



Manufacturing Data Science



Data Preprocessing

(第 5 章 數據預處理)

Chia-Yen Lee, Ph.D. (李家岩 博士)

Department of Information Management (資訊管理學系)
National Taiwan University (國立台灣大學)

- 第一章 製造數據科學
- 第二章 製造系統分析與管理
- 第三章 數據科學基礎與模型評估
- 第四章 數據科學分析架構與系統運算決策
- 第五章 數據預處理與製造數據特性
- 第六章 線性分類器
- 第七章 無母數迴歸與分類
- 第八章 決策樹與集成學習
- 第九章 特徵挑選與維度縮減
- 第十章 類神經網路與深度學習
- 第十一章 集群分析
- 第十二章 特徵工程、數據增強與數據平衡
- 第十三章 故障預測與健康管理
- 第十四章 可解釋人工智慧
- 第十五章 概念漂移
- 第十六章 元啟發式演算法
- 第十七章 強化學習

藍：老師課堂講授

綠：學生自學

- 附錄A 線性迴歸
- 附錄B 支持向量機
- 附錄C 統計製程管制與先進製程控制
- 附錄D 超參數最佳化

- 應用涵蓋

產能規劃、瑕疵檢測、製程監控與診斷、機台保養、需求預測、生產排程、電腦視覺、自動光學檢測、原料價格預測與採購等

- 本章我們探討「製造數據科學」中，如何運用製造的**領域知識**強化「**數據預處理**」(data preprocessing)以提升數據的品質，以及製造現場存在哪些重要特性是我們需要特別留意的。
- 數據預處理將原始數據經過處理，其目的在於使得數據完善、精確、易於解釋、且轉換成適當的格式以利後續機器學習／數據科學建模。
- 此章節「數據預處理」我們依序討論
 - 「**數據整合**」(data integration) (學生自學)
 - 「數據清理」(data cleaning)
 - 「數據轉換」(data transformation)
 - 「數據品質」(data quality)

□ 數據整合

- 指將數據庫中不同的數據表單做整併或串接的動作，以形成一個考量全面資訊的數據大表。
- 數據整合在串接過程中常用的「主要特徵」
 - 「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)〕
 - 「時序特徵」（例如時間、時間區間）
- 兩者的配對可以分為四種情形
 - 「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」

□ 純特徵對純特徵

- 「純特徵對純特徵」的情形是純粹依據主要特徵就可進行合併，如圖5.1所示。
- 分別收集到兩個數據集分別有「產品編號」與其他「特徵」，因此我們僅需依照「產品編號」進行串接即可。
- 可能面臨到串接後因兩數據集的共同「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。

產品 ID	特徵 A		產品 ID	特徵 B		產品 ID	特徵 A	特徵 B
產品 01	11	+	產品 01	17	=	產品 01	11	17
產品 02	12		產品 02	16		產品 02	12	16
產品 03	13		產品 04	15		產品 03	13	NA
						產品 04	NA	15

圖 5.1 純特徵間的表單串接

□ 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。
- 串接方法將對時序特徵進行特徵工程（萃取出平均數、標準差、偏態、峰態等），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（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，因而使得兩數據集能進行合併。

產品 ID	特徵 A		產品 ID	時間	訊號 A		產品 ID	特徵 A	訊號 A-平均
產品 01	11	+	產品 01	12:00	17	=	產品 01	11	16
產品 02	12		產品 01	12:30	16		產品 02	12	200
產品 03	13		產品 01	13:00	15		產品 03	13	39
			產品 02	12:00	232				
			產品 02	12:30	200				
			產品 02	13:00	168				
			產品 03	12:00	35				
			產品 03	12:30	44				
			產品 03	13:00	38				

□ 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」對應到單一純特徵的形式，因而使得兩數據集能進行合併。

產品 ID	時間	訊號 A		產品 ID	特徵 A		產品 ID	訊號 A	特徵 A
產品 01	12:00	17	+	產品 01	11	=	產品 01	17	11
產品 01	12:30	16		產品 02	12		產品 01	16	11
產品 01	13:00	15		產品 03	13		產品 01	15	11
產品 02	12:00	232					產品 02	232	12
產品 02	12:30	200					產品 02	200	12
產品 02	13:00	168					產品 02	168	12
產品 03	12:00	35					產品 03	35	13
產品 03	12:30	44					產品 03	44	13
產品 03	13:00	38					產品 03	38	13

□ 時序特徵對時序特徵

- 兩數據集皆含有時間序列特性，時序特徵可分為兩種類型，「事件型」(event-based)與「週期型」(period-based)的紀錄方式。
- 可分別基於「事件型」或「週期型」的數據集為基準串接。
- 其中最重要的一點在於我們期望的**目的**為何
 - 若目的為「故障排除」(troubleshooting)，我們關注於「事件」發生的根因，因此以事件型的表單作為主表單（基準）進行串接。
 - 若目的為「製程監控」(monitoring)，則我在乎固定時間「週期」間隔數據所產生的變化，將以週期型的表單作為主表單。

	以「事件型」特徵為基準串接	以「週期型」特徵為基準串接
記錄方式	「事件」發生時才會記錄 (例如機台換模、停機、人為調機)	「週期」的固定間隔記錄 (例如半小時記錄一次)
串接前特性	樣本數較少且稀疏	樣本數較多且完整
串接後優點	數據較為完整，可能的遺漏值較少	保留數據週期性的變化
串接後缺點	「事件」間的數據將被忽略	串接後將使得「事件型」特徵產生大量的遺漏值
目的與使用時機	探討事件之間的關聯 故障排除 (Troubleshooting)	探討長期的變化 (例如磨耗) 監控 (Monitoring)

◻ 時序特徵對時序特徵

- 「時序特徵」串接方法
 - 「最近時間」(nearest time)
 - 由某一觀測值找尋另一數據集中時間點最相近的樣本進行串接。
 - 「往前／往後追溯」(rolling forward / backward)
 - 但由某一觀測值找尋另一數據集中最相近且較早／較晚的樣本進行串接。
- 應採用哪種串接方法比較合適視情況而定(it depends)，並沒有標準答案
- 以「事件型」表單做基準串接為例，做故障排除在發生前會有一些徵兆，因此在探討因果關係時串接「事件」發生後的數據串接較不合理
 - 若以「事件型」數據基準串接一個「週期型」數據，其中「週期型」數據每半小時才抽取一個樣本（抽樣間隔較大），因而我們採用「往前追溯」法。

事件型			週期型					
時間	訊號 A		時間	訊號 B		時間	訊號 A	訊號 B
12:23	23	+	12:00	17	=	12:23	23	17
12:33	32		12:30	25		12:33	32	25
12:35	33		13:00	34		12:35	33	25
13:02	35		13:30	42		13:02	35	34
13:33	52		14:00	17		13:33	52	42

□ 視覺化

- 實際上，數據整合在數據結構關係複雜時盡可能以視覺化呈現，這樣有利於後續不論是「數據清理」或是「特徵工程」時理解特徵之間的關係
- 火車運輸相關的多個數據集整合案例

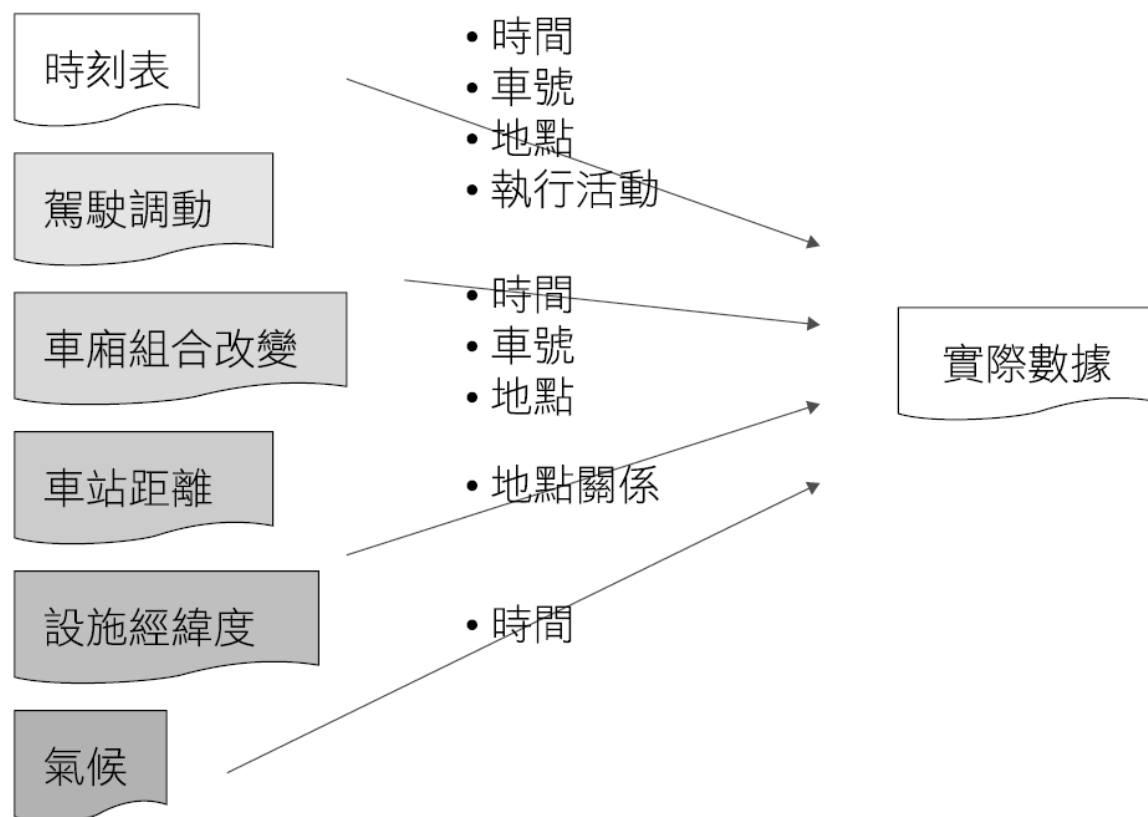


圖 5.5 火車運輸多個數據集的串接

□ 數據清理主要課題為

- 「數據縮減」 (data reduction)
- 「離群值偵測」 (outlier detection)
- 「遺漏值填補」 (missing value imputation)

數據縮減

- 「數據縮減」是將冗餘(redundant) (有相同數值、流水編號) 的「特徵」或重複(duplicate)記錄的「觀測值」進行移除。
 - 然而這些數據原則上不具資訊量，並且它們將增加額外的計算負擔，因此須將它們適當的移除。

產品 ID	特徵 A	特徵 B		產品 ID	特徵 A
產品 01	17	10		產品 01	17
產品 02	16	10		產品 02	16
⋮	⋮	⋮		⋮	⋮
產品 99	21	10		產品 99	21
產品 99	21	10			
產品 99	21	10			

數據縮減

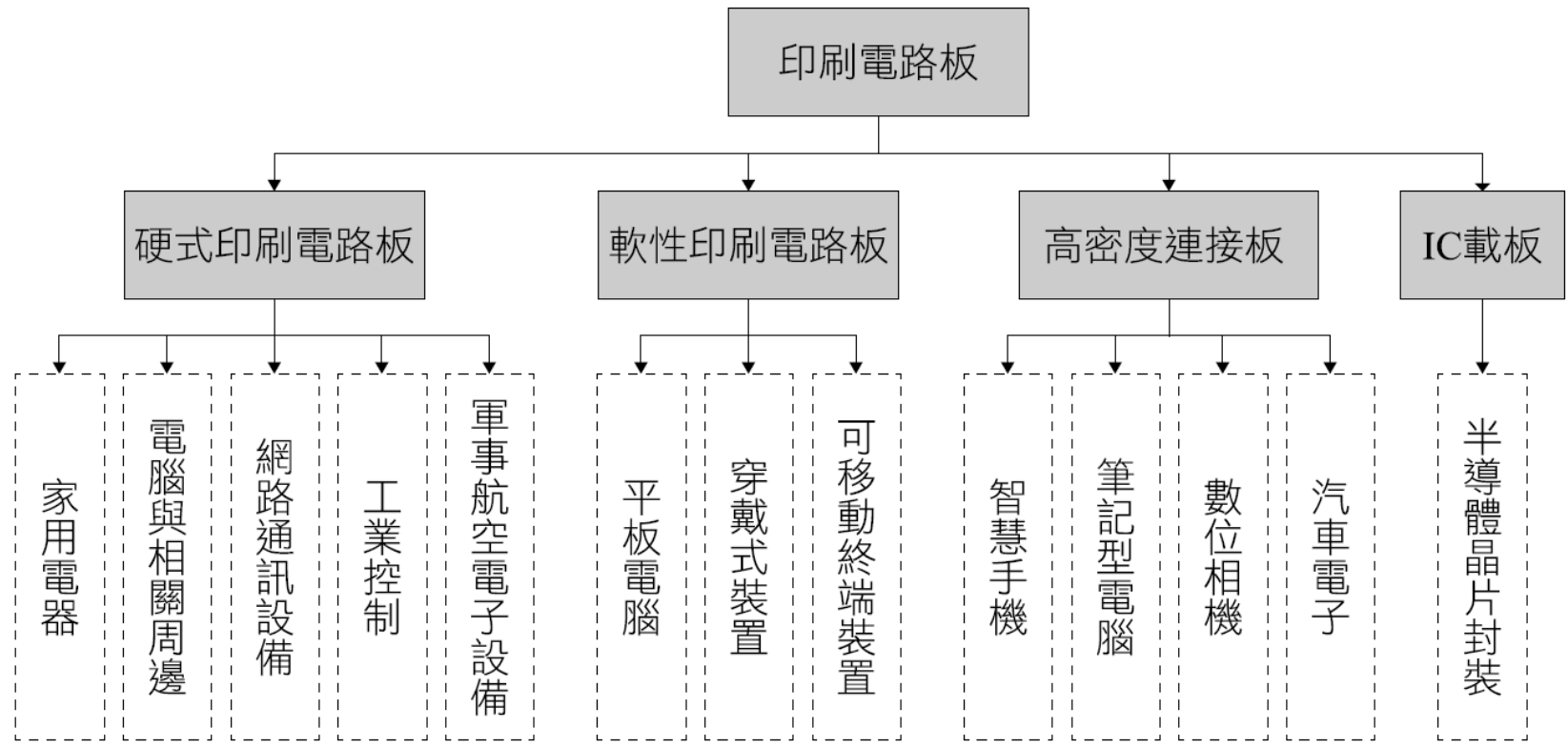
- 冗餘是指某一特徵可以由另外一個特徵透過數學關係或常理推導而出
 - 例如「地區」 (北中南東離島) 與「地址」兩個變量中，地區是冗餘的
- 實務上未必能直接將這種冗餘的特徵直接刪除，因為此冗餘特徵可能強化了模型預測的準確度
- 同時須留意該冗餘特徵所帶來「維度的詛咒」或「共線性」等潛在問題

□ 數據縮減方法

- 數據立方聚合(data cube aggregation)：
 - 從某一個特徵角度，將其不同類別的「觀測值」聚合
 - 例如將機台 (tool) 聚合成機台群組 (tool group) 計算統計量 (例如加總)
- 特徵挑選(feature selection)或維度縮減(dimension reduction)
 - 此為**特徵或維度**的縮減，透過監督式學習挑選重要特徵或透過非監督式學習以線性／非線性組合將特徵合併或刪除
- 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)的樣本以減少數據量
 - 如果為離散型數據：(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。
 - 如果為連續數據，可透過離散化方法 (discretization) 將連續數據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)，可透過下抽樣 (downsampling)、移動平均法 (moving average)、滑動時窗 (sliding window) 等方法，縮減數據樣本。

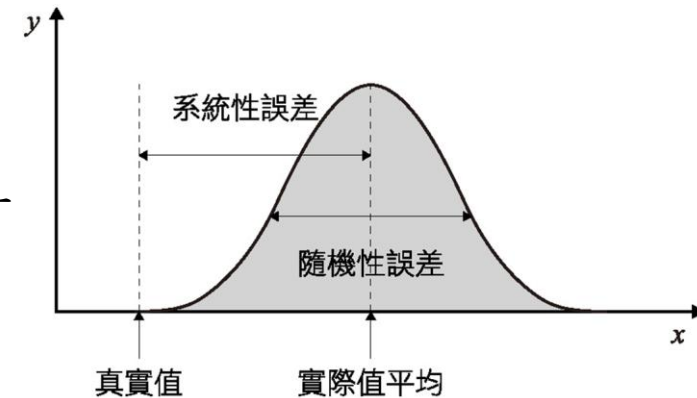
□ 概念層級

● MECE



離群值偵測與處理

- 「系統性」(systematic)與「隨機性」(random)
 - 「系統性」：結構化、有跡可循產生的異常。
 - 「隨機性」：自然、任意產生的異常，其隨機性可能服從某個分配。



- 在製造業中常見的「系統性離群值」

- 「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數據將產生明顯的離群值，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，因此有時須保留這些稀少且珍貴的異常數據。

- 「實驗貨／工程貨」(experiment/engineering product)


- 由於產品研發及機台調機為快速調整、改善良率所產生出的試驗數據。

- 「人為錯誤／干擾」(human error/interference)

- 可能是人為數據填寫key-in 錯誤或是環境干擾到感測器對數據的收集（例如堆高機從機台旁邊開過使感測器收到異常訊號）

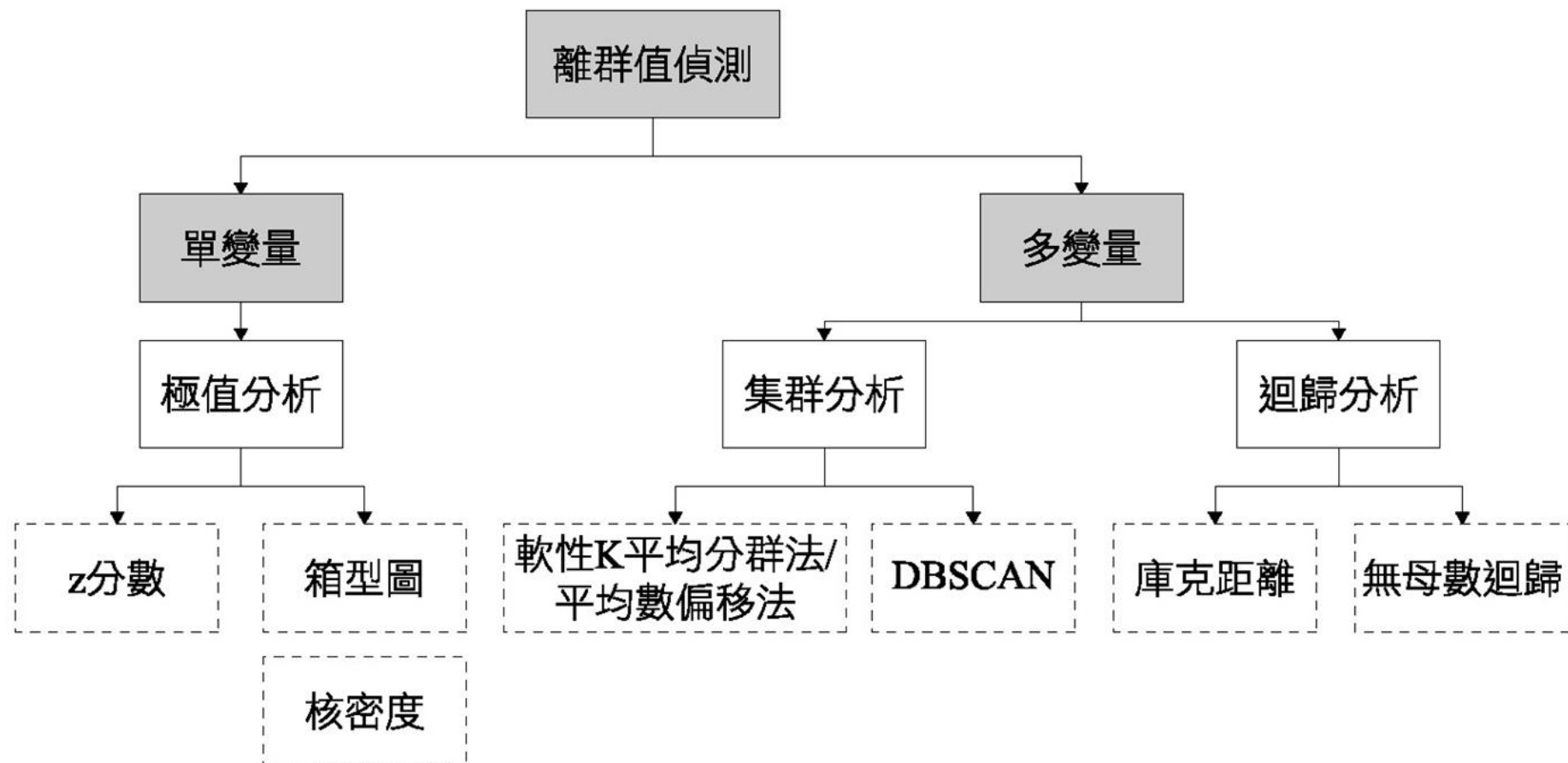
- 領域知識協助判定系統性與隨機性離群值，也作為偵測與處理離群值的依據

- 採取相對應的「治療」(treatment)，進行「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）。

表 5.2 實驗貨與一般貨的數據比較 

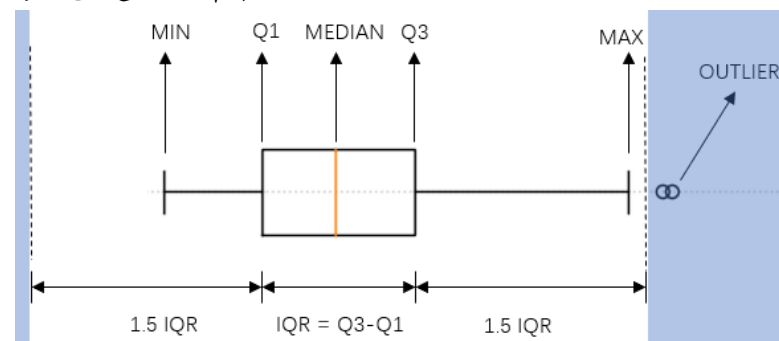
	實驗貨 / 工程貨 ($n \ll p$)	一般正常貨 ($n \gg p$)
數據量	較少（僅測試用，有實驗成本）	較多（大量生產）
數據數值	參數較分散（實驗中）	許多參數已成定值（穩定）
成本	較高（需要反覆實驗、調機）	較低（大量生產）
良率	較低	較高且穩定
分析方法	實驗設計 / 田口方法、最佳化方法、反應曲面法、LASSO、無母數方法	深度學習、梯度提升機、有母數方法

離群值的偵測方法



極值分析(單變量)


- 有母數的「z 值」(z-score) : $z_i = \frac{x_i - \bar{x}}{s}$
 - 假設特徵服從常態分配，其中 \bar{x} 為樣本平均數、 s 則為樣本標準差，因此對「z 值」取絕對值排序後可偵測出樣本為離群值(例如z值的絕對值大於2)
- 無母數的「箱型圖」(boxplot)
 - 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, $Q3 - Q1$)，代表著數據的變異程度（與標準差相似），再進一步以1.5倍的「四分位間距」推算「最大值」與「最小值」，而在極值外的便是離群值。



- 「核密度」(kernel density)
 - 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。

數據清理

□ 多變量

- 「**集群分析**」 (MDS textbook Ch. 11)
 - 「軟性K 平均分群法」 (soft K-means clustering)
 - 「平均數偏移法」 (mean shift)
 - 「DBSCAN」 (density based spatial clustering of applications with noise)
- 「**迴歸分析**」
 - 「庫克距離」 (Cook 's distance) 
 - https://en.wikipedia.org/wiki/Cook%27s_distance
 - 以樣本角度觀之，在線性迴歸的假設下，計算每個樣本對迴歸係數的影響力，影響力越大的越可能為離群值

Cook's distance D_i of observation i (for $i = 1, \dots, n$) is defined as the sum of all the changes in the regression model when observation i is removed from it^[5]

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{ps^2}$$

where $\hat{y}_{j(i)}$ is the fitted response value obtained when excluding i , and $s^2 = \frac{\mathbf{e}^\top \mathbf{e}}{n - p}$ is the mean squared error of the regression model.^[6] $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

- 「無母數迴歸」 (non-parametric regression)



- 將資料分割成 K 個互不相交的群集，當資料點與該群集中心的相似度高於其他群集時，則歸類於該群集中，反之，歸屬於新群集，再以新群集的平均值為中心，反覆計算直到結果收斂
- 目的：使各資料點到所屬群集中心的總距離變異平方和最小

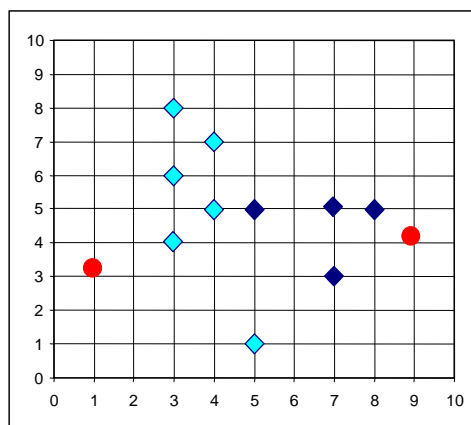
$$E = \sum_{l=1}^K \sum_{i=1}^{n_k} (\mathbf{x}_{il} - \mathbf{m}_l)^T (\mathbf{x}_{il} - \mathbf{m}_l)$$

■ 步驟

1. 隨機選取 K 筆資料點作為 K 個起始群集中心值
2. 將剩下的每一筆資料分配到離群集中心最近的群集中，並根據群集中的資料點，重新計算各群集的平均值
3. 計算資料點到群集中心的距離，若發現總距離變異平方和下降，則表示群集中心有所改變，需將資料點重新分配到新的群集
4. 直到總距離變異不再下降或達到所設定的計算次數為止

K平均法群集分析過程 ($k=2$)

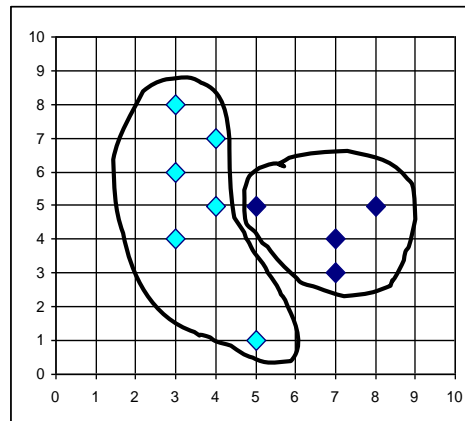
■ Example



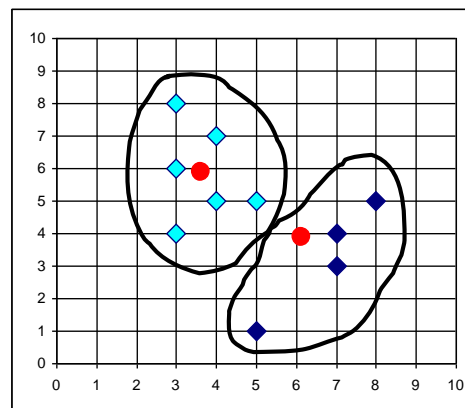
$K=2$

Arbitrarily choose K object as initial cluster center

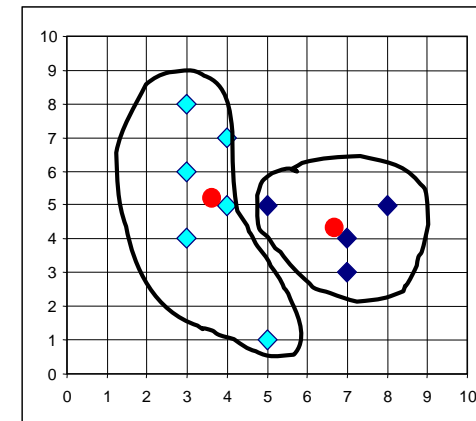
Assign each objects to most similar center



reassign

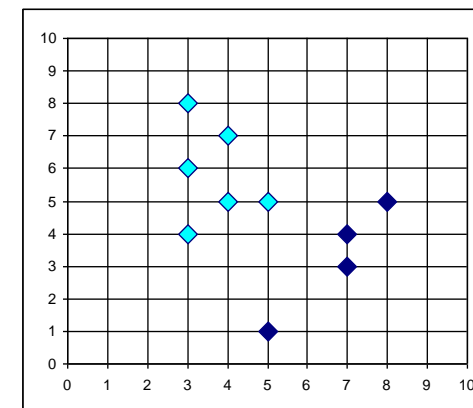


Update the cluster means

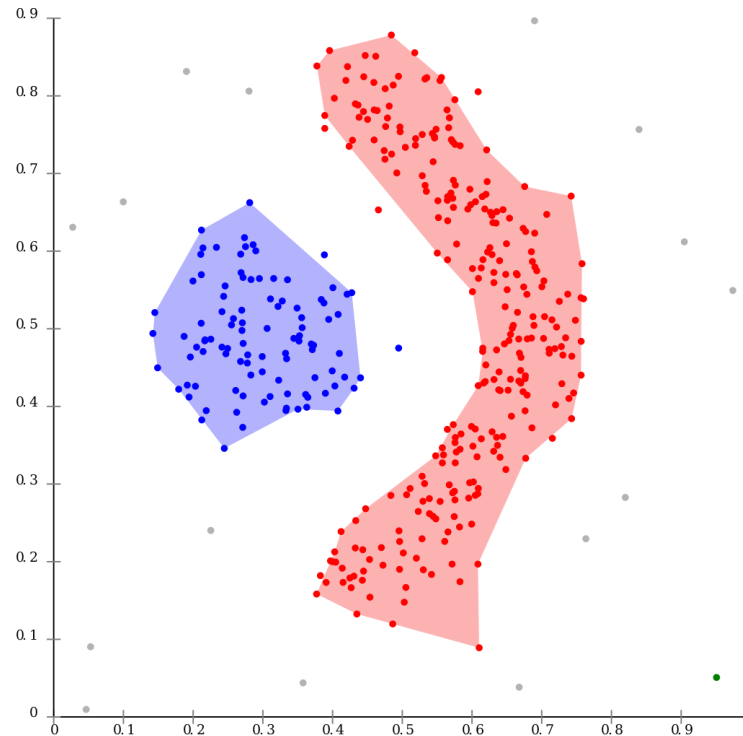


reassign

Update the cluster means



- DBSCAN (Density-based spatial clustering of applications with noise)
 - clustering algorithm proposed by Ester, Kriegel, Sander and Xu (1996).
 - 目的是找到密度相連對象的最大集合。



- DBSCAN can find non-linearly separable clusters. This dataset cannot be adequately clustered with k-means or Gaussian Mixture EM clustering

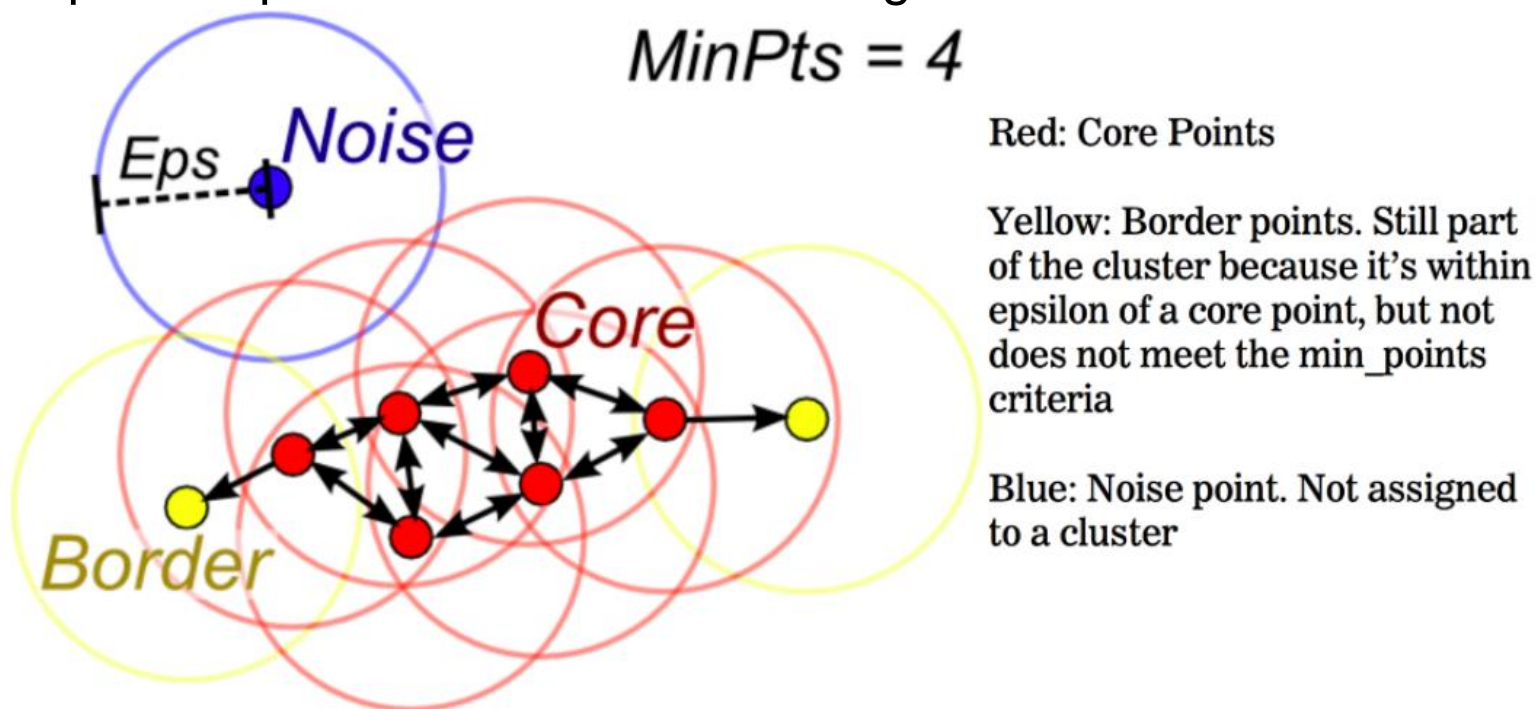
DBSCAN

□ Epsilon

- The **maximum distance** (euclidean distance) between a pair of points. The two points are considered as neighbors if and only if they are separated by a distance less than or equal to epsilon ϵ .

□ MinPoints

- The minimum number of points required to form a dense cluster
- The MinPts = 4 means minimum 4 points are required to form a dense cluster. Also, a pair of points must be separated by a distance of less than or equal to Eps to be considered as neighbors.



▣ Based on the above two parameters, data points are classified into 3 categories as follows:

▣ Core point

- A selected point is a core point if it has **at least minimum number of points** (MinPts) including itself **within its epsilon-neighborhood**. In figure 1, **red** points are core points that have at least MinPts=4 in their neighborhood. If we've a core point, it means it is a dense region.

▣ Border point

- A selected point that is **within a neighborhood of a core point** but it itself cannot be a core point. In the figure 1, **yellow** points are identified as border points. If we've a border point, it means the point is in a vicinity or at the border of dense region.

▣ Noise point

- A selected point that is neither a core point nor a border point. It means these points are **outliers** that are not associated with any dense clusters. In the figure 1, **blue** point is identified as noise point.

□ Pros

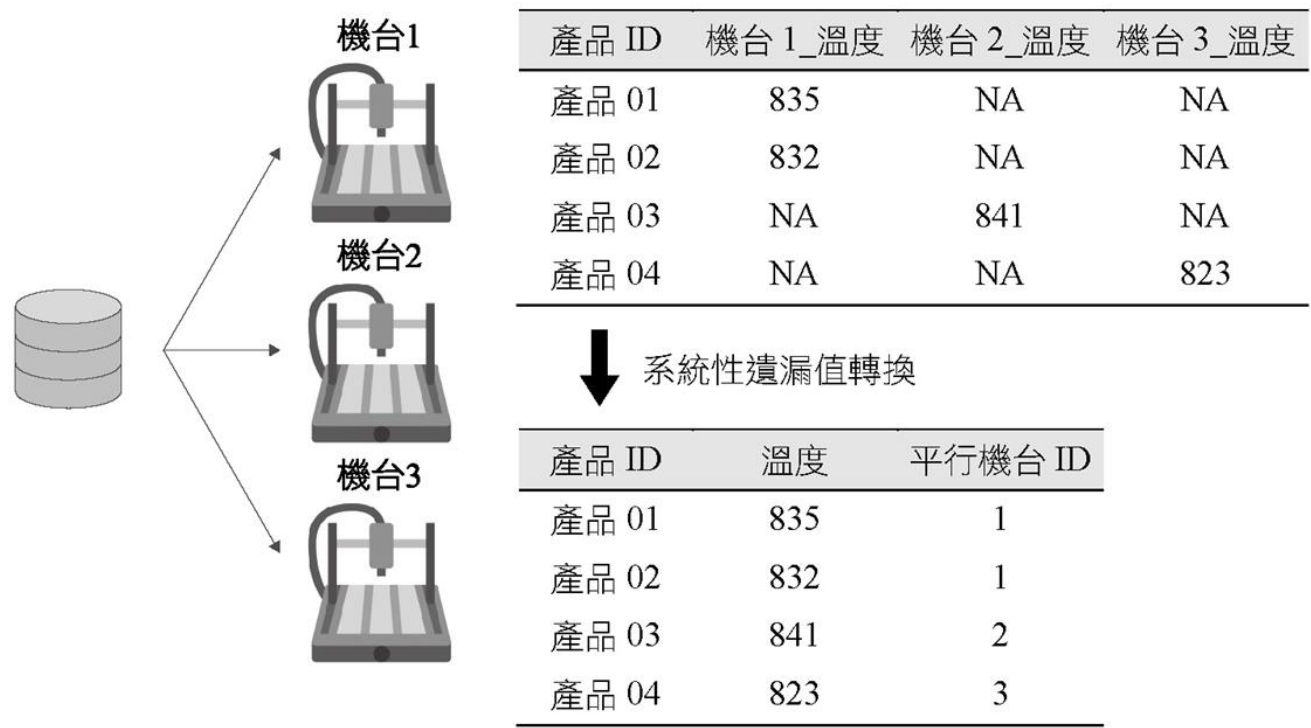
- 1. 與K-means方法相比，DBSCAN不需要事先知道要形成的群集的數量。
- 2. 與K-means方法相比，DBSCAN可以發現任意形狀的群集。
- 3. 同時，DBSCAN能夠識別出noise。對離群點有較好的robustness。
- 4. 對於樣本的順序不敏感，即Pattern的輸入順序對結果的影響不大。
- 5. 被設計與資料庫一同使用，可以加速區域的查詢，例如 使用R* tree

□ Cons

- 1. 不能很好反映高維數據。
- 2. 在邊界的觀測值可屬於任一個群，取決於數據處理的順序。
- 3. 由於DBSCAN算法直接對整個數據集進行操作，並且在聚類之前需要建立相應的R*樹，並繪制k-dist圖，因此算法所需的內存空間和I/O消耗都相當可觀。
- 4. 由於DBSCAN算法使用了全局性表征密度的參數，因此當各個類的密度不均勻，或類間的距離相差很大時，聚類的質量較差。
- 5. 分群品質取決於regionQuery(P, ϵ)函數中距離的測量。最常用的距離度量是歐式距離，尤其是在高維數據會發生“維度的詛咒”，這種度量基本上是无用的，很難為 ϵ 找到一個恰當的值。

遺漏值填補

- 系統性遺漏值(systematic missing value)
- 隨機性遺漏值(random missing value)
- 使用者應根據製造「數據特性」以及「分析目的」，來決定填補遺漏值的方法，以避免忽略原本應有的資訊。
- 「系統性漏值」的填補需要領域知識的輔助
 - 常見「平行機台」(parallel machine)加工所產生出的數據。



遺漏值填補是...補「資料」?

	English	Math
Student_A	80	76
Student_B	80	91
Student_C	80	83
Student_D	80	62
Student_E	80	?
Avg.	80	

Max: 91

Min: 62

Avg: 78

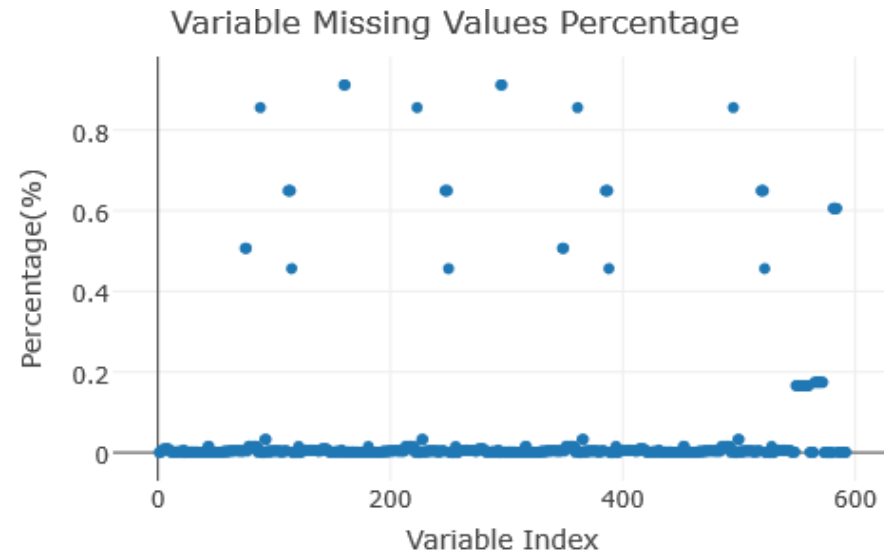
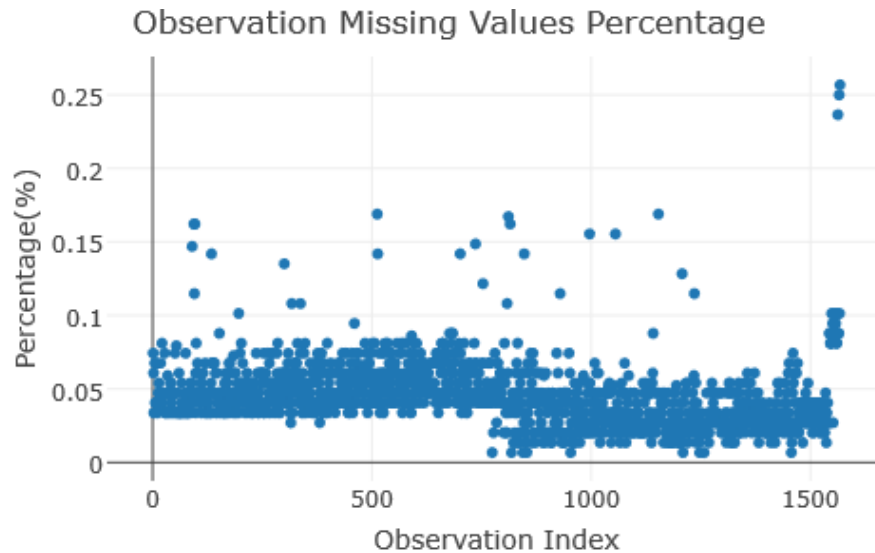
迷思：補遺漏值？

□ 不偏估計量 vs. 變異程度

觀測值	原始資料值	第 11 筆遺漏	利用平均數估計	利用標準差估計
1	0.0886	0.0886		
2	0.0684	0.0684		
3	0.3515	0.3515		
4	0.9874	0.9874		
5	0.4713	0.4713		
6	0.6115	0.6115		
7	0.2573	0.2573		
8	0.2914	0.2914		
9	0.1662	0.1662		
10	0.44	0.44		
11	0.6939	?		
平均值	0.4023	0.3731		
標準差	0.2785	0.2753		
誤差值				

(簡禎富、許嘉裕，2014)

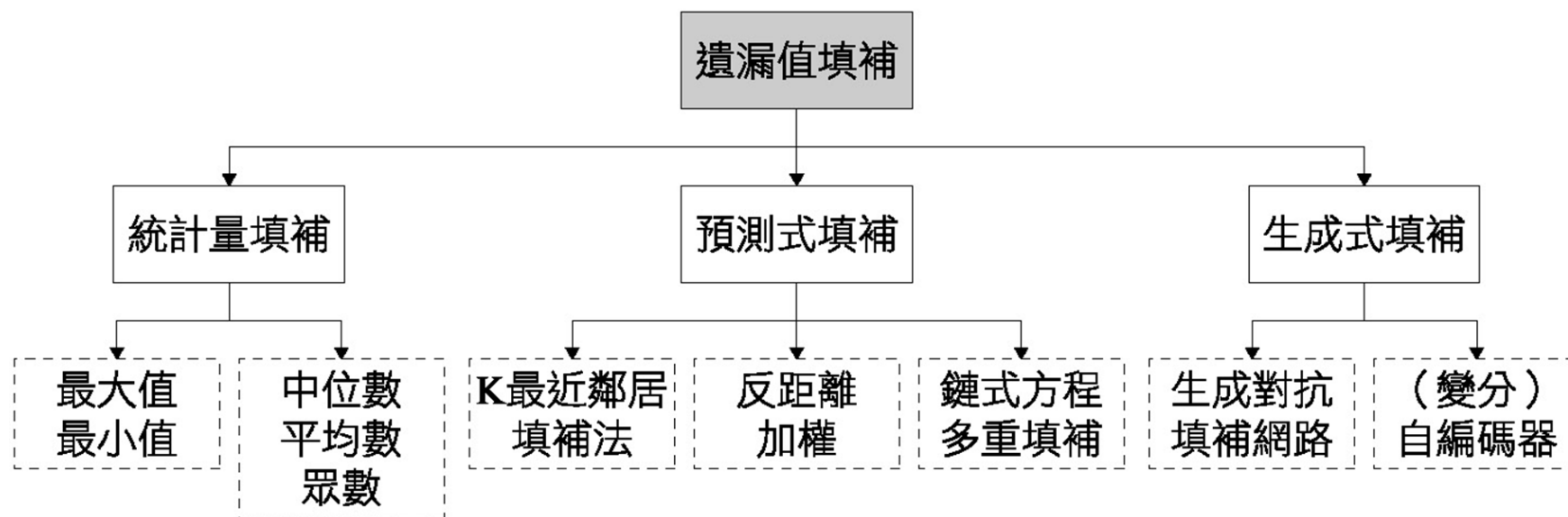
Missing Value Imputation



- Remove the column or row with too many missing values (>40%)
- Remove the row without label/target value/Y (or **unsupervised learning**)
- Impute: Mean and Median (by the same category), Mode, “Others”, “N/A”, “NaN”
- Impute by **Model**: K nearest neighbor (KNN), Multivariate Imputation by Chained Equations (MICE), Inverse Distance Weighting (IDW), etc.
- Ignore: some algorithms can handle the missing values, eg. **LightGBM** and **XGBoost** (ignore is different from imputing “NA” or “0”).

遺漏值填補

- 「隨機性遺漏值」的填補主要方法
 - 「統計量填補」 (statistics imputation)
 - 「預測式填補」 (predictive imputation)
 - 「生成式填補」 (generative imputation)
- **補值一定要透過其他特徵找關係**，而非單純用該特徵的統計量填補。
- 並不是所有的遺漏值都一定需要填補，對於後續所使用的某些預測模型或演算法（例如隨機森林、梯度提升機），其可忽略遺漏值依然繼續模型訓練（忽略的概念不同於填補N/A）。



□ 遺漏值填補

● 統計量填補

- 最大值、最小值、中位數、平均數、眾數

● 預測式填補

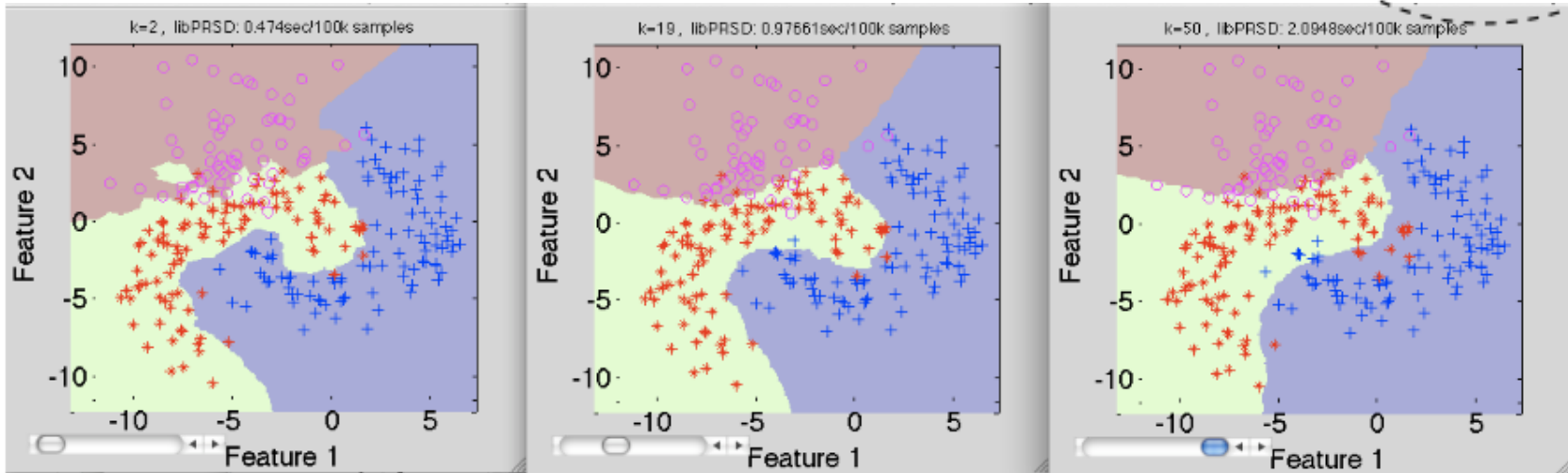
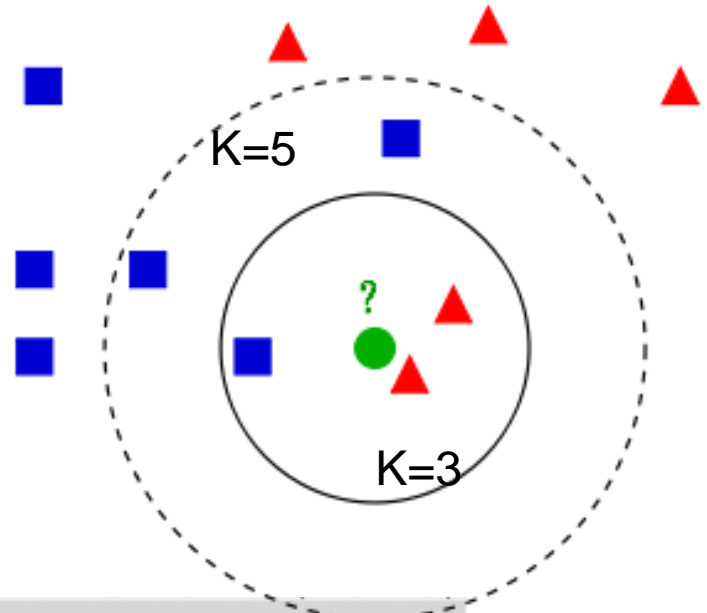
- 「K最近鄰居填補法」(K-nearest neighbor imputation, KNNI)
- 「反距離加權」(inverse distance weighting, IDW)
- 「鏈式方程多重填補法」(multiple imputation by chained equations, MICE)

● 「生成式填補」的主要方法

- 「生成對抗填補網路」(generative adversarial imputation network, GAIN)
 - 是將「生成對抗網路」(generative adversarial network, GAN) 的模型應用在填補上，以「生成模型」生成填補數據，再以「判別模型」判別生成後的數據為實際還是被生成的。
- 「自編碼器」(autoencoder, AE)
- 「變分自編碼器」(variational autoencoder, VAE)

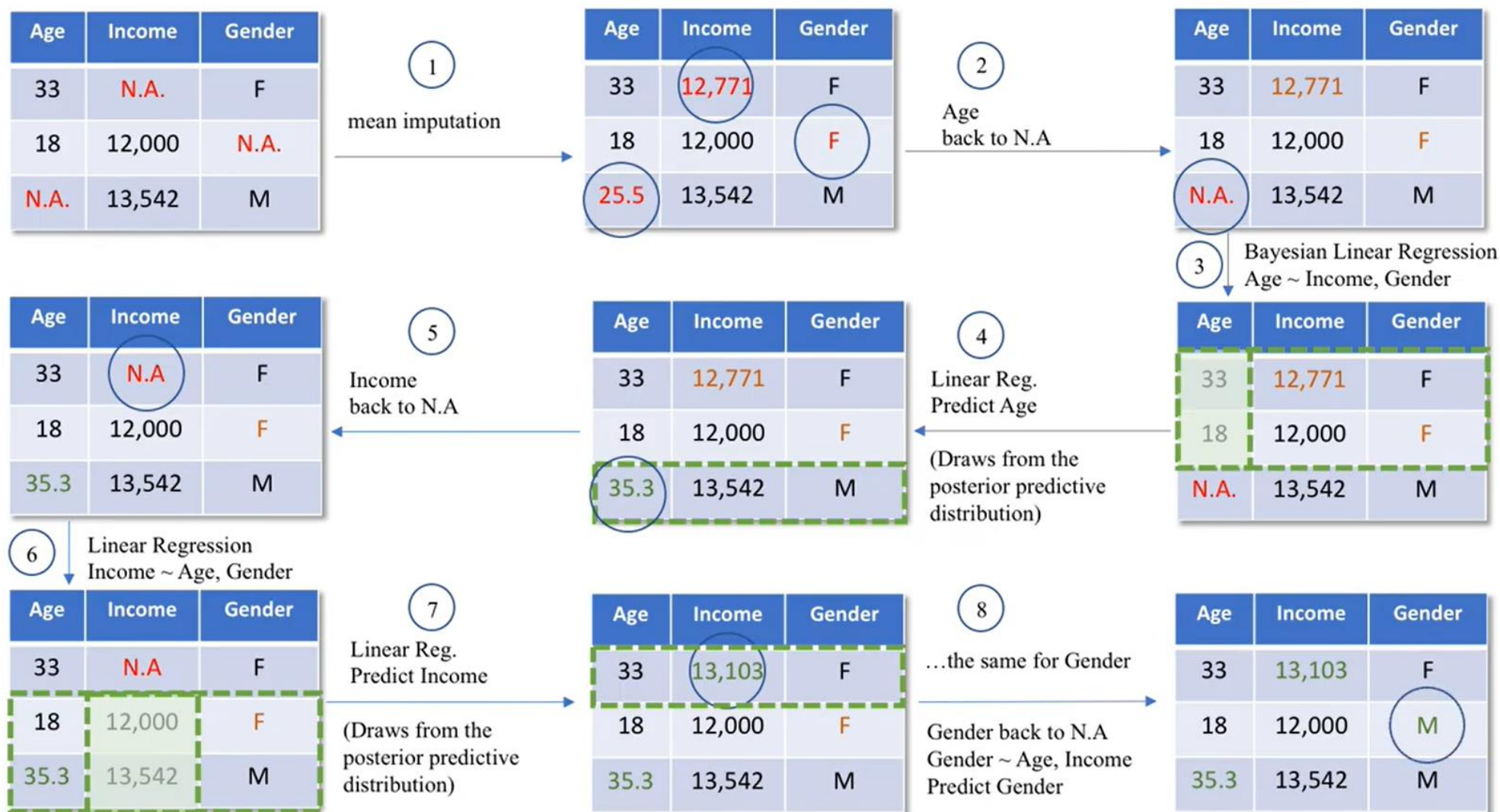
■ 假設現有資料庫中發現某一顧客其購買反應態度為一遺漏值

顧客	性別	年齡	薪水	購買反應
A	女	27	\$19,000	No
B	男	51	\$64,000	Yes
C	男	52	\$105,000	Yes
D	女	33	\$55,000	Yes
E	男	45	\$45,000	No
F	女	45	\$100,000	?



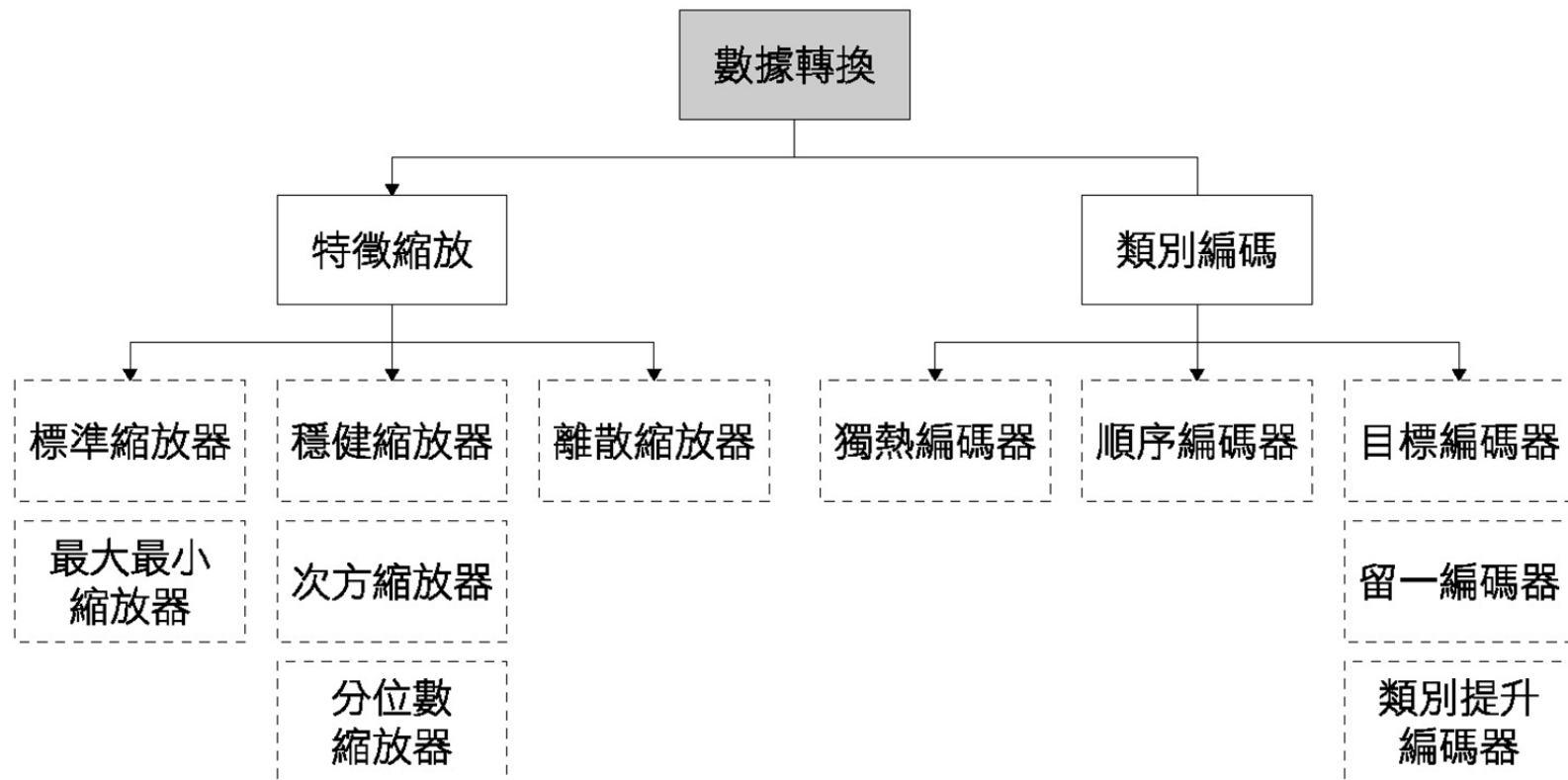
鏈式方程多重填補法(MICE)

Multiple Imputation by Chained Equations (MICE) – Single Iteration



□ 數據轉換

- 是針對**特徵**的轉換，是為了使得特徵之間或特徵與目標值間的分布範圍與結構更加接近，平衡自身分布與分布間的關係將使得模型易於建構。
- 「連續型」特徵的轉換 → 「縮放」(scaling)
- 「類別型」特徵的轉換 → 「編碼」(encoding)

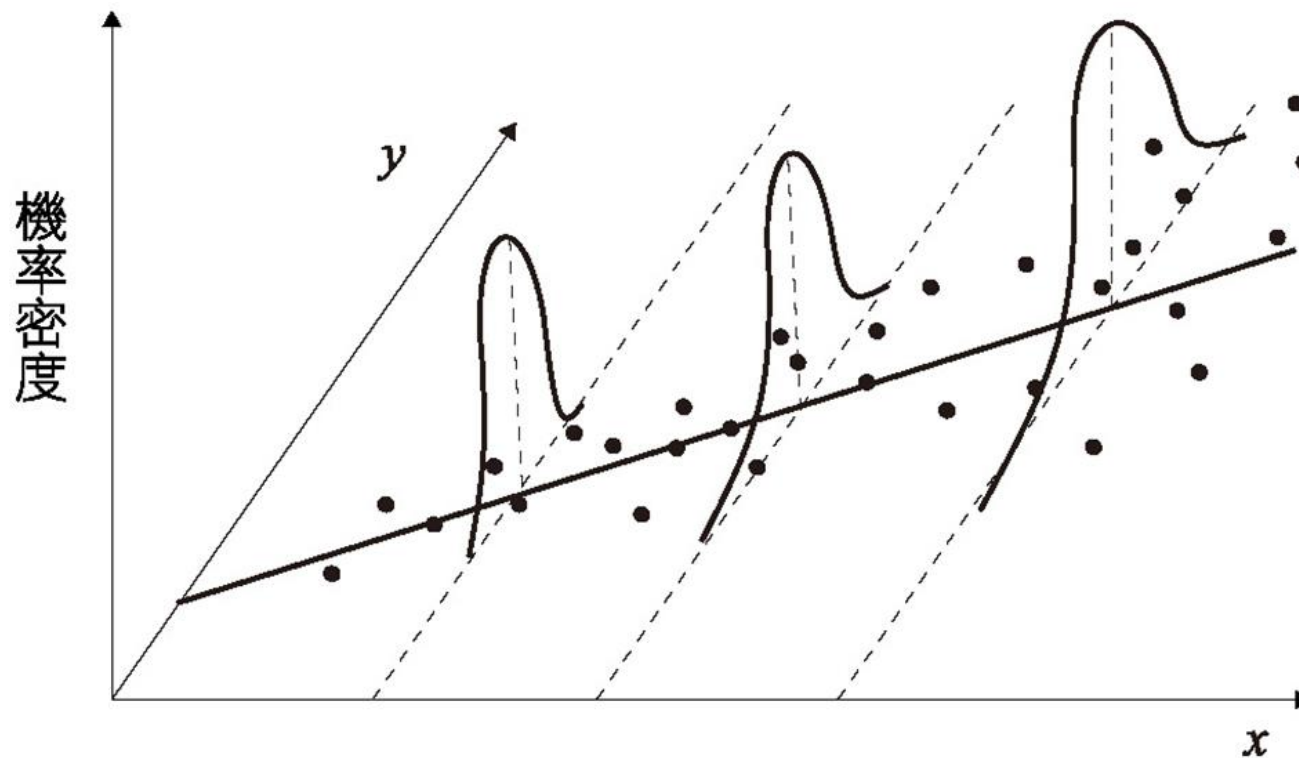


□ 特徵縮放方法

- 「標準縮放器」 (standard scaler)
 - 「標準化」 (standardization) : $x' = \frac{x - \bar{x}}{s}$
 - 其中 \bar{x} 為樣本平均數而 s 則為樣本標準差，轉換後平均數為0且標準差為1
- 「最小最大縮放器」 (min-max scaler)
 - 「歸一化」 (normalization) : $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$ or $x' = \frac{x}{\max(x)}$
 - 轉換後介於0與1之間，其中 $\max(x)$ 與 $\min(x)$ 可以是目前數據集中的最大值與最小值，或可以是歷史數據中的最大值與最小值。
- 「穩健縮放器」 (robust scaler)
 - 處理離群值的箱型圖方法 : $x' = \frac{x - \text{median}(x)}{\text{IQR}(x)}$
 - 以中位數作為中心、四分位間距作為變異進行縮放，對於離群值相對穩健
- 「次方轉換器」 (power transformer)
- 「分位數轉換器」 (quantile transformer)
- 「離散轉換器」 (discrete transformer)

□ 特徵縮放方法：次方轉換器

- 特徵以特定非線性函數進行轉換，並期望轉換後分配近似於「**常態分配**」，常用於具有「異質性」(heteroscedasticity)特性的數據分布，也就是特徵隨著另一特徵在不同數值大小下產生變異數不同的情形。



□ 特徵縮放方法：次方轉換器

- 能減緩具有極端偏態分布或離群值所帶來的不平衡
- 「Box-Cox 轉換」
- 「Yeo-Johnson轉換」

$$\text{Box - Cox: } x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln x & \text{if } \lambda = 0 \end{cases} \quad (5.5)$$

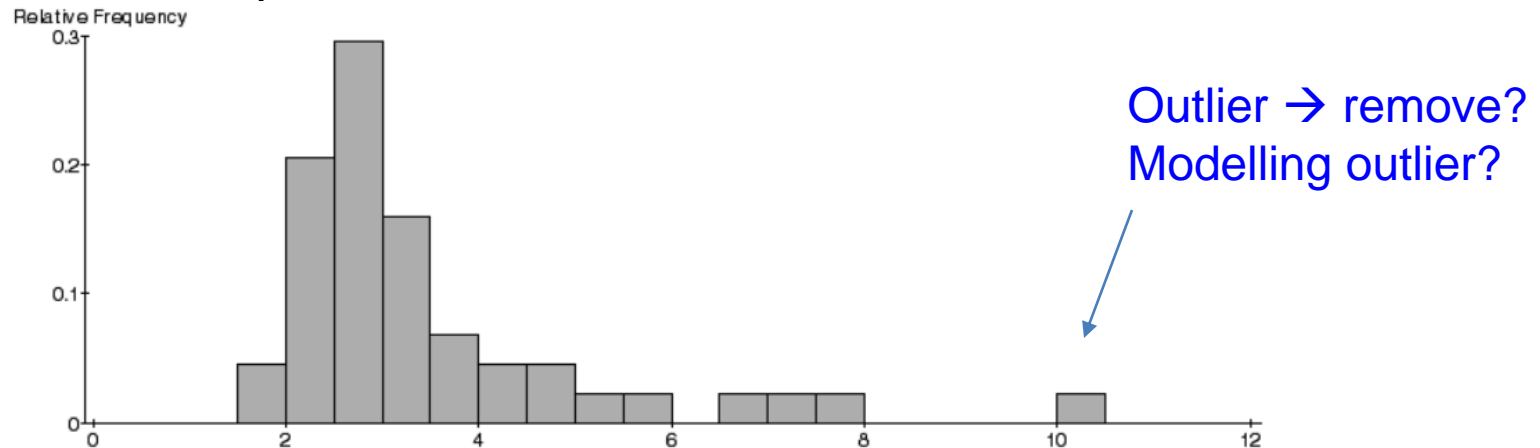
$$\text{Yeo - Johnson: } x^{(\lambda)} = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, x \geq 0 \\ \ln(x+1) & \text{if } \lambda = 0, x \geq 0 \\ -\frac{(-x+1)^{(2-\lambda)} - 1}{2-\lambda} & \text{if } \lambda \neq 2, x < 0 \\ -\ln(-x+1) & \text{if } \lambda = 2, x < 0 \end{cases} \quad (5.6)$$

□ Rounding

- Replacing a number with a different number at the **specified number of digits** that is approximately equal to the original.
- Even you can $\text{round}(\text{value} * m)$, $\text{round}(\log(\text{value}))$, or round the value as categorical feature.

□ Log/SquareRoot/Exp Transformation (**Log for long tail**)

- $\log(x)$ slowly increase over x , that is, log can compress a large number and expand a small number. Eg. $\text{Log}([100, 1000], 10) = [2, 3]$. Also, we may use $\log(1 + x)$ or $\log(x / (1 - x))$. Similarly to square root or cube root.
- Exp is use to disperse the data.



□ 特徵縮放方法：分位數轉換器

- 依據「分位數」使特徵分布（累積機率密度）去近似「常態分配」（或「均值分配」）。其思維與次方轉換器相似，但更直接地改變分配結構
- 不論是「次方轉換器」或「分位數轉換器」都可能會破壞特徵間的線性關係，因此須留意欲轉換的特徵與其餘特徵原有的關係。

□ 特徵縮放方法：離散轉換器

- **離散化(discretization)**: 將連續資料分配到數個小區間
 - 包括了依據「領域知識」、「二值化」(binarization)、「分位數」、「分箱法」以及「分群法」（例如單一特徵的K平均分群法）
- Binarization 二值化
 - If x is larger than a **threshold**, then transform x to 1; otherwise 0.
 - For categorical color, if you don't care what color is, then you can use 1 representing color; 0 for no color (transparent or white).
- Binning and Bucketization 裝箱 (for discretization 離散化)
 - Bucketizer transforms a column of continuous features to a column of feature **buckets**, where the buckets are specified by users.
 - Eg. age: [0-19] for 1; [20-39] for 2; [40-59] for 3; [60 above] for 4.
 - For categorical variable, you may use **Group By** or **Select Count()**, and then transform x to "**Other**" if the frequency less than a threshold.

數據轉換

□ 特徵縮放方法：離散轉換器

- Question: age: [0-19] for 1; [20-39] for 2; [40-59] for 3; [60 above] for 4.

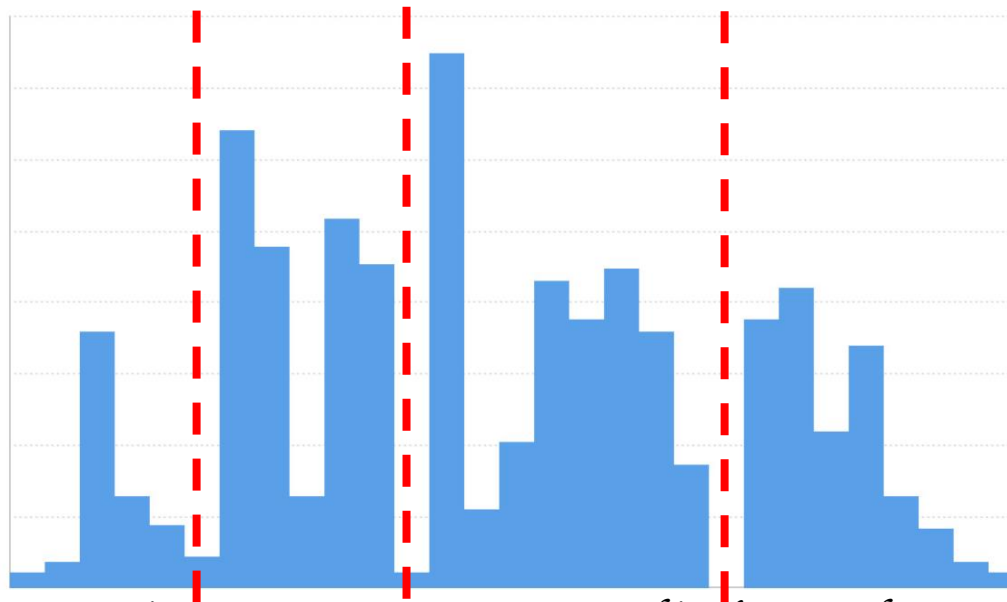
— Is this a good method with Likert scale? Misleading? (WHY?)

➤ Half-day	1	\$100
➤ Day	2	\$200
➤ Half-week	3	\$500
➤ Week	4	\$1000
➤ Half-month	5	\$2000
➤ Month	6	\$4000

類別!! 不可 $\pm \cdot \div$

為啥“月”是“半天”的40倍?需解釋!

- Use distribution for binning, eg. quantization or **quantile binning**.
- **Draw distribution and then binning**



<https://antv-2018.alipay.com/zh-cn/vis/chart/histogram.html>

□ 類別編碼

- 「獨熱編碼器」 (one-hot encoder, or dummy encoder)
 - 以多個二元特徵取代單一類別特徵
- 「順序編碼器」 (ordinal encoder)
 - 具有順序特性的特徵轉換
- 「貝氏編碼器」 (Bayesian encoder)
 - 以目標值將類別特徵轉為連續特徵
 - 「目標編碼器」 (target encoder)
 - 「留一編碼器」 (leave-one-out encoder)
 - 「類別提升編碼器」 (catboost encoder)

類別編碼

- 獨熱編碼器(one-hot encoding)
 - 將一個類別的每個「水準」轉為各別一個「二元變數」(binary variable)又被稱為「啞變數」或虛擬變數(dummy variable)

	↓ 獨熱編碼	↓ 順序編碼
藥水類型	藥水 A 藥水 B 藥水 C	藥水類型
藥水 A	1 0 0	1
藥水 B	0 1 0	2
藥水 C	0 0 1	5
藥水 C	0 0 1	5
藥水 A	1 0 0	1
藥水 B	0 1 0	2

- Given N levels, the method will generate **N-1 dummy variables**. (共線性問題)
- 順序編碼器
 - 將具有「程度」之分的類別特徵轉為連續的方法

□ 類別編碼：獨熱編碼器

- 三種建議方式可以縮減水準數目
- 類別整合(grouping)
 - 把相似的類別合併，例如產品轉為成產品群組、機台轉為機台群組、處方轉為處方群組、模具轉為模具群組、藥水轉為藥水群組等。
- 縮小數據收集的時間區間
 - 例如收集一年的數據可能造成某類別特徵（例如處方）有大量的水準，然而若現場做故障排除時，可能只需要一個月的數據收集，這時可大幅減少欄位裡的類別數目。
- 刪除只出現一次的水準
 - 因為該水準缺少「再現性」，以科學的角度不易「驗證」，例如只開一次模的模具。

類別編碼：貝氏編碼器(Bayesian encoder)

- 當「領域知識」不足情況下，同時也不想增加數據維度時，可採用目標值(label)為依據的編碼方式
- 目標編碼器：將類別特徵在某水準下目標值的統計量（平均數）作為其轉換後的數值，每一水準以其目標值「良率」的平均作為編碼後的數值。
- 留一編碼器：為目標編碼器的延伸，而其不同之處在於取平均數時將該樣本的目標值排除在外，以減低離群值所帶來的影響。
- 類別提升編碼器：為了改善前兩種編碼器的「數據洩漏」問題，引入了「時間」的概念，先將樣本進行隨機重新排列，對於某一樣本進行數據轉換時，僅使用這一系列數據前的觀測值進行目標值（標籤）的平均值計算。

		↓ 目標編碼	↓ 留一編碼	↓ 類別提升編碼
藥水類型	良率	藥水	藥水	藥水
藥水 A	0.75	0.7	0.65	-
藥水 B	0.78	0.8	0.82	-
藥水 C	0.87	0.9	0.93	-
藥水 C	0.93	0.9	0.87	0.9
藥水 A	0.65	0.7	0.75	0.7
藥水 B	0.82	0.8	0.78	0.8

數據理解就是選取資料(範圍)、資料異質、資料品質

- 檢視資料三項目：樣本個數(n)、變數/特徵個數(p)、異質性(資訊量/格式)
 - 一個母體平均數的單尾檢定時，當樣本數佔母體總數的比例R=80%時，顯著水準0.05 要調整為0.0001. (馬瀾嘉，2019)
 - 樣本大到接近母體，不需要檢定，直接敘述統計視覺化
- 變數太多造成維度過高，使分析時間過長→ 維度詛咒

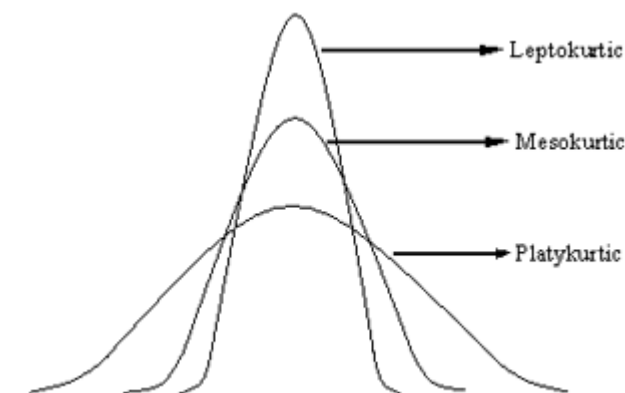
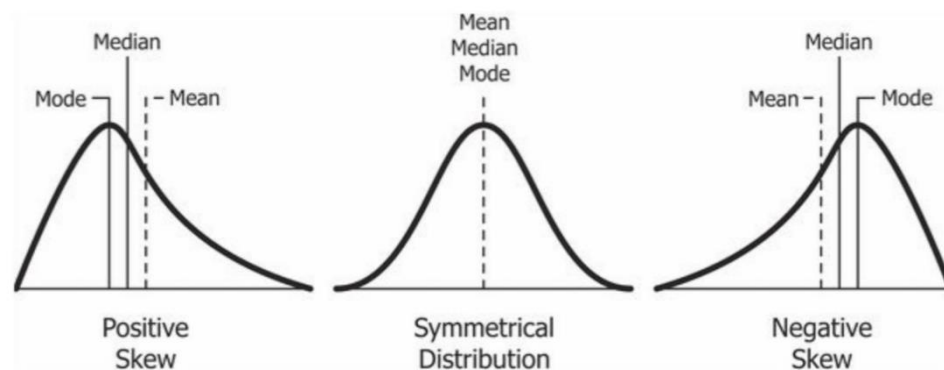
範圍	預期的樣本比例在範圍內	近似預期頻率超出範圍	近似頻率 (假設每天實驗一次)
$\mu \pm 0.5\sigma$	0.382 924 922 548 026	3次中發生2次	每星期四至五次
$\mu \pm \sigma$	0.682 689 492 137 086	3次中發生1次	每星期兩次
$\mu \pm 1.5\sigma$	0.866 385 597 462 284	7次中發生1次	每星期
$\mu \pm 2\sigma$	0.954 499 736 103 642	22次中發生1次	每三個星期
$\mu \pm 2.5\sigma$	0.987 580 669 348 448	81次中發生1次	每三個月
$\mu \pm 3\sigma$	0.997 300 203 936 740	370次中發生1次	每年
$\mu \pm 3.5\sigma$	0.999 534 741 841 929	2 149次中發生1次	每六年
$\mu \pm 4\sigma$	0.999 936 657 516 334	15 787次中發生1次	每43 年 (約一生兩次)
$\mu \pm 4.5\sigma$	0.999 993 204 653 751	147 160次中發生1次	每403 年 (近代以來僅1次)
$\mu \pm 5\sigma$	0.999 999 426 696 856	1 744 278次中發生1次	每4776年 (人類記錄歷史以來僅1次)
$\mu \pm 5.5\sigma$	0.999 999 962 020 875	26 330 254次中發生1次	每72 090年 (智人出現以來僅4次)
$\mu \pm 6\sigma$	0.999 999 998 026 825	506 797 346次中發生1次	每138萬年 (直立人出現以來僅1-2次)
$\mu \pm 6.5\sigma$	0.999 999 999 919 680	12 450 197 393次中發生1次	每3400萬年 (恐龍滅絕以來僅2次)
$\mu \pm 7\sigma$	0.999 999 999 997 440	390 682 215 445次中發生1次	每10.7億年 (地球誕生以來僅4次)

馬瀾嘉（2019），巨量資料分析下如何調整顯著水準，智慧科技與應用統計學報，16（2），19-36。

□ Data Quality Investigation

● Univariate Statistic

- Statistics: mean(median, mode), variance, skewness, kurtosis
- benchmarking with Engineer's intuition



● Univariate Interpretation

- Linear Regression (with response variable Y).
- $y = \beta_0 + \beta_i x_i + \varepsilon$
- If the estimate of the β_i **violates the engineering experience** (sign changed)
 - Eg. Etching time (x) and thickness (y) should be with negative β_i .
- Then the variable may have quality issue.

□ Data Quality Investigation

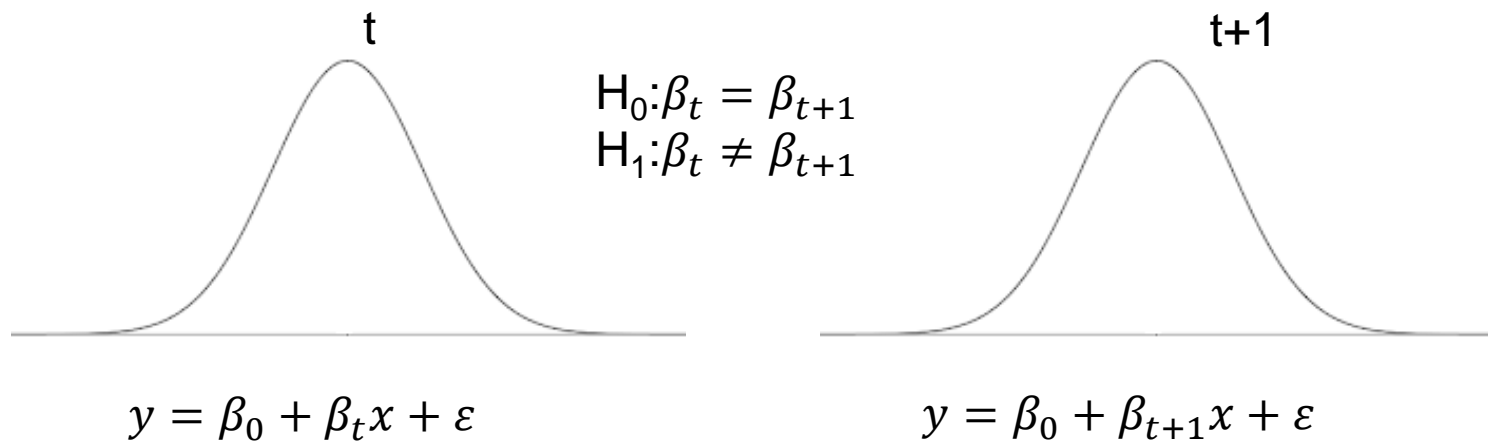
● Multiple Column

— Correlation or Covariance between two variables

- Eg. Height and weight should be positive correlation.
- Eg. Grades of MATH and SCIENCE should be positive correlation.
- Eg. Lithography: Exposure latitude (EL) versus depth of focus (DOF) should be **negative**.

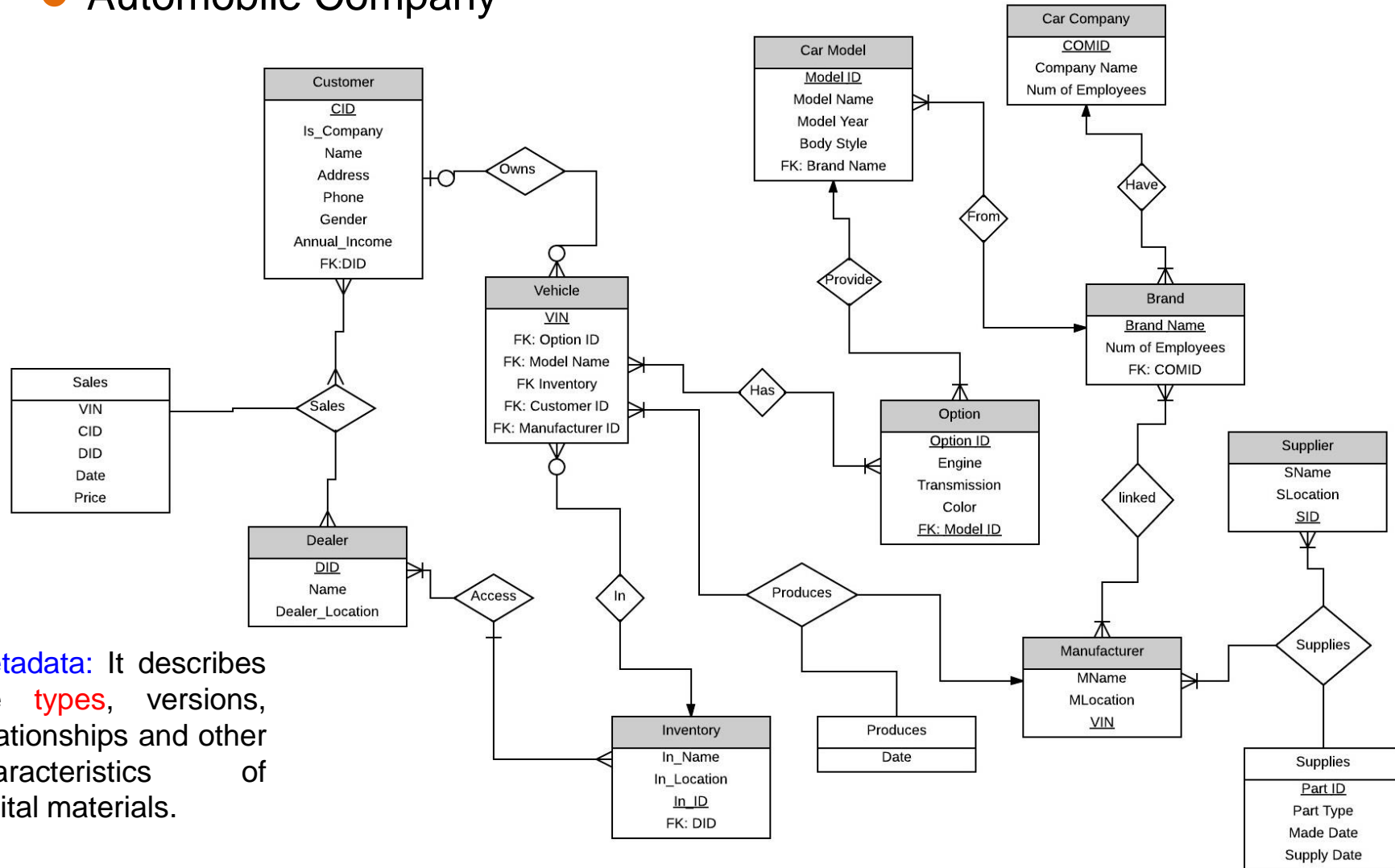
● Validity and Reliability (信度與效度)

- 敘述性統計大部分做的是checking within distribution
- 也可根據不同時間點對同一母體收集資料，做between distribution checking



Data Quality Investigation- Entity-Relationship (ER) Model

Automobile Company

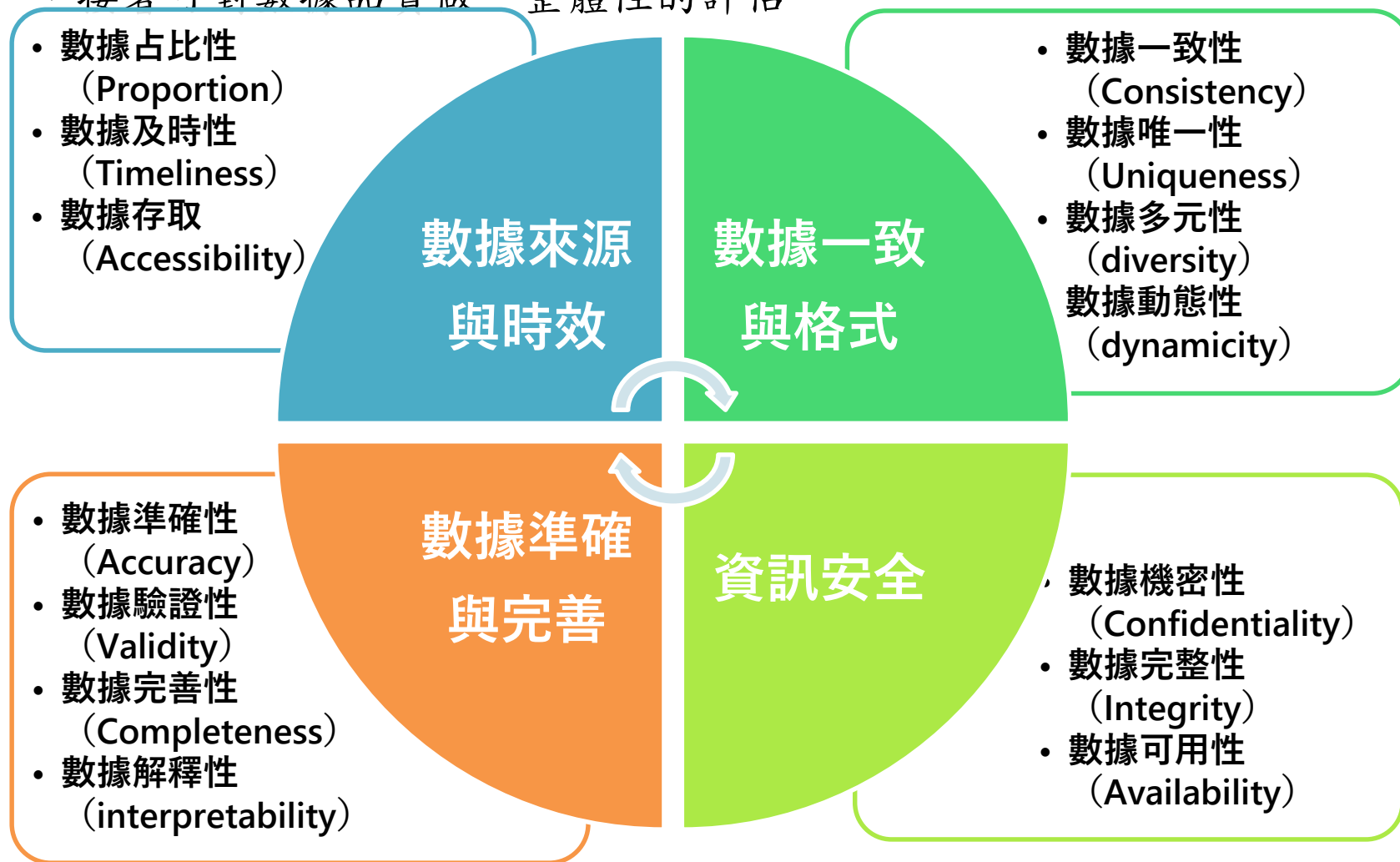


Metadata: It describes the **types**, versions, relationships and other characteristics of digital materials.

數據品質評量

□ 數據品質的評量構面

- 除了透過視覺化(visualize)確認數據集的噪音、離群值、遺漏值等狀況，接著可對數據品質做一整體性的評估



□ 數據來源與時效

- 數據占比性 (proportion)
 - 收集數據樣本相對於母體所佔的百分比。
- 數據及時性 (timeliness)
 - 數據的年齡或收集時間是否適用於分析目前的問題。例如數據是否即時反應實際系統的真實狀況。
- 數據存取性 (accessibility)
 - 數據可獲得性，是否使用者有適當權限存取數據且容易快速檢索 (retrievable)

□ 數據準確與完善

● 數據準確性 (accuracy)

- 從資料庫中收集到的數據是否是正確、可靠且驗證過，能夠真實地反應實際狀況。可從元數據 (metadata)，並同時確認數據表單之間的關係 (entity-relation model)。或進行迴歸分析，例如研磨時間 (x) 與產品的厚度量測 (y) 應為負相關。

● 數據驗證性 (validity)

- 對於輸入資料正確性的檢查工作，目的在確定資料處理時所使用資料正確無誤且符合需求。

● 數據完善性 (completeness)

- 表達所收集數據其深度與廣度對於特定問題提供資訊量的完整度。數據資訊量是否收集完善，滿足「互無遺漏」 (collectively exhaustive) 的原則可協助決策議題，通常可藉由魚骨圖分析展開或預測績效指標加以評估。

● 數據解釋性 (interpretability)

- 數據特徵具有物理特性或因果關係上可闡述的能力。

□ 數據一致與格式

● 數據一致性 (consistency)

- 數據不可違反語義規則的定義或資料庫資料型態的定義。包含數據值與工程物理特性是否一致（例如數值上下界區間規範）、同樣變數欄位在不同資料庫間是否有一致的定義或資料類別 (type)（也就是元資料metadata 的一致性）、新舊數據的相容性 (compatible)。

● 數據唯一性 (uniqueness)

- 對於特定字串、記錄或數據集，系統內部或跨系統間不需要有重複的存在。

● 數據多元性 (diversity)

- 也稱多樣性 (variety)，與完善性互補。數據樣本應有一定的代表性，且樣本間的多元變異對決策議題呈現足夠與不同視角的資訊量。或可與異質性 (Heterogeneity) 有關，其數據收集格式的來源可為異質，可包含結構化（例如數據表單）與非結構化的數據（例如影片、文字、聲音、圖像等）。

● 數據動態性 (dynamicity)

- 衡量數據易變 (variability) 的特質，數據是否時間相關 (time-related) 或概念漂移 (concept drift)。同一特徵在相同(相似)樣本在不同時間點的機率分配是否一致，或兩個以上的特徵間在不同時間下的相關性（例如兩特徵的相關係數）是否一致。

□ 資訊安全

● 數據機密性 (confidentiality)

- 資訊安全三要點之一，保護企業的敏感數據，確保資料傳遞與儲存的隱密性，可透過加密 (encryption) 或認證 (authentication) 等方式，避免未經授權 (authorization) 的使用者有意或無意的揭露資料內容。

● 數據完整性 (integrity)

- 資訊安全三要點之一，在傳輸、儲存資訊或資料的過程中程中有其正確性與一致性，確保資訊或資料不被未授權的修改、刪除、損毀，或在篡改後能夠被迅速發現。

● 數據可用性 (availability)

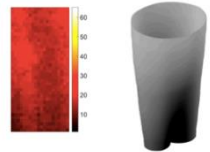
- 資訊安全三要點之一，當使用者需透過資訊系統進行操作時，資料與服務須保持可用狀況並能滿足使用需求。

□ Note

- 對於數據管控，應建立存取機制與培訓人員對於數據品質與資安的意識。例如盡量減少組織內人為活動，所導致的數據品質問題，同時應設計系統存取與權限等機制或審核流程。此外，所有開發人員和資料庫管理員都必須對業務流程有很好的詮釋與認知，並且在開發和設計資料庫和應用程序有統一且一致的參考模式。

表 5.3 製造現場不同的數據類型

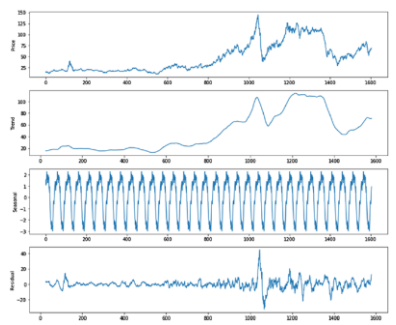
數據種類	特性	主要特徵	常見的特徵工程與分析模型
一般數據	特徵之間相關	領域知識相關特性	<ul style="list-style-type: none">以領域知識進行的數據轉換與整合統計量（平均值、標準差、偏態、峰態等）特徵學習迴歸分析、決策樹、符號迴歸（symbolic regression）
時間序列	<ul style="list-style-type: none">時間相關（趨勢性、週期性、季節性）函數結構低維度	時域、領域知識相關特徵	<ul style="list-style-type: none">時間標記特徵滑動時窗特徵（sliding window based statistics）時間序列分解差分整合移動平均自迴歸模型（autoregressive integrated moving average, ARIMA）
高頻訊號	<ul style="list-style-type: none">時間相關（趨勢性、週期性）振動訊號（高頻）函數結構低維度高取樣頻率	時域、頻域、時頻域特性	<ul style="list-style-type: none">滑動時窗特徵（sliding window based statistics）傅立葉轉換（Fourier transform, FT）短時距傅立葉轉換、小波轉換（wavelet transform, WT）希爾伯特-黃轉換（Hilbert-Huang transform, HHT）
圖片影像	<ul style="list-style-type: none">空間相關多通道訊號高維度（3D 影像）	影像特性	<ul style="list-style-type: none">卷積神經網路（convolutional neural network, CNN）自編碼器（autoencoder, AE）生成對抗網路（GAN）
影片	<ul style="list-style-type: none">時間與空間相關多通道訊號高維度	上述所有特性	上述所有方法



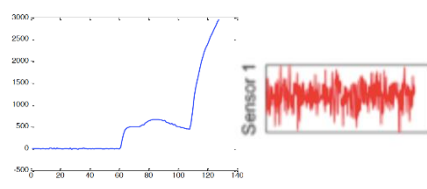
Paynabar, K. (2019). Data science for manufacturing automation: low dimensional learning from high dimensional data. IEEE CASE 2019 Data Science Workshop.

Value x
(one-to-one)

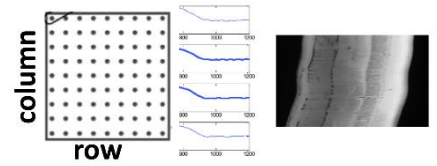
Vector $x = f(t)$
(many-to-one)



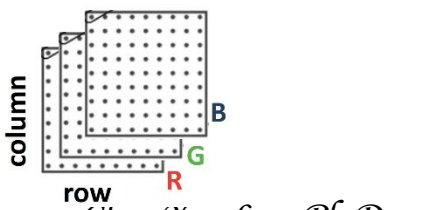
Vector $x = f(t)$
(many-to-one)



Matrix X
(many-to-one)



Tensor \mathcal{X}
(many-to-one)



Data Source	Scale	Issues
Production data (MES)	Categorical/continuous /time	High dimension, multicollinearity, class imbalance, missing value
Equipment data	Categorical/continuous	High dimension, too many categorical levels, time series, missing value
Parts/Supplier data	Categorical	Too many categorical levels
Transportation data	Categorical/continuous	too many categorical levels, time series, missing value
Maintenance/Repair Data	Binary/categorical/continuous	Typing error, text, missing value, Choosing “others” or “NA”
Testing/Inspection Metrology data	Binary/continuous/figure	Sampling data, time series, multi-response, metrology delay

Revised from Chen (2015)

表 5.4 製造現場特性、管理挑戰與數據分析議題		
現場特性	管理挑戰	數據分析議題
瓶頸站	瓶頸為某項製程加工速度最慢，影響整體系統績效。其挑戰包含瓶頸鑑別、流動變異性、產線平衡、從 WIP 中選擇優先貨批 (lot) 加工、在製品的堆積、瓶頸站轉移 (產品組合改變) 等。	Little's Law 是常用的分析手法之一。考慮流動變異性下的多目標進行生產排程 (考慮每日需求生產目標、產線平衡、換線換模次數、產品交期、能源使用等)。
批量生產	產品批量加工需釐清批量的階層關係、最小生產批量、批量大小、產品組合優化等。	產品加工具階層關係；經過混批、拆批、併批等，批量回溯追蹤不易，數據分散。
平行機台	新舊機台的機差問題、排程的複雜度 (等效非等效、綁機、機台能力限制、迴流等)。	產品在平行機台加工的數據中，有時會存在大量的系統性遺漏值、與其他變數組合造成高維度類別特徵、變數間共線性問題。
黃金機台 / 機台能力差異	製程能力、產出率、良率特別好的機台，使得大量產品由該機台加工，挑戰包含工作負荷不平衡、零組件更換頻率、產能支援 (backup)。	數據選擇偏誤、推論偏誤、數據不平衡、類別特徵呈現製程能力等。
少量多樣	小批量加工、換線換模、產能損失、良率不穩品管不易。	數據不平衡、某類別樣本少、快速換模技術 (內部換模作業移到外部換模作業)。
處方與零件	處方種類繁多製程管理不易、零件庫存管理不易、許多模具零組件僅一次性使用等	由於處方與零件均屬類別特徵且水準數高，須留意類別編碼方式，常造成高維度的類別特徵 (例如使用 one-hot encoding)

表 5.4 製造現場特性、管理挑戰與數據分析議題		
現場特性	管理挑戰	數據分析議題
抽樣品檢	由於未全檢將產生量測值的遺漏，量測過程可能產生延遲或干擾。抽樣頻率可依製程能力改變（Cp 與 Cpk）而調整。	遺漏值（有抽才有值）、多反應值（multi-response）、量測延遲、數據不平衡（良品多不良品極少）。
維修保養	保養成本與維修成本的權衡、修機的產能損失、機台可靠度分析（MTTR 與 MTBF）、虛擬量測預測異常後需留意保養決策。	手寫表單數據（雙周保、月保、季保、年保）、文字描述、打字（typing）錯誤、故障排除下拉式選單「其他」類比例多。
工程與實驗貨	研發產品或機台測試，借機測試的產能損失、整備時間產能損失、機台汙染等。	其數據樣本少（小數據）、其值有時為整體中的特例（離群值、機台參數範圍大）。
等候時間限制	產品在特定製程加工完成後，若未在規範時間進入下一製程加工將產生缺陷。製程路徑、傳送、機台閒置等會增加批次加工機台的排程複雜度。	產品缺陷、產線良率、缺陷的隨機性、排程限制、等候時間限制的分析與驗證等。
換線與換模	為生產不同的產品換模具，造成整備時間的產能損失或影響產品品質。	換線換模時間的變異、加工順序相依（sequence-dependent）的整備時間、產品缺陷等。
存貨	經濟訂購批量、價格折扣、易腐壞（perishable）產品、存貨與缺貨風險權衡、供應鏈長鞭效應。	存貨種類、類別型特徵、數據關聯於供應商、存貨=產能+前置時間+不確定性。

資料來源：Lee and Chien (2020).

□ Duplicate/Redundant Entries Detection

- means that the observations having the same value of features show **different target value**. (同樣一組x有不同的y 怎辦?)
- Two rows with the same feature values but their labels are different.

□ Treatment

- Remove the “**old/out-of-date**” one (from time aspect)
- Cause: some observation having outlier y (有些觀測值是outlier在y)
 - For binary class in y, in the same x we can find the majority class in y and remove the minority class. (找到這組x在y上表現較多majority的反應，把另一反應較少的刪除)
- Need one more additional/**new variable** to discriminate these two (Y as vector)
 - Eg. Furnace with the same lot QC, adding Furnace_location (low, median, high) or wafer_location (1 to 25)
 - Use feature engineering to generate new feature
- Cause: smaller noise
 - Consider the **stochastic model** with noise (eg. regression) when prediction/classifier modelling (eg. noise caused by particles)
- Cause: larger noise
 - Denoise: use moving average to smooth the noise

□ Inconsistency between training data and test data (Brainstorming)

- Label in training but not shown in testing

- X

- Y (label)

- Label in testing but not shown in training

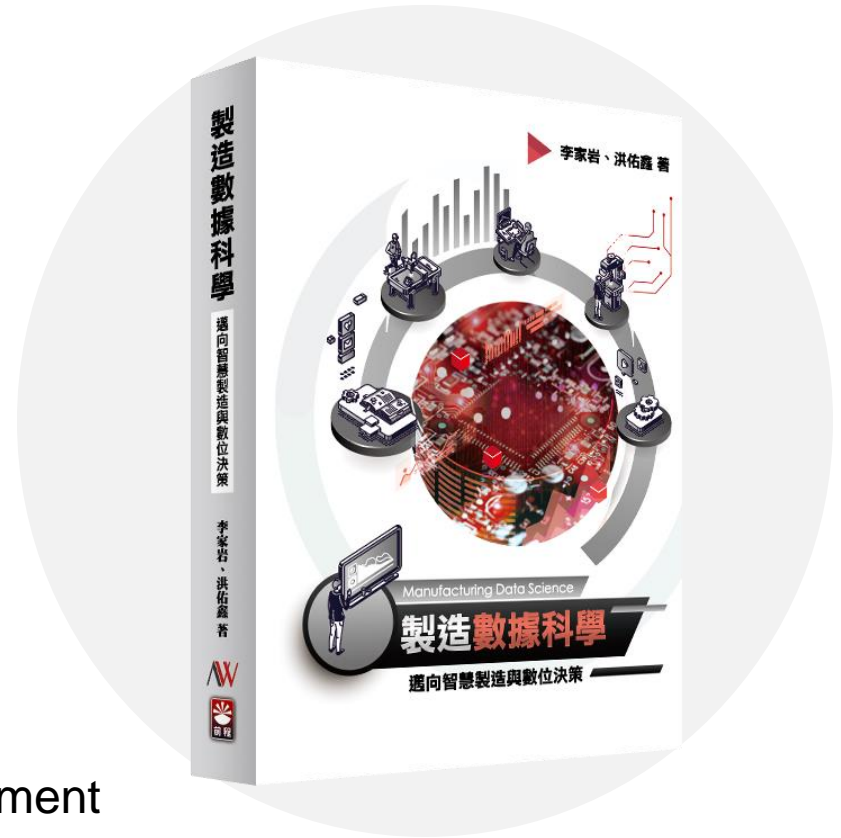
- X

- Y (label)

□ 結語

- 為避免garbage-in garbage-out，數據預處理為關注的焦點，花費大量的時間與人力，常佔整個數據分析專案所花費時間的七成以上。
- 數據才是提升預測準確度的有效辦法，而非所使用的模型。
- 透過探索式資料分析(EDA)與視覺化的工具，除了協助改善資料品質外，也協助重新檢視問題的釐清與定義。
- 數據預處理的確在分析過程扮演關鍵性的階段，達到「承先（問題定義）啟後（特徵工程與挑選）」的功用。
- 開放數據庫(public dataset)
 - 作為預處理與建模的練習，此處列舉數據科學相關領域常見的開放數據庫做為參考。
 - Kaggle(<https://www.kaggle.com/>)
 - UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/>)
 - Kdnuggets(<http://www.kdnuggets.com/datasets/index.html>)
 - NASA Prognostics Center of Excellence(<https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/>)
 - 政府資料開放平台(<https://data.gov.tw/>)

Thanks for your attention



NTU Dept. of Information Management
name: 李家岩 (FB: Chia-Yen Lee)
phone: 886-2-33661206
email: chiayenlee@ntu.edu.tw
web: <https://polab.im.ntu.edu.tw/>