

Day 4

資料清理數據前處理

# EDA資料類型介紹



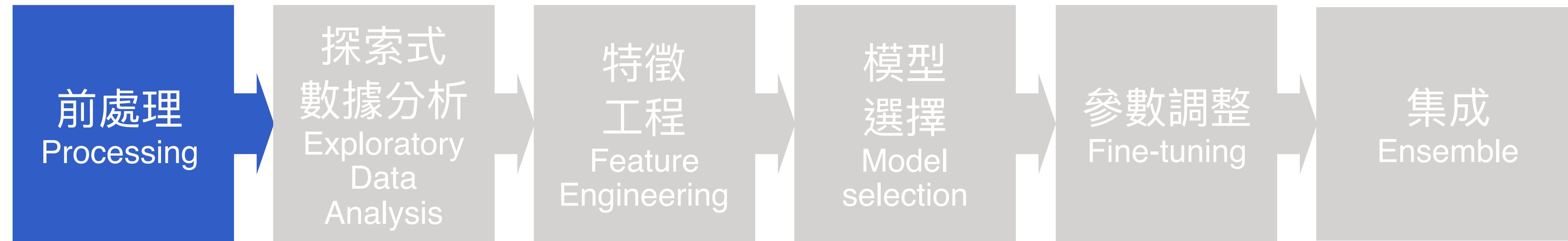
出題教練

游為翔 / 杜靖愷

# 知識地圖 機器學習前處理 欄位的資料類型介紹及處理

## 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

了解 pandas dataframe 欄位的基本資料類型

## 資料的欄位變數一般可分

- 離散變數: 只能用整數單位計算的變數
  - 房子的房間數量、性別、國家
- 連續變數: 在一定區間內可以任意取值的變數
  - Ex: 測量的身高、飛機起飛到降落所花費的時間、車速

當然還有日期、boolean 等等不同的格式，實務中遇到再 google 就好

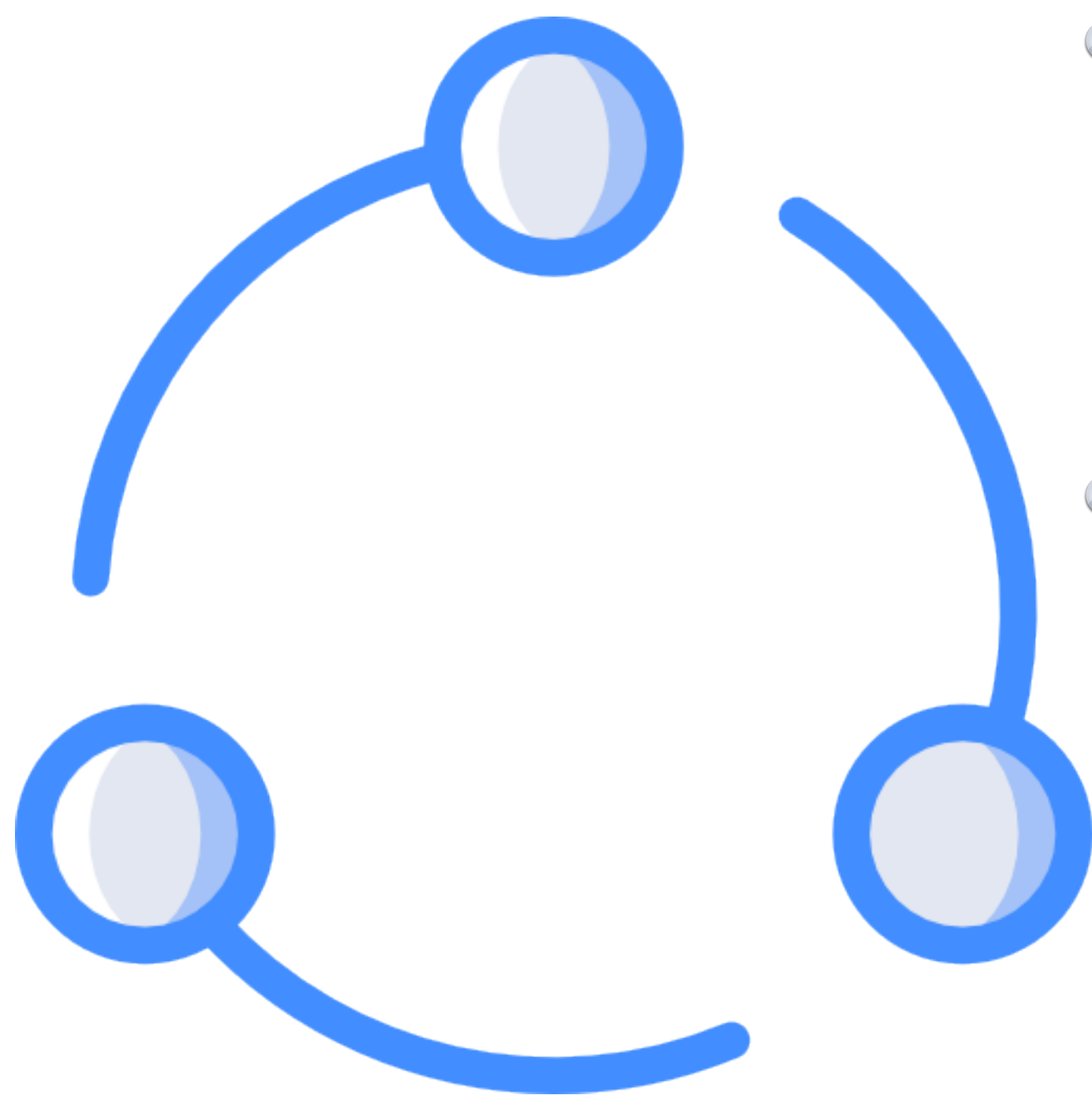
01

02

## Pandas DataFrame中 最常見的欄位資料類型有三種

- float64 : 浮點數，可表示離散或連續變數
- int64 : 整數，可表示離散或連續變數
- object : 包含字串，用於表示類別型變數

03



- 拿到資料的第一步，通常就是看我們有什麼，觀察有什麼欄位，這些欄位代表什麼意義、以什麼樣的資料類型來儲存
- 資料原來是字串/類別的話，如果要做進一步的分析時（如訓練模型），一般需要轉為數值的資料類型，轉換的方式通常有兩種
  - Label encoding：使用時機通常是該資料的不同類別是有序的，例如該資料是年齡分組，類別有小孩、年輕人、老人，表示為 0, 1, 2 是合理的，因為年齡上老人 > 年輕人、年輕人 > 小孩
  - One Hot encoding：使用時機通常是該資料的不同類別是無序的，例如國家

# 解題時間

## It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

