

Day 7

# 資料清理數據前處理

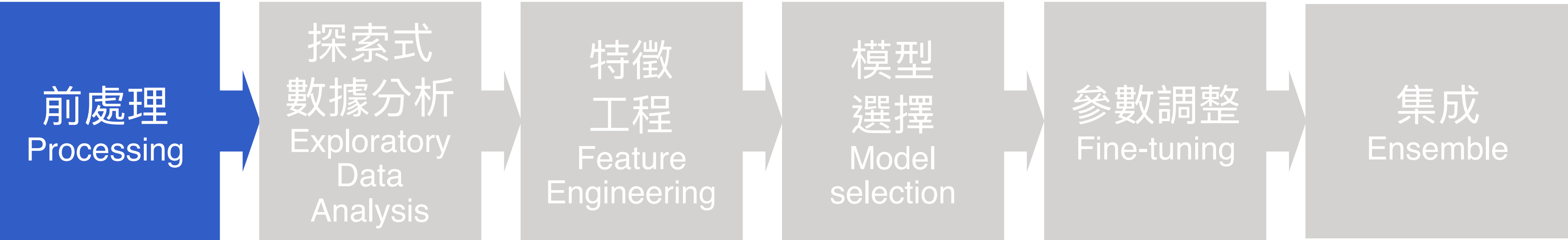
常用數值取代：中位數與分位數  
連續數值標準化



# 知識地圖 機器學習前處理 中位數與分位數連續數值標準化

## 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

1

如何處理例外值

2

如何進行數據標準化

# 常用以填補的統計值

---

## 常用以填補的統計值

## 方法

中位數 (median)

`np.median(value_array)`

分位數 (quantiles)

`np.quantile(value_array, q = ...)`

眾數 (mode)

`scipy.stats.mode(value_array)`: 較慢的方法  
dictionary method: 較快的方法

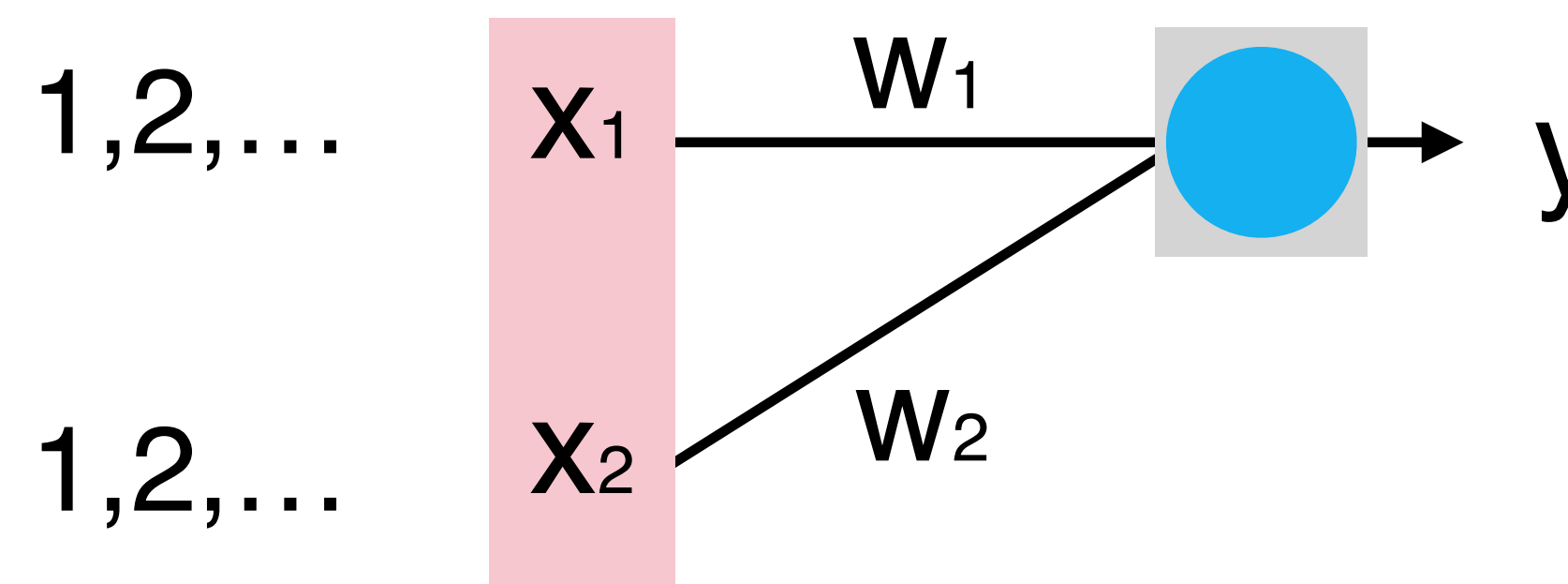
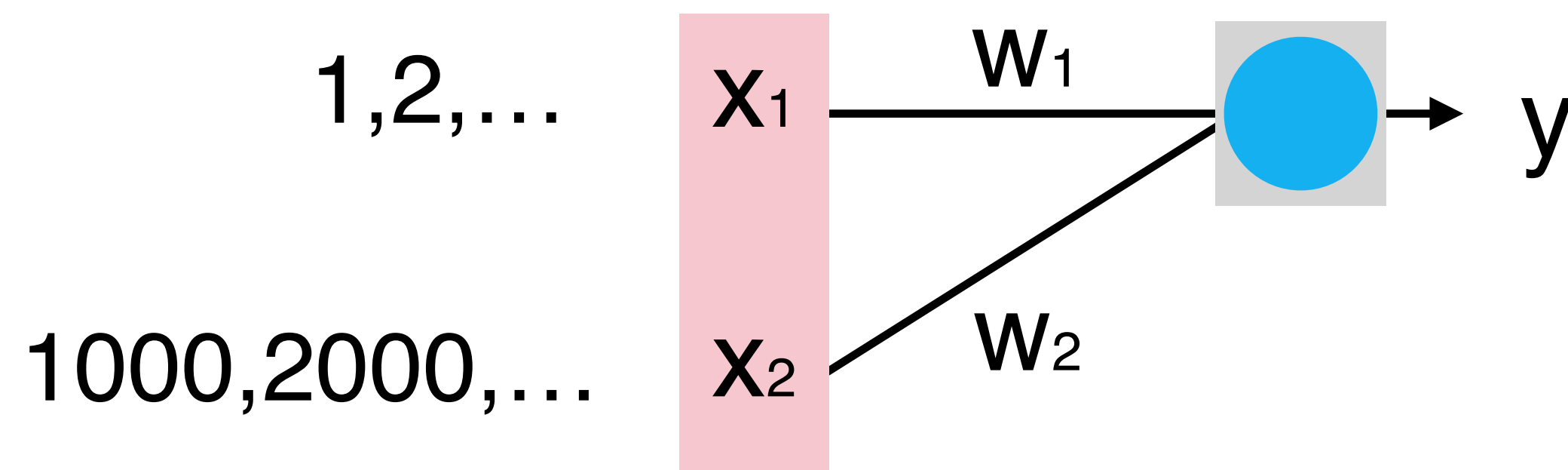
平均數 (mean)

`np.mean(value_array)`

# 連續型數值標準化

- 為何要標準化

改變一單位的  $x_2$  對  $y$  的影響完全不同



- 是否一定要做標準化 (有沒有做有差嗎)

看使用的模型而定

- Regression model : 有差
- Tree-based model : 沒有太大關係

**Requires little data preparation.** Other techniques often require data normalization. Since trees can handle qualitative predictors, there is no need to create dummy variables.

# 連續型數值標準化

## 常用的標準化方法

## 公式

Z 轉換

$$\frac{(x - \text{mean}(x))}{\text{std}(x)}$$

空間壓縮

$$Y = 0 \sim 1, \quad \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$Y = -1 \sim 1, \quad \left( \frac{x - \min(x)}{\max(x) - \min(x)} - 0.5 \right) * 2$$

$$Y = 0 \sim 1, \text{ (針對特別影像)}, \quad \frac{x}{255}$$

## 特殊狀況

有時候我們不會使用 min/max 方法進行標準化，而會採用 Qlow/Qhigh normalization (如將空間壓縮第一例中的 min 改為 q1, max 改為 q99)

# 解題時間

## It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

