

Day 28

## 特徵工程

# 特徵組合 - 數值與數值組合



出題教練

陳明佑

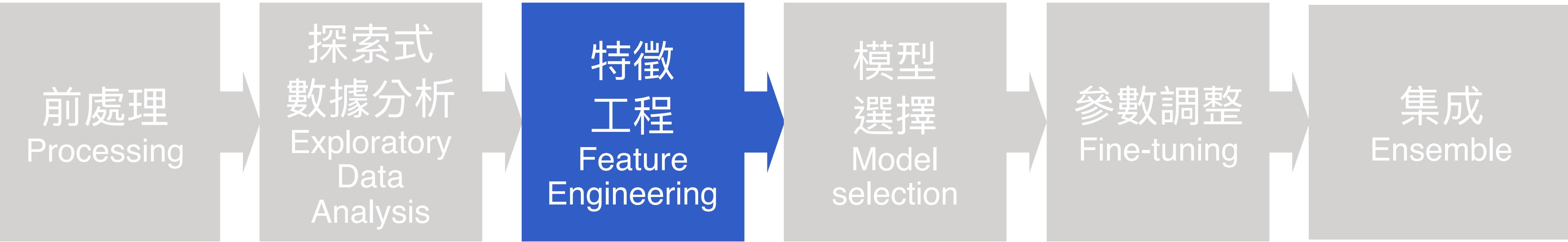


# 知識地圖 特徵工程 特徵組合 - 數值與數值組合

## 特徵工程

### 監督式學習

#### Supervised Learning



### 非監督式學習

#### Unsupervised Learning



## 特徵工程 Feature Engineering

### 概論

數值型特徵

類別型特徵

時間型特徵

填補缺值

去離群值

類別型特徵處理

時間型特徵處理

去偏態

特徵縮放

特徵組合

特徵篩選

特徵評估

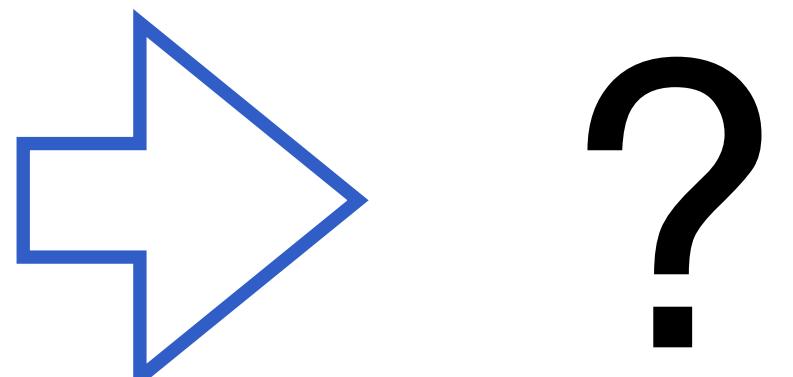
# 本日知識點目標

- 數值與數值的特徵組合，除了基礎的加減乘除等四則運算，最關鍵的部分是什麼？
- 機器學習的關鍵又是什麼？

# 特徵組合 (1 / 3)

在計程車費預估中，有四個欄位分別表示起終點的經緯度  
想想看，是否可以用這些組合出與車費更有相關的特徵？

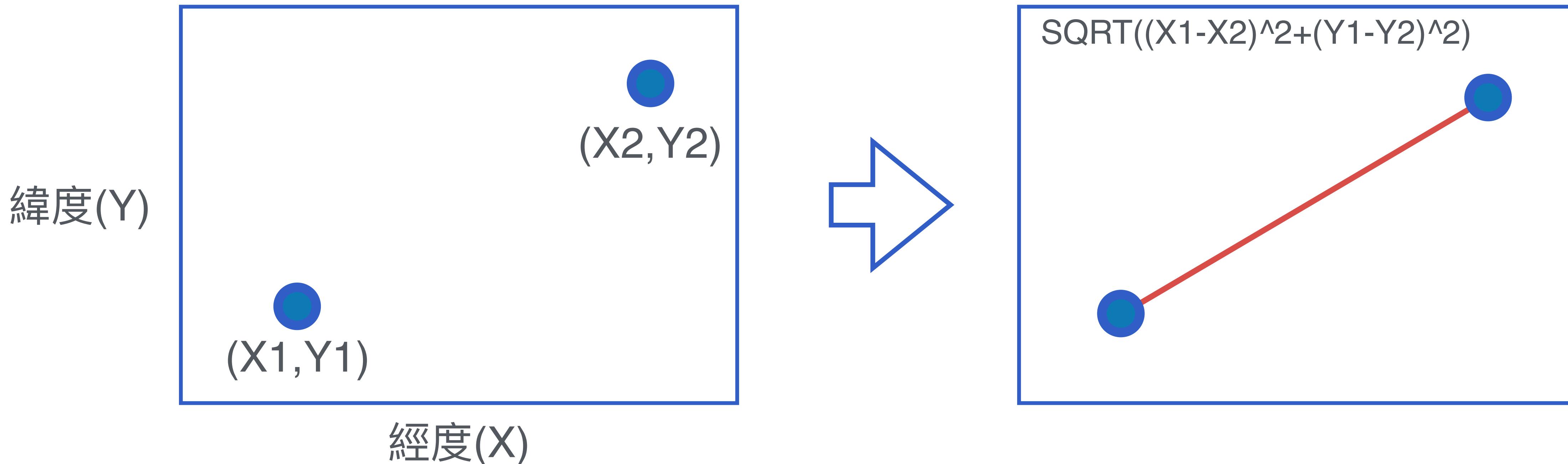
起點經度	起點緯度	終點經度	終點緯度
-73.99058	40.76107	-73.98112	40.75863
-73.98840	40.72343	-73.98964	40.74169
-74.01578	40.71511	-74.01202	40.70788
-73.97732	40.78727	-73.95803	40.77883



## 特徵組合 (2 / 3)

合理的想法是：將這四個特徵看成座標

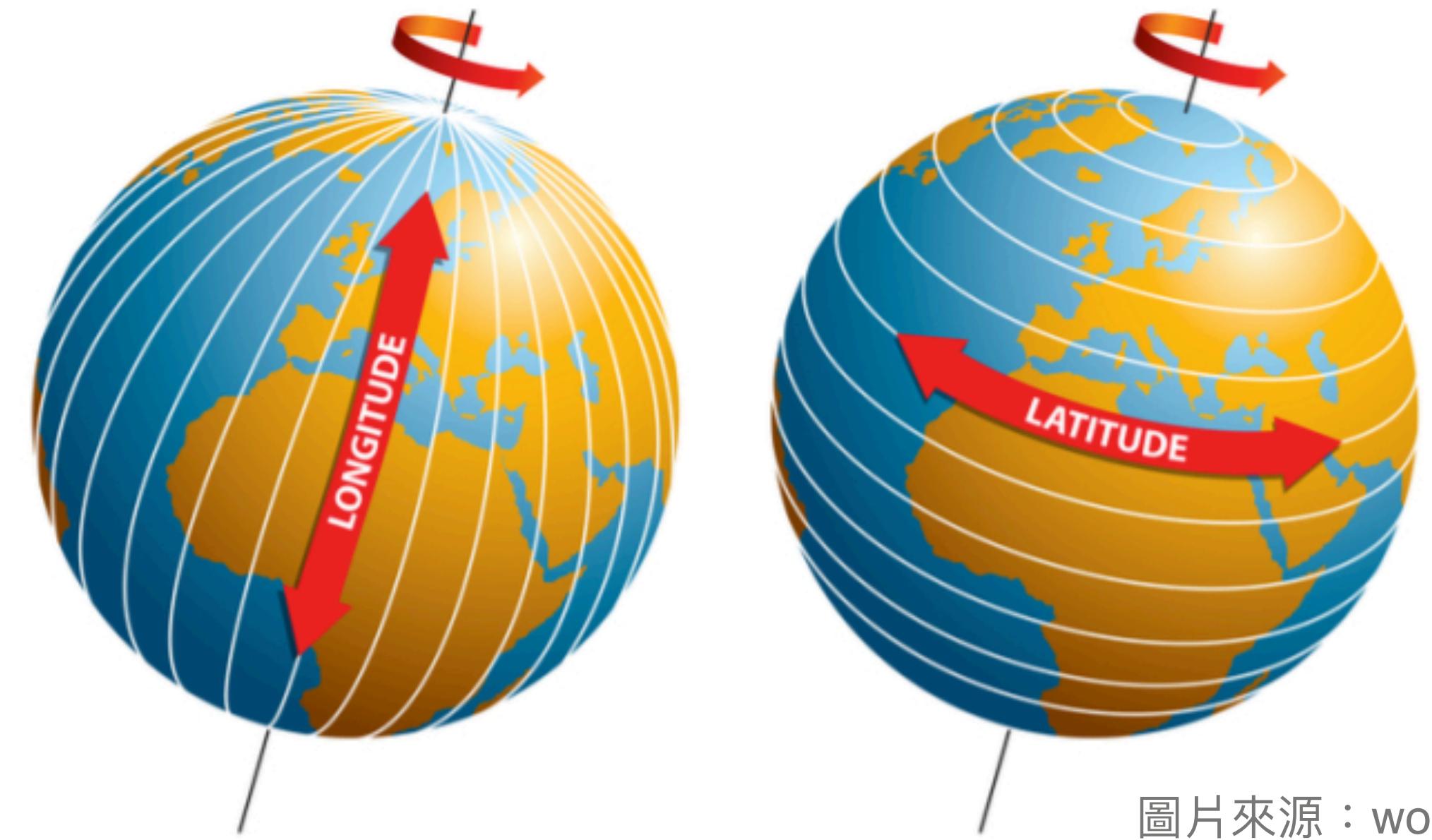
因此用平面座標距離組合出來的特徵，更有預測力也非常合理



想一想：還有沒有可能合成更強力的特徵呢？

# 特徵組合 (3 / 3)

事實：經緯度每一度並不一樣長



圖片來源：[worldatlas](#)

觀察資料緯度集中在 40.75 度附近

可以算得經度與緯度代表的長度比為  $\cos(40.75\text{度}) : 1 = 0.75756 : 1$

由此校正後的兩地距離，預測正確度更高

# 特徵工程的核心概念：領域知識

- 機器學習的關鍵在特徵工程
- 特徵工程的關鍵在領域知識

回想一下：

只是知道有四個數值欄位，預測力有限

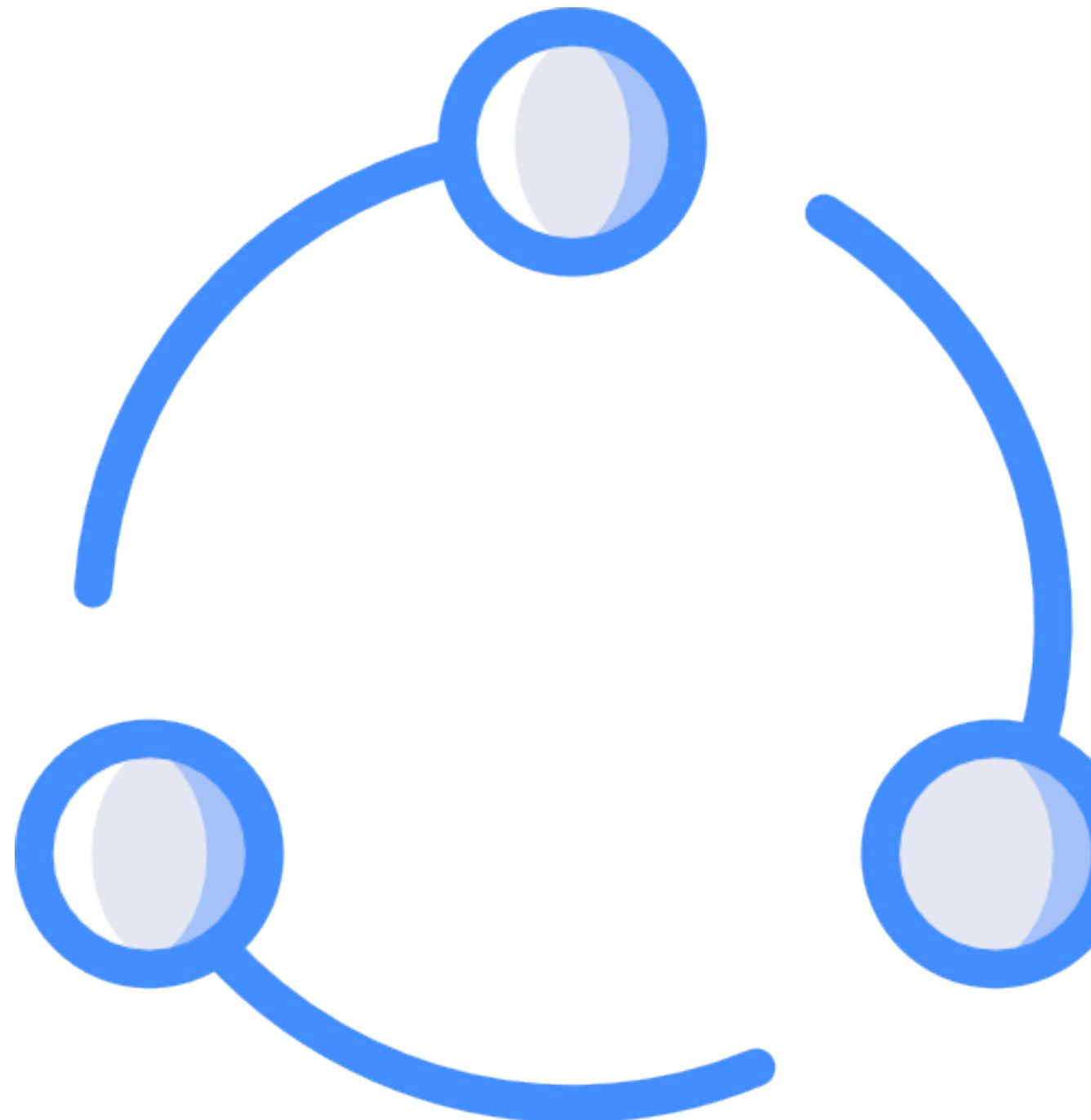
加上知道這四個數值是座標相關，就可以使用高斯距離合成特徵

再加上知道這是經緯度，就可以得到更精確的結果

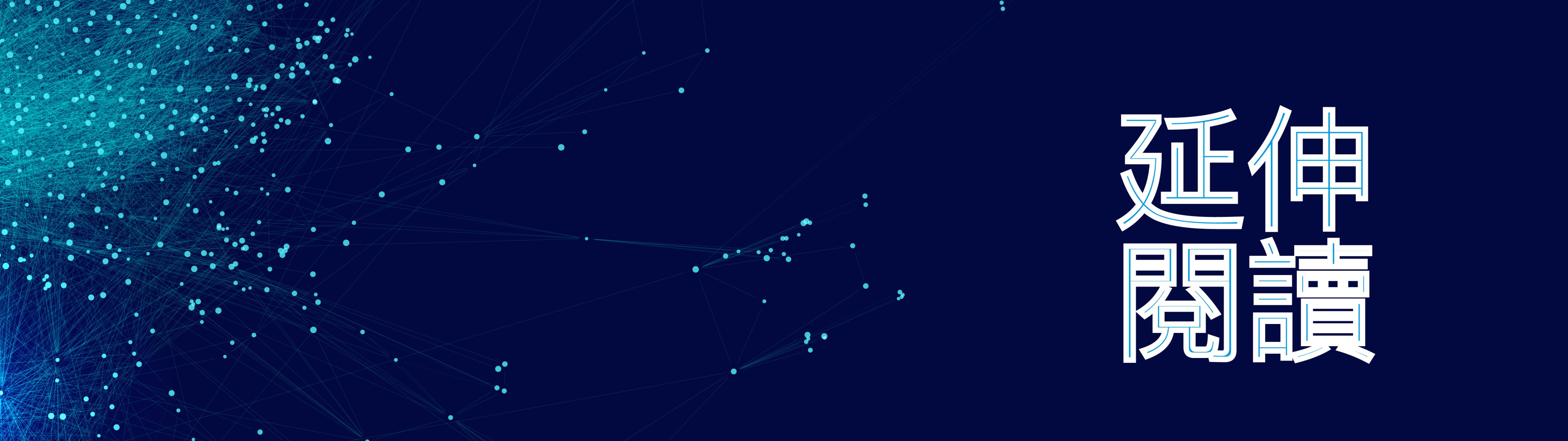
因此，對問題領域知識的了解，才是特徵工程最重要的環節

\*Day26所說時間的幾種週期，也可視為我們對「時間」的知識

# 重要知識點複習



- 數值與數值的特徵組合，最關鍵的部分是領域知識
  - 機器學習的關鍵是特徵工程，當然其餘部分仍然很重要，但是各部分都熟悉之後，最有效提升模型預測力的部分就是特徵工程
- \*\*註：好的資料能夠更有效提升預測力，特徵工程最有效的前提是資料集固定時(例如競賽)



# 延伸 閱讀

除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有  
多餘時間，可再補充延伸閱讀文章內容。

# 推薦延伸閱讀

## 特徵組合&特徵交叉 (Feature Crosses)

SegmentFalut 網頁連結

- 這裡有一些延伸的特徵組合方式，例如講義中提到的運算組合方式，或者使用綜合特徵的離散化/分箱，或者將兩種獨熱編碼綜合...等有趣的特徵組合方式。這些方式提供同學在合成特徵時參考，但這些只是方法，合成特徵比較有效的方式還是參考領域知識(如果有的話)。

## 目錄

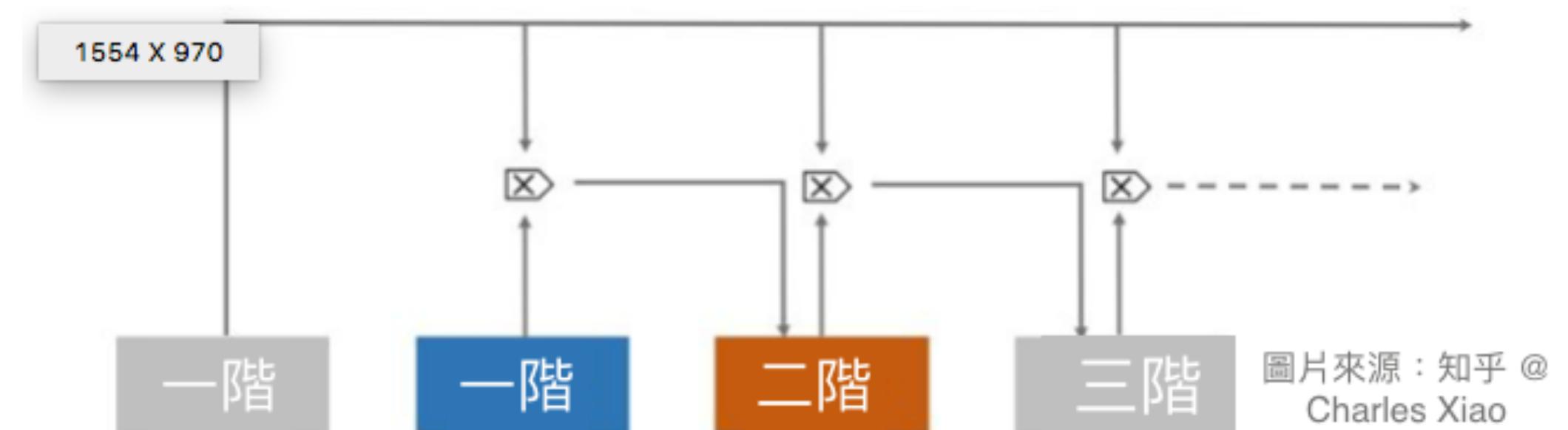
- 合成特徵(synthetic Feature)
- 特徵組合(feature cross)：對非線性規律進行編碼
- 特徵組合的種類
- 特徵組合(Feature Crosses)：組合獨熱矢量
- 代碼部分練習 學習目標：
- 使用分桶特徵列訓練模型
- 特徵組合
- 使用特徵組合訓練模型

# 推薦延伸閱讀

## 簡單高效的組合特徵自動挖掘框架

壹讀 (原始來源知乎) 網頁連結

- 這裡討論的「自動學習交叉特徵」，講的是當資料中類別種類多的類別特徵，這類問題我們一般的做法是使用LR(邏輯斯回歸) 或者FM(分解機器)...而本文則是討論FM進階的FFM。
- 不過這是在特徵都是屬於種類多的類別特徵時，才只好做的處理方式，如果這類特徵很少，或者在競賽中跳過這類特徵就可以有不錯準確度了，建議可以跳過這部分內容。



圖片來源：知乎 @  
Charles Xiao



解題時間

It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

