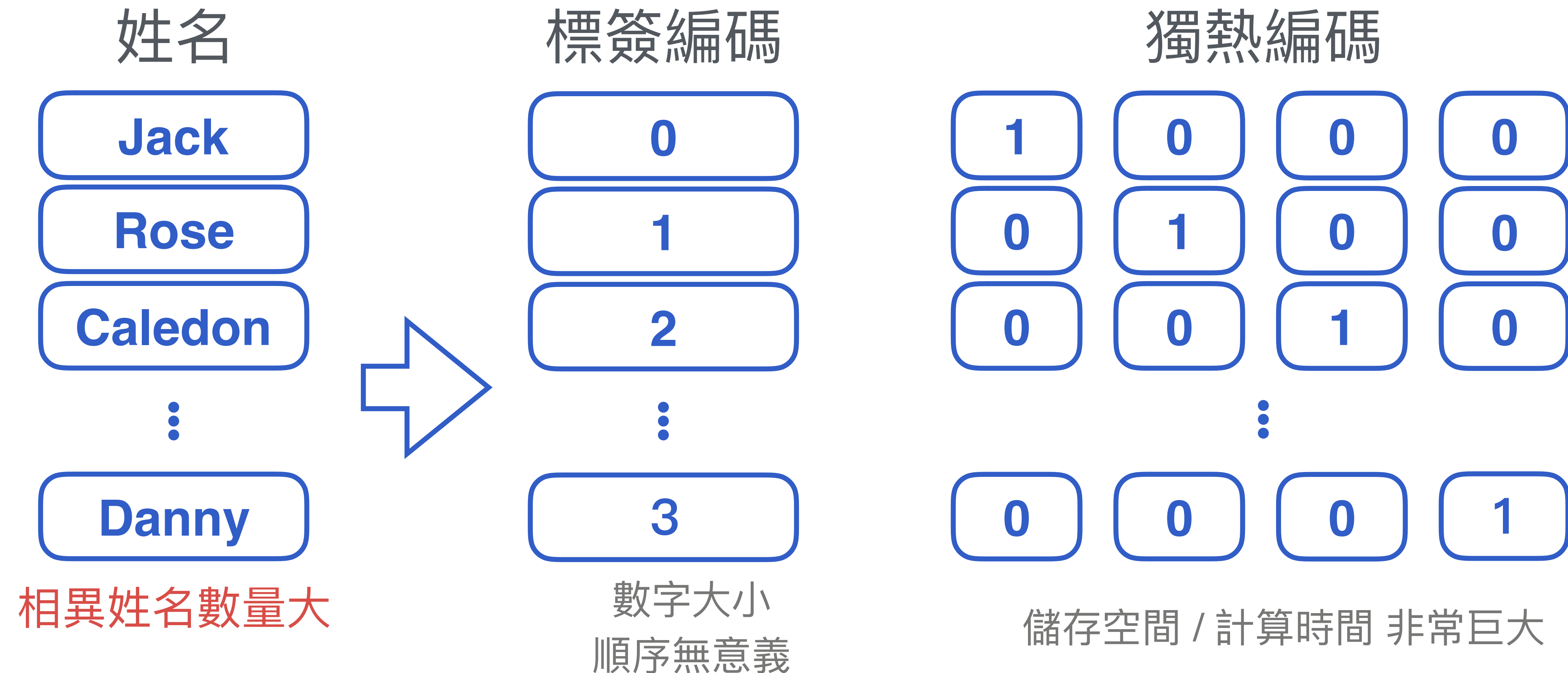


\*自然語言處理時，字詞的計數編碼又稱詞頻，本身就是一個很重要的特徵

# 特徵雜湊 ( Feature Hash ) ( 1 / 2 )

類別型特徵最麻煩的問題：相異類別的數量非常龐大, 該如何編碼？

\*舉例：鐵達尼生存預測的旅客姓名

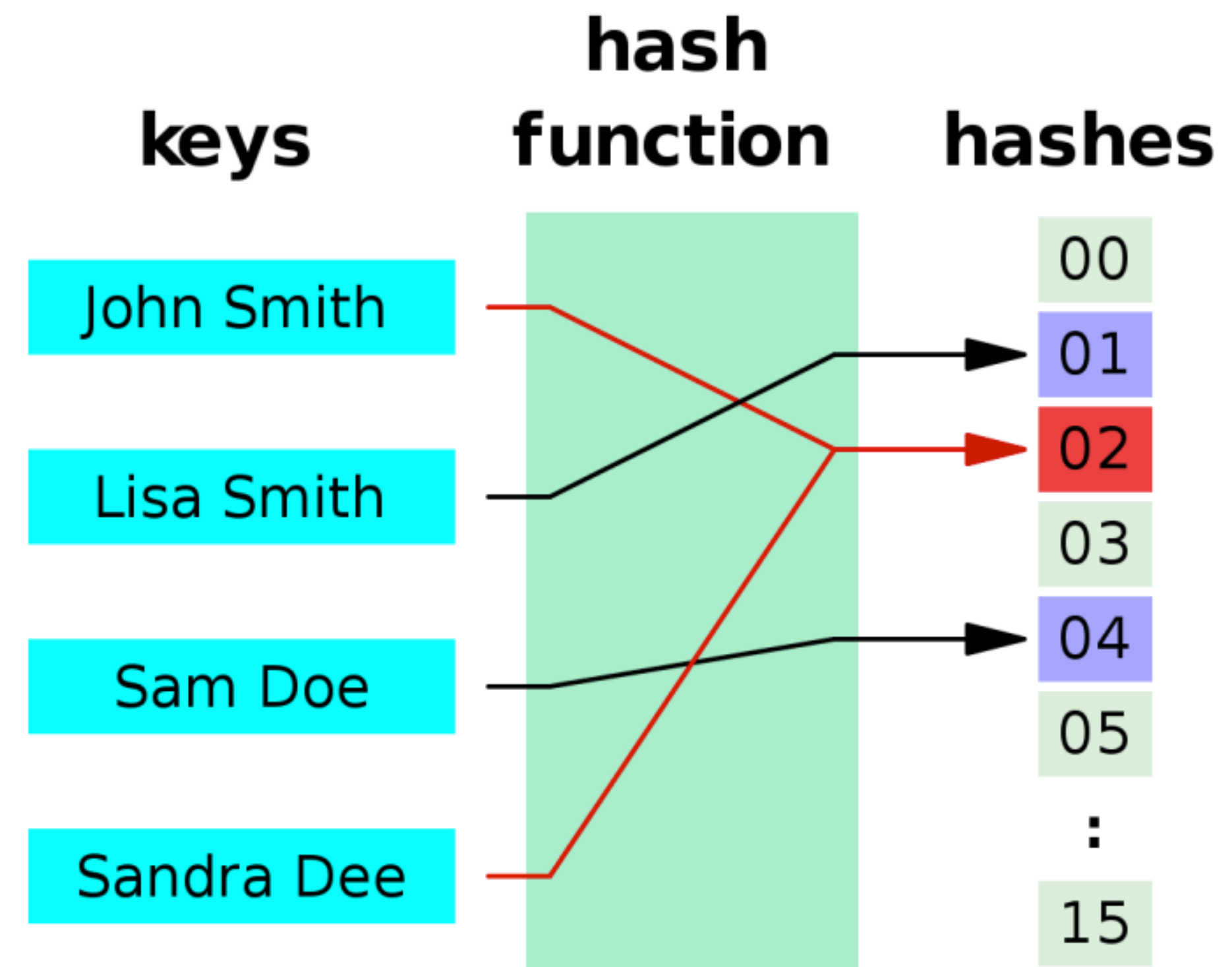


# 特徵雜湊 ( Feature Hash ) ( 2 / 2 )

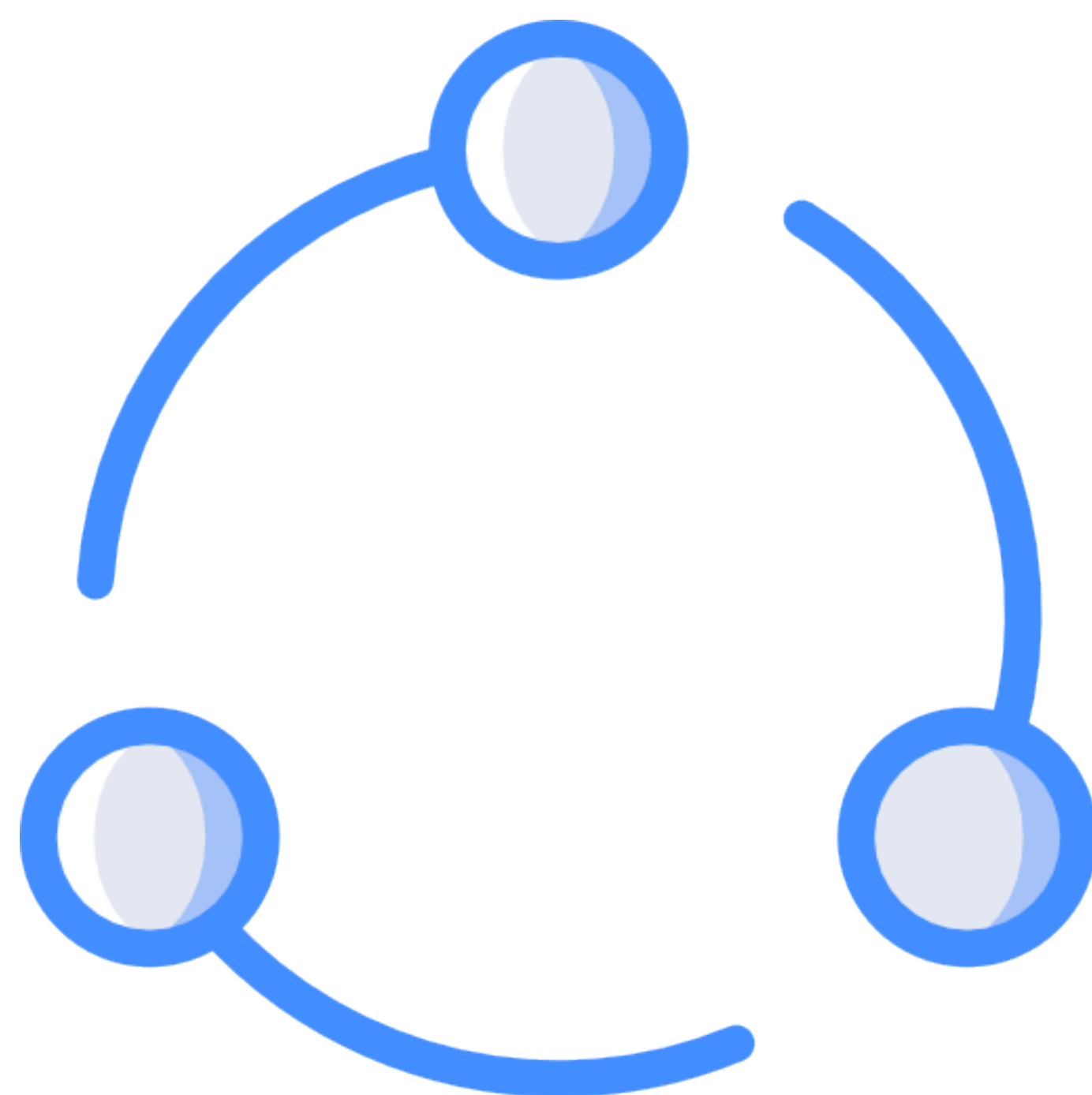
這個問題沒有很好的通用解法...只能採折衷方案或個別情況解決

## 特徵雜湊

- 特徵雜湊是一種折衷方案
- 將類別由雜湊函數定應到一組數字
- 調整雜湊函數對應值的數量
- 在計算空間/時間與鑑別度間取折衷
- 也提高了訊息密度, 減少無用的標籤



圖片來源：維基百科 [https://en.wikipedia.org/wiki/Hash\\_function](https://en.wikipedia.org/wiki/Hash_function)



- 計數編碼是計算類別在資料中的**出現次數**，當**目標平均值**與**類別筆數**呈正/負相關時，可以考慮使用
- 當**相異類別數量**相當**大**時，其他編碼方式效果更差，可以考慮雜湊編碼以節省時間

\*註：雜湊編碼效果也不佳，這類問題更好的解法是嵌入式編碼 (Embedding)，但是需要深度學習並有其前提，因此這裡暫時不排入課程



# 解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

