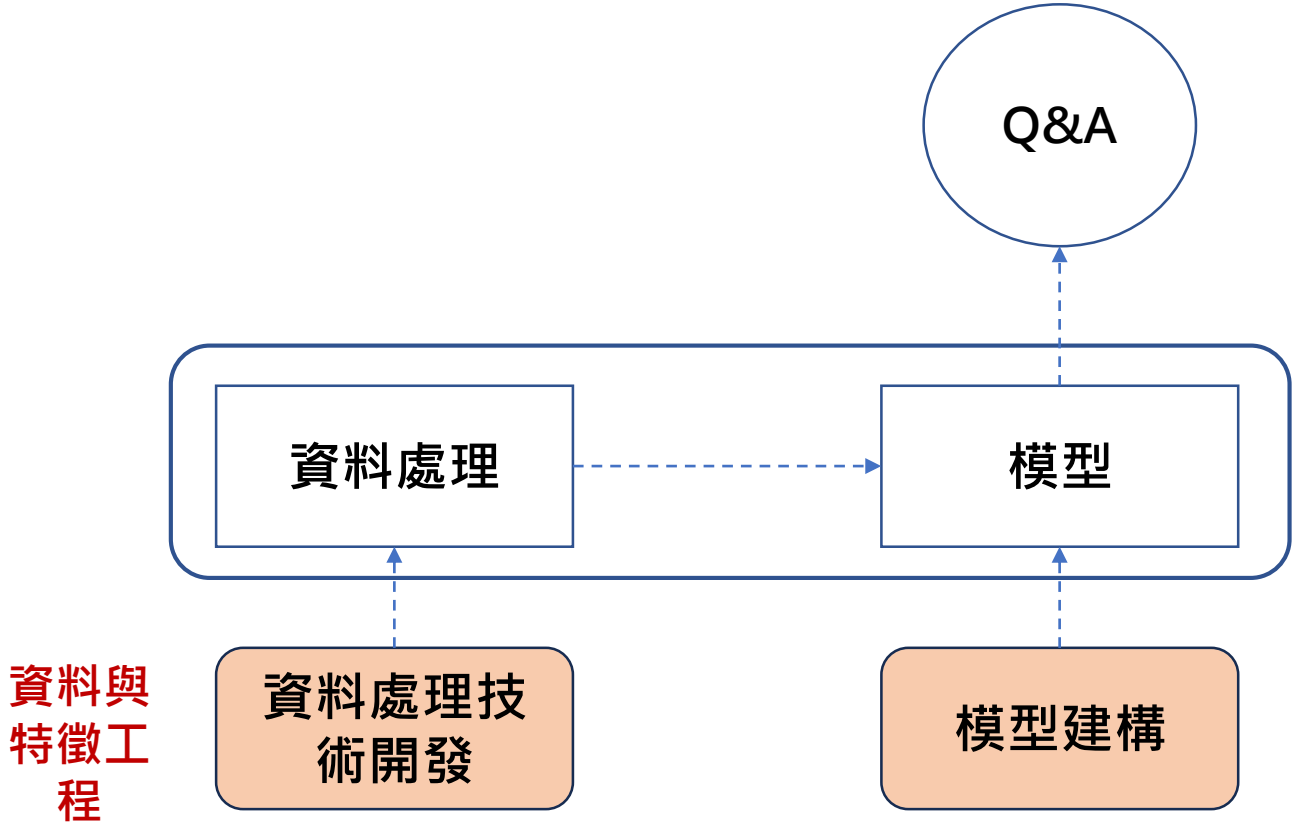
資料與特徵工程
Data & Feature Engineering



資料來源 (資料分析)
資料蒐集
資料生成
特徵選擇
特徵擷取
資料品質
資料平衡
資料不足

模型選擇 (Selection)
模型訓練 (Training)
模型驗證 (Validation)
模型測試 (Testing)
模型評量 (Evaluation)

Contents

Whats

特徵
特徵工程
資料種類/型態
資料工程

How

1. 資料採集
- 實體與屬性分析
- 要因分析法
- 情境分析法
- 專家訪談法
- 關聯分析法

2. 資料清理
- 資料縮減
- 離群值偵測
- 漏值填補

3. 資料整合
- 純特徵對純特徵
- 純特徵對時序特徵
- 時序特徵對純特徵
- 時序特徵對時序特徵

4. 特徵提取
- 主成分分析 (PCA)
- 線性判別分析 (LDA)
- 獨立成分分析 (ICA)
- 特徵雜湊 (Feature Hashing)

5. 特徵選擇
- 過濾方法
- 包裝方法
- 嵌入方法

6. 特徵編碼
- Label Encoding
- One-hot Encoding

7. 特徵縮放
- 正規化 (Normalization)
- 標準化 (Standardization)
- Min-Max 縮放

Whats in Feature Engineering

特徵
特徵工程
資料種類/型態
資料工程

特徵是甚麼

A typical quality or an important part of something

事物異於他事物的特點

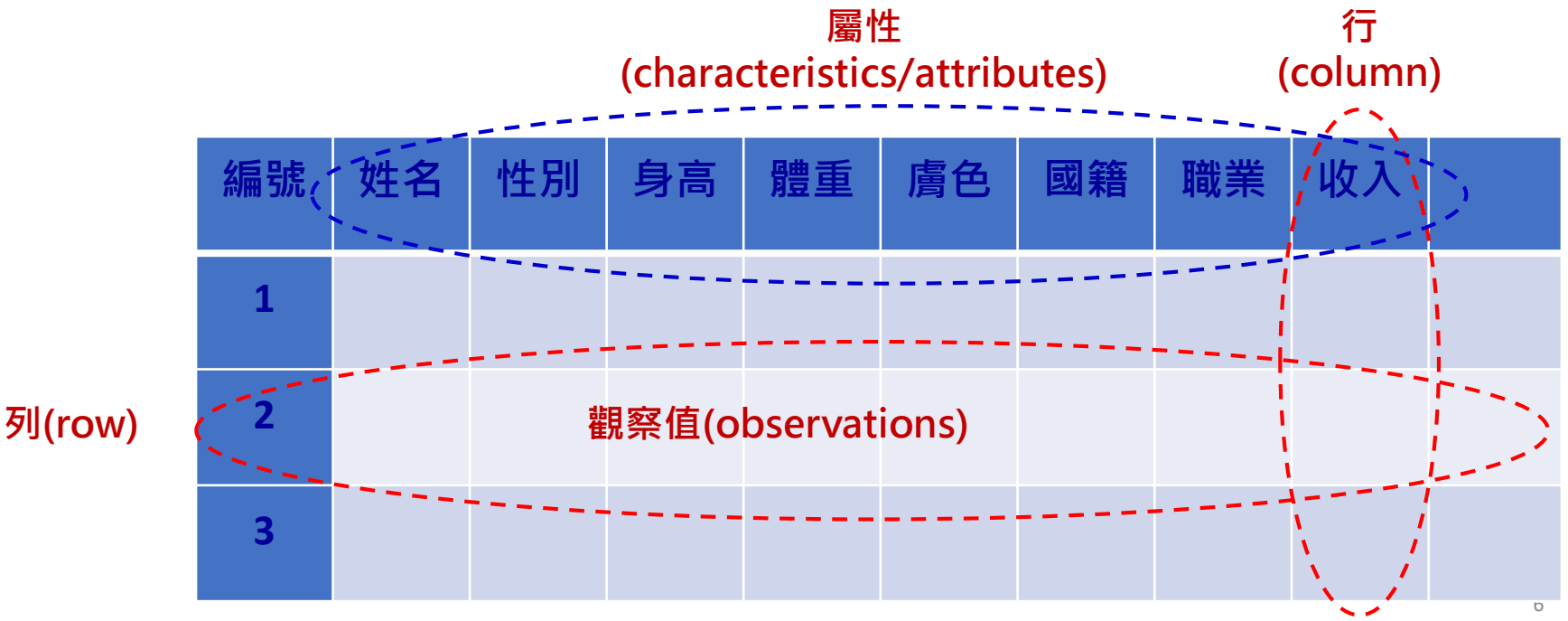
特徵可以是任何能夠幫助我們理解、描述或區分資料的資訊。

在機器學習和模式識別中，特徵是被觀測對象的可測量性能或特性。

特徵 Features

- 特徵是對「機器學習」過程有意義的資料屬性，即為從原始數據中所擷取、選擇出的輸入變數

- 一般來說，我們處理的資料都是表格形式的(tabular data)，按行列組織。資料的每一列又稱為觀察值(observation)，代表問題的實例



特徵 Features

- 監督式學習常執行「預測」一個「值」的任務，而這個「值」通常是資料中的一個「屬性」；我們通常會使用資料中的「其他屬性」，來預測「這一個屬性」。

- 被預測的「屬性」稱作反應（response），其餘屬性則叫作特徵(feature)。

何謂特徵工程

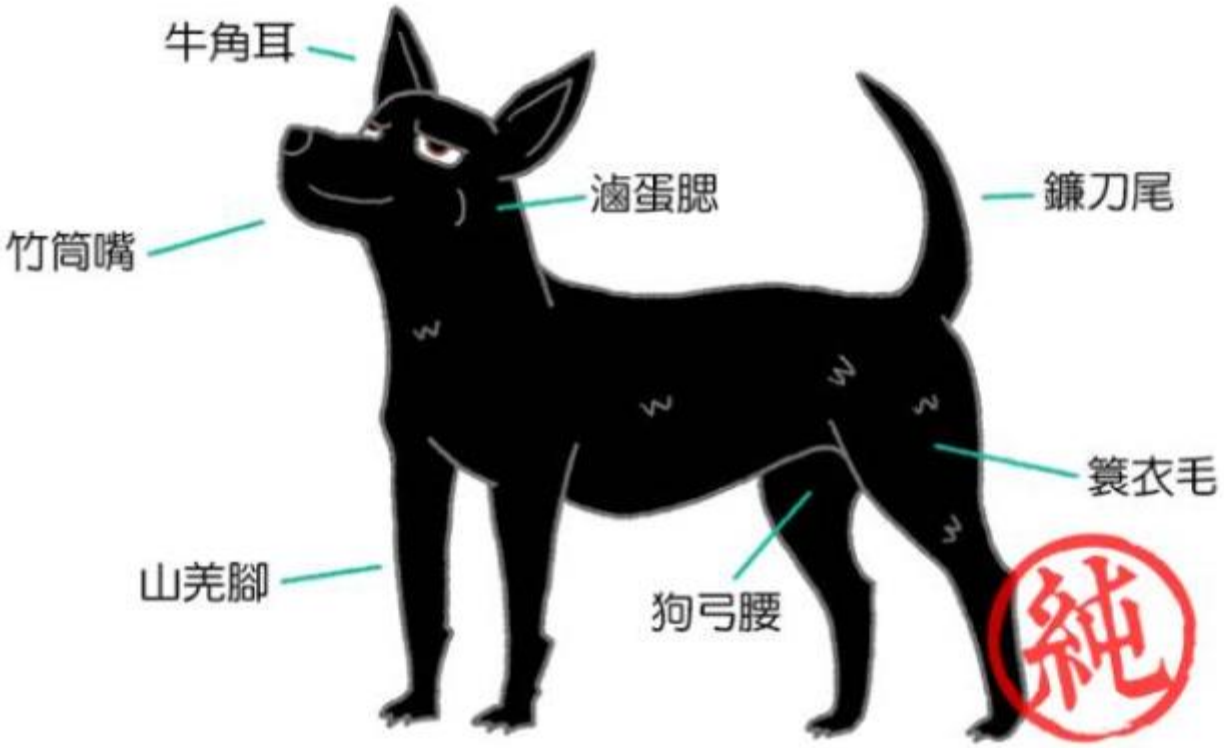
特徵工程是指的是將原始數據轉換成更適合機器學習算法使用的特徵或屬性的過程。

特徵工程的目標是提取、轉換和選擇特徵，以提高機器學習模型的性能，並使模型能夠更好地理解數據。

特徵工程包括以下主要步驟：

1. 特徵提取：將原始數據轉換為可供機器學習模型使用的數值表示的過程。
2. 特徵轉換：通過數學轉換將原始特徵轉換為新的特徵的過程，以改善模型的性能。
3. 特徵選擇：選擇最具信息量的特徵，並且丟棄冗餘或無關的特徵的過程。這有助於降低模型的複雜性，提高模型的訓練速度，並減少過擬合的風險。

正港的「台灣犬」要有這些特徵！



外貌特徵
行為特徵
生理特徵
生物特徵
運動特徵

Why特徵工程

特徵工程是機器學習中的一個重要步驟，因為好的特徵可以使模型更容易理解數據並提高預測性能。

不同的問題和數據集可能需要不同的特徵工程技術，因此特徵工程通常需要根據具體情況進行調整和優化。它需要對領域知識和數據的深刻理解，以選擇和設計適當的特徵，以實現最佳的模型性能。

資料種類/型態
Data Type

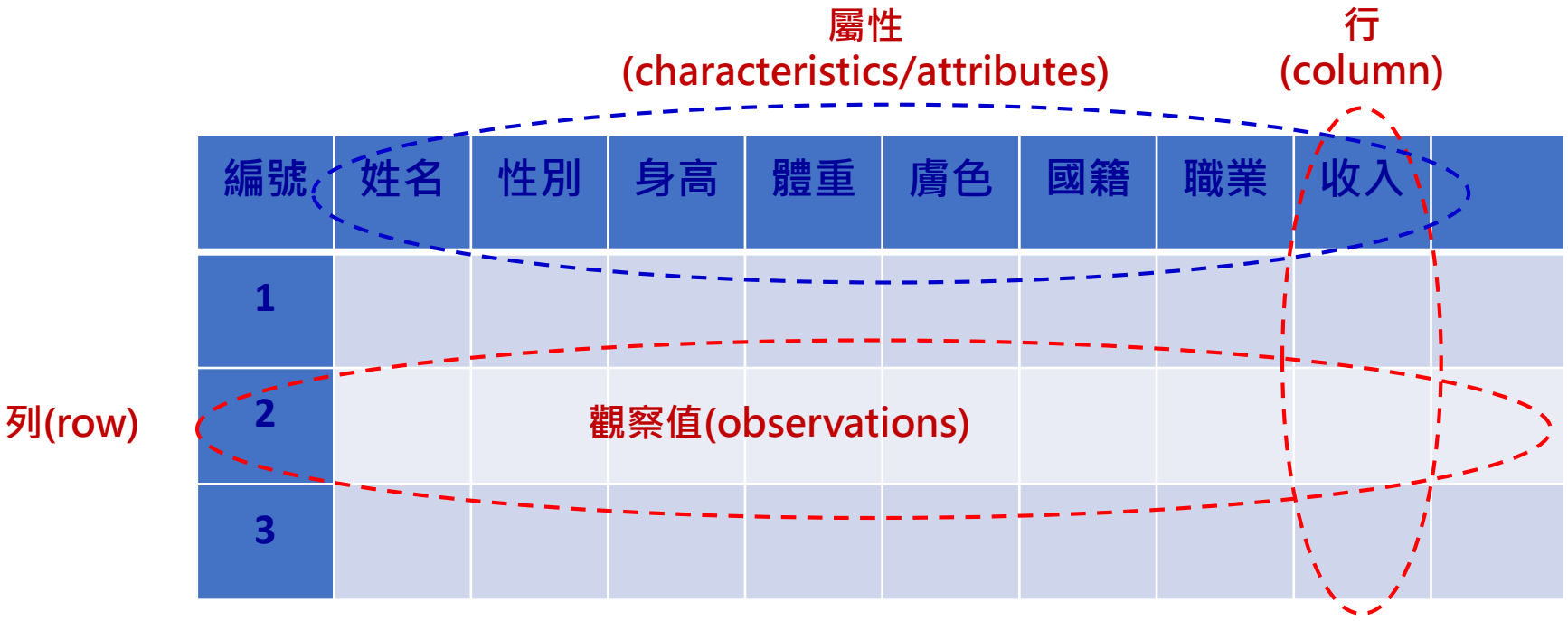
資料種類有很多，依據不同觀點有不同的分類方法

結構/非結構/ 半結構資料	依據資料是否能以資料庫table條列呈現區分
定量資料/定性資料	依據資料是否可以計量，是否具有單位區分
類別資料/連續資料	依據資料是否可以排序與比較區分
截面資料/時間序列 資料/面板資料	依據資料是否帶有時間戳記(time stamp)區分
數值、數字、符號、 文字、圖像、影音	依據資料本身型態區分

結構化/非結構化/半結構化資料

結構化資料：

- 有組織的資料，資料可以被呈現在資料庫table的行、列
- 一列(row)代表一筆紀錄，統計的術語稱為觀測(observation)
- 每個欄位/行(column)則稱為表徵/屬性(characteristic/attribute)



結構化/非結構化/半結構化資料

非結構化資料：

- 沒有組織的資料、未經整理的資料
- 無一定形式且不遵循標準組織結構規範（如表格）的資料
- 例如圖像、音頻、視頻、電子郵件、文字處理文檔
- 非結構化數據往往比結構化數據更大，佔用更多存儲空間

半結構化資料:

- 結構與非結構資料的組合
- 例如履歷表，表格具一定結構，但欄位內容極為多樣

結構化/非結構化/半結構化資料

	結構化	非結構化	半結構化
資料收集	固定欄位 固定格式 固定順序	不遵循標準組織結構規範的資料	具有欄位 但不一致
資料庫	關聯式資料庫	非關聯式資料庫	
舉例	銷售資料	影像、影音、圖檔、office、網頁等	履歷表

定量資料/定性資料

定量/量化資料(Quantitative data)

- 是以數值(numerical)形式表現出來的資料，為衡量某樣東西的量
- 對觀察對象測量指標的數值大小所得的資料，表現為數值大小
- 又稱計量資料(measurement data)、數值資料(numerical data)或尺度資料(scale data)
- 如：身高(cm)，體重(Kg)，細胞數(個)，人口數(人)

定性/質性資料(Qualitative data)

- 本質上是分類(categorical)
- 描述某樣東西的性質，而非進行測量
- 如：對某個人的印象、意見與看法

連續型資料/類別型資料

類別資料(Categorical data)

- 以類別來區分的資料
- 如男女性別、教育程度、職業別、區域別、滿意程度、偏好程度、品質好壞等

連續資料(Continuous data)

- 可以無限制細分的資料
- 即在任意兩個數值間可插入無限多個數值
- 資料依其發生時間則可分為橫斷面資料和時間序列資料

定量/定性 & 連續型/類別型

	連續型資料	類別型資料
	定量/量化資料	定性/質性資料
描述	以數值形式衡量觀察對象的量（數值大小），且數值可無限的細分	以類別來區分的資料，描述某樣東西的性質，而非進行測量
舉例	華氏溫度或攝氏溫度 捐血的血量	陰天或晴天 姓名

* 資料可以同時是定量和定性的

例如，餐廳的評分（１到５顆星）是數值，但是這個數值也可以代表星級類別。使用「定量」的星級系統打分數，並且公布帶有小數的平均分數值比如4.71顆星，那麼這個資料就是「定量」的

截面資料/時間序列資料/面板資料

以時間做為區分，考慮資料是否與蒐集時間有關聯性

1. 截面資料(cross-sectional data)

截面資料是指在某一時點收集的不同物件的資料。它對應同一時點上不同空間(物件)所組成的一維資料集合，特點就是離散性高。截面資料體現的是個體的差異，通常橫截面資料表現的是無規律的，通常用於分析比較被選擇的主體的差異。

2. 時間序列資料(time-series data)

時間序列資料是指對同一物件在不同時間連續觀察所取得的資料，著眼於研究物件在時間順序上的變化，尋找空間(物件)歷時發展的規律。

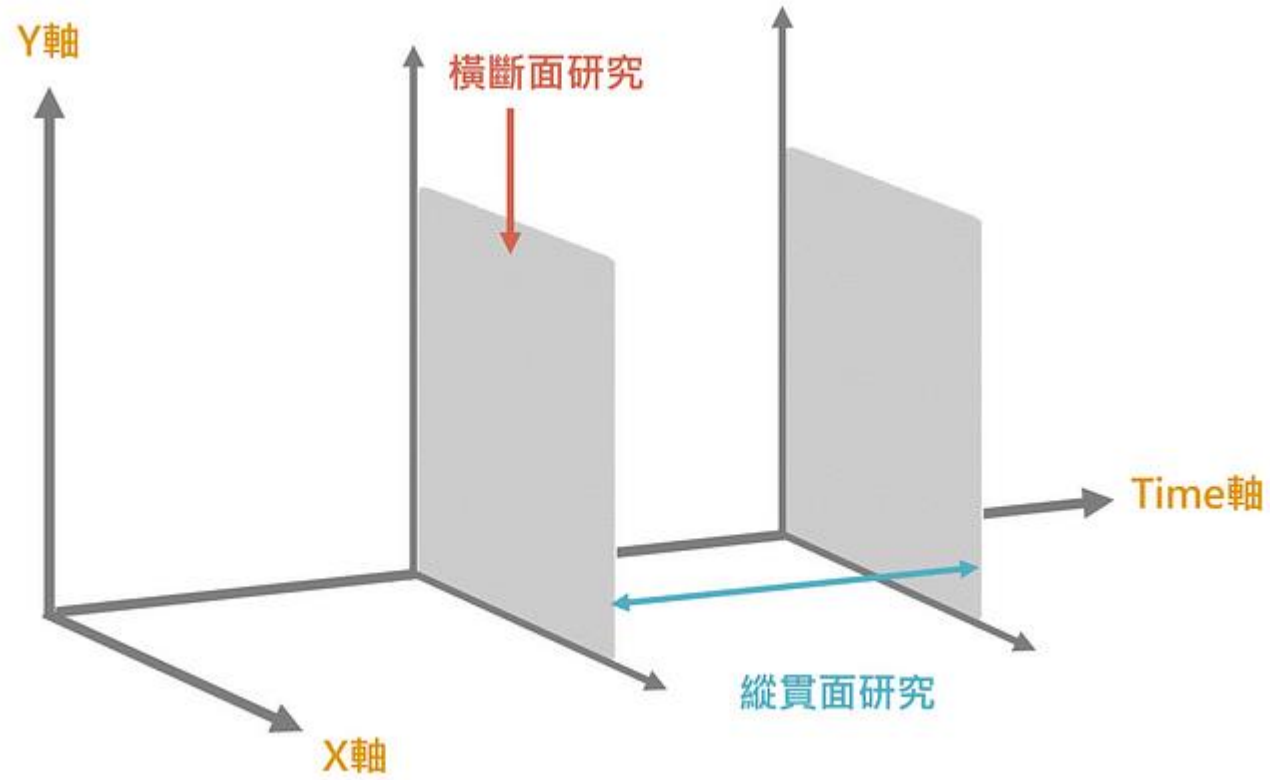
它是一組按照時間發生先後順序進行排列的數據點序列，通常一組時間序列的時間間隔為一恆定值（如1秒，5分鐘，12小時，7天，1年）

3. 縱橫資料(longitudinal data)或面板資料(panel data)

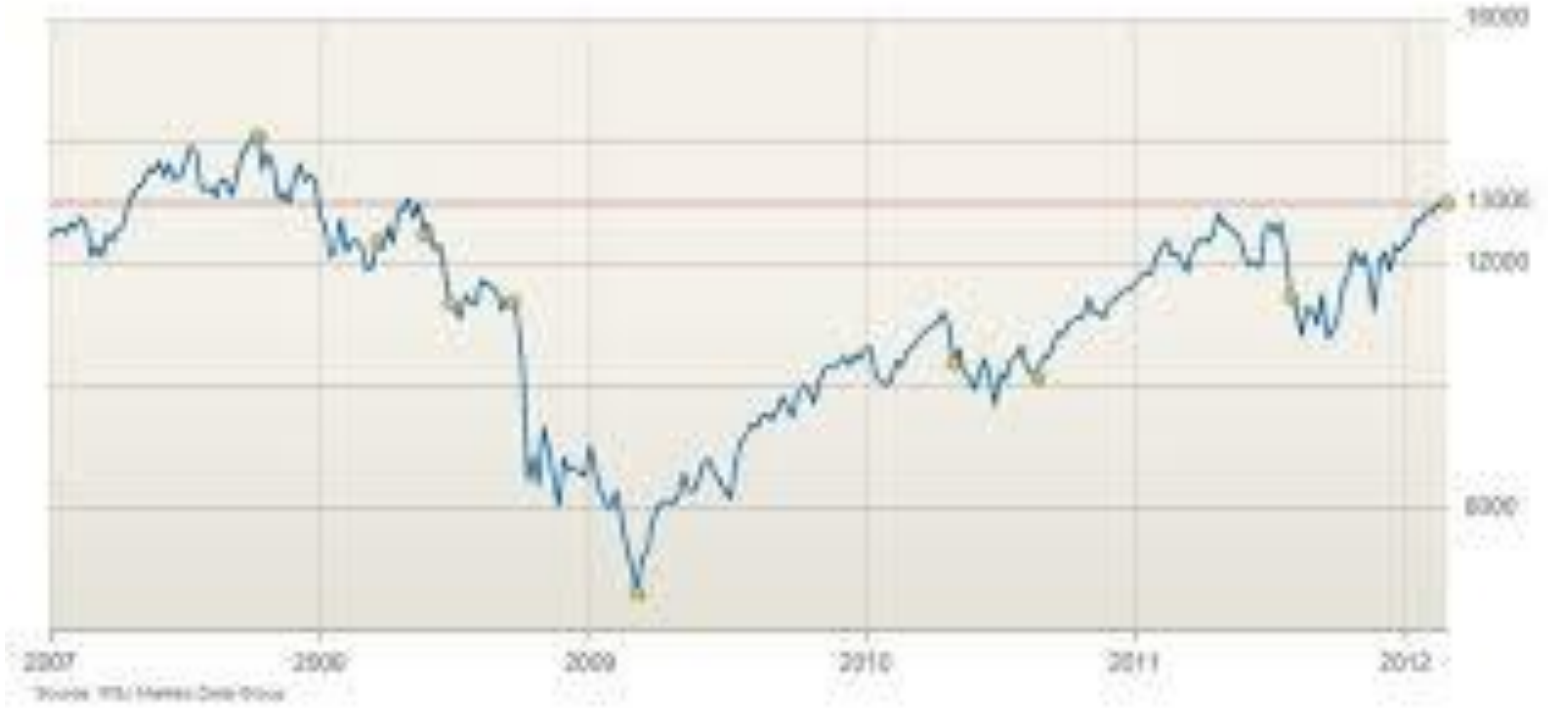
面板資料是截面資料與時間序列綜合起來的一種資料資源。 它可以用於分析各樣本在時間序列上組成的資料的特徵，它能夠綜合利用樣本資訊，通過模型中的引數,既可以分析個體之間的差異情況，又可以描述個體的動態變化特徵。

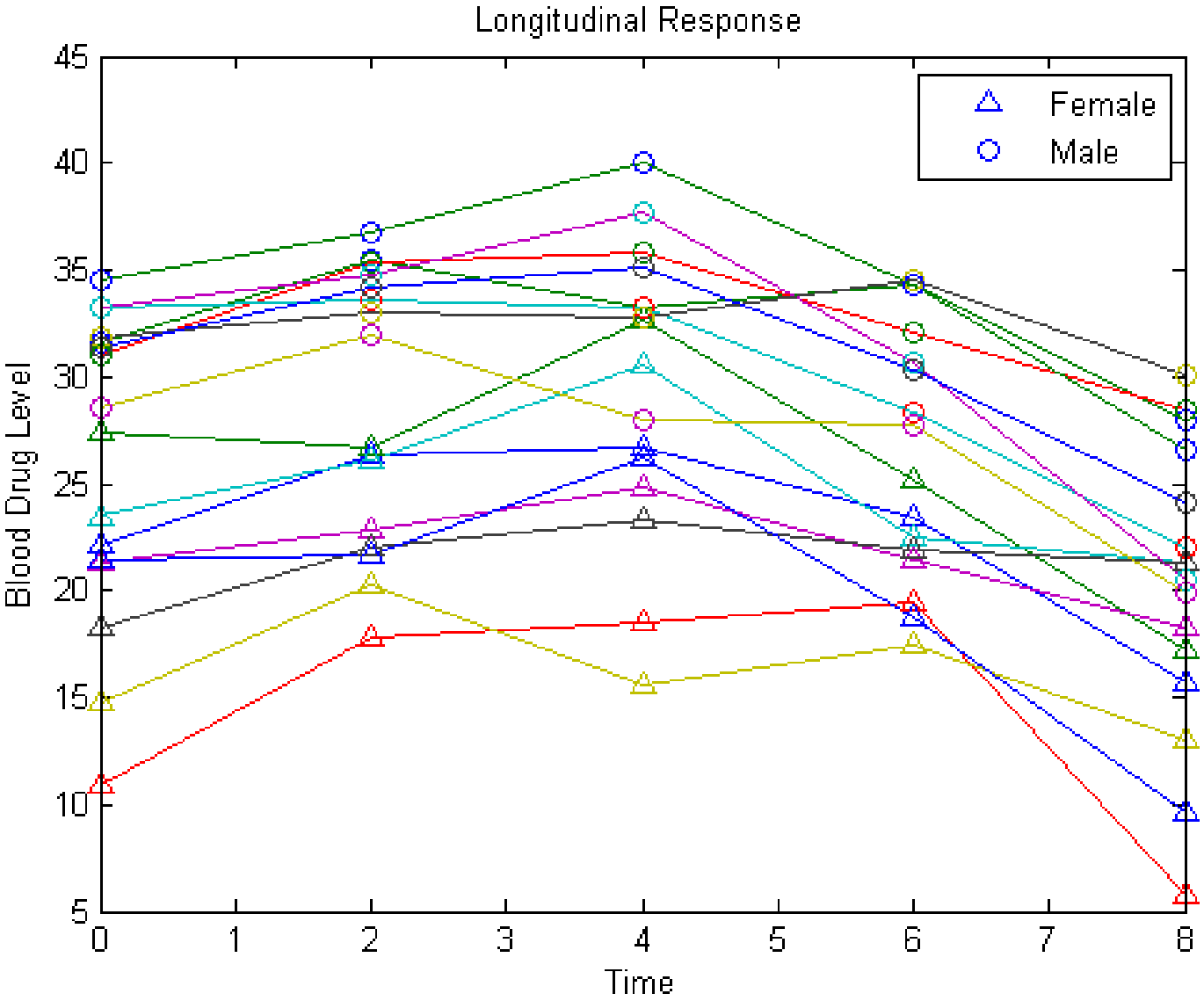
截面資料/時間序列資料/面板資料

	適用範圍	分析方法	舉例
截面資料	許多個體在同一個時間下由於個體不同而產生的資料	線性回歸、主成分分析等	台北、台中、台南某一天的平均溫度
時間序列資料	某一個個體隨時間變化產生的資料	自回歸、滑動平均法等	台南一年來每天的平均溫度
面板資料	綜合兩者，許多個體由於個體不同以及時間變化所產生的資料	綜合兩者	台北、台中、台南一年來每天的平均溫度



（繪圖者：陳靖宜）

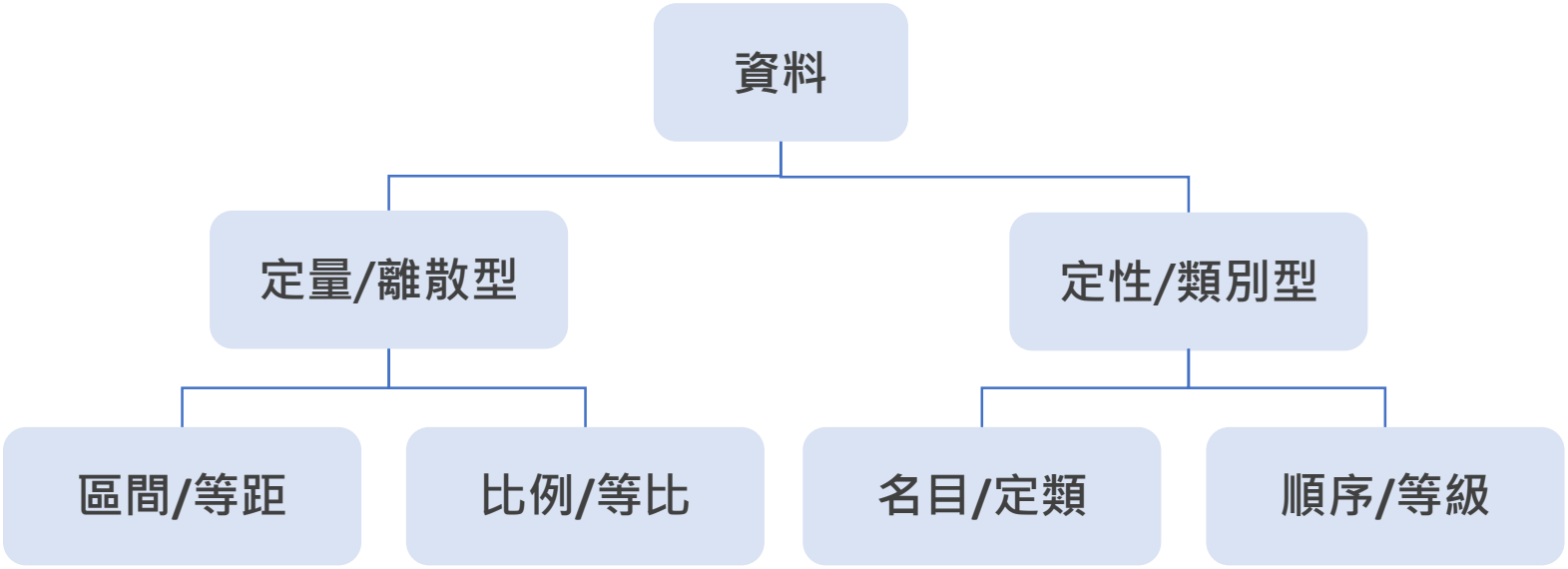

18000
13000
12000
8000
2007
2008
2009
2010
2011
2012


Longitudinal Response
Female
Male
Blood Drug Level
Time
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8

資料的四個形態

統計學家一般把資料分成 四種型態：

- 區間/等距資料（interval data）
- 比例/等比資料（ratio data）
- 名目/定類資料（nominal data/ categorical data）
- 順序/等級資料（ordinal data/ rank data）



名目資料
Nominal Data

名目資料又稱「定類資料」
通常會將一個集合分成互斥且能完全分辨的類別

例如：將「性別」區分成「男」、「女」
名目內容（如：男、女）本身具有意義，但編碼後 （如「男」為「1」、「女」為「0」） 的數字大小，並不代表任何意義

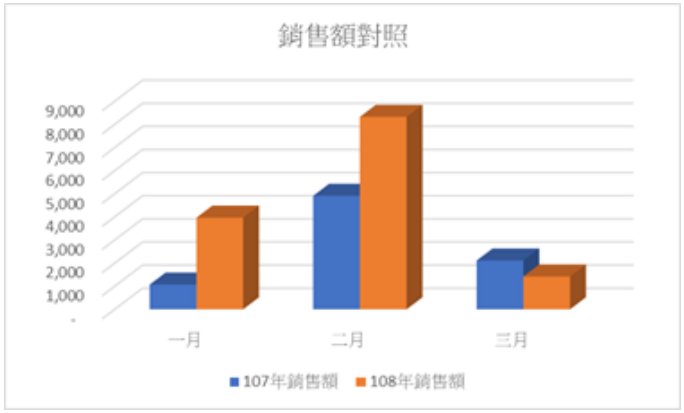
這些資料都是「定性」的不能執行任何「定量」數學操作或排序，但在統計處理時，可以累加次數(頻率數)，例如男性156人、女性182人，或者按次數多寡依序排列

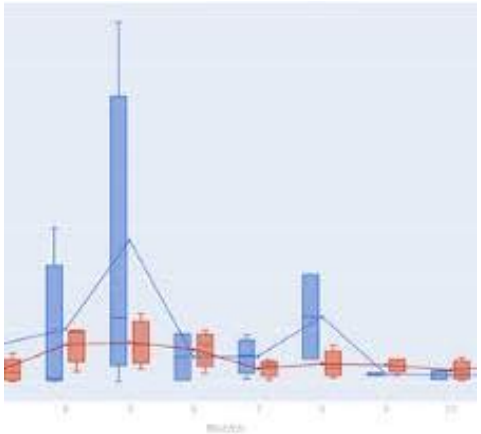
順序資料
Ordinal Data

順序資料又稱「定序資料」或「等級資料」
順序資料可以自然排序 (naturally ordered)，能區分等級或順序，例如：
教育程度裡，從小到大依序為：國小、國中、高中、大學、研究所。

順序資料亦可以像定類資料那樣進行計數，也可以比較和排序
編碼後的數字有大小的分別，但無法進行加減，無法說明大多少

例如：第一名比第二名好，第二名又比第三名好，但第一名第二名之間，
與第二名第三名之間無法確認是等值的。





區間資料
Interval Data

區間資料又稱等距資料，具有次序與距離。
編碼後的數值資料不僅可以排序，而且編碼後的數字為等距，「值」之間的差異也有意義。這意味著，區間資料不僅可以對「值」進行排序和比較，而且可以加減，但因為不具絕對原點，所以不能乘除。

例如:
1. 溫差計算，如果美國德州的溫度是華氏90℉(攝氏32℃)，阿拉斯加州的溫度是華氏40℉（攝氏4℃），可以計算出90-40=50℉(32-4=28℃)的溫差。

2. 年份2000年/2並不具意義

3. 問卷調查，從「非常滿意7分、很滿意6分、滿意5分、普通4分、不滿意3分、很不滿意2分、非常不滿意1分」等選項中圈選出符合的。

比例資料
Ratio Data

比例資料又稱等比資料，具有次序、距離與唯一原點、無負值，而且各數值間具有等差與比率的關係。

比例資料也是「定量資料」，不僅繼承了區間資料的加減運算，而且有了一個絕對零點(true zero)的概念，可以做乘除運算。

例如：
價格為100元、200元、300元...等。

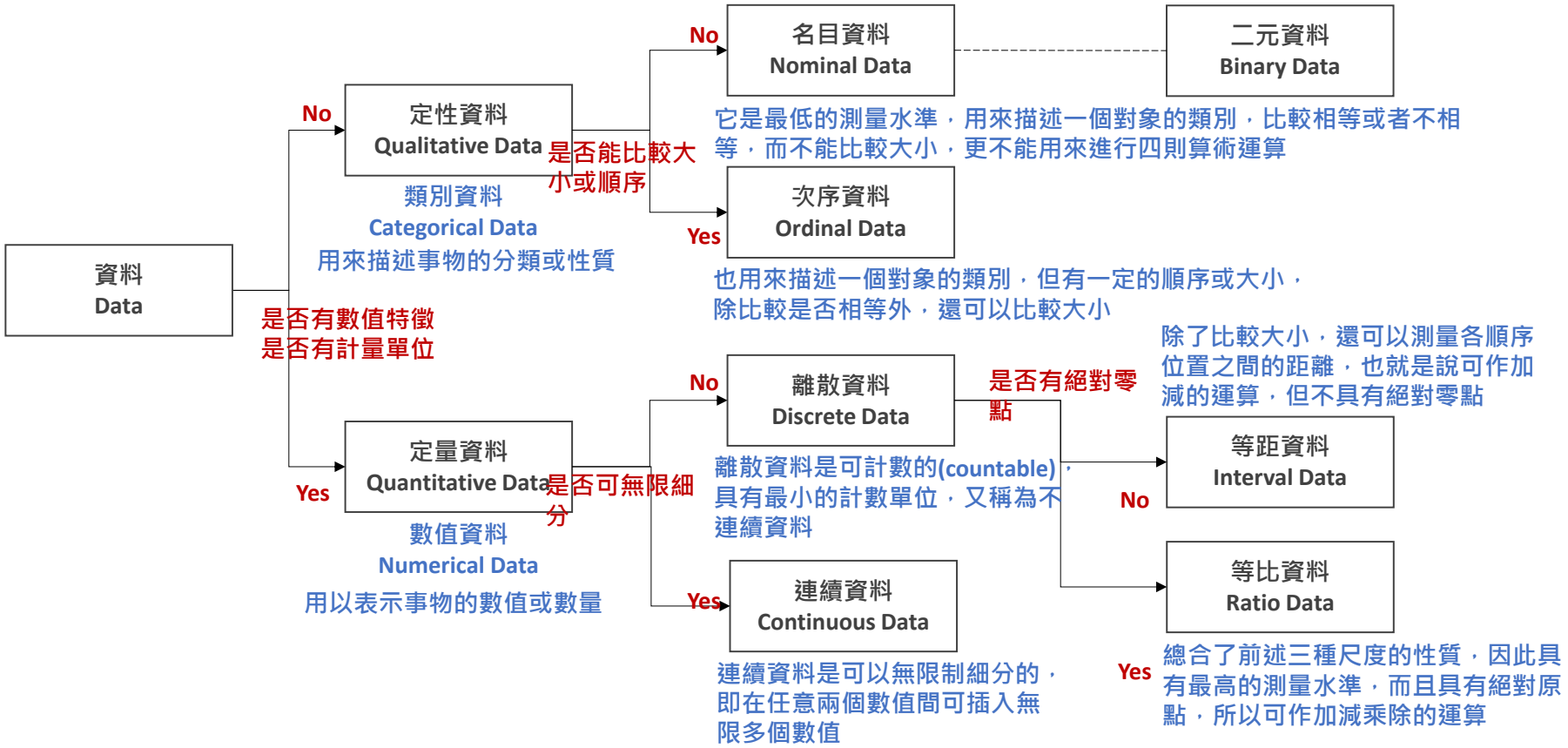
因為等距，所以能夠加減（如：價格200元與價格100元之間差100元）

具絕對原(零)點，所以能乘除（如：價格200元/2=價格100元）

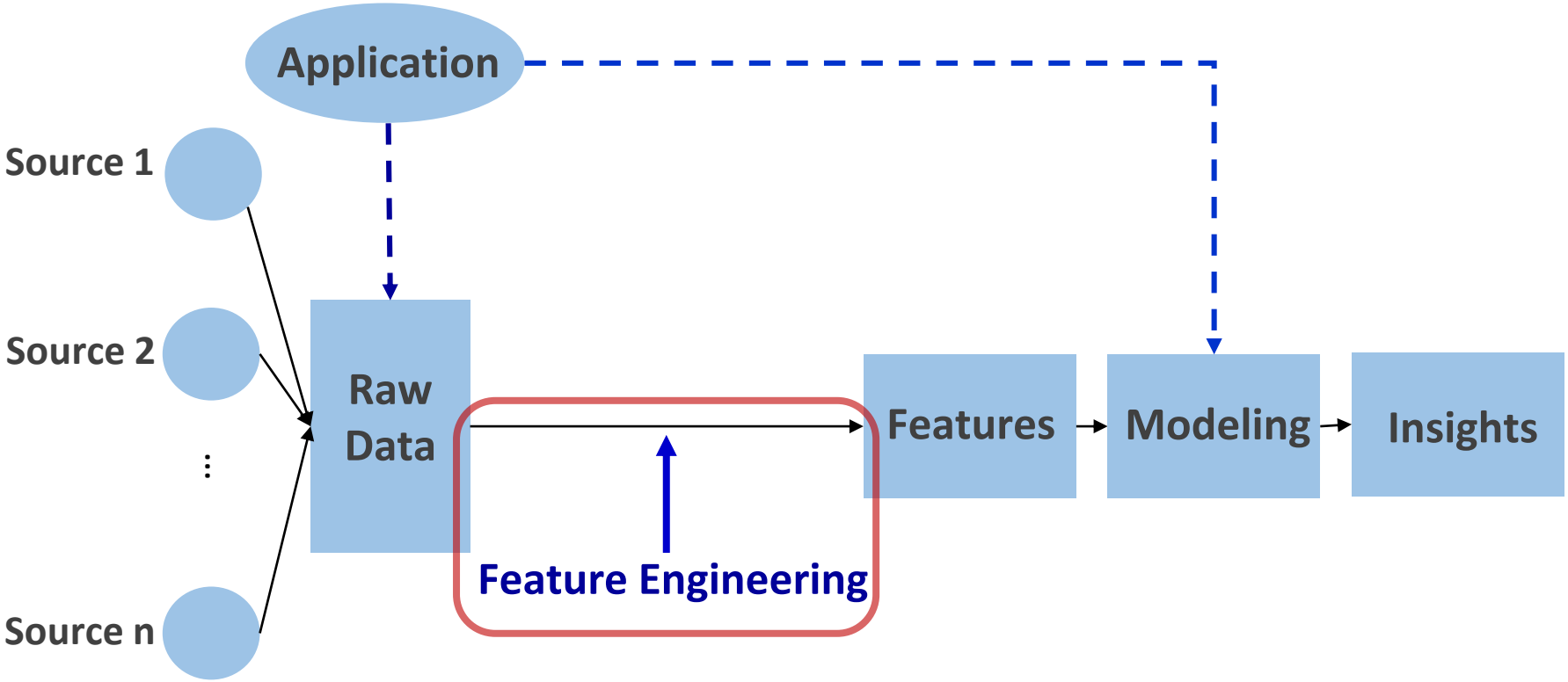
資料四種形態之比較

	次序	距離	原點	應用
名目	無	無	無	決定是否相等
順序	有	無	無	決定大於小於
區間	有	有	無	決定區間或差異的相等
比例	有	有	有	決定比率的相等

資料分類方式


資料
Data
是否有數值特徵
是否有計量單位
No
定性資料
Qualitative Data
類別資料
Categorical Data
用來描述事物的分類或性質
Yes
定量資料
Quantitative Data
數值資料
Numerical Data
用以表示事物的數值或數量
是否能比較大小或順序
No
名目資料
Nominal Data
二元資料
Binary Data
它是最低的測量水準，用來描述一個對象的類別，比較相等或者不相等，而不能比較大小，更不能用來進行四則算術運算
Yes
次序資料
Ordinal Data
也用來描述一個對象的類別，但有一定的順序或大小，除比較是否相等外，還可以比較大小
是否可無限細分
No
離散資料
Discrete Data
離散資料是可計數的(countable)，具有最小的計數單位，又稱為不連續資料
Yes
連續資料
Continuous Data
連續資料是可以無限制細分的，即在任意兩個數值間可插入無限多個數值
是否有絕對零點
No
等距資料
Interval Data
除了比較大小，還可以測量各順序位置之間的距離，也就是說可作加減的運算，但不具有絕對零點
Yes
等比資料
Ratio Data
總合了前述三種尺度的性質，因此具有最高的測量水準，而且具有絕對原點，所以可作加減乘除的運算

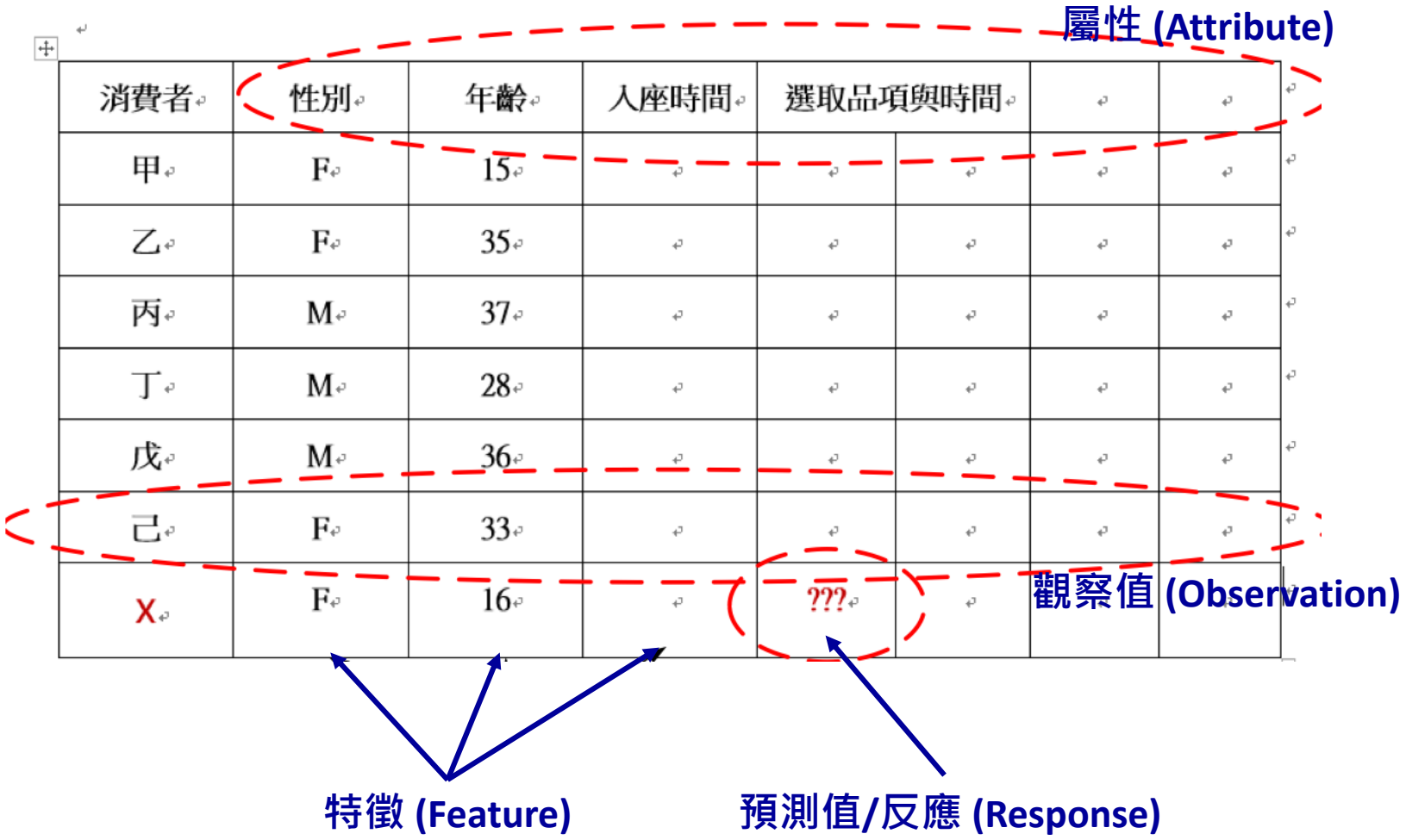
資料與特徵工程
Data & Feature Engineering



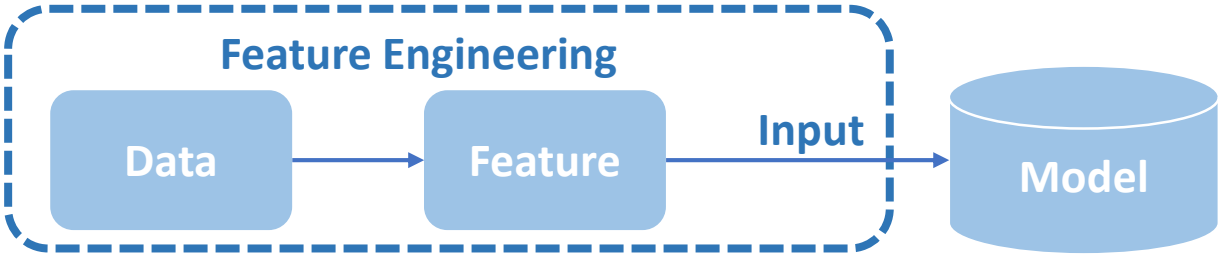
特徵 Feature

- 特徵是對「機器學習」過程有意義的資料屬性，即為從原始數據中所擷取、選擇出的輸入變數
- 一般來說，我們處理的資料都是表格形式的(tabular data)，按行列組織。資料的每一列又稱為觀察值(observation)，代表問題的實例
- 監督式學習常執行「預測」一個「值」的任務，而這個「值」通常是資料中的一個「屬性」；我們通常會使用資料中的「其他屬性」，來預測「這一個屬性」。
- 被預測的「屬性」稱作反應（response），其餘屬性則叫作特徵(feature)。

特徵 Feature


屬性 (Attribute)
消費者
性別
年齡
入座時間
選取品項與時間
甲 F 15
乙 F 35
丙 M 37
丁 M 28
戊 M 36
己 F 33
X F 16 ???
觀察值 (Observation)
特徵 (Feature)
預測值/反應 (Response)

特徵工程
Feature Engineering



- 特徵工程(Feature Engineering)是一系列將原始數據轉換為可用於機器學習模型特徵的預處理步驟

- 模型需要能夠最佳地描述資料的特徵組合，對原始資料進行清洗、提取、選擇和轉換等過程，以使用精確的特徵組合進行訓練，提升模型的性能

- 適當的特徵工程可以顯著影響機器學習項目的成功，使模型能夠從數據中提取有意義的模式和關係

特徵工程步驟

資料採集	確保你已經獲取到足夠的資料，包括需要的特徵和目標變數
資料清理	包括處理缺失值、異常值(outliers)、重複數據等，以確保資料質量
資料整合	整合不同來源的資料，將它們合併成一個統一的數據集
特徵提取	根據問題需求，從原始資料中提取新的特徵，以幫助模型更好地理解數據
特徵選擇	透過統計方法或機器學習技巧，選擇一部分最重要的特徵，以減少維度和降低雜訊，同時提高模型效能
特徵編碼	將資料轉換為機器學習模型可以理解的形式，這可能包括編碼類別變數（例如獨熱編碼或標籤編碼）和標準化連續變數（例如歸一化或標準化）
特徵縮放	將特徵縮放到相同的範圍或標準，通常使用技巧如標準化（歸一化）或最小-最大縮放等方法

不同資料類型之特徵工程

資料類型	常用特徵工程方法、技術
數值型	標準化、歸一化、特徵縮放、特徵交互等
類別型	獨熱編碼、標籤編碼、嵌入式編碼、目標編碼、特徵雜湊、順序編碼等
時間序列型	滑動窗口、指數加權移動平均、季節性分解、時間特徵提取
文本型	詞袋模型、TF-IDF、Word2Vec、N-grams、主題建模、情感分析、文本嵌入
圖像型	顏色直方圖、SIFT、SURF、HOG、CNN 特徵提取、遷移學習、圖像增強
混合型	將不同類型的特徵結合，例如堆疊、融合、多模型集成

資料需求分析

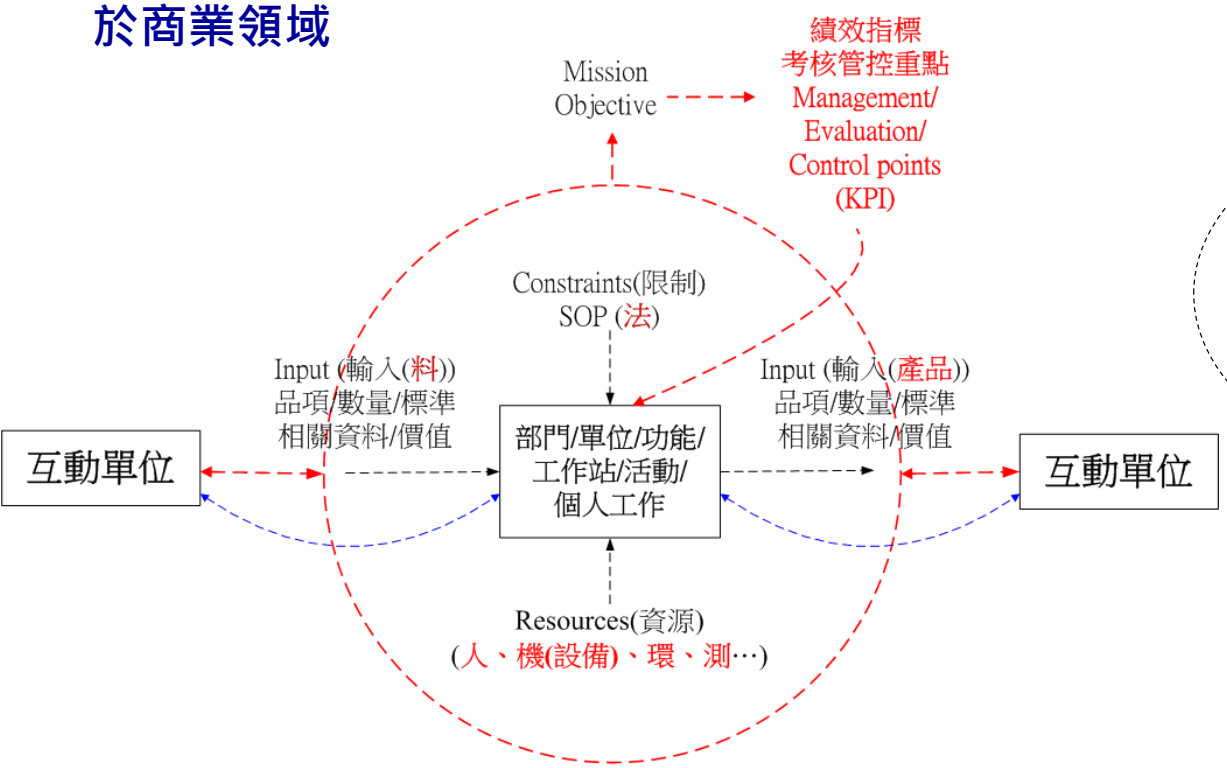
(1)實體(Entity)與屬性(Attributes)分析
(2) 要因分析法
(3) 情境分析法
(4) 專家訪談法
(5) 關聯分析法

(1) 實體(Entity)與屬性(Attributes)分析

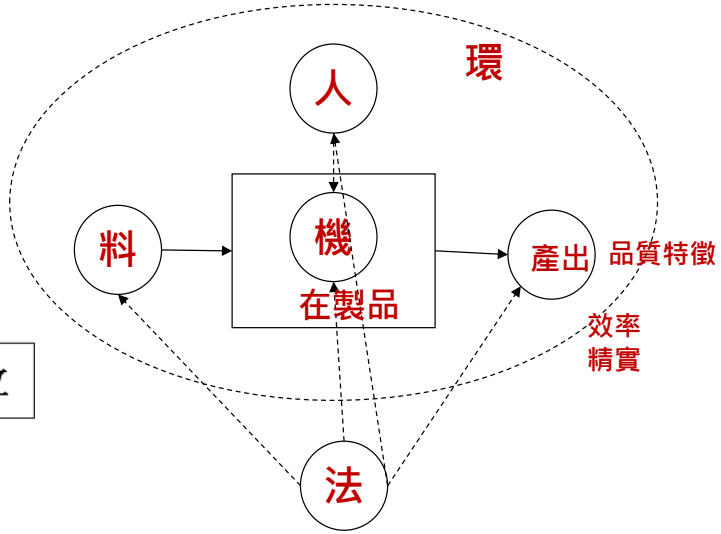
考慮數據中的不同實體（對象或個體）以及與這些實體相關的屬性（特徵）。這包括確定哪些實體和屬性可能對問題有影響，以及如何有效地表示它們。

例如：
在房地產領域，實體可以是房屋，屬性可以是房屋大小、價格、位置等。

於商業領域



於製造領域

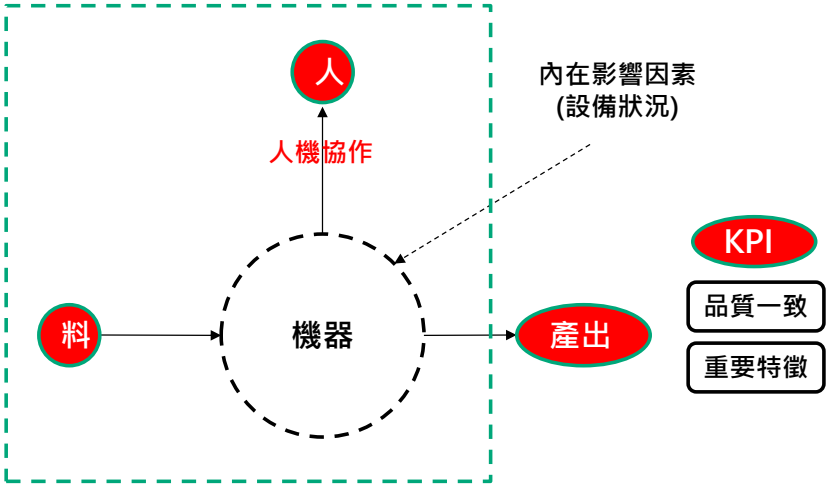


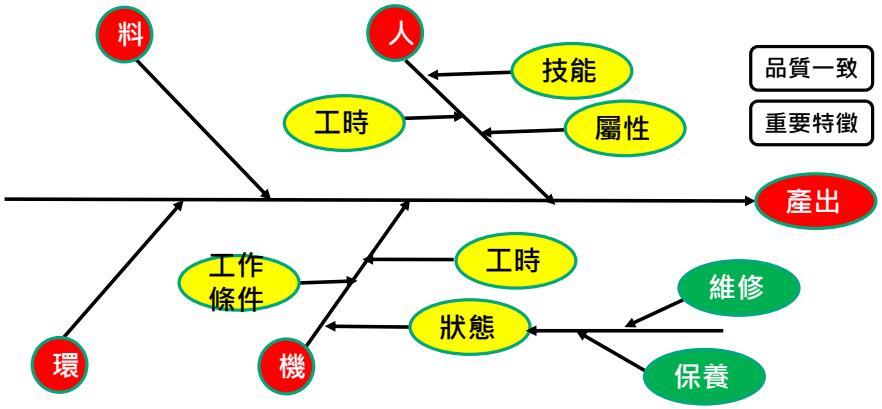
(2) 要因分析法

根(要)因分析：

根(要)因分析是一種統計方法，用於確定與目標變數之間的因果關係最密切的特徵。通過分析數據，它可以幫助確定哪些特徵對目標變數的變化具有最大的影響。

例如，在醫療研究中，根(要)因分析可以用來確定哪些因素最可能影響患者的健康狀態。



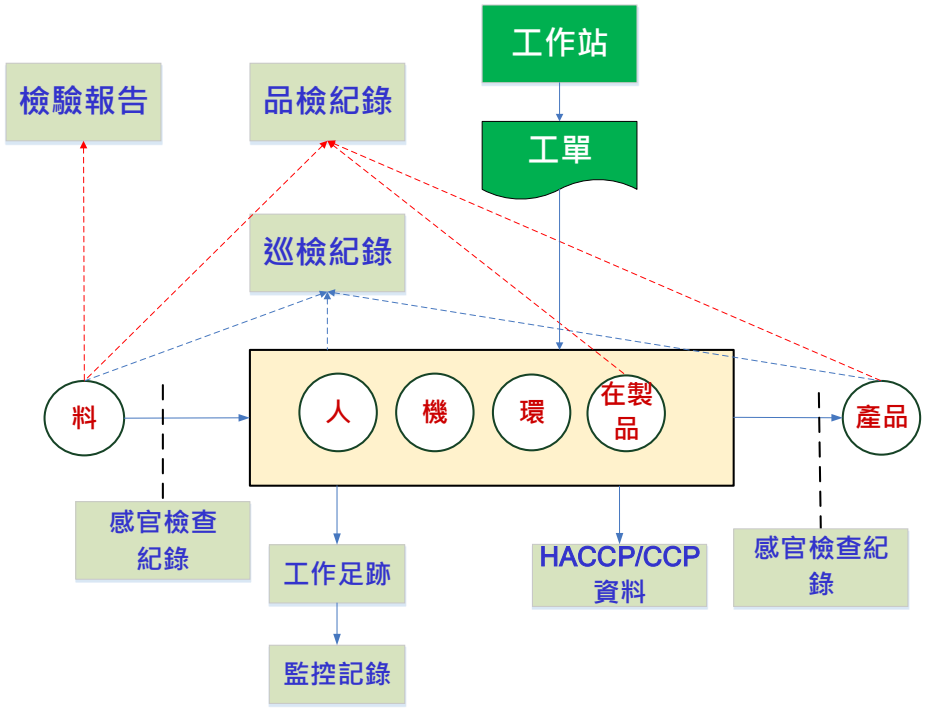


(3) 情境分析法

情境分析：

情境分析是根據特定問題和領域的知識，考慮哪些特徵在特定情境下可能具有更大的預測價值。這需要對問題背景和領域有深入的理解。

例如，在自然語言處理中，分析文本數據時，特定詞語可能在特定情境下具有更多的情感或主題相關性。



(4) 專家訪談法

專家訪談：

專家訪談是通過諮詢領域專家的意見和建議，以確定關鍵特徵的方法。專家可以提供有關哪些特徵可能對問題有影響的寶貴見解。

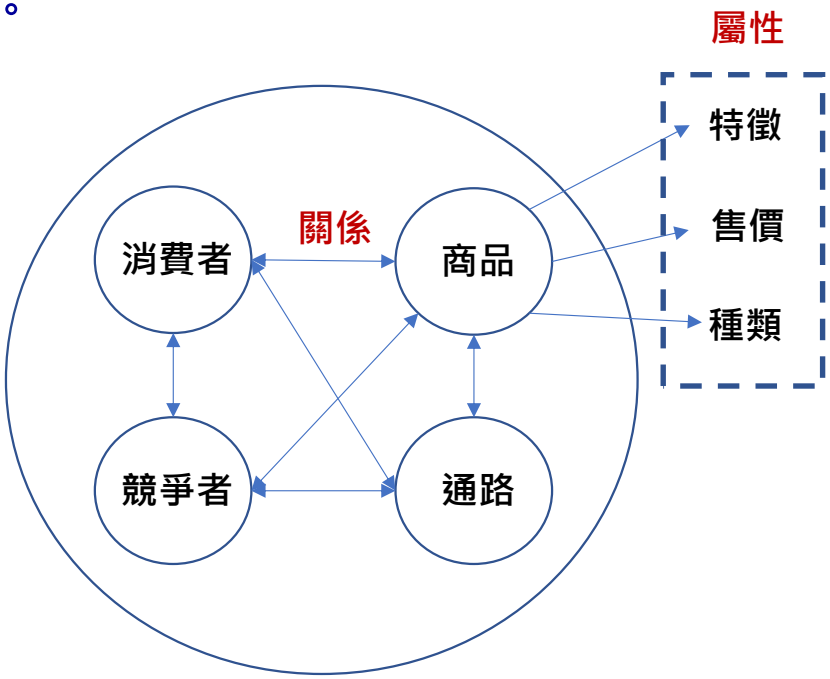
例如，在金融領域，金融分析師可能提供有關金融市場特徵和趨勢的見解。

(5) 關聯分析法

關聯分析：

關聯分析是用於發現數據中不同特徵之間的相互關聯性的方法。這有助於確定哪些特徵經常一起出現，以便將它們組合成更有信息價值的特徵。

例如，如果市場銷售數據顯示某兩個產品經常一起購買，可以考慮將它們作為一個特徵。



資料採集

(1)問卷調查與訪談 (Surveys and Interviews)
(2)網頁抓取 (Web Scraping)
(3)感測器與物聯網 (IoT Sensors)
(4)API接口 (APIs)
(5)實驗與觀察 (Experiments and Observations)
(6)公開資料集 (Public Datasets)
(7)企業內部資料 (Internal Data)
(8)資料錄製與日誌 (Data Logging and Logs)
(9)社交媒體分析 (Social Media Analysis)
(10)圖像和視覺數據 (Image and Visual Data Collection)

問卷調查與訪談 (Surveys and Interviews):

- 利用問卷或訪談的方式直接從受訪者處收集資料。
- 可選擇線上或線下的方式進行調查。
- 主要應用於社會科學、市場調查等領域。

網頁抓取 (Web Scraping):

從網頁上提取資料，這樣的方法通常是自動化的，使用爬蟲技術來抓取公開的網頁內容。
適用於收集各種公開資訊，如新聞、社交媒體數據、商品價格等。

感測器與物聯網 (IoT Sensors):

通過安裝在物理環境中的感測器收集資料，如溫度、濕度、位置等實時數據。
這在智能城市、健康監控、製造業等領域中非常常見。

API接口 (APIs):

通過與其他系統、應用程式或服務的API接口，獲取結構化數據。
這些數據來源可以是社交媒體平台、金融數據庫、天氣服務等。

實驗與觀察 (Experiments and Observations):

- 在受控的實驗環境中收集資料，這通常是針對特定的變數進行觀察。
- 這種方法在科學研究中比較常見，例如生物學、心理學、物理學等領域。

公開資料集 (Public Datasets):

利用政府或研究機構提供的公開資料集來進行分析。
許多政府機構、學術機構和公司會提供可公開使用的數據，如政府數據門戶網站、學術資料庫等。

企業內部資料 (Internal Data):

組織或公司內部的各種資料，如銷售數據、客戶資料、交易記錄等。
通常這些資料會通過企業系統自動收集並進行存儲。

資料錄製與日誌 (Data Logging and Logs):

通過自動化系統記錄事件和過程數據。例如伺服器、網站和應用程序的日誌記錄。
用於收集運營資料、監控資料、錯誤追蹤等。

社交媒體分析 (Social Media Analysis):

- 收集來自社交媒體平台（如Twitter、Facebook、Instagram等）的資料。
- 通常使用API來抓取公開的貼文、評論、標籤等數據，並進行情感分析或趨勢分析。

圖像和視覺數據 (Image and Visual Data Collection):

- 通過拍攝或視頻錄製收集資料，這通常用於人臉識別、物體識別等領域。
- 在自駕車、醫療影像分析等領域尤其重要。

資料清理

(1) 資料縮減 (data reduction)
(2) 離群值偵測 (outlier detection)
(3) 漏值填補 (missing value imputation)

(1)資料縮減

- 「資料縮減」是將冗餘(redundant)的「特徵」或重複(duplicate)記錄的「觀測值」進行移除
- 這些資料原則上不具資訊量，並且它們將增加額外的計算負擔
- 冗餘是指某一特徵可以由另外一個特徵透過數學關係或常理推導而出

- 例如「地區」(北中南東離島) 與「地址」兩個變量中，地區是冗餘的因為可由地址推得:例如「分數」與「及格與否」及格與否是冗餘的，可由分數推得(例如及格為大於等於60分)

➔ 實務上未必能直接將這種冗餘的特徵直接刪除，因為此冗餘特徵可能強化了模型預測的準確度但同時須留意該冗餘特徵所帶來的「維度災難(詛咒)(curse of dimensionality)」或「共線性」等潛在問題。

維度災難與共線性問題

維度災難(詛咒) (curse of dimensionality)

- 在機器學習和數據挖掘時，當維數提高，空間的體積提高太快，造成可用數據變得很稀疏， 致無法獲得可靠的分析結果

共線性 (collinearity)

- 兩個(含)以上的自變項的相關性過高時，會導致主要的影響因子受到不良影響，導致目標變數受到不良影響
- 當線性迴歸模型、羅吉斯迴歸模型，甚至是層數低等非樹狀的神經網路有許多變數時，每個變數都會造成潛在造成共線性，因此辨識每個變數是否可能有共線性是很重要的

資料縮減方法

資料聚合、特徵挑選、維度縮減、數量縮減

1. 資料聚合(data aggregation)

從某一個特徵角度，將其不同類別的「觀測值」聚合

例如可將機台(tool)聚合機台群組(tool group)計算統計量(例如加總、平均值、異數)，或可將每日生產批量聚合成每月生產批量的統計量

2. 特徵挑選(feature selection)或 3. 維度縮減(dimension reduction)

- 特徵合併

例如從溫度、壓力、濕度、濃度流量中挑選出影響良率的重要工程參數，或透過與變數(Y)計算與特徵(X)的相關係數並將相關係數接近0的特徵刪除

- 特徵刪除

例如將國語、數學、社會、自然的考科四科加總變「總分」(也就是維度縮減從四變一)或可將量測值長、寬、高的三個量測值加總來判斷「良品與否」(也就是維度縮減從三變一)

資料縮減方法

資料聚合、特徵挑選、維度縮減、數量縮減

4. 數量縮減(numerosity reduction)

- 此為觀測值或樣本數量的縮減，直覺上可透過抽樣(sampling) (i.e. 分層抽樣)來抽取有代表性(representative)的樣本以減少資料量。

- 離散型資料: (1)樣本某重要特徵中「水準」只出現一次的予以刪除，這是由於該樣本在此特徵沒有再現性(reproducibility); (2)可透過概念階層(concept hierarchy)，找出較高層次的概念(也就是一般化generalization，將部分類別合併計算。

- 連續資料: 可透過離散化方法(discretization) 將連續資料依其特徵值分為若干區間(裝箱分類)，並計算每個區間的統計量。如果為高頻時序數據(例如振動訊號)，可透過 (1)下抽樣(down sampling)、(2)移動平均法(moving average)、(3)滑動時窗(sliding window)等方法，縮減資料樣本。

下抽樣(down sampling)

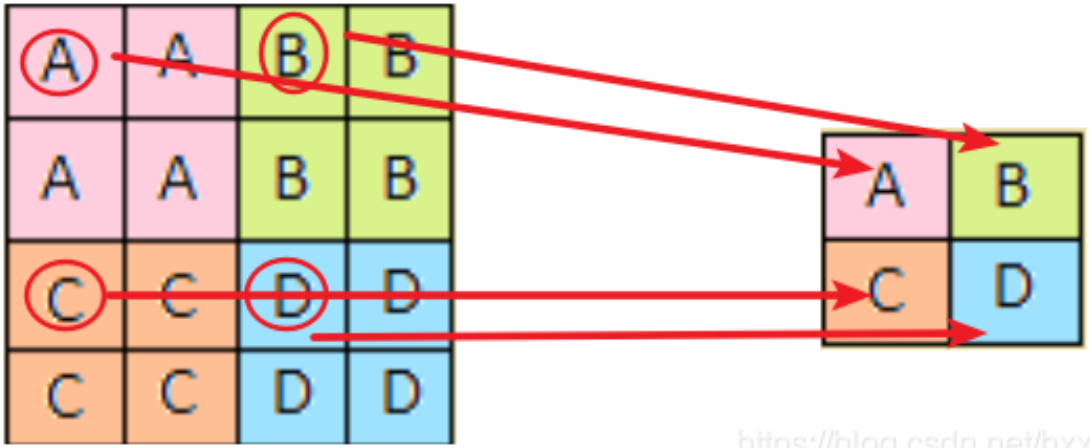
(1)下抽樣(down sampling)

降低資料取樣率或解析度的過程。單個時間序列在一個時間範圍內的多個數據點在一個對齊的時間戳中與數學函式一起聚合成單個值。

向下採樣至少需要兩個元件：

- 時間間隔(Interval): 一個時間範圍，間隔以「大小」「 單位」格式指定，例如，可以將多個值匯總1分鐘或1小時

- 聚合函數: 確定如何合併區間中的值的數學函數



下抽樣（downsampling）是一種在數據處理中減少數據點數量的方法，通常用於縮小數據集的大小，以減少計算成本或改善數據可視化。抽樣可以以不同的方式執行，最常見的方法是從原始數據集中選擇一個子集，以保留重要的信息，同時減少數據點數量。

以下是一些抽樣的示例：

1.隨機抽樣：從原始數據集中隨機選擇一定比例的數據點，以形成較小的子集。例如，如果有一個包含1000個數據點的數據集，可以隨機抽樣10%的數據點，得到一個包含100個數據點的子集。

2.等間距抽樣：以固定間隔從原始數據集中選擇數據點，以減少數據點的數量。例如，如果有一個時間序列數據集，每天記錄一個數據點，可以進行等間距抽樣，每隔一週保留一個數據點，從而減少數據點的密度。

3.聚合抽樣：將原始數據按某種方式進行聚合，例如取平均值或最大值，以減少數據點數量。例如，如果有一個包含每小時溫度數據的數據集，可以將其聚合為每日平均溫度，從而減少數據點的數量。

下抽樣的目的是在減少數據點數量的同時，盡量保留原始數據的關鍵特徵和趨勢，以便在後續分析中仍能提供有價值的信息。但需要謹慎選擇抽樣方法，以確保不會失去重要的數據。

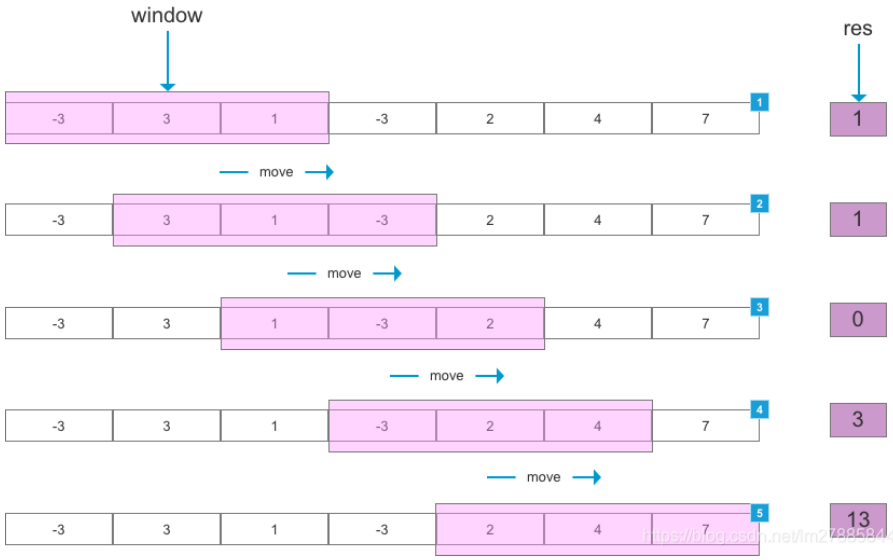
移動平均法與滑動時窗

(2)移動平均法(moving average，簡稱MA）

- 利用統計分析的方法，將一定時間內的數值加以平均並將不同時間的平均值連接起來，便得到了移動平均線
- 常見方法分為算術平均法、加權平均法和指數平滑移動法三種

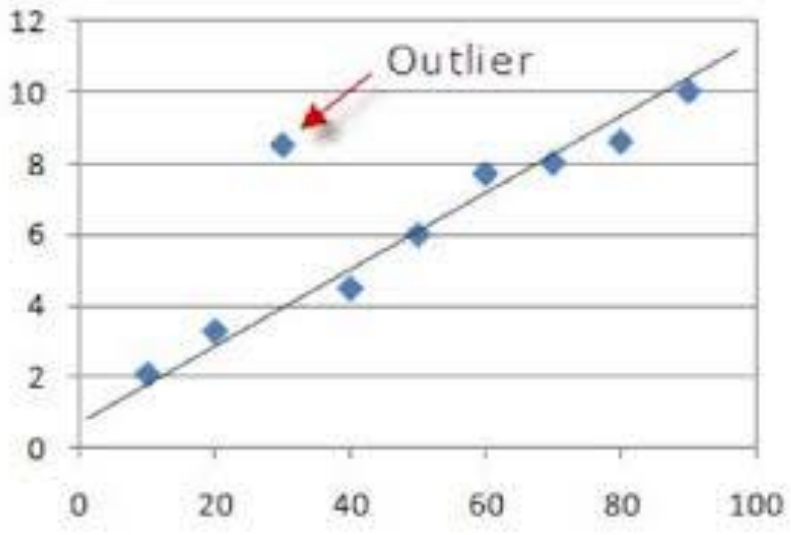
(3)滑動時窗(sliding window)

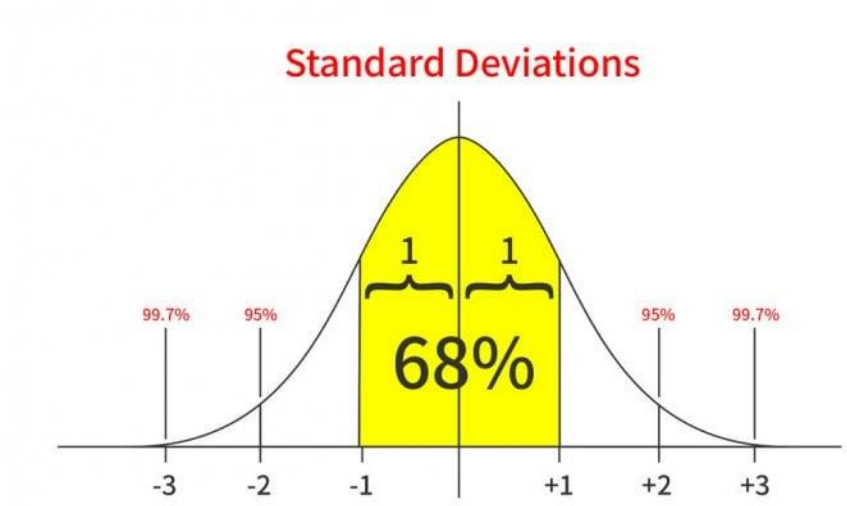
- 通常應用於時間序列數據或空間數據，藉由一次移動一個固定大小的窗格，來對數據進行平滑或抽樣，以減少數據點的數量，並提高計算效率



(2)離群值偵測與處理

- 離群值是超出其類型預期常態的極端數據點，基本上是指一個或一組與預期樣本和模式有顯著差異的數據點
- 可能是整個數據集太過混雜所致，也可能是某個數據集的極端，如下圖，離群值指最右側和最左側的數據
- 離群值會破壞整個樣本的一致性，影響機器學習的效果與分析(預測)的準確性，導致分析的結果難以解釋





(2)離群值偵測與處理

離群值可能代表檢測異常的情況，也可能是測量錯誤、實驗問題、或是單次無意義的新突變。

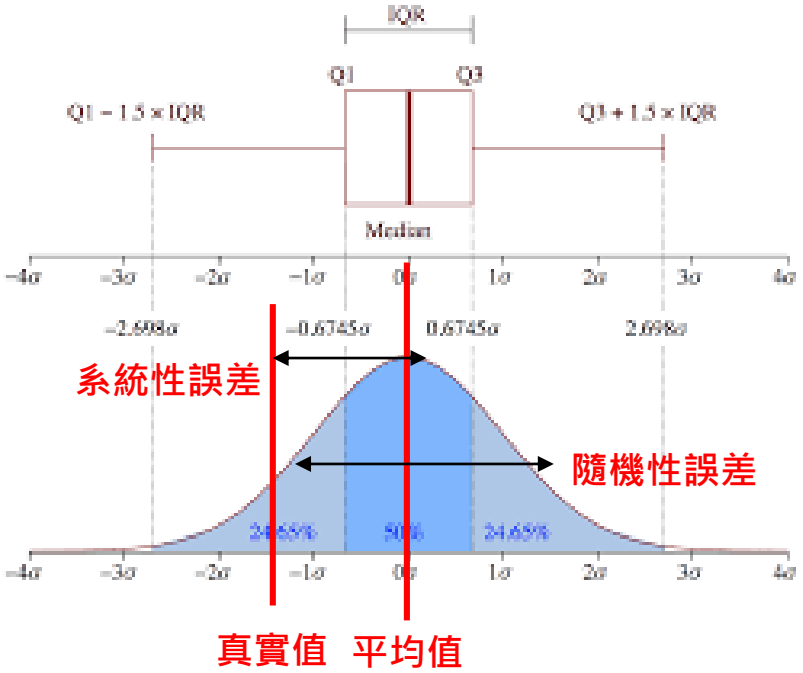
離群值除了會影響到預測的準確性之外，它在某些特殊的應用場景扮演了非常重要的角色，可用於異常偵測與預測。整體來說，就是有一小群特別的資料，但會對整個群體產生重大影響的異常資料，所以離群值的偵測和處理，無論在統計或資料科學，都是一個重要的議題。

例如：
疾病檢測
網路詐騙
信用卡盜刷
網路攻擊

系統性離群值&隨機性離群值

系統性離群值 (systematic outlier)
指的是結構化、有跡可循產生的異常，是非隨機因子造成的離群值，偵測與處理上一般需要領域知識的輔助

隨機性離群值 (random outlier)
指的是自然、任意產生的異常，其隨機性可能合乎某個分配(例如常態分配、伯努力分配)，處理上可用敏感度分析檢查變數的穩健性，以及環境噪音所帶來的影響。



製造業中常見之系統性離群值

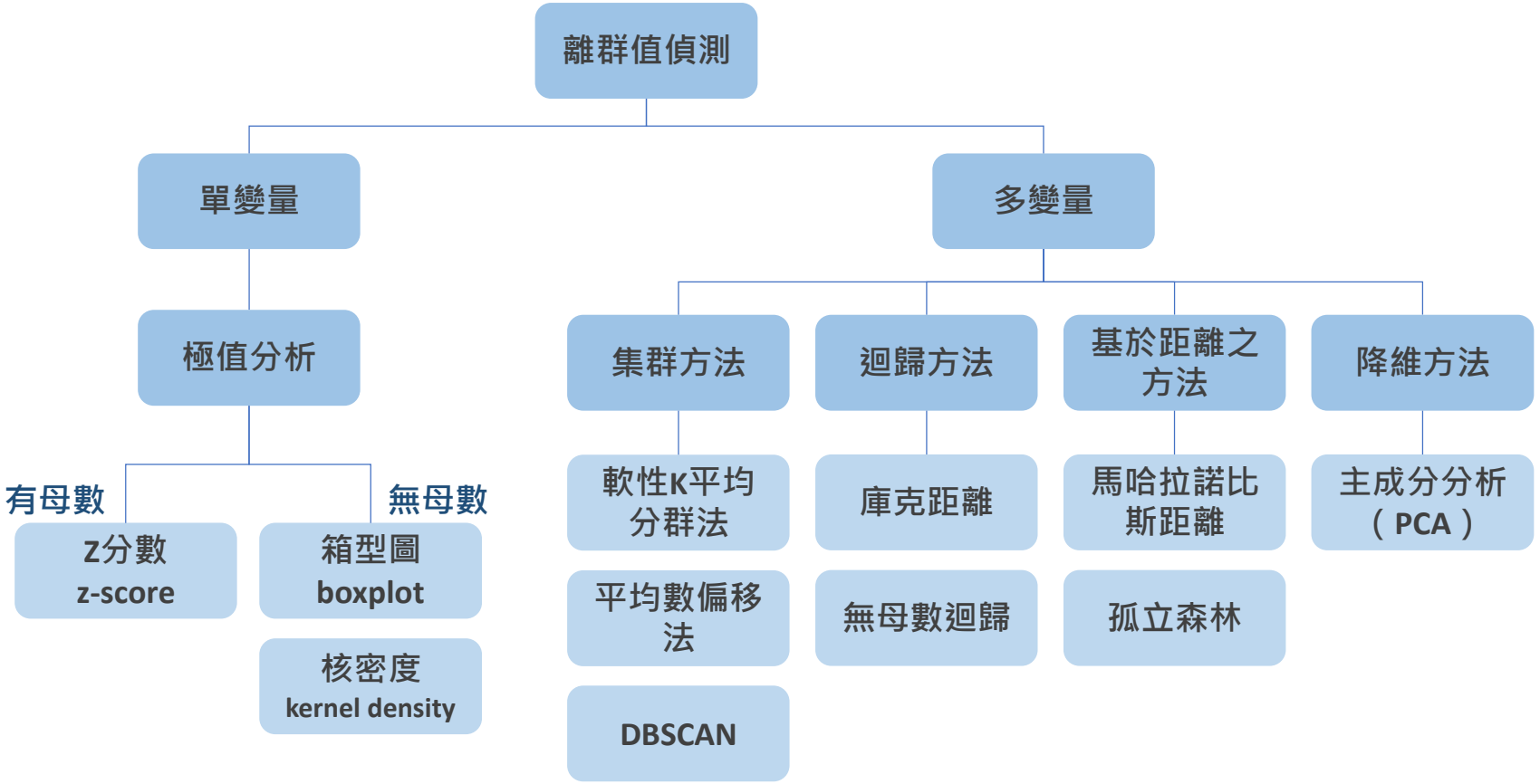
(1) 機台異常 (machine anomaly)
發生於機台異常時，機台製程參數、控制系統以及感測器的資料將產生明顯的離群值，這樣的「系統性離群值」對於製程控制、良率預測以及機台壽命預測與異常診斷等議題上常為重要的觀測值，因此有時並不刪除這些離群值而須保留這些稀少且珍貴的異常資料

(2) 實驗數據/工程數據 (experiment/engineering Data)
研發所產生出的試驗資料與穩定量產的資料差異很大，通常存在顯而易見的離群值。例如實驗貨配方(recipe)的參數設定與一般產品設定差異大，因此在機器學習過程，會將這樣的「系統性離群值」樣本移另做分析較為適

(3) 人為錯誤/干擾 (human errors/interference)
可能是人為資料填寫key-in錯誤或是環境干擾到感測器對資料的收集，例如堆高機從機台旁邊開過，使感測器收到異常訊號，這樣的「系統性離群值」通常不易判別，須深入瞭解製造現場的作業流程與環境等情形(有時會將它們視為遺漏值處理)

離群值偵測方法

離群值的偵測主要方法包含了單變量的極值分析(Extreme Value Analysis)以及多變量的集群分析(Clustering)、迴歸分析(Regression)、基於距離的方法(Distance-Based)以及降維方法(Dimensionality Reduction)



有母數分析 & 無母數分析

有母數分析方法(檢定)

- 資料是假設取樣資料遵循的機率分佈是根據一組參數
- 最常見的假設是資料為常態分佈
- 大部分的一般統計方法為有母數，t 檢定、變異數分析、線性迴歸、Pearson相關係數等皆屬之

無母數分析方法(檢定)

- 無法符合有母數分析所設計的方法
- 常使用符號(正負)或排序(大小順序)取代測量數值，或使用各分類的次數以進行統計分析
- 適用於類別、序位尺度資料分析與資料分布未知的情況

無母數分析之優缺點

無母數分析方法優點

- 母群體分布未知或不是常態分布，或是樣本數不夠大時皆可使用。是無母數分析方法的最大優點。
- 計算簡單且快速。
- 雖然在母群實際上為常態分配時，較有母數分析方法不易得到顯著結果；但在母群體不是常態分布時，無母數分析方法之檢力較有母數分析方法高。

無母數分析方法缺點

- 只使用資料的符號、排序等特性，浪費了數值之集中趨勢、分散性及分佈所提供的資訊。
- 針對常態分布資料如果仍進行無母數分析，將使檢力降低。
- 當欲檢定的資料不符合有母數分析法之假設前提時才建議使用無母數分析法，為一種互補的統計方法，而非用於取代有母數分析法。

Z分數 (Z-Score)

Z分數是一種將原始分數以「在平均數之上或之下幾個標準差」的方式表示分數，意即我們可以透過Z分數知道個體位於群體中的相對位置，Z分數的算式為：

$$z = \frac{x - \mu}{\sigma}$$

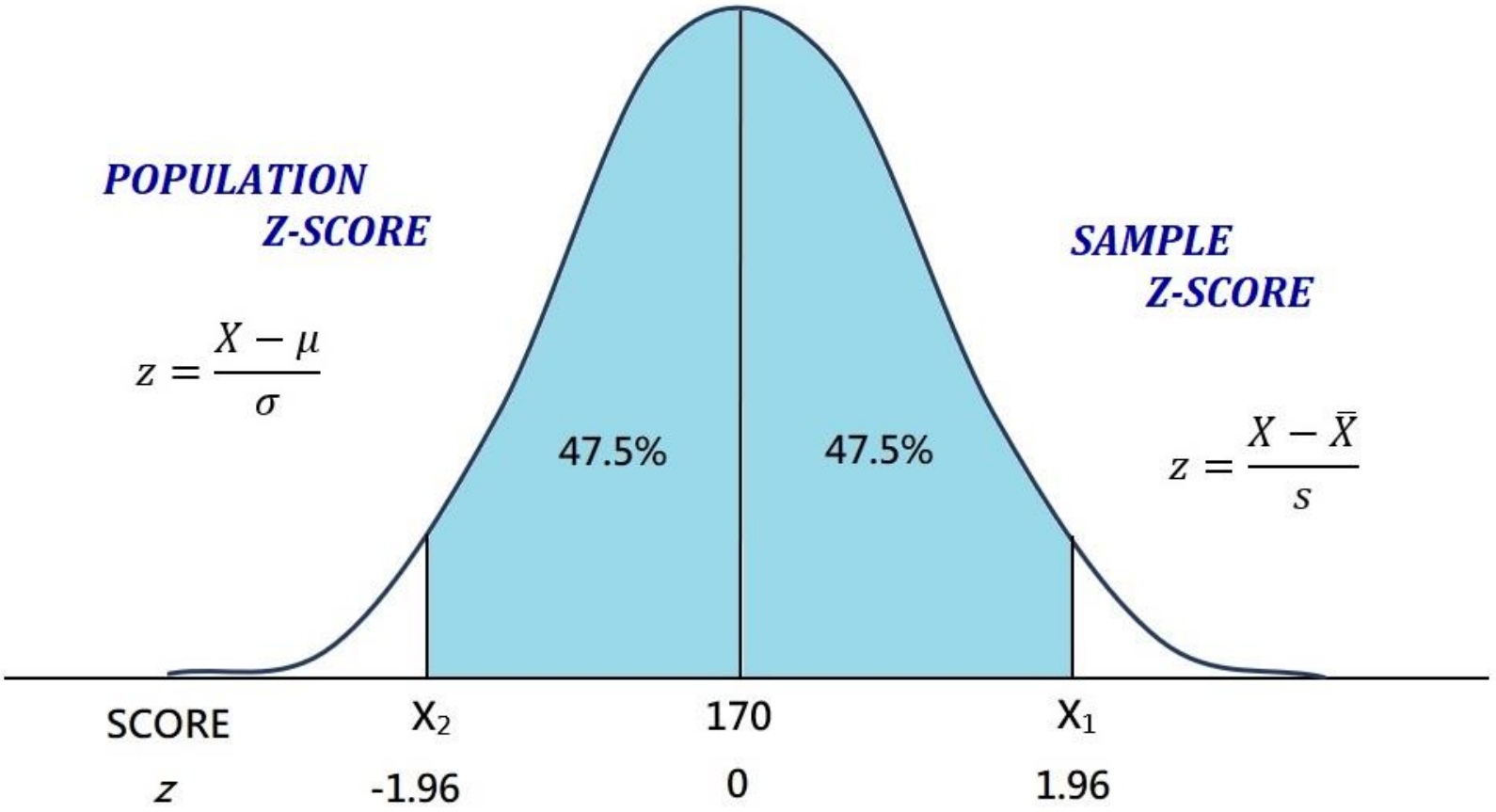
註：*μ為平均數, σ為標準差* (將原始分數減去平均數後除以標準差就能夠得出Z分數)

舉例：
小明這次數學考了90分，班上平均為80分，標準差是10分
那麼小明的Z分數就是1，意思是小明的分數比全班平均多了1個標準差
小美這次考70分，Z分數就是-1，也就是比平均數還少"1"個標準差

小明 $\frac{90-80}{10} = 1$　　小美 $\frac{70-80}{10} = -1$

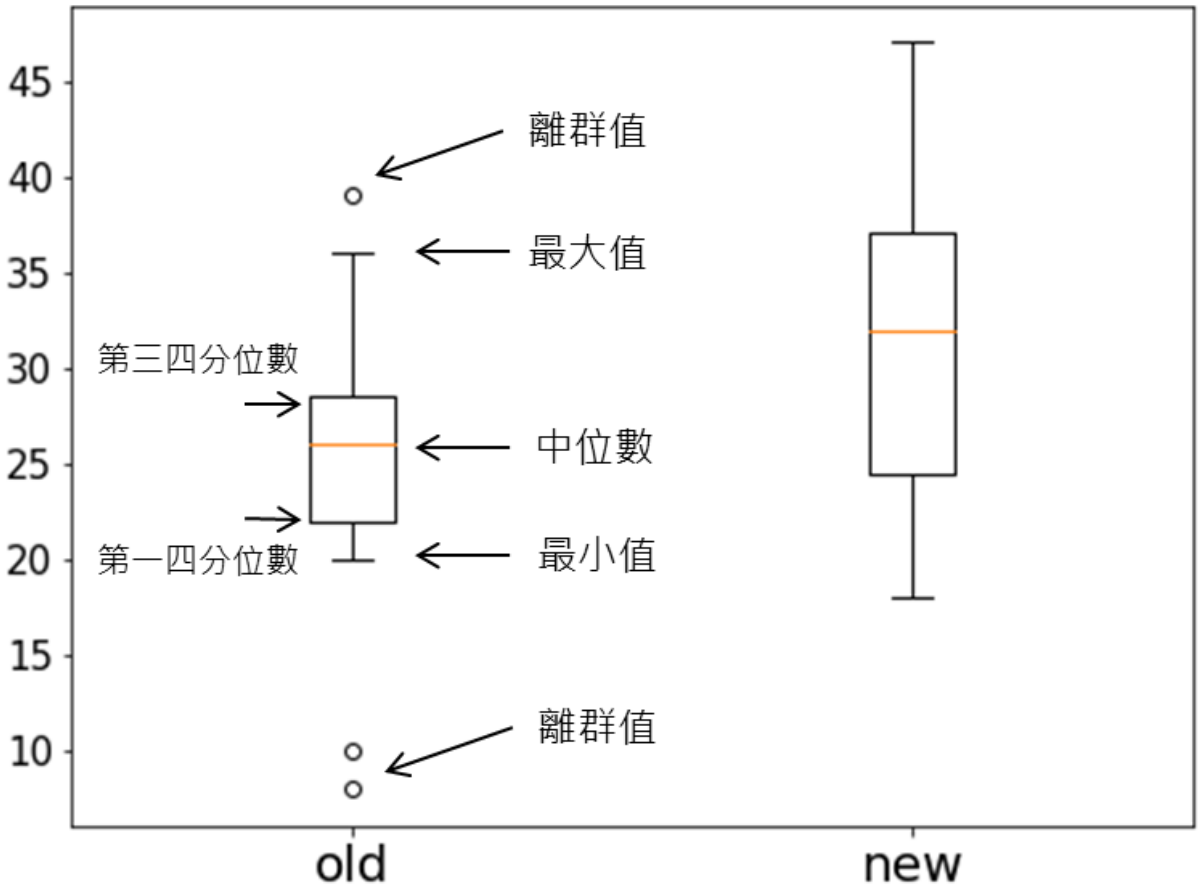
➔ Z分數的正負可以判斷個人的成績是否高於平均數
➔ Z分數的絕對值可以判斷距離平均數的差距有多遠

Z分數 (Z-Score)


POPULATION
Z-SCORE
$z = \frac{X - \mu}{\sigma}$
SAMPLE
Z-SCORE
$z = \frac{X - \bar{X}}{s}$
47.5%
47.5%
SCORE
X_2
170
X_1
z
-1.96
0
1.96

箱型圖 (Boxplot)

箱型圖（英文：Boxplot），又稱為盒鬚圖，是一種用作顯示一組數據分散情況資料的統計圖。因圖形如箱子，且在上下四分位數之外常有線條像鬍鬚延伸出去而得名。

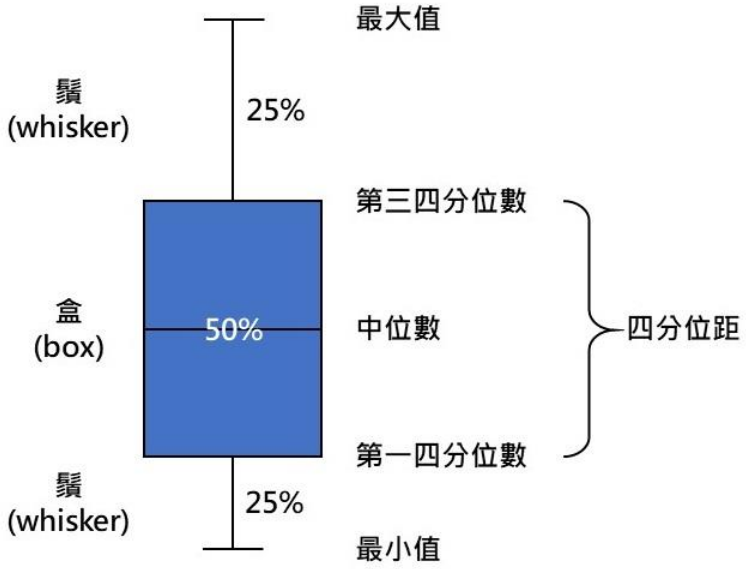


箱型圖 (Boxplot)

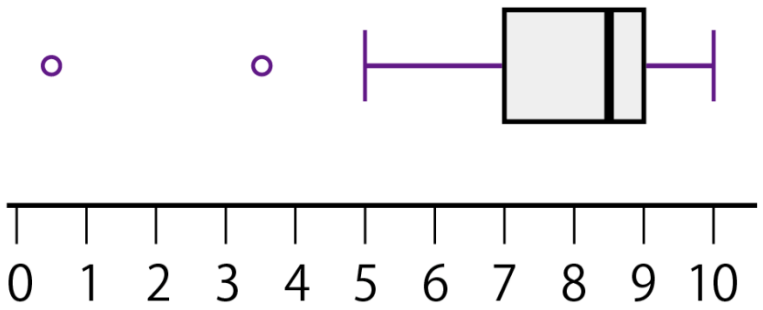
「箱型圖」未假設任何分配(但較適用於集中的分配)，能顯示出一組數據的最大值、最小值、中位數、四分位數

先找出樣本的中位數 (median)後，推算25%的「第一分位數」(first quantile, Q1)與 75%的「第三分位數」(third quantile, Q3)，兩者相減後便可得到「四分位間距」(interquartile range，IQR)，Q3-Q1，代表著資料的變異程度(與標準差相似)

進一步以1.5倍的「四分位距」推算「最大值」與「最小值」，而在極值外的便是離群值



箱型圖 (Boxplot)



這組數據顯示出：

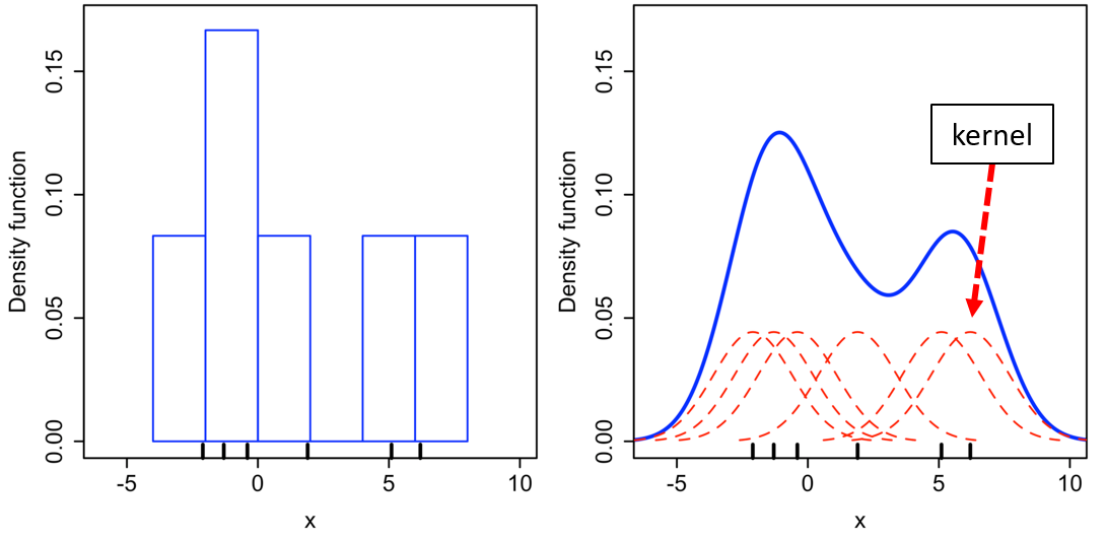
- 下邊界=5
- 第1四分位數（Q1）=7
- 中位數、第2四分位數（median、Q2）=8.5
- 第3四分位數（Q3）=9　上邊界=10　平均值=8
- 四分位間距（interquartile range，簡稱IQR）= Q3-Q1=2 （即ΔQ）

當數值與第1與第3四分位數的範圍差距1.5×IQR以上時，該值為離群值(outlier)
數值位於範圍外1.5×IQR到3×IQR範圍的數值，稱作適度離群值(mild outlier)
數值位於範圍外3×IQR以上的數值，稱作極端離群值(extreme outlier)

核密度 (Kernel Density)

核密度則同樣未假設任何分配可適用於多峰分配，並更進一步以無母數的方式估計完整的分配，並可依照樣本的密度來排序離群值的可能性。

核密度估計其實是對直方圖的一個延伸應用。把直方圖轉畫成折線圖，資料分布從離散型的資料轉為連續型的資料分佈，可以觀察數據的分布與趨勢。



集群分析 (Cluster Analysis)

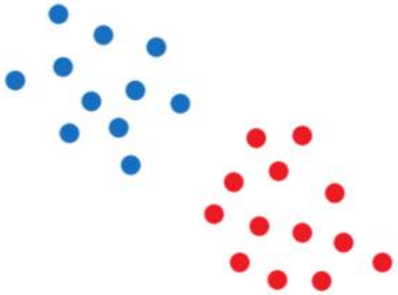
集群分析(Cluster Analysis) 在離群值偵常用「軟性K平均分群法(soft k-means clustering)、「平均數偏移法」(mean shift)以及「DBCAN」(density based spatial clustering of applications with noise)三種方法，使同時考慮多個特徵的聯合分布來分析離群值。

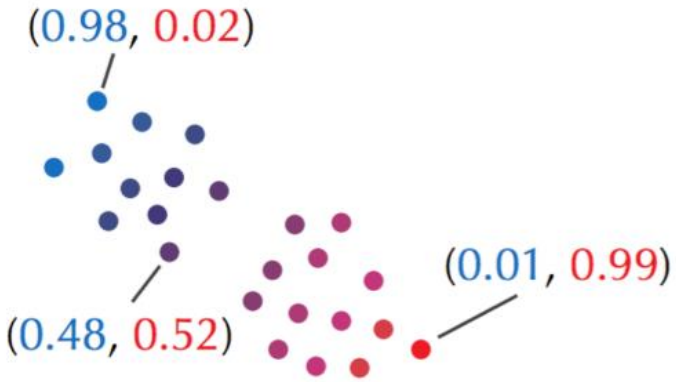
前兩者均假設了每個集群各別服從一個機率分配，並以密度大小排序離群值的可能性

後者「DBSCAN」則無假設任何分配，以樣本間的密度與連結性進行分群，並將未連結的樣本視為離群值。

軟性K平均分群法 (Soft k-means clustering)

- Soft k-means clustering 是 K-means 聚類的變體
- 傳統的 K-means 聚類將每個資料點分配到唯一的最近中心，而軟 K-means 允許資料點以不同程度屬於多個群集
- 這種多重歸屬性是根據每個資料點到每個中心的距離（或相似性分數）來計算的，通常使用高斯分布函數



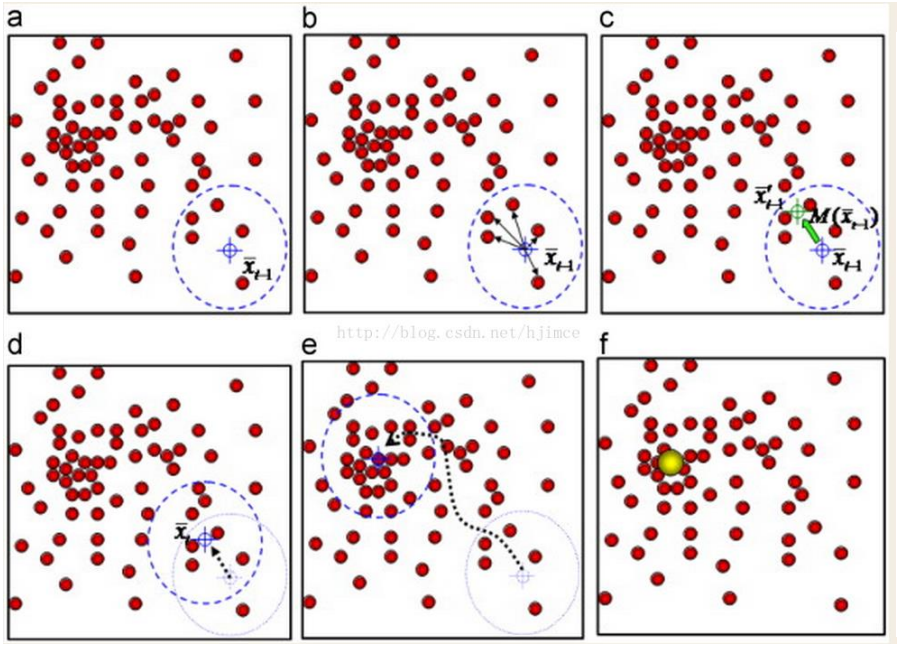


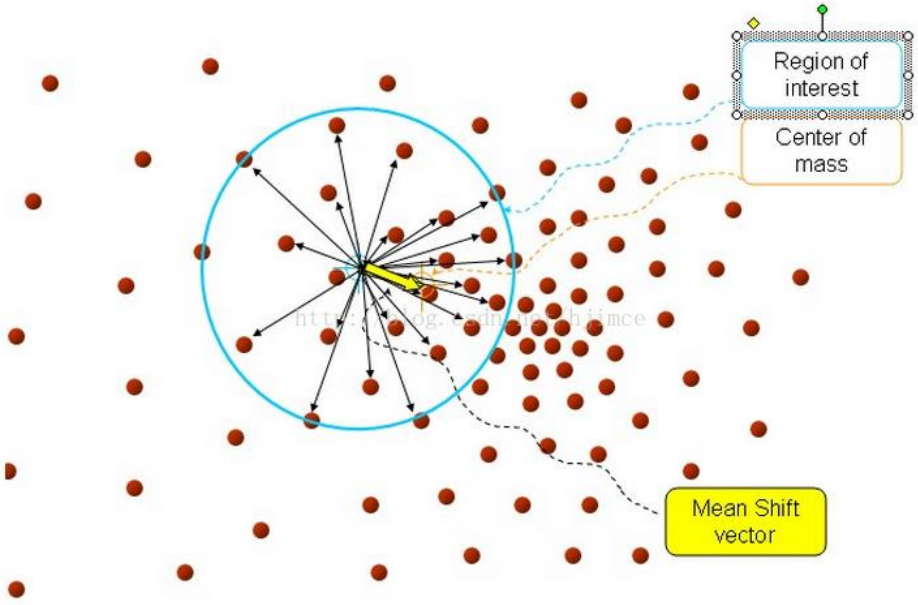
Hard choices: points are colored red or blue depending on their cluster membership.

Soft choices: points are assigned "red" and "blue" *responsibilities* r_{blue} and r_{red} ($r_{blue} + r_{red} = 1$)

平均數偏移法 (mean shift)

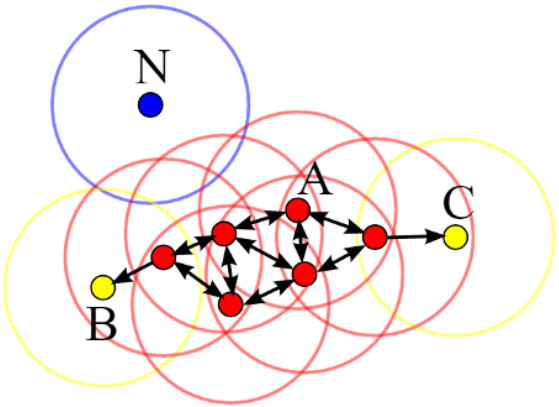
Mean Shift 是基於密度的聚類演算法，其演算法思想是假設不同聚類的資料集符合不同的概率密度分佈，找到任一樣本點密度增大的最快方向（最快方向的含義就是Mean Shift），樣本密度高的區域對應於該分佈的最大值，這些樣本點最終會在局部密度最大值收斂，且收斂到相同局部最大值的點被認為是同一聚類的成員。





DBSCAN
(Density-Based Spatial Clustering of Applications with Noise)

DBSCAN演算法，係根據資料點在特徵空間中的密度進行聚類
給定某特徵空間裡的一個資料點集合，演算法會把附近的點分成一組，並標記出位於低密度區域的局外點，就能把一些離聚集點比較遠的點自動被當做雜訊(noise) /離群點(outliers)剔除



- eps: neighborhood radius
- min_samples: 4

- A: Core
- B, C: not core
- N: noise

從任意一個點出發，以上圖而言假設從A出發，搜尋A周圍eps範圍以內的「資料數量」，當前的eps範圍裡有超過min_samples個資料時，我們就認為A是一個Core，然後開始去對A的eps範圍內的其他資料做一樣的事情，直到現在某一個點的eps範圍內不具備min_samples數量的點了停止。而如果今天出發的點是N，則在最一開始周圍就找不到足夠數量的點，所以N就會被判斷為Noise。

迴歸分析 (Regression Analysis)

迴歸分析包括了「庫克距離」(Cooks distance)以及「無母數迴歸」(non-parametric regression)兩種方法。

「迴歸」的目的是找到依變數或是應變數(outcome)Y 與一個或是一個以上的自變數X的函數關係。

一般表示為 $Y=f(X)+\epsilon$

- Y是依變數
- X=(X1,X2,…,Xp)為自變數
- f代表固定但是未知的函數，也就是Y與X1,X2,…,Xp之間的關係
- ϵ是隨機變數，包含沒有測量到或無法測量的變數，且獨立於X

庫克距離 (Cook's Distance)

- 一種在線性迴歸分析中用於識別影響迴歸模型的個別數據點的方法
- 用於檢測對迴歸係數估計具有重要影響的離群值或極端觀察值
- 可評估數據中的異常點，它衡量了每個數據點對迴歸模型的影響程度。距離越大，代表該觀察值對於迴歸模型的擬合影響越大，可能被視為離群值

ei: 第 i 個觀察值的殘差
p: 回歸模型中，自變數的數目
MSE: 殘差的均方和
hii: 第 i 個觀察值的槓桿值

$$D_i = \frac{e_i^2}{(p+1) \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

庫克距離的計算方式是通過刪除每個單個觀察值，然後計算在刪除觀察值後迴歸模型參數的變化。如果刪除某個觀察值導致迴歸模型的係數發生重大變化，則該觀察值的庫克距離較大

無母數迴歸 (non-parametric regression)

無母數回歸（Nonparametric Regression）是一種統計建模方法，用於探索和建立變數之間的關係，相對於傳統的參數化回歸方法，無母數回歸具有更大的靈活性，可以更好地處理複雜的數據結構，包括離群值。

- 無需事先假設數學模型：

不要求在建模之前明確指定數學模型，適合處理數據中的非線性或複雜關係，以及未知的數據結構

- 基於局部信息的估計：

通常基於局部信息，關注每個數據點周圍的局部結構，而不是全局趨勢，因此可捕捉數據中的細微變化

- 較好處理異常值：

對於數據中的離群值更具魯棒性，不易受到單個異常值的干擾

- 典型方法：

局部加權線性回歸 (Local Weighted Linear Regression，LWLR)、核回歸 (Kernel Regression)、K-最近鄰回歸 (K-Nearest Neighbors Regression)等

馬哈拉諾比斯距離 (Mahalanobis Distance)

馬哈拉諾比斯距離（Mahalanobis Distance），也稱為馬氏距離，是一種多變量距離度量，它用於衡量兩個多維數據點之間的相似性或距離。

- 考慮變數之間的相互關係：

考慮了數據中不同變數之間的協方差結構，能夠捕捉多變量數據的相互關係，適用於具有不同變異性和共變異性的數據集

- 多變量離群值檢測：

常用於識別多變量數據中的離群值。當一個數據點的馬哈拉諾比斯距離相對較大時，它可能被視為離群值，因為這表示該數據點在多變量空間中遠離數據集的中心。

- 可用於分類和聚類：

可衡量不同類別之間的相似性或區分度，因此在多類別問題中特別有用。

- 特徵選擇和降維：

可用於特徵選擇和降維，以確定哪些變數對於區分不同數據點最具價值。

馬哈拉諾比斯距離 (Mahalanobis Distance)

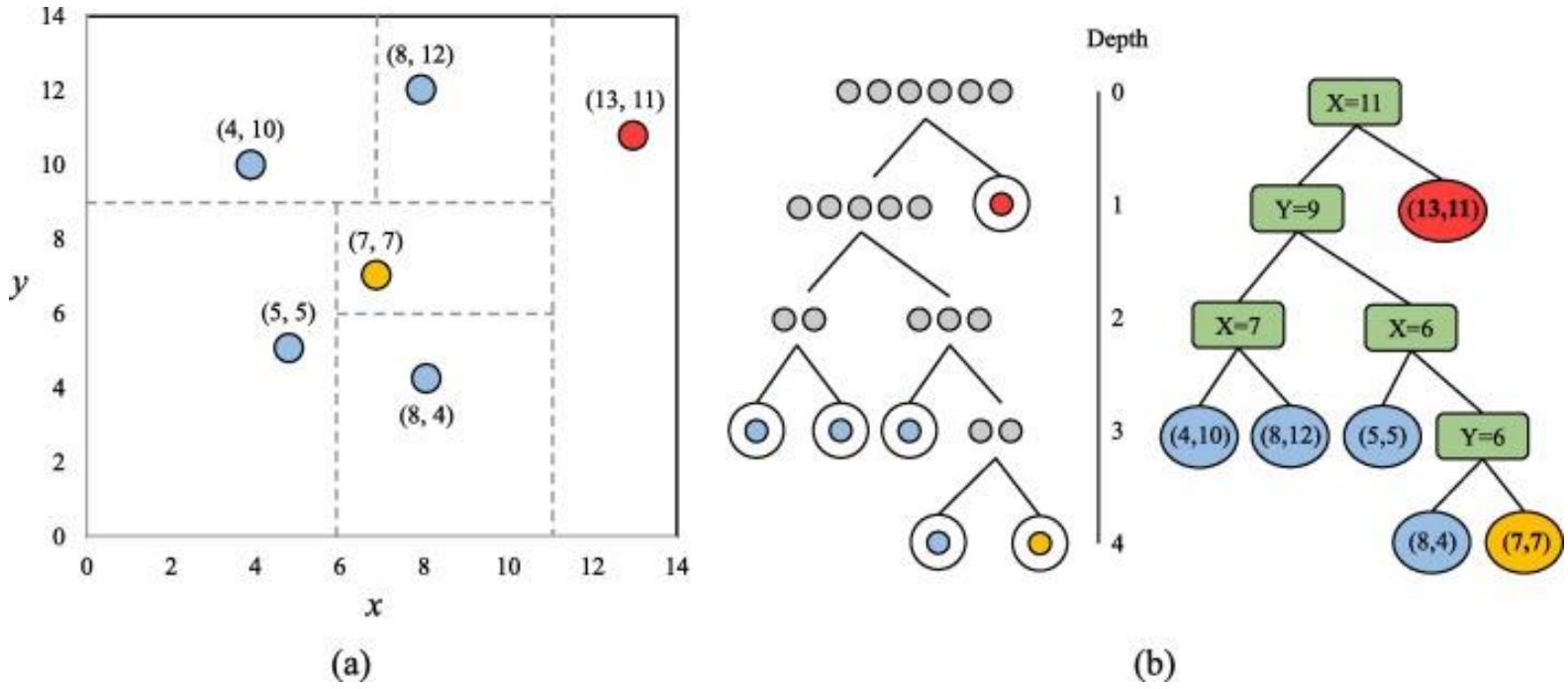
馬哈拉諾比斯距離的計算方式如下：

$$\mathrm{Dis}_{\mathrm{mahalanobis}}(\mathrm{x}_1, \mathrm{x}_2) = \sqrt{(\mathrm{x}_1 - \mathrm{x}_2)\Sigma^{-1}(\mathrm{x}_1 - \mathrm{x}_2)^{\mathrm{T}}}$$

對於具有n個變數的多變量數據，假設x1和x2分別是兩個數據點（觀察值）的向量，每個向量包含n個變數的值。
協方差矩陣Σ表示數據的協方差結構，而μ表示數據的平均值向量。

孤立森林 (Isolation Forest)

- Isolation Forest是一種基於樹狀結構的離群值檢測方法
- 透過隨機切割數據，將數據分割為子集，然後隔離離群值
- 由於離群值通常需要更少的切割操作才能被隔離，因此Isolation Forest能夠相對較快地識別多變量數據中的離群值
- 特別適用於高維資料集。



孤立森林 (Isolation Forest)

孤立森林運作步驟：

- 隨機選擇特徵和數據點

隨機選擇一個特徵，在該特徵的值範圍內隨機選取一個閾值，再隨機選擇一個數據點，比較它的特徵值和閾值，然後重複此過程，直到數據點被隔離或達到最大樹深度。

- 樹狀結構的構建

隔離森林使用二叉樹結構，每個節點表示一個特徵和閾值的組合。樹的深度通常是一個超參數，可以根據需求調整。

- 隔離過程

在樹的構建過程中，數據點被隨機分配到不同的子樹。通過比較特徵值和閾值，樹的分支將數據點分為兩個子集，其中一個子集包含較少的數據點，而另一個子集包含較多。異常值通常需要較少的分支步驟才能被隔離，因此在樹的葉節點處的深度相對較小。

- 異常值得分計算

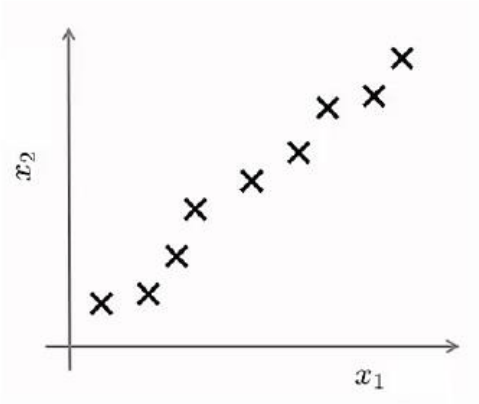
計算每個數據點的得分，該得分表示從根節點到達該數據點所需的平均分支數。得分越低，表示數據點越容易被隔離，因此越可能是異常值。

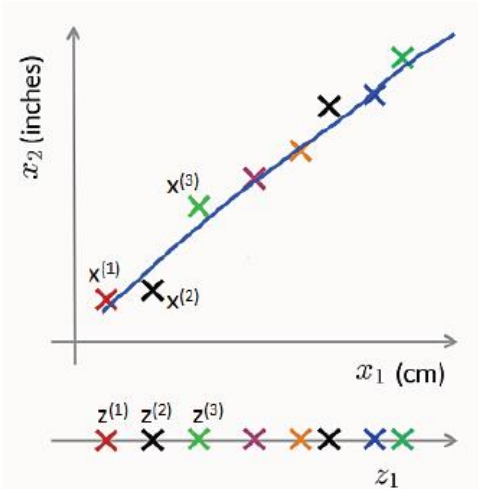
- 閾值設定

根據應用需求設定閾值，得分高於閾值的數據點被視為正常數據，而得分低於閾值的數據點被視為異常值。

降維方法 – 主成分分析

- 主成分分析 (Principal Component Analysis，PCA)是一種用於多變量數據分析的降維技術
- 透過尋找數據中的主成分(特徵向量)，來將高維數據映射到低維空間
- PCA可以保留最重要的變異性，幫助識別多變量數據中的模式，包括離群值
- 當數據點在新的低維空間中偏離正常分佈時，它們可能被視為離群值
- 可視化多變量數據，幫助識別數據中的異常模式





- 壓縮資料可進而組合新的、抽象化的特徵，減少冗余的資訊
- 左圖的x1與x2高度相關，因此可合併為一個特徵（右圖）
- 把x(i)投影到藍色的線，從二維降低為一維

降維方法 – 主成分分析

主成分分析的關鍵思想包括以下幾點：

- **特徵值分解**

PCA基於線性代數的特徵值分解方法，通過對數據協方差矩陣進行特徵值分解，找到數據中的主成分。

- **主成分的選取**

特徵值分解產生了一組特徵向量，稱為主成分。這些主成分按照特徵值的大小排列，最大的特徵值對應的主成分包含了最多的數據變異性。因此，選擇前幾個主成分可以實現數據的降維，同時保留大部分信息。

- **數據轉換**

通過將原始數據投影到所選的主成分上，可以將高維數據轉換為低維表示。這種轉換使我們可以更容易地視覺化數據或對數據進行後續分析。

- **降維和信息保留**

PCA的一個主要應用是降維。通過選擇較少的主成分，我們可以實現數據的降維，同時盡量保留原始數據的變異性。這有助於減少計算複雜性，刪除噪聲，並提高數據分析的效率。

- **數據壓縮**

PCA還可以用於數據壓縮，特別是在存儲和傳輸大量數據時。通過將數據轉換為主成分表示，我們可以實現數據的壓縮而不失去重要信息。

離群值處理方法

- 在實務上，通常在偵測到離群值後，需先以領域知識判別是否為「系統性離群值」或為「隨機性離群值」
- 若離群值存在時，依據領域知識進一步採取相對應的「治療」(treatment)，行「刪除」(delete)(離群值與目標母體不一致)或「保留」(retain)
- 離群值為潛在重要特徵可能提供重要資訊

觀測值若保留，處理的方式可以如下：

- 保留原始數值
- 資料轉換:對有離群值的特徵取對數(log)進行重新縮放(rescale)，或裝箱分類(binning)
- 視為遺漏值(regard as missing value):可使用遺漏值填補方法或填補為「其他」類別
- 增加一個新的特徵(欄位)並標註離群值，例如新增二元變數欄位，該新增欄位以原特徵欄位為基準，若為離群值則標1，反之則標註0

(3)遺漏值填補

「遺漏值填補」（Missing Value Imputation; MVI）
當資料集包含一個或多個屬性缺失值必須適當將其填補，使成為完整資料集以利後續的資料分析。

根據 Strike et al.（2001）和 Raymond and Roberts（1987）的說法，當不完整資料集包含非常少量的缺失數據時，例如遺漏率小於 10%或 15%，可以直接從資料集中刪除不完整的資料樣本（即案例刪除法），而不會對最終的資料挖掘或分析結果產生顯著影響，但並非每個領域問題資料集都遵循這種規則。

「遺漏值」可分為「系統性遺漏值」(systematic missing value)與「隨機性遺漏值」(random missing value)。前者遺漏值有特定組型或傾向，後者是指遺漏現象完全是隨機發生的，和自身或其它變數的值無關。

(3)遺漏值填補

1. 系統性遺漏值（Systematic Missing Values）：

系統性遺漏值是由於某種固定的系統性原因而出現的遺漏值。這些原因可能與數據的收集過程、記錄方式或數據來源的特性有關

例如：

- 人為錯誤：數據收集者犯了錯誤，導致某些數據未被記錄。
- 數據格式問題：數據來源的格式或規範不一致，導致某些數據無法被提取。
- 硬體或軟體故障：數據收集或儲存過程中的硬體或軟體故障可能導致數據丟失。

2. 隨機性遺漏值（Random Missing Values）：

隨機性遺漏值是出現在數據中的遺漏值，其出現似乎是無規律的，無法歸因於固定的系統性原因。這些遺漏值可能是由於隨機的事件或無法預測的因素導致

例如：

- 訪問數據遺漏：在一個調查中，某些受訪者可能未回答某些問題，這是由於個人選擇或遺漏，無法預測。
- 儀器測量誤差：在科學實驗中，儀器的測量可能存在誤差，導致某些測量值無法獲得。

(3)遺漏值填補

產品ID	機台1-溫度	機台2- 溫度	機台3-溫度
產品01	835	NA	NA
產品02	832	NA	NA
產品03	NA	841	NA
產品04	NA	NA	823

產品ID	溫度	平行機台
產品01	835	1
產品02	832	1
產品03	841	2
產品04	823	3

資料整合

(1) 純特徵對純特徵
(2) 純特徵對時序特徵
(3) 時序特徵對純特徵
(4) 時序特徵對時序特徵

資料整合

「資料整合」是指將資料庫中不同的資料表單整併或串接的過程，以形成一個考量全面資訊的資料大表。

資料整合在串接過程中常用的「主要特徵」可分為兩種：

1. 純特徵

例如產品ID、機台ID)通常為主鍵(primary key)，在包含多個特徵時，稱為組合鍵(composite key)

2. 時序特徵

例如時間、時間區間

上述兩者的配對可以分為四種情形：

	純特徵	時序特徵
純特徵	純特徵對純特徵	純特徵對時序特徵
時序特徵	時序特徵對純特徵	時序特徵對時序特徵

純特徵對純特徵

「純特徵對純特徵」依據主要特徵進行合併

如下例:
僅需依照「產品編號」進行串接。然而，可能面臨到串接後因兩資料集的共同「產品編號」非完全一對一而產生遺漏值，因此事後需進行遺漏值填補(若某一產品串接後漏值過多可直接刪除)

產品ID	特徵A
01	11
02	12
03	13

+

產品ID	特徵B
01	17
02	16
04	15

=

產品ID	特徵A	特徵A
01	11	11
02	12	12
03	13	NA
04	NA	15

純特徵對時序特徵

把純特徵當主表單以串接時間序列特性的特徵(或訊號)而轉出純特徵大表

此時的串接方法將對時序特徵進行特徵工程(例如萃取出平均數、標準差、偏態、峰態等手法)，將原本的「產品編號」一對多個時間的特徵，透過設定時間區間計算統計量(例如三個時間點樣本計算平均值)，轉成對應到單一數值的特徵的形式，因而使得兩資料集能進行合併

產品ID	特徵A
01	11
02	12
03	13

+

產品ID	時間	訊號A
01	12:00	17
01	12:30	16
01	13:00	15
02	12:00	232
02	12:30	200
02	13:00	168
03	12:00	35
03	12:30	44
03	13:00	38

特徵工程

=

產品ID	特徵A	訊號A平均
01	11	16
02	12	200
03	13	39

時序特徵對純特徵

把時序特徵當主表單以串接純特性而轉出時序特徵大表

串接方法是將原本的「產品編號」與「時間」進行多對一的特徵串接，每個「產品編號」對應到單一純特徵型式，使兩資料能合併。

產品ID	時間	訊號A
01	12:00	17
01	12:30	16
01	13:00	15
02	12:00	232
02	12:30	200
02	13:00	168
03	12:00	35
03	12:30	44
03	13:00	38

+

產品ID	特徵A
01	11
02	12
03	13

=

產品ID	訊號A	特徵A
01	17	11
01	16	11
01	15	11
02	232	12
02	200	12
02	168	12
03	35	13
03	44	13
03	38	13

時序特徵對時序特徵

兩資料集皆含有時間序列特性的特徵(或訊號)

「時序特徵」可分為兩種類型：

- 事件型 (event-based) :「事件」發生時才會記錄，例如機台換模、停機、人為調機)

- 週期型 (period-based) :固定間隔記錄 (例如半小時記錄一次)

「時序特徵」串接方法有兩種：

- 最近時間 (nearest time) 的串接是以主要的數集為核心，由某觀測值找另一資料集中時間點最相近的樣本進行串接。

- 往前/往後追溯 (rolling forward/backward) 的串接則同樣以主要的資料集為核心，但由某一觀測值找尋另一資料集中最相近且較早/較晚的樣本進行串接。

時序特徵對時序特徵

事件型

時間	訊號A
12:23	23
12:33	32
12:35	33
13:02	35
13:33	52

週期型

時間	訊號B
12:00	17
12:30	25
13:00	34
13:30	42
14:00	17

時間	訊號A	訊號B
12:23	23	17
12:33	32	25
12:35	33	25
13:02	35	34
13:33	52	42

特徵提取
Feature Mining/Extraction

特徵挖掘（Feature Mining），又稱為特徵工程或特徵提取，是數據科學中關鍵的一步，旨在將原始數據轉換為對機器學習模型有意義且具有資訊價值的特徵。有效的特徵挖掘能顯著提升模型的準確性和效能。

1. 基本特徵提取

•原始特徵：直接從原始數據中提取特徵（例如，文字、圖像、時間序列）。

手動特徵選擇：根據領域知識識別相關的特徵。

2. 統計特徵

描述性統計：計算平均值、中位數、標準差、變異數、偏態和峰態等。

聚合操作：如總和、最大值、最小值、計數、分位數，特別適用於分組數據或時間序列數據。

3. 領域特定的轉換

- 文字數據：
  - 詞袋模型（Bag-of-Words, BoW）
  - TF-IDF（詞頻-逆文檔頻率）
  - 詞嵌入（例如 Word2Vec、GloVe、BERT）
- 圖像數據：
  - 條形方向梯度直方圖（HOG）
  - SIFT、SURF（尺度不變特徵變換）
  - 從神經網絡的卷積層提取深層特徵。
- 時間序列數據：
  - 趨勢和季節性分解。
  - 傅立葉或小波變換。
  - 滯後特徵（Lag Features）。

4. 特徵轉換

- 正規化：將數據縮放到統一範圍內（例如最小-最大縮放）。
- 標準化：將特徵轉換為均值為 0，標準差為 1。
- 對數、平方根或 Box-Cox 轉換：減少數據偏態。
- 主成分分析（PCA）：降維的同時保留數據變異性。
- t-SNE 和 UMAP：非線性降維，用於可視化。

5. 特徵交互

多項式特徵：通過組合現有特徵生成高次特徵。

交叉特徵：結合分類特徵（例如特徵哈希）。

6. 特徵選擇

- 過濾方法（Filter Methods）：
  - 統計測試（例如卡方檢驗、ANOVA）。
  - 相關性分析。
- 包裝方法（Wrapper Methods）：
  - 遞歸特徵消除（RFE）。
  - 前向選擇與後向選擇。
- 嵌入方法（Embedded Methods）：
  - 正則化技術，如 L1（Lasso）或 L2（Ridge）。
  - 基於樹模型的特徵重要性（例如隨機森林、XGBoost）。

7. 自動化特徵工程

特徵工具：例如 Feature Tools，用於自動特徵提取。

深度學習：通過神經網絡層自動提取高維特徵。

8. 高級技術

•無監督特徵挖掘：

- 聚類（例如 k-means、DBSCAN）用於將數據分組為潛在特徵。
- 自編碼器（Autoencoder）生成潛在表示。

•時間序列嵌入：將序列表示為固定長度的向量（例如序列嵌入、相似度度量）。

9. 特徵編碼

分類特徵：

一熱編碼（One-Hot Encoding）。
標籤編碼（Label Encoding）。
目標編碼（Target Encoding）。

日期與時間特徵：

提取年份、月份、日期、時刻、工作日等。
對週期數據進行週期性轉換。

10. 特徵增強

外部數據：整合其他數據集以豐富特徵集。

知識圖譜：利用圖結構推導特徵之間的關係。

(1) 主成分分析 (PCA)
(2) 線性判別分析 (LDA)
(3) 獨立成分分析 (ICA)
(4) 特徵雜湊 (Feature Hashing)

特徵提取之主要類別

數值特徵	這類特徵是連續型的，通常包括數值或浮點數值。常見的數值特徵提取方法包括標準化、正規化、離散化等
類別特徵	這類特徵通常包含有限的離散值，如性別、國家。特徵提取方法包括獨熱編碼(One-Hot Encoding)、標籤編碼(Label Encoding)、嵌入(Embedding)等
文本特徵	當處理文本數據時，需要將文本轉換為數值特徵。常見的文本特徵提取方法包括詞袋模型(Bag of Words)、詞嵌入(Word Embedding)、N-gram特徵(考慮相鄰單詞的組合)、主題建模(Latent Dirichlet Allocation, LSA)等
時間序列特徵	處理時間序列數據時，可以提取各種統計特徵(如均值、標準差、最大值)、滯後特徵(Lag Features)、滑動窗口特徵(Moving Window Features)等
圖像特徵	當處理圖像數據時，可以使用卷積神經網絡(CNN)來提取特徵，或預先訓練好的卷積神經網絡(如VGG、ResNet)的中間層作為特徵提取器
音頻特徵	音頻特徵提取可以包括Mel-Frequency Cepstral Coefficients (MFCCs)和音頻的Spectrogram等

特徵提取

- 將原始數據轉換為可供機器學習算法處理的特徵的過程
- 特徵提取的方法和技術取決於數據的性質和問題的要求，並可能組合多種不同的特徵提取技術以獲得更好的性能
- 主要的特徵提取技術簡述如下：

主成分分析 (PCA)	• PCA 是無監督學習方法，用於減少數據維度 • 它將數據投影到新的主成分，是原始特徵的線性組合 • 主成分按方差排序，可用於數據可視化和減少冗餘信息
線性判別分析 (LDA)	• LDA 是監督學習方法，主要用於分類 • 它尋找區分不同類別的最佳線性組合，降低數據維度 • 目標是最大化類別之間的區分度，同時最小化類別內方差
獨立成分分析 (ICA)	• ICA 是盲源分離技術，用於分離混合信號中的獨立成分 • 可用於找到原始特徵的獨立組合，以解釋數據 • 廣泛應用於信號處理和語音識別等領域
特徵雜湊 (Feature Hashing)	• 特徵雜湊用於處理高維稀疏數據 • 它透過雜湊函數將原始特徵映射到固定數量的特徵，降低維度 • 常用於處理大型數據集，減少內存使用，尤其適用於文本處理和分類

主成分分析 (PCA)

Principal Component Analysis, PCA	
技術	PCA通過對協方差矩陣的特徵值分解來找到主成分。這意味著它將數據投影到新的坐標系統，其中主成分表示了原始特徵的變異性
運作流程	• 計算特徵的協方差矩陣 • 使用特徵值分解（Eigenvalue Decomposition）或奇異值分解（Singular Value Decomposition）找到協方差矩陣的特徵向量 • 選擇主成分（特徵向量），它們對應著最大的特徵值
用途	PCA主要用於降維，以減少冗餘特徵，並在數據分析和可視化中提供線性結構
原理	PCA的核心原理是最大化變異性，它通過找到投影方向，使得變異性最大，進而降低數據維度
應用	用於圖像壓縮、數據壓縮、維度減少等

線性判別分析 (LDA)

Linear Discriminant Analysis, LDA	
技術	LDA通過找到最佳投影方向，以最大化類內差異性，同時最小化類間差異性，以提高分類性能
運作流程	• 計算類內散佈矩陣和類間散佈矩陣 • 使用特徵值分解找到這些矩陣的特徵向量 • 選擇特徵向量以實現最大類內類間差異性比
用途	LDA主要用於分類，以提高不同類別的可區分性
原理	LDA的原理在於尋找一個投影，以在新的特徵空間中最大程度地分離不同類別的樣本。
應用	用於人臉識別、文本分類、生物信息學等

獨立成分分析 (ICA)

Independent Component Analysis, ICA	
技術	ICA是一種盲源分離技術，旨在找到線性變換，使得投影後的特徵盡可能相互獨立
運作流程	• ICA使用盲源分離算法，如FastICA，估計混合信號的獨立成分
用途	ICA主要用於信號處理，盲源分離，如語音分離和生物信號分析
原理	ICA的核心原理在於找到一組線性變換，使得投影後的特徵盡可能相互獨立
應用	用於語音處理、腦電圖（EEG）信號分析、MRI信號處理等

特徵雜湊 (Feature Hashing)

Feature Hashing	
技術	特徵雜湊是一種用於處理高維稀疏數據的方法，通過雜湊函數將高維特徵映射到低維空間，其中特徵數量有限
運作流程	對每個特徵應用雜湊函數，將其映射到特徵桶
用途	特徵雜湊用於處理高維稀疏數據，減少計算和存儲成本
原理	特徵雜湊的原理在於通過哈希函數將原始特徵映射到有限數量的桶中，不同特徵可能映射到相同位置，並以稀疏表示方式保存
應用	用於自然語言處理中的文本分類、大規模數據處理、推薦系統等

特徵選擇

(1) 過濾方法
(2) 包裝方法
(3) 嵌入方法

特徵選擇

- **特徵選擇是從原始特徵集中挑選出最重要的特徵，以減少維度並提高模型性能的過程。**
- **主要的特徵提取技術如下：**

過濾方法 (Filter Methods)	• 過濾方法是在特徵選擇之前獨立於機器學習模型的方法 • 它使用統計測量來評估每個特徵的重要性，例如方差或相關性 • 過濾方法簡單且高效，但不考慮特徵之間的相互關係
包裝方法 (Wrapper Methods)	• 包裝方法是一種使用機器學習模型的方法，來評估特徵的重要性 • 它通過不斷訓練模型，選擇不同特徵子集，並根據模型性能進行評估 • 包裝方法較為計算密集，但可以提供更準確的特徵選擇結果
嵌入方法 (Embedded Methods)	• 嵌入方法是將特徵選擇嵌入到機器學習模型的訓練過程中 • 它使用模型性能指標來評估特徵的重要性，通常透過正規化或特徵權重調整 • 嵌入方法可同時優化模型性能和特徵選擇

過濾方法 (Filter Methods)

過濾法是一種簡單且有效的特徵選擇方法，通常在數據預處理階段進行。它的主要思想是使用統計方法或相似性來評估每個特徵與目標變量之間的關聯性，然後基於這些評估進行特徵選擇

常見的過濾法包括以下幾種：

- **方差過濾：刪除方差過小的特徵，因為它們對目標變量的預測能力有限**
- **相關係數過濾：計算每個特徵和目標變量之間的相關性，保留相關性較高的特徵**
- **互信息過濾：使用互信息分數來評估特徵的信息增益，選擇互信息較高的特徵**

步驟1	對每個特徵計算某種統計度量，例如方差、相關性、互信息等。這些統計度量用來評估每個特徵的重要性
步驟2	根據統計度量的值，篩選出最重要的特徵，通常是高於一個特定閾值的特徵
步驟3	將所選特徵用於後續的機器學習任務，丟棄未選特徵

包裝方法 (Wrapper Methods)

包裝法是一種特徵選擇方法，通常與特定的機器學習模型結合使用。它通過評估在不同特徵子集上訓練模型的性能來選擇最佳特徵子集。

常見的包裝法包括以下幾種：

- **正向選擇(Forward Selection)：從空特徵集合開始，逐步添加最具信息性的特徵，以最大化模型性能。**
- **後向消除(Backward Elimination)：從包含所有特徵的集合開始，逐步刪除對模型性能貢獻較小的特徵。**
- **逐步選擇(Stepwise Selection)：結合正向選擇和後向消除的方法，進行特徵選擇。**

步驟1	創建一個候選特徵子集，可以是所有特徵的子集或一個初始空子集
步驟2	訓練機器學習模型，使用每個候選特徵子集進行模型訓練。
步驟3	根據模型的性能指標(如交叉驗證分數)評估每個候選特徵子集的性能
步驟4	根據性能指標選擇最佳特徵子集，重複步驟1到4，直到選擇的特徵子集達到滿意的性能水平。

嵌入方法 (Embedded Methods)

嵌入法是在機器學習模型的訓練過程中進行特徵選擇的方法。它通常使用特徵的重要性分數，這些分數是在訓練過程中由模型自動計算的。

常見的嵌入法包括以下幾種：

- **決策樹和隨機森林：這些模型可以計算特徵的重要性分數，然後可以根據這些分數選擇最重要的特徵。**
- **L1 正則化(Lasso)：L1 正則化將特徵的權重稀疏化，促使一些特徵權重變為零，從而實現特徵選擇。**
- **基於梯度的方法：某些機器學習算法，如梯度提升樹和支持向量機，可以通過梯度下降來調整特徵的權重，實現特徵選擇。**

步驟1	在模型訓練過程中，使用所有特徵進行初始訓練。
步驟2	根據模型的性能指標(如損失函數或正則化項)評估每個特徵的重要性
步驟3	調整特徵的權重或進行特徵選擇，以優化模型性能
步驟4	重複步驟1到3，直到模型性能達到滿意的水平

特徵編碼

(1) Label Encoding
(2) One-hot Encoding
(3) Embedding

特徵編碼(Feature Encoding)

- **特徵編碼是將原始數據轉（如類別、文本、時間、數值等）換為模型能夠理解的數字形式的過程**

基礎	標籤編碼 (Label Encoding)	適用於有序類別特徵，將每個類別映射為一個整數編碼值
	獨熱編碼 (One-Hot Encoding)	用於處理無序類別特徵，將每個類別轉換為一個二進制特徵，其中每個類別都有一個對應的二進制位
	順序編碼 (Ordinal Encoding)	類似於標籤編碼，但保留類別之間的順序信息
	計數編碼 (Count Encoding)	將類別特徵編碼為該類別在數據集中的出現次數
	目標編碼 (Target Encoding)	用於分類問題，將類別特徵轉換為目標變量的統計摘要（例如，平均值）
	均值編碼 (Mean Encoding)	類似於目標編碼，但使用其他特徵的統計資訊（例如，平均值）來編碼類別特徵
	二進制編碼 (Binary Encoding)	將類別特徵轉換為二進制碼，減少維度，適用於高基數特徵
	時間特徵編碼 (Time Feature Encoding)	針對時間數據，使用週期性編碼、時間間隔編碼等方法

標籤編碼 (Label Encoding)

- 類似流水號，將新出現的類別依序邊上新代碼，已出現的類別邊上已使用的代碼
- 可轉換為分數，但分數的大小順序沒有意義

獨熱編碼 (One-hot Encoding)

- 為了改良標籤編碼中數字大小沒有意義的問題，將不同的類別分別獨立為一欄
- 需要較大的記憶空間與計算時間，且類別數量越多時越嚴重

嵌入編碼 (Embedding)

Embedding（嵌入）在數學上指的是一個數學結構經映射包含到另一個結構中，而在ML 、DL 領域，則是指將實體(entity)高維離散的特徵映射到相對低維的連續向量空間中。

Embedding（嵌入）是一種在機器學習和自然語言處理（NLP）中廣泛使用的技術。它是將離散型數據映射到連續型向量空間的方法。

嵌入的目標是將數據表示為具有數值特徵的向量，以便計算機可以更好地理解和處理這些數據。這個概念不僅適用於自然語言處理，還可以用於圖像處理、推薦系統等領域。

嵌入技術的核心思想是通過學習數據的表示，使得相似的數據在嵌入空間中更加接近。

常見的詞嵌入模型包括Word2Vec、GloVe（全局向量）和FastText等，它們通過處理大量的文本語料庫來學習詞語之間的關係，並生成詞嵌入向量。

這些向量可以用於訓練神經網絡模型，如循環神經網絡（RNN）和卷積神經網絡（CNN），以執行各種自然語言處理任務。

詞嵌入已經成為NLP領域的標準技術，它能夠在文本數據中捕獲豐富的語義信息，從而提高了模型的性能和效果。

特徵縮放

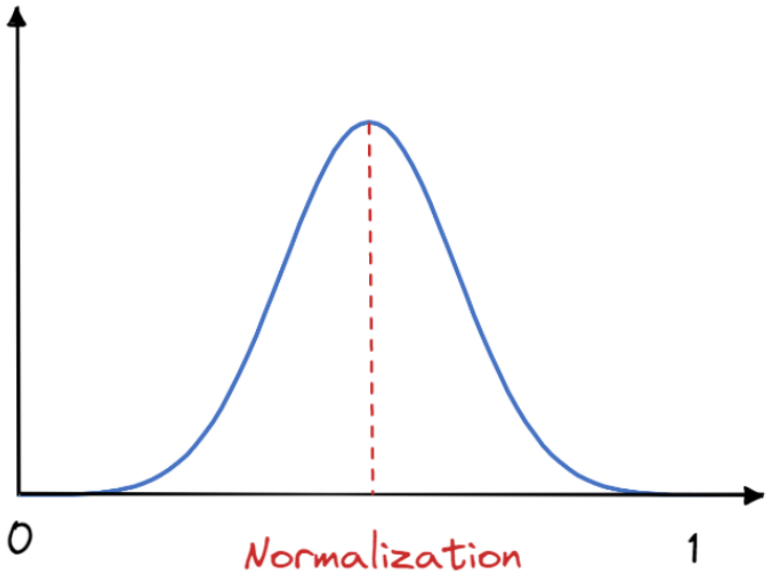
(1) 正規化 (Normalization)
(2) 標準化 (Standardization)
(3) Min-Max Scaling 縮放

正規化 (Normalization)

- 正規化是一種縮放特徵值的方法，使它們落在特定範圍內，通常是[0, 1] 或[-1, 1]
- 目的是確保不同特徵的值處於相似的範圍，以便它們可以在模型中進行比較

- 運作流程：
  - 選取正規化的範圍，通常是[0, 1] 或[-1, 1]
  - 對每個特徵值 X 執行以下轉換：

$$X_{norm} = \frac{X - \min(x)}{\max(x) - \min(x)}$$

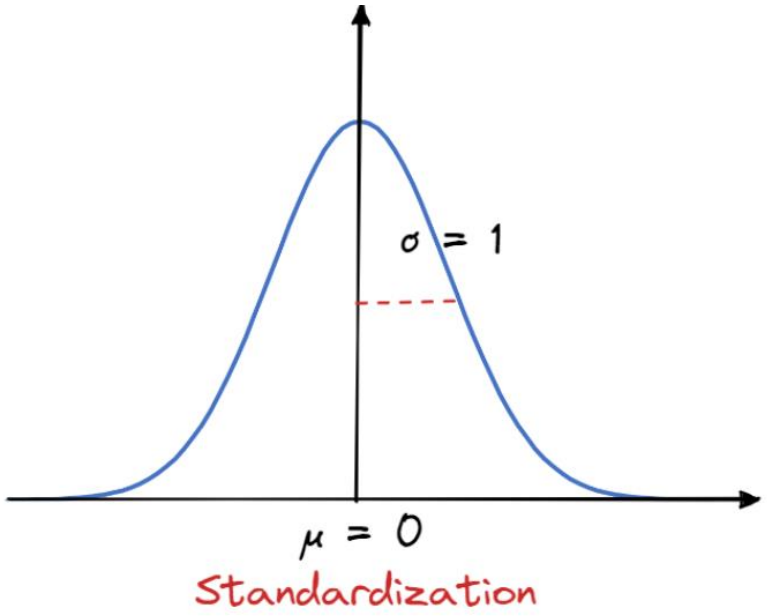


標準化 (Standardization)

- 標準化是一種將特徵值轉換成均值為0，標準差為1的分佈的方法
- 使特徵值更接近正態分佈，有助於機器學習演算法的性能

- 運作流程：
  - 計算每個特徵的平均值 μ 和標準差 σ
  - 對每個特徵值 x 執行以下轉換
  - 使所有特徵值的均值為0，標準差為1

$$X_{stand} = \frac{X - \text{mean}(x)}{standard\ deviation(x)}$$



最小-最大縮放方法 (Min-max Scaling)

- 最小-最大縮放是一種將特徵值縮放到特定範圍（通常為[0, 1]）的方法，以便它們可以在相同的範圍內進行比較
- 此方法保留了原始特徵值的相對關係

- 運作流程：
  - 選取縮放的目標範圍，通常是[0, 1]
  - 對每個特徵值 x 執行以下轉換
  - 使所有特徵值都落在[0, 1]的範圍內

$$X_{Scaled} = (\mathrm{X} - X_{min}) \div (X_{max} - X_{min})$$