

*If you can't measure it, you can't manage it.* -彼得·杜拉克

# 資料科學產業應用 Data Science & Business Application Data Analysis

郭俊良 博士

逢甲大學-勞動部雲端運算與數位轉型培訓班

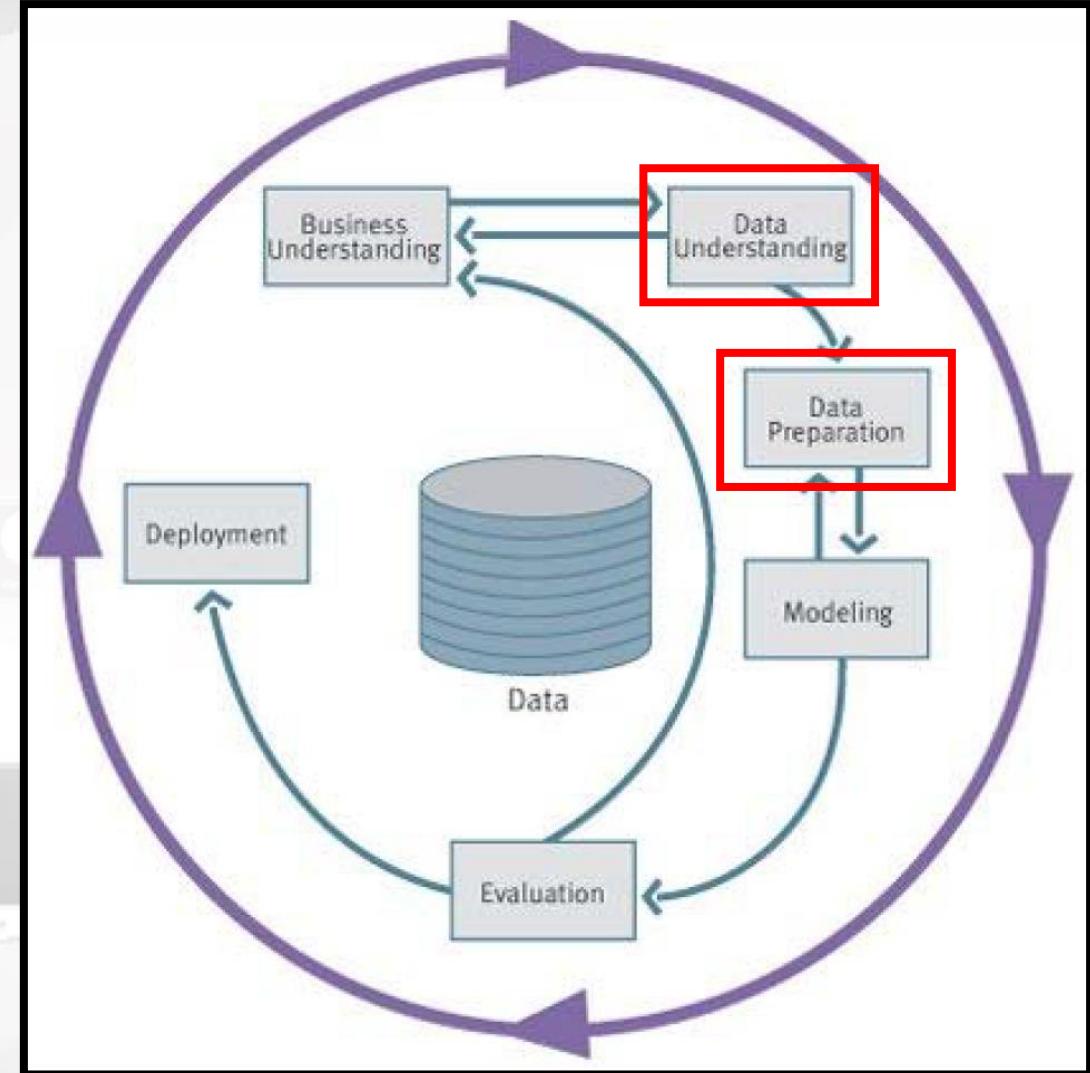
MCNU Logistics Big Data Analytics Research Center

# Outline

- 資料探勘建模SOP
- Data Understanding
- Data Preparation
  - 資料的萃取-轉換-載入 (Data Extract-Transform-Load, Data ETL)
  - 資料檢視 (Data Inspection)
  - 探索性資料分析 (Exploratory Data Analysis, EDA)
  - 資料視覺化 (Data Visualization)
  - 資料調整處理 (Data Manipulation)
    - ◆ 資料清理 (Data Cleaning)
    - ◆ 資料化約 (Data Dimension Reduction)
    - ◆ 資料分割 (Data Partition)
- Workshop 2 – 資料分析操作基礎 / 資料轉換練習

# 資料探勘建模SOP - CRISP-DM週期

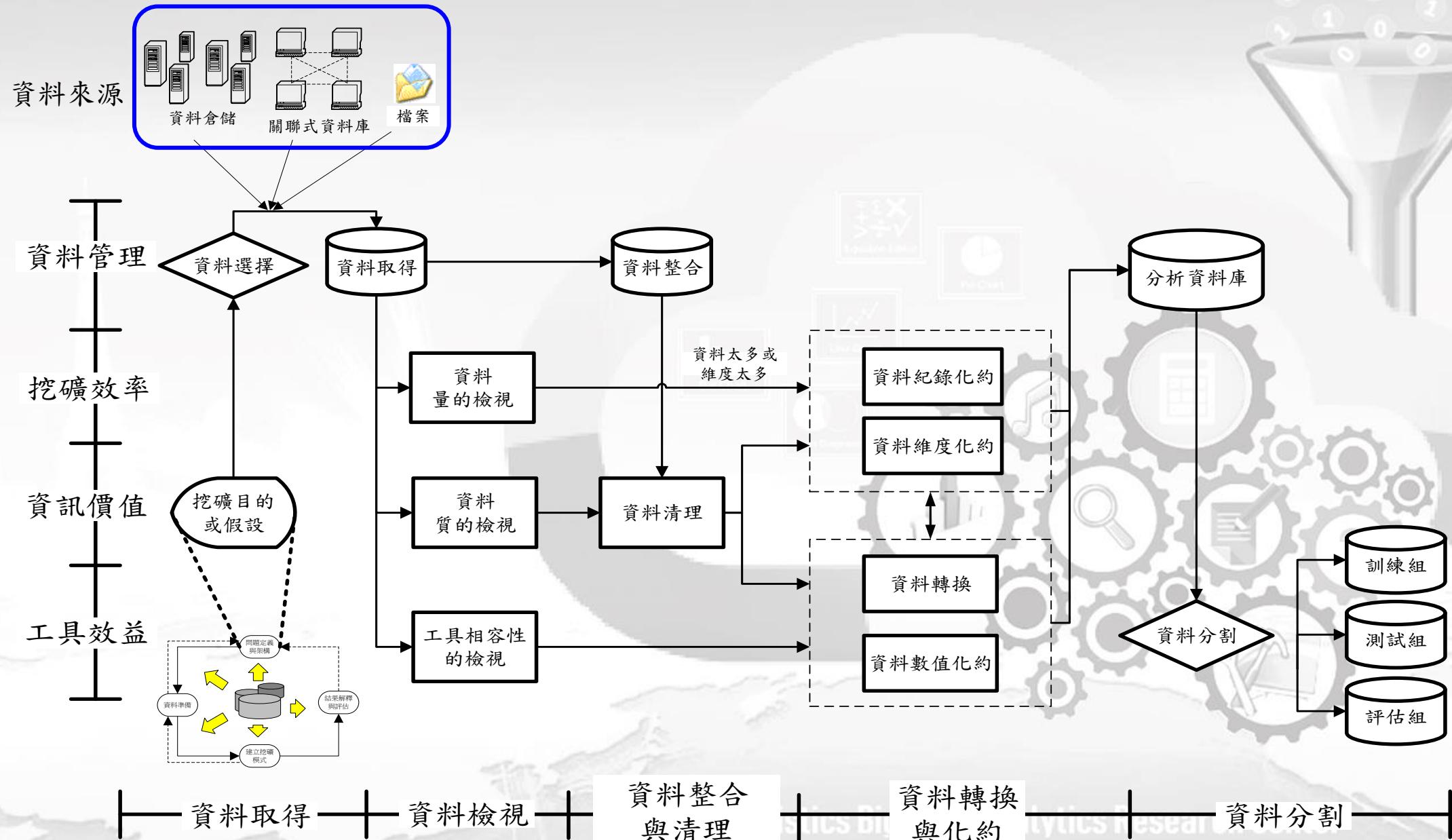
1. 理解商業問題 (Business Understanding)
2. 理解資料特性 (Data Understanding)
3. 資料整備 (Data Preparation)
4. 模型建立 (Modeling)
5. 模型評估 (Evaluation)
6. 模型部署 (Deployment)



# Introduction

- 資料科學建模效益(Performance)良窳的關鍵因素：資料品質(Data Quality)和資料完整性(Data Integrity)
- 資料蒐集方式與工具多元化導致資料庫或資料倉儲存在資料雜訊、資料遺漏及資料格式或尺度不一致的情形
- 巨量數位資料具有4V(Volume、Velocity、Variety、Veracity)特性，若直接分析原始資料，很可能因資料品質不佳而導致事倍功半的結果或推導出具偏誤的結論
- 資料整備(Data Preparation)是指在瞭解問題與目的之後，進行資料分析與建立模式前，為確保分析品質和結果正確性所進行的資料蒐集、資料預處理、資料轉換及資料分割等一連串過程

# 資料整備流程 - 蒐集、理解到結構化



# Data Understanding

Data Structure

Data Source & Acquisition

Data Type

Data Set

Data Scale

# Data Structure

- Structure Data
  - Dataset retrieved from Relational Database System
  - Data Warehouse / Data Mart / Data Lake
- Semi-structured Data
  - CSV
  - Xml
  - Json
  - NoSQL databases
- Unstructured Data
  - Text file
  - Web Scraping data
  - Image File
  - Video File
  - Audio File
  - Signals
  - Others ...



# Data Source & Acquisition

- 檔案(File Data)

- 本機資料：取得快速且閱讀容易，缺點是一旦建立後，後續再做資料處理就不太容易，同時若資料量太大時會增加存取及處理的難度。例如：**Excel**、**純文字資料檔**等
- 線上資料：使用爬蟲程式(Web Crawler)在網際網路的網頁內容爬疏(scrape)下來的資料 (資料多半屬於**非結構化資料**)

- 關聯式資料庫(Relational Databases)

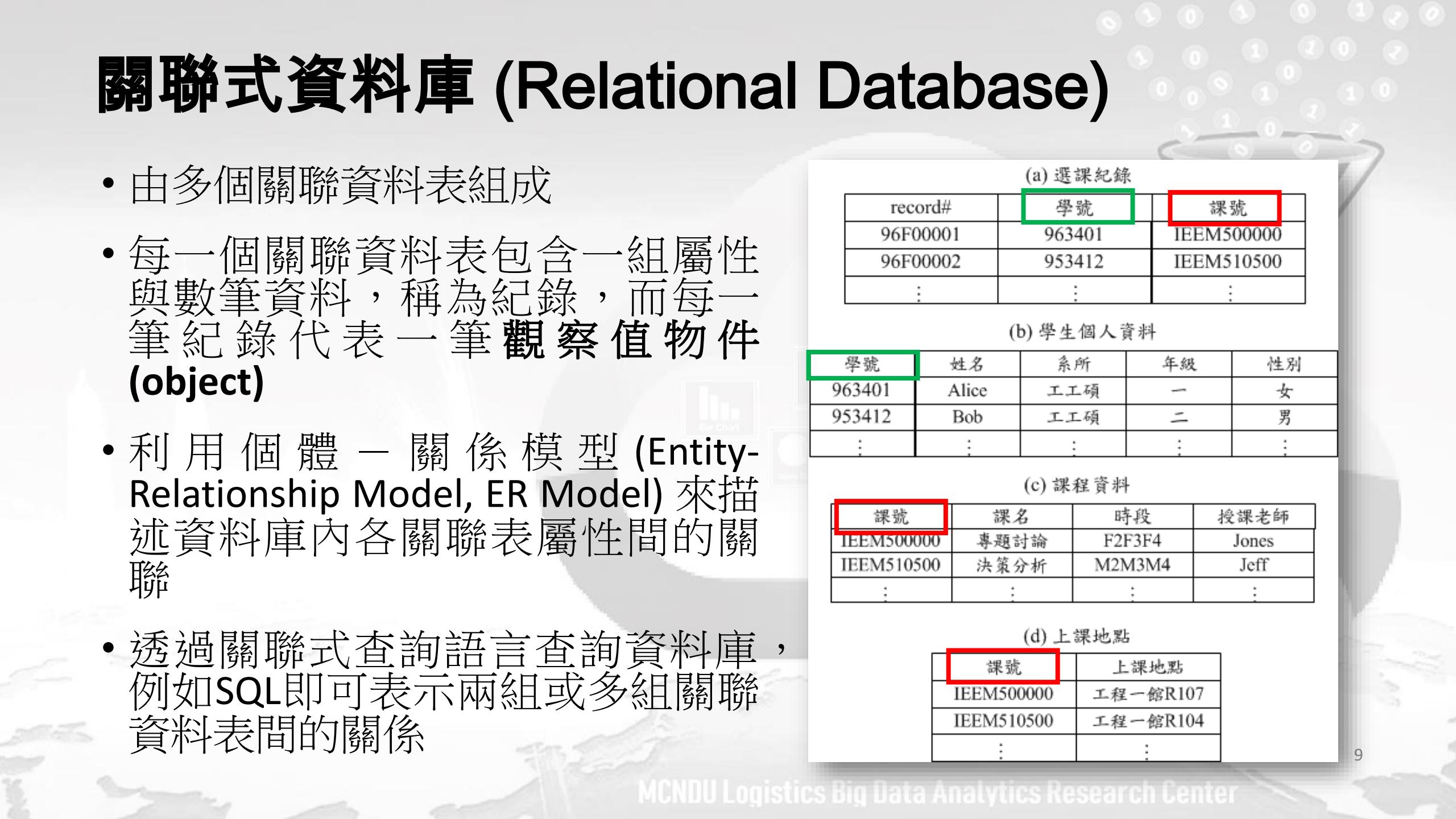
- 資料是以**結構化建模**方式連結資料間的關係，進而產生關聯式資料庫
- 透過查詢語言工具 (Select Query Language, SQL)進行資料存取與操作並將結果輸出，如：**Mircosoft SQL Server**、**mySql**、**Oracle**等

- 資料倉儲(Data Warehouse)與資料超市(Data Mart)

- 大型彙整具**多維度結構**的資料庫，並以「切割」的觀念來讀取資料，資料量可達數百GB甚至TB
- 並非以關聯資料庫透過連結表格的方式處理

# 關聯式資料庫 (Relational Database)

- 由多個關聯資料表組成
- 每一個關聯資料表包含一組屬性與數筆資料，稱為紀錄，而每一筆紀錄代表一筆觀察值物件 (object)
- 利用個體—關係模型 (Entity-Relationship Model, ER Model) 來描述資料庫內各關聯表屬性間的關聯
- 透過關聯式查詢語言查詢資料庫，例如SQL即可表示兩組或多組關聯資料表間的關係



The figure illustrates four relational tables representing student records:

- (a) 選課紀錄 (Grade Record):

record#	學號	課號
96F00001	963401	IEEM500000
96F00002	953412	IEEM510500
:	:	:
- (b) 學生個人資料 (Student Personal Information):

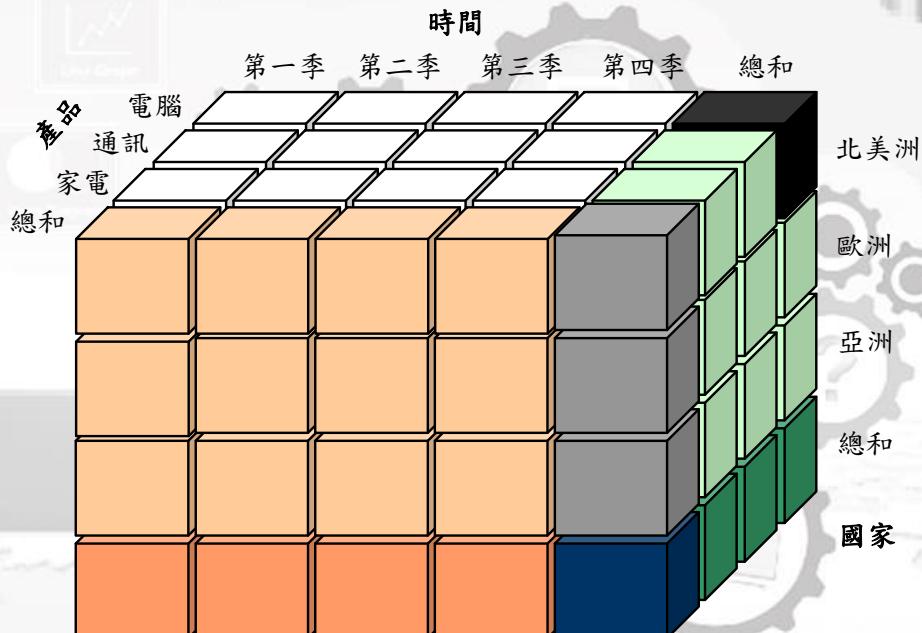
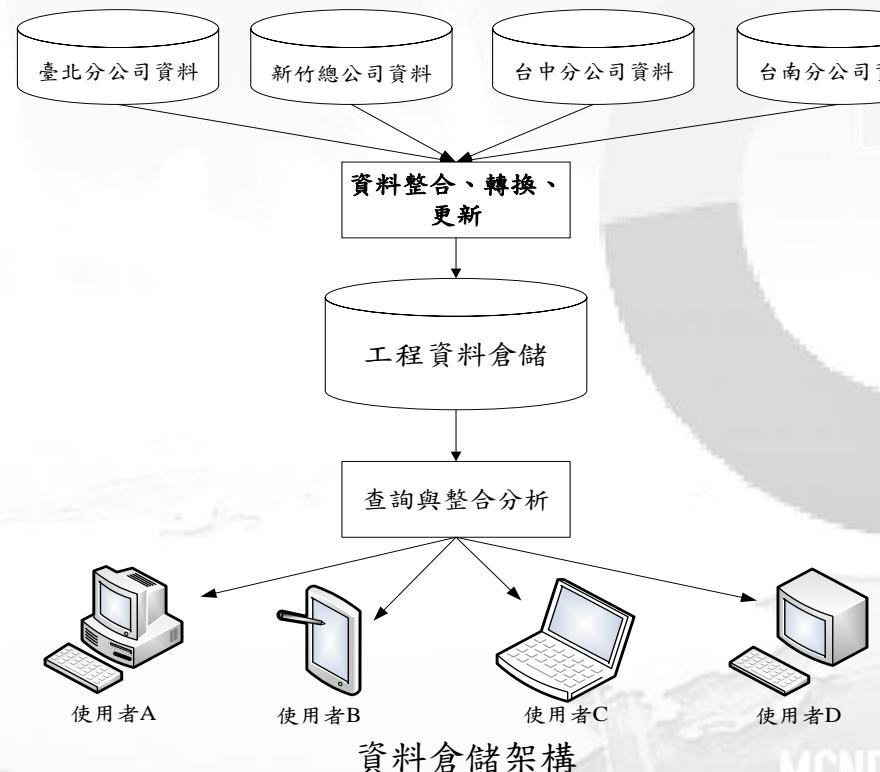
學號	姓名	系所	年級	性別
963401	Alice	工工碩	一	女
953412	Bob	工工碩	二	男
:	:	:	:	:
- (c) 課程資料 (Course Information):

課號	課名	時段	授課老師
IEEM500000	專題討論	F2F3F4	Jones
IEEM510500	決策分析	M2M3M4	Jeff
:	:	:	:
- (d) 上課地點 (Classroom Location):

課號	上課地點
IEEM500000	工程一館R107
IEEM510500	工程一館R104
:	:

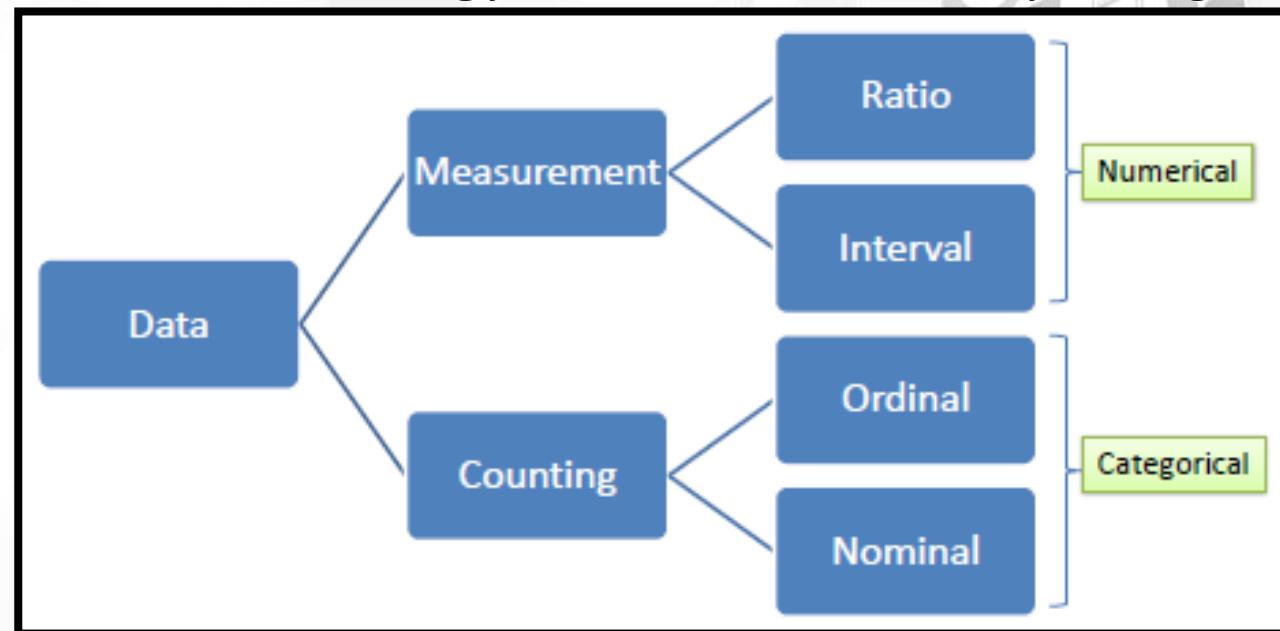
# 資料倉儲 (Data Warehouse)

- 儲存來自不同來源的資料，可由單一或多個資料庫所組成
- 資料大多經過資料處理，並以「切割」觀念來讀取資料
- 利用**多維度資料方塊** (multidimensional data cube) 檢視多維度的資料，以提供分析所需的關聯分析或概念階層的關係



# Data Type – General Concept

- **Numerical Data** (Measurement, **Continuous Variable**)
  - **Numerical**: any value within a finite or infinite interval (e.g., height, weight, temperature, blood glucose, ...)
  - Two types of numerical data: *interval* and *ratio*
- **Categorical Data** (Counting, **Discrete Variable**)
  - Nominal data: "gender" with two categories, male and female
  - Ordinal data: "level of energy" with three orderly categories (low, medium and high)



# Data Type – 基於統計分析區分(1/4)

- 橫斷面資料 (Cross-sectional data)
  - 同一時間、同一議題、針對不同樣本群體(年齡、種族、區域等)所進行蒐集的資料
- 縱貫面資料 (Longitudinal data)
  - 又可稱為 Panel Data
  - 同一組樣本對象，在不同時間進行施測所蒐集的資料，以瞭解資料隨時間所產生的變化，例如：
    - **時間性資料**：資料本身或資料庫中含有時間前後或順序相關的特性，例如某個時間點每一位顧客購買的產品
    - **時間序列資料(Time Series Data)**：記錄著一段時間區間的結果，例如某一檔股票的每日股價
  - 縱貫性資料可進行三種研究：趨勢研究(Trend Study)、世代研究(Cohort Study)、固定連續樣本研究(Panell Study)

# Data Type – 基於軟體應用類型區分(2/4)

- **空間資料**：為資料中包含空間相關的屬性，如亞洲區域的氣溫資料、Google Map、地理資料庫、積體電路設計規劃等
- **文字資料**：常見如專利報告、診斷報告、筆記、產品規格書等
  - 結構化資料、半結構化資料、非結構化資料
  - 文字資料的處理稱為**文字探勘**(Text Mining)，常見的應用包括文件分群、摘要擷取
- **多媒體資料**：包括圖片、聲音及視訊等，因檔案大小一般都非常龐大，在資料的儲存與搜尋上均需要特殊的方法，例如資料壓縮(data compression)

# Data Types – 基於電腦儲存格式區分(3/4)

- 數值

- 最常用的資料類型，儲存內容為數值型態
  - 包含整數(Integer)、浮點數(float)

- 字元與字串

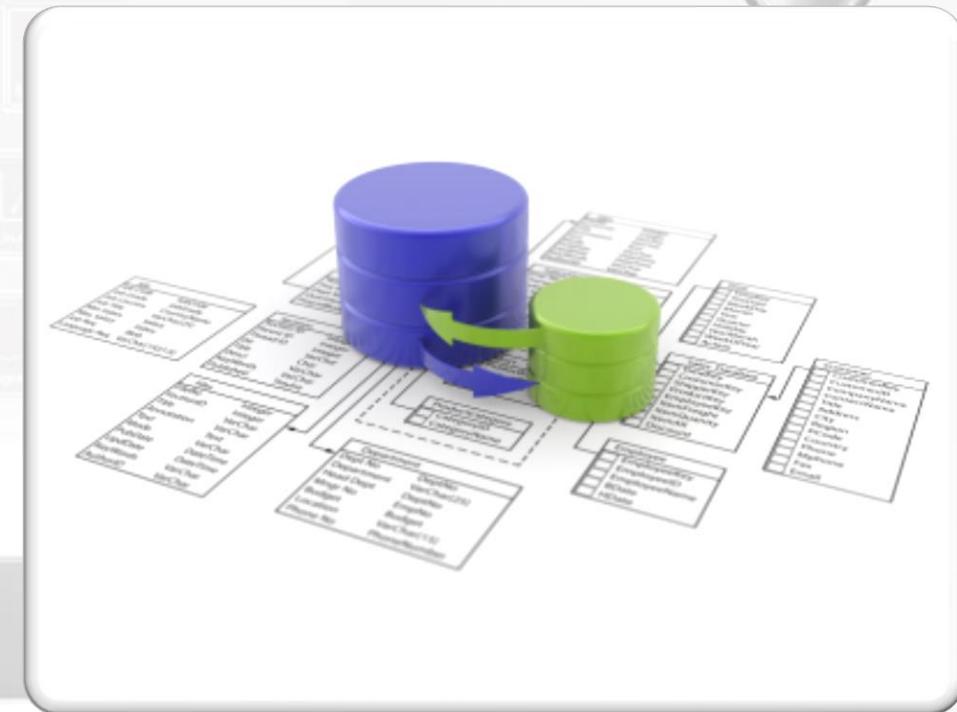
- 字元(char)：電腦將任一鍵盤上的任一符號均視為字元，例如以「男」與「女」字元記錄性別
- 字串(string)：是字元的陣列組合；字串資料型態即是儲存一串互為相同或不同的字元，例如姓名「王大功」是「王」、「大」與「功」等三個字串的陣列組合

- 布林值 (Boolean)

- 布林資料只有兩種值—真(True)與偽(False)，通常用來儲存一些可供程式判斷的條件結果，例如以「真」記錄滿18歲的人，以「偽」記錄未滿18歲的人

# Data Type – 基於程式語言區分(4/4)

- 原始資料型態
  - 數值資料 (numeric)
  - 字元資料 (character)
  - 邏輯資料 (Boolean)
- 衍生資料型態
  - 字串 (string)
  - 陣列 (array)
  - 矩陣 (matrix)
  - 時間/日期 (date : R, Python)
  - 串列 (list : R, Python)
  - 資料框 (data frame : R, Python)



# Data Set

- A **collection of data** presented in a **tabular form**
- Alternatives for columns, rows and data
  - **Columns**: Also called Fields, Attributes, Variables
  - **Rows**: Also called Observations, Records, Objects, Cases, Instances, Examples, Vectors, Tuple (值組)
  - **Data**: Also called Values, Cell
- In predictive modeling, “columns” are also called:
  - **predictors or attributes** are the input variables
  - **target or class attribute** is the output variable whose value is determined by the values of the predictors and function of the predictive model

# Table of Dataset

ID	Outlook	Temp	Humidity	Windy	Play Golf
1	Rainy	85	92	False	No
2	Rainy	80	88	True	No
3	Overcast	83	86	False	Yes
4	Sunny	70	80	False	Yes
5	Sunny	68	?	False	Yes
6	Sunny	65	56	True	No
7	Overcast	64	62	True	Yes
8	Rainy	72	95	?	No
9	Rainy	?	70	False	Yes
10	Sunny	75	72	False	Yes
11	Rainy	75	74	True	Yes
12	?	72	78	True	Yes
13	Overcast	81	66	False	Yes
14	Sunny	71	79	True	No

# Data Scale

- 資料都有對應的屬性(attribute/data type)及其衡量尺度(scale)，以量化和衡量不同資料在該因子的水準(level)
  - 自然量化尺度(Natural Quantitative Scale)
    - 指欲衡量的對象具有自然形成的公認尺度
    - 例：衡量時間可以使用分鐘、小時，衡量距離可以使用公里、海浬等
  - 質化尺度(Qualitative Scale)
    - 適用的衡量對象並沒有自然公定的標準，必須以人為方式來制定尺度衡量(scale construction)
    - 例：用李克尺度(Likert Scale)訂定問卷衡量尺度
- 當某個因子不易找到對應屬性時，可以代理屬性(proxy attributes)作為衡量

# Data Scale

衡量的層次	內容說明
名目尺度 (nominal Scale)	衡量的數字僅是作為代碼，數字大小不具任何意義，也不能做數學運算
類別尺度 (categorical Scale)	衡量的數字僅是用來表示歸屬的類別，因此類別尺度的資料可以重複
順序尺度 (ordinal Scale)	衡量的數字表示方案之間的大小順序關係
間距尺度 (interval Scale)	衡量的數字可有意義地描述並比較數字之間的差距大小。無固定原點，也可以調整分隔的間距大小
比率尺度 (ratio Scale)	衡量的數字可做比率倍數的比較。有固定原點
絕對尺度 (absolute scale)	所衡量的數字具有絕對的意義，無法再做其他有意義的轉換

# Data Scale

- **Scale (純量/標量)**

- Wikipedia: 只有大小，沒有方向，可用實數表示的一個量，實際上純量就是實數，純量這個稱法只是為了區別與向量的差別。純量可以是負數，例如溫度低於冰點。與之相對，向量（又稱矢量）既有大小，又有方向。在物理學中，純量是在座標變換下保持不變的物理量。

- **Vector (向量)**

- 有序排列的資料內容(數值、字元/串、日期等)

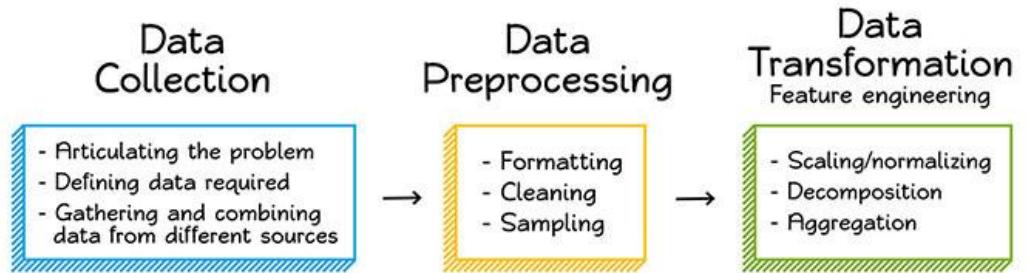
- **Matrix (矩陣)**

- 二維的數組

- **Tensor (張量)**

- 二維以上的數組

## Data Preparation Process



# Data Preparation

Data Collection - ETL

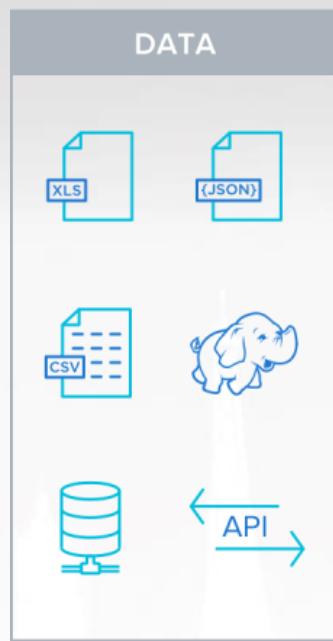
Data Inspection

Data Preprocessing

Data Transformation

Exploratory Data Analysis

# Data Preparation



**Data Preparation** is the process of cleaning, structuring and enriching raw data into a desired output for analysis.

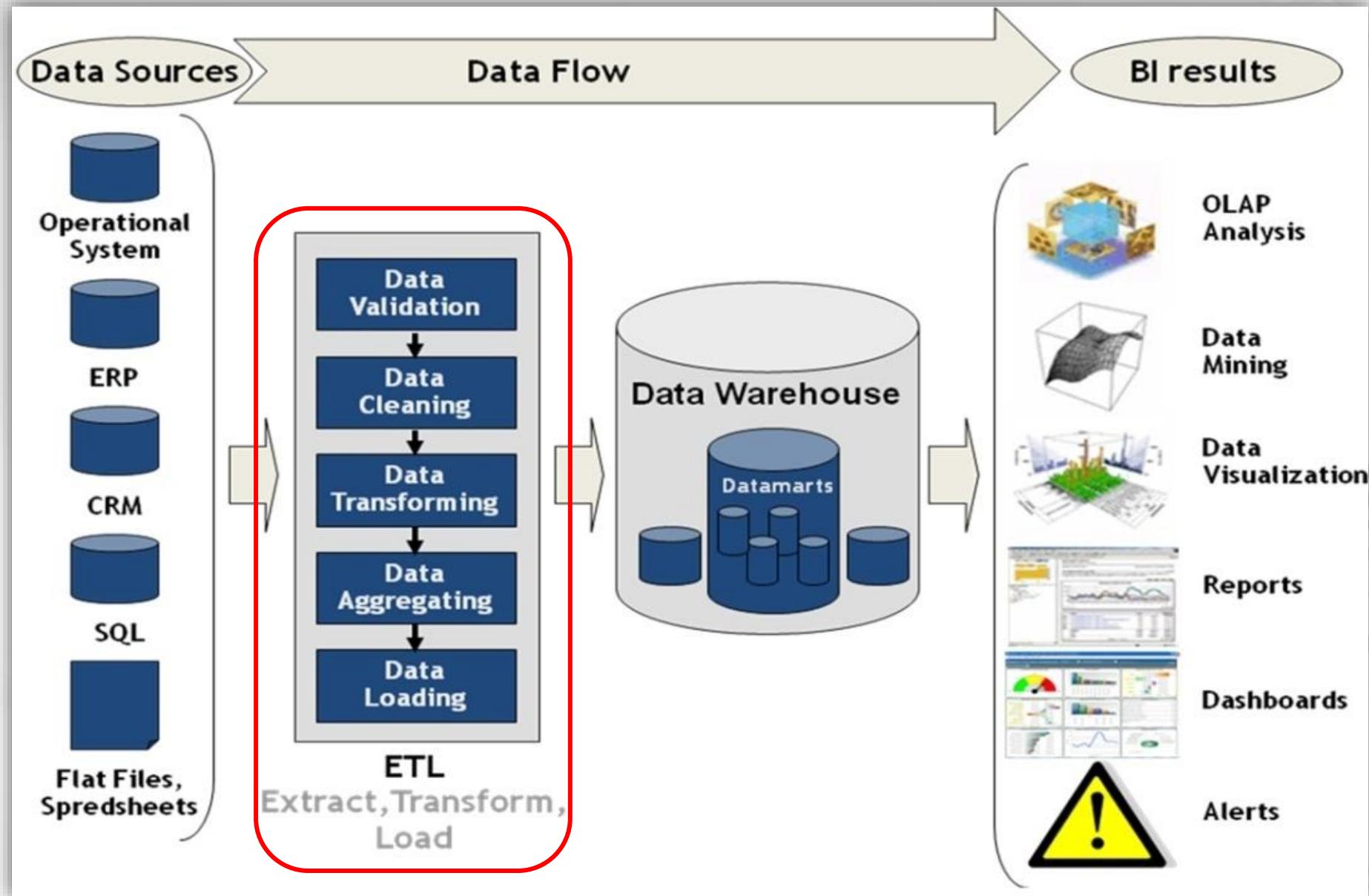


Source: <https://www.trifacta.com/data-preparation/#:~:text=Data%20Preparation%20Challenges>

# Data ETL

- **ETL (Extraction, Transformation and Loading)**
- Extracts data from data sources and loads it into data destinations using a set of transformation functions
  - **Data extraction** proceeds the task to extract data from a variety of data sources, such as flat files, relational databases, streaming data, XML files, and ODBC/JDBC data sources
  - **Data transformation** proceeds the task to clean, convert, aggregate, merge, and split data
  - **Data loading** proceeds the task to load data into destination databases via update, insert or delete statements, or in bulk

# Process of Data ETL



# 資料的萃取-轉換-載入 (Extract-Transform-Load, ETL)

問題	原因	資料ETL步驟
不正確的資料	資料的值超出合理範圍	資料萃取
不一致的資料	不同來源資料整合後所出現的分歧	
重複的資料	重複記錄的欄位或數值	
冗餘的資料	出現相同意義的資料或欄位	
遺漏值	量測設備或人為因素所造成的資料遺漏	資料清理
雜訊	資料本身的誤差或資料輸入的偏差	
離群值	資料本身的特性、不當量測或資料輸入錯誤	
資料尺度不合適	資料格式不符合挖礦工具的假設	資料轉換
資料太多	資料或維度過高	資料化約

# Data Inspection

- 目的：以不同維度來檢視所獲得的資料，以便能事先觀測出其中的錯誤，並與領域專家討論以決定是否修正或刪除其中資料
- 聚焦在資料的數量與品質兩方面
  - 資料數量
    - 檢視量化資料的三個維度：**樣本個數**、**變數或特徵個數**、**不同的資料值**
    - 樣本個數太少會影響結果的解釋程度；當個數太多時，則統計上的顯著不見得有實質意義
    - 變數個數太多則會造成資料維度過高，使得分析時間過長
  - 資料品質
    - 檢視資料的**集中趨勢**(平均數、中位數、眾數等)以及**變異程度**
    - 以不同圖來檢視**資料遺漏**、**資料雜訊**等

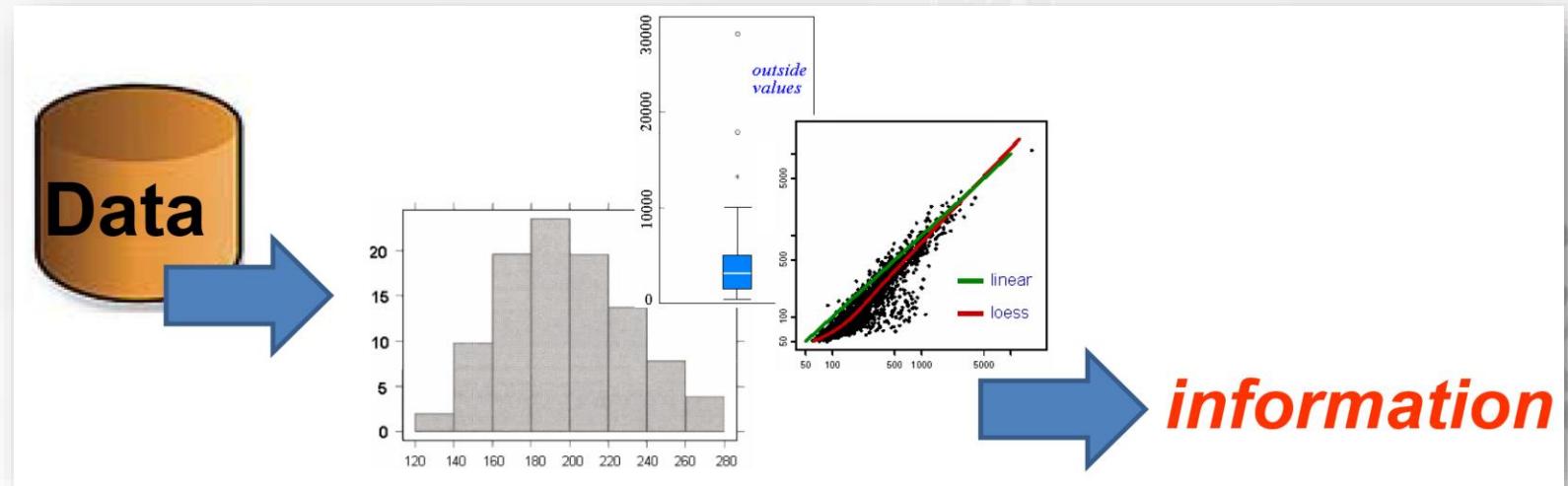
# 資料探索與資料視覺化

- 統計表：原始資料經整理後，按特定規則製成的表格
- 統計圖：以不同圖形樣式來表示統計資料各項特徵的圖形
- 類型：
  - 呈現資料次數分布：直方圖、條形圖與圓餅圖
  - 表現資料分布離勢情形：莖葉圖、盒鬚圖與常態機率圖
  - 顯示時間序列變化：趨勢圖
  - 顯現兩變數相對變化關係：散佈圖
- 資料型態的不同，所適用的統計圖亦有所不同
  - 離散資料：適用長條圖
  - 連續資料：適用直方圖與圓餅圖



# Data Visualization

- Purpose of Data Visualization
  - Statistically transfer quantitative information into graphics to provide visual representations of characteristic of data



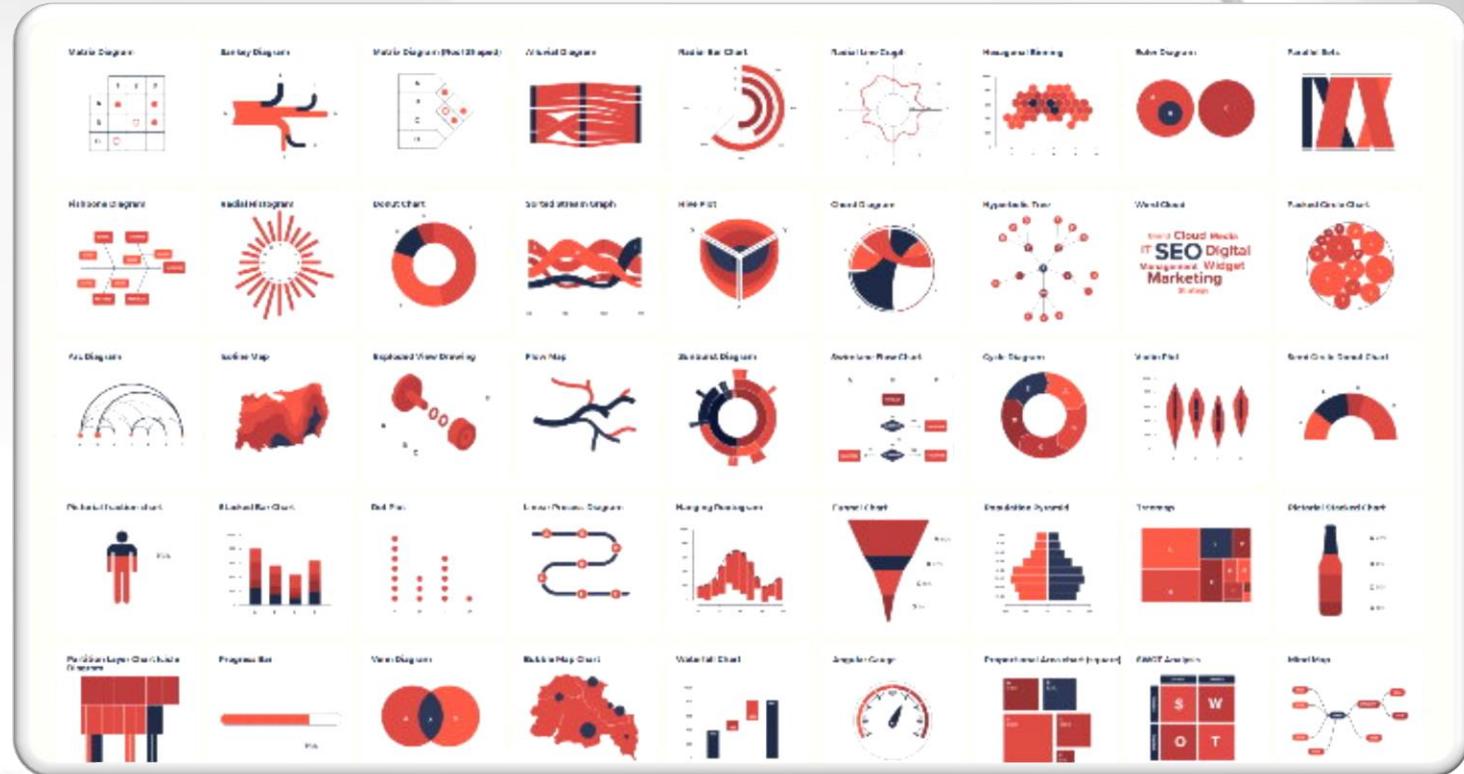
- Act as an Entry of *Exploratory Data Analysis*
  - A strategy and technique to initially uncover vital patterns under a pile the dataset

# 資料視覺化案例 - Dr. Hans Rosling



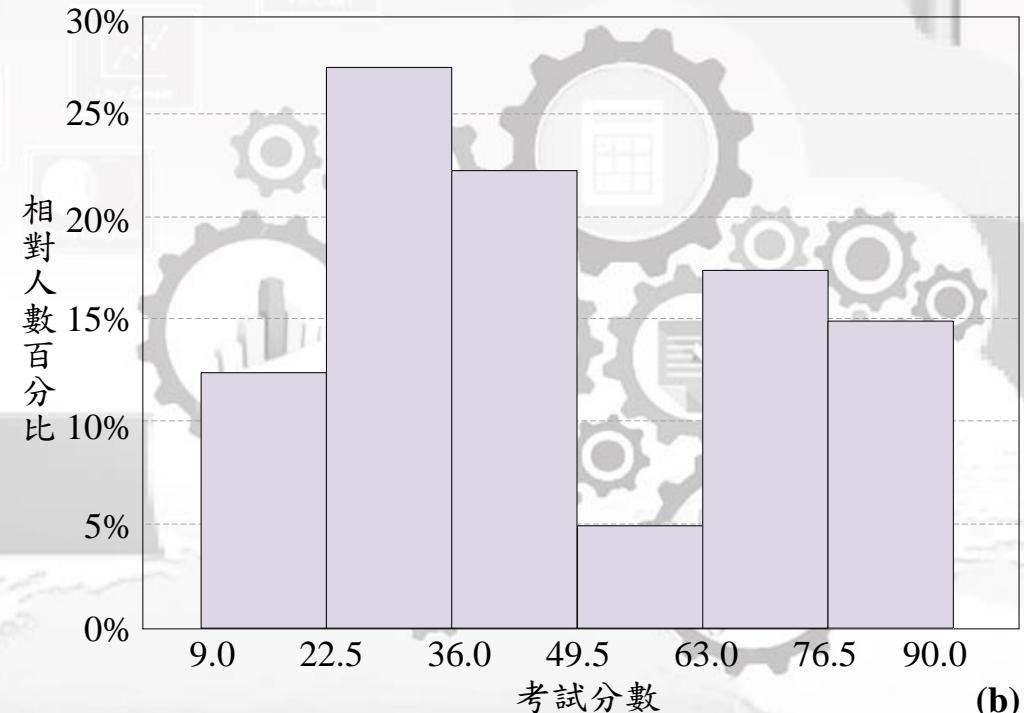
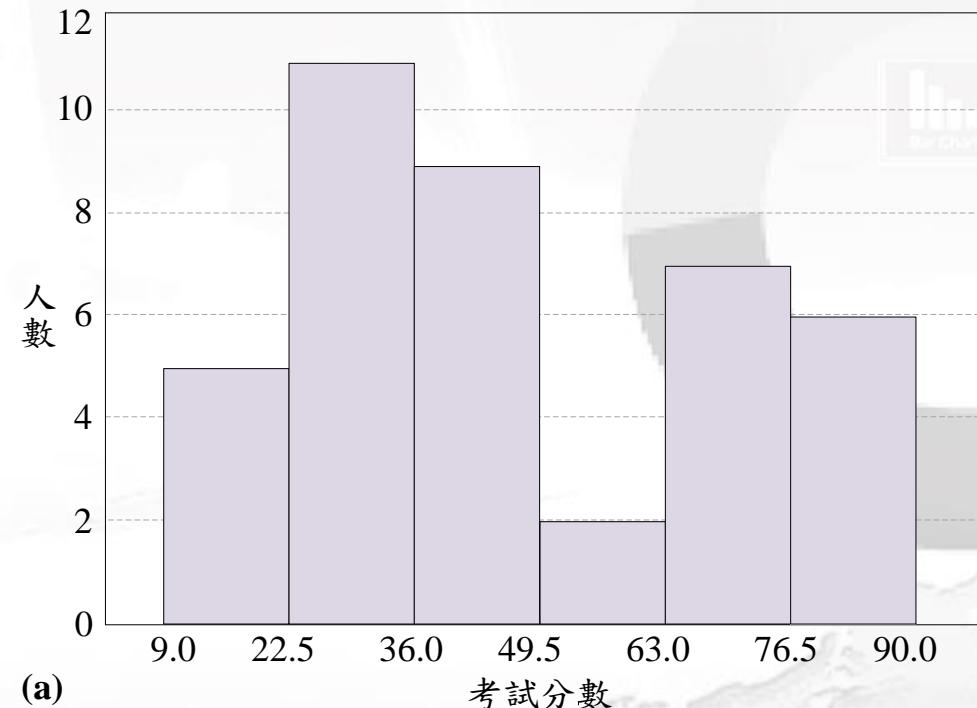
# 資料視覺化 - 常用統計圖形

- 直方圖
- 長條圖
- 柏拉圖
- 圓餅圖
- 盒鬚圖
- QQ圖
- 折線圖
- 散佈圖
- 平行座標圖



# 直方圖 (Histogram)

- 將一組數據分成數組後，依照各組距的範圍與次數，繪製成連續型資料之次數分布圖，通常可呈現等距及比率變數的資料
- 通常將各組組中點標記於橫軸、次數記於縱軸，以各組之組距為底，次數為高，依次將各組繪成長方形並緊靠在一起



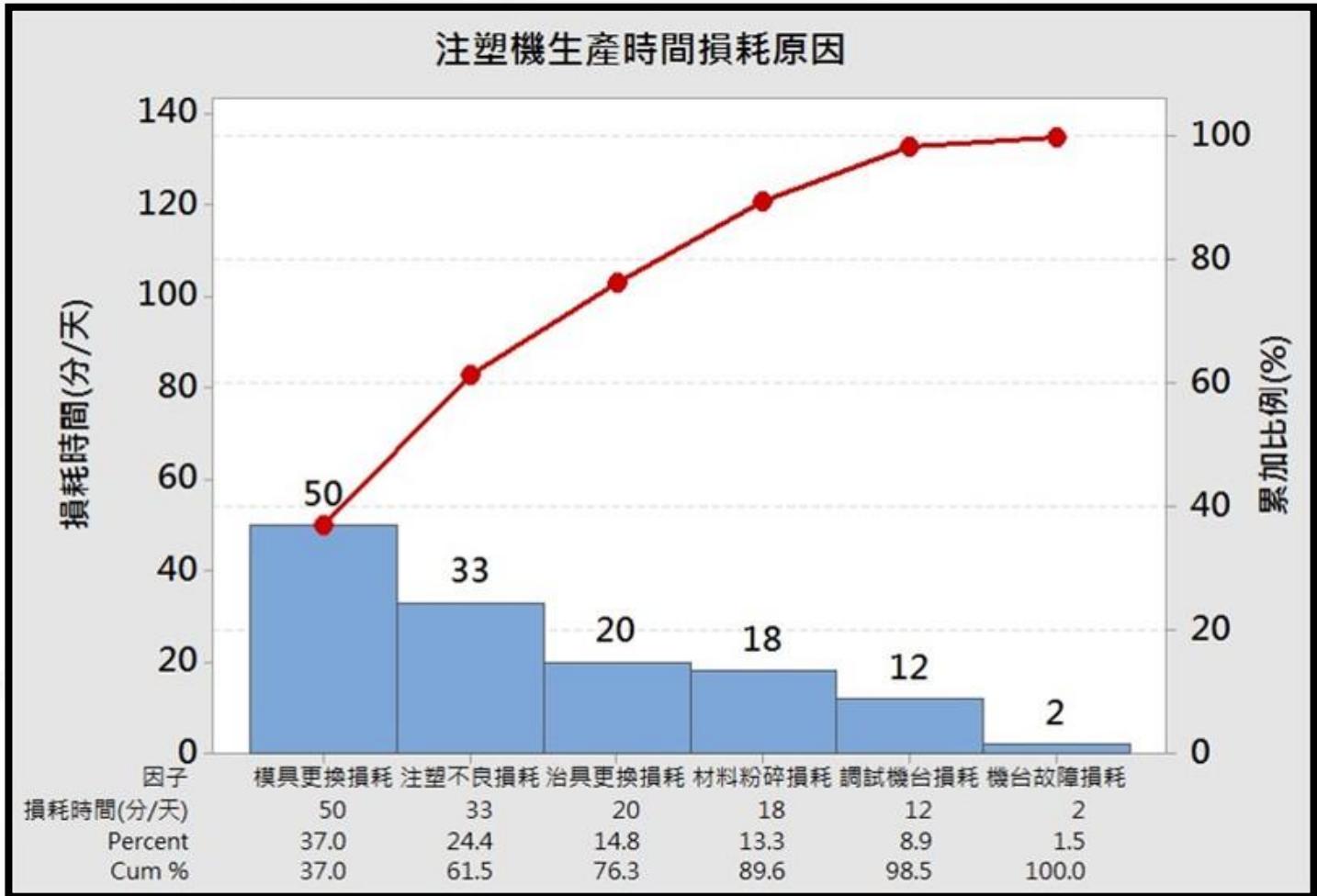
# 長條圖 (Bar Chart)

- 與直方圖類似，但主要用以顯示類別資料的分布及排序情形，因此條柱不相連。藉由將各組的標誌放在圖形的橫軸上，縱軸則為次數尺度或累積次數尺度等
- 各條柱寬度相等，以條柱高度表示各類別次數的多寡



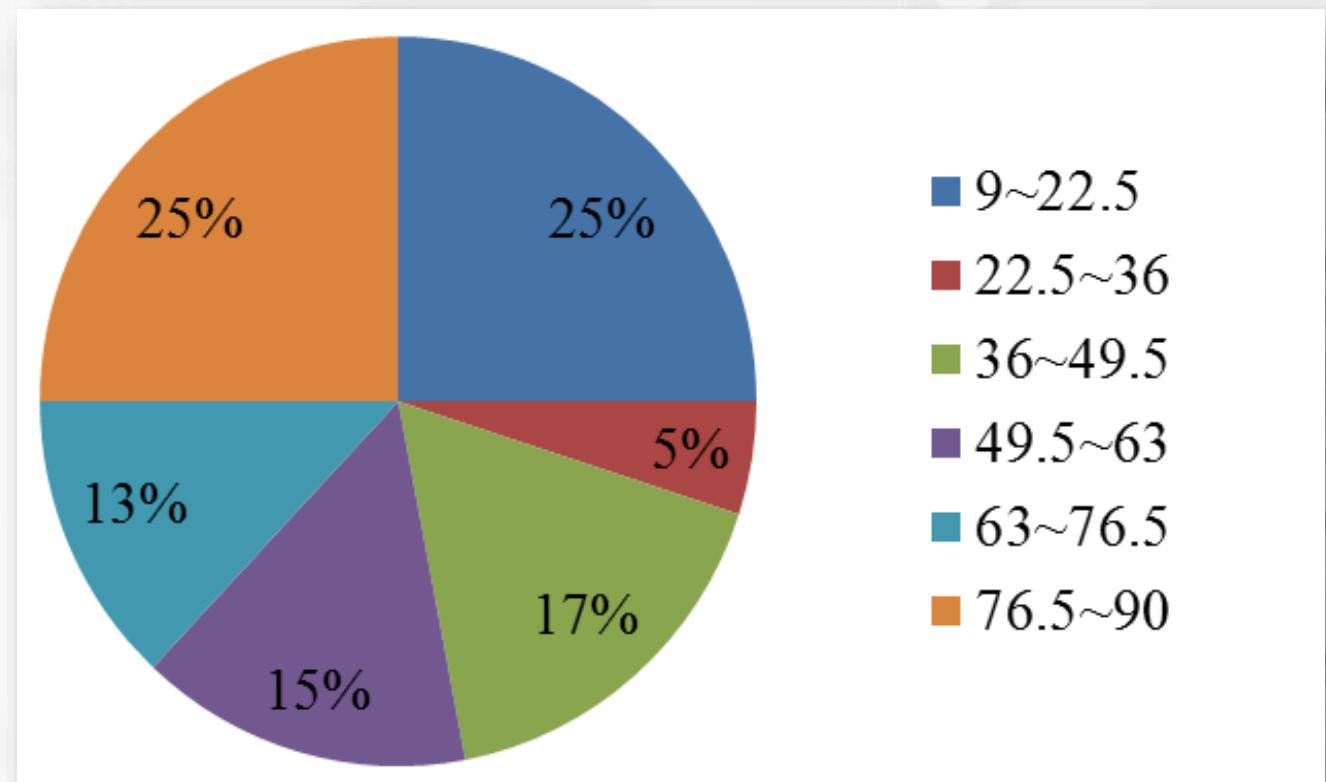
# 柏拉圖 (Pareto Diagram)

- 以發生的頻率累計排序的呈現，大多應用於80/20



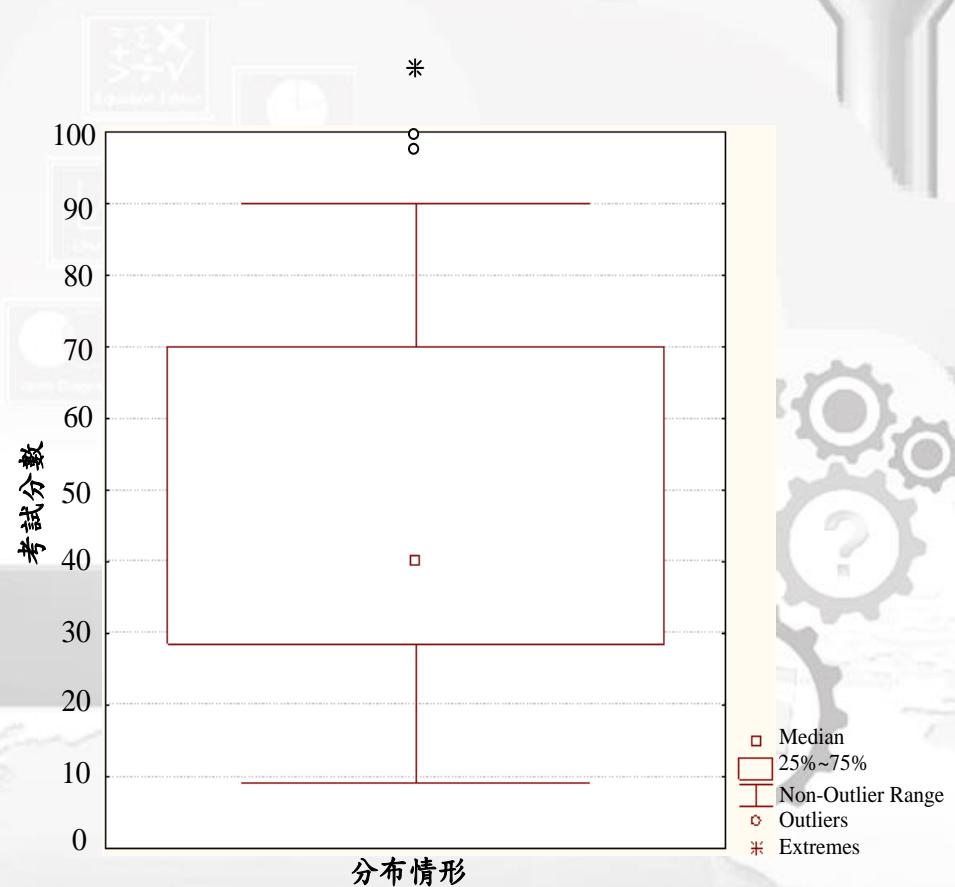
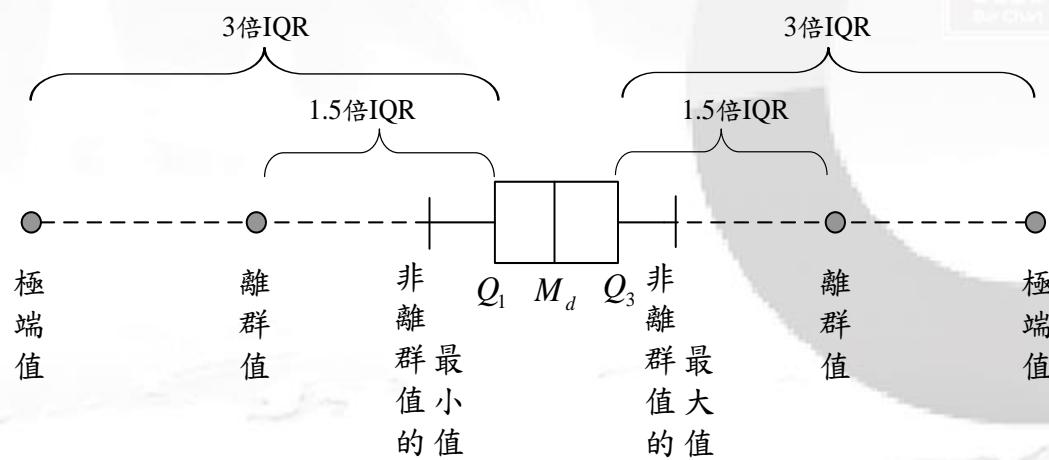
# 圓餅圖 (Pie Chart)

- 適用於呈現類別型的資料分布，依各組的相對次數或比率將該圓餅劃分為扇形，常用於表示整體與部分資料之間的比率關係。其中圓面積代表整體，扇形面積則代表部分
- 以學生考試成績範圍為例，將分組資料依比例繪製成如下



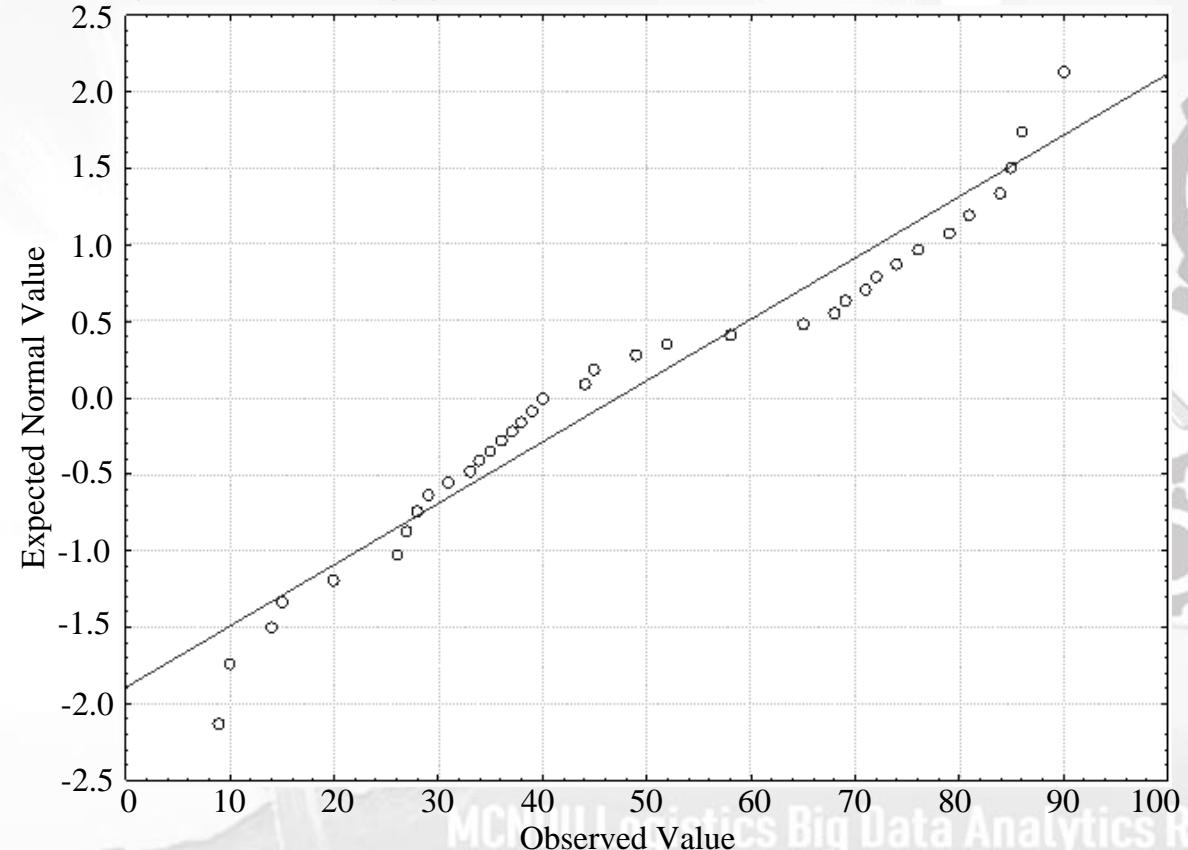
# 盒鬚圖 (Box-and-whisker Plot)

- 亦稱箱型圖 (Box-plot)，利用圖形呈現資料的中央趨勢與離散程度，不需繪製出實際的觀察值即可顯示所分配的總計統計量
- 可用以檢驗資料的極端量數及分配型態



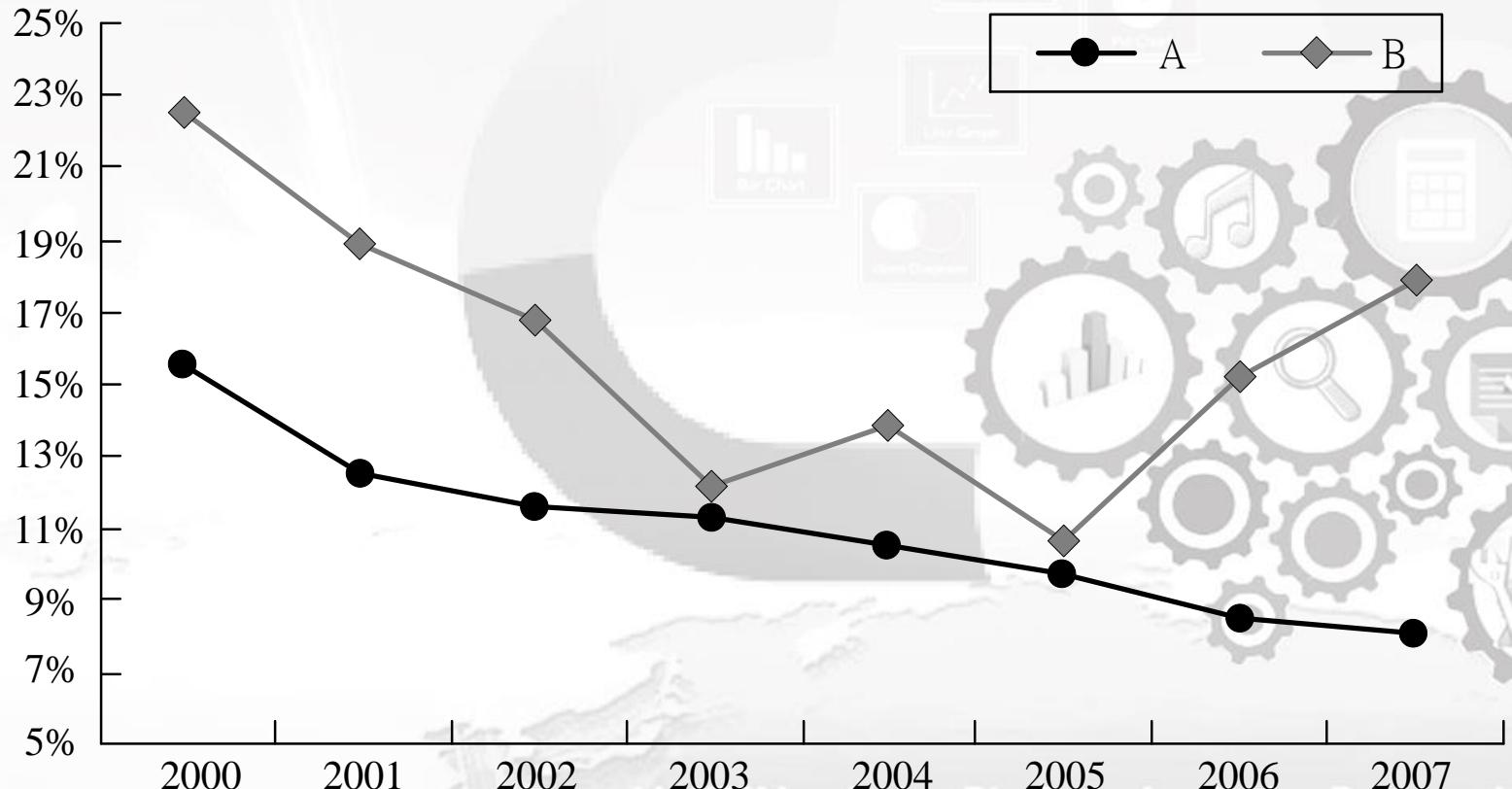
# 分位數圖 (Quantile-Quantile Plot)

- 又稱**Q-Q圖**，通過分位數來**比較兩個機率分布**的圖形方法。主要是**用以檢視資料由一個分布對應至另一個分布是否有位移**
- 如果被比較的兩個分布比較相似，則其**QQ圖**近似地位於 $y = x$ 上
- 常態機率圖（normal probability plot）是**Q-Q圖**的一個特例



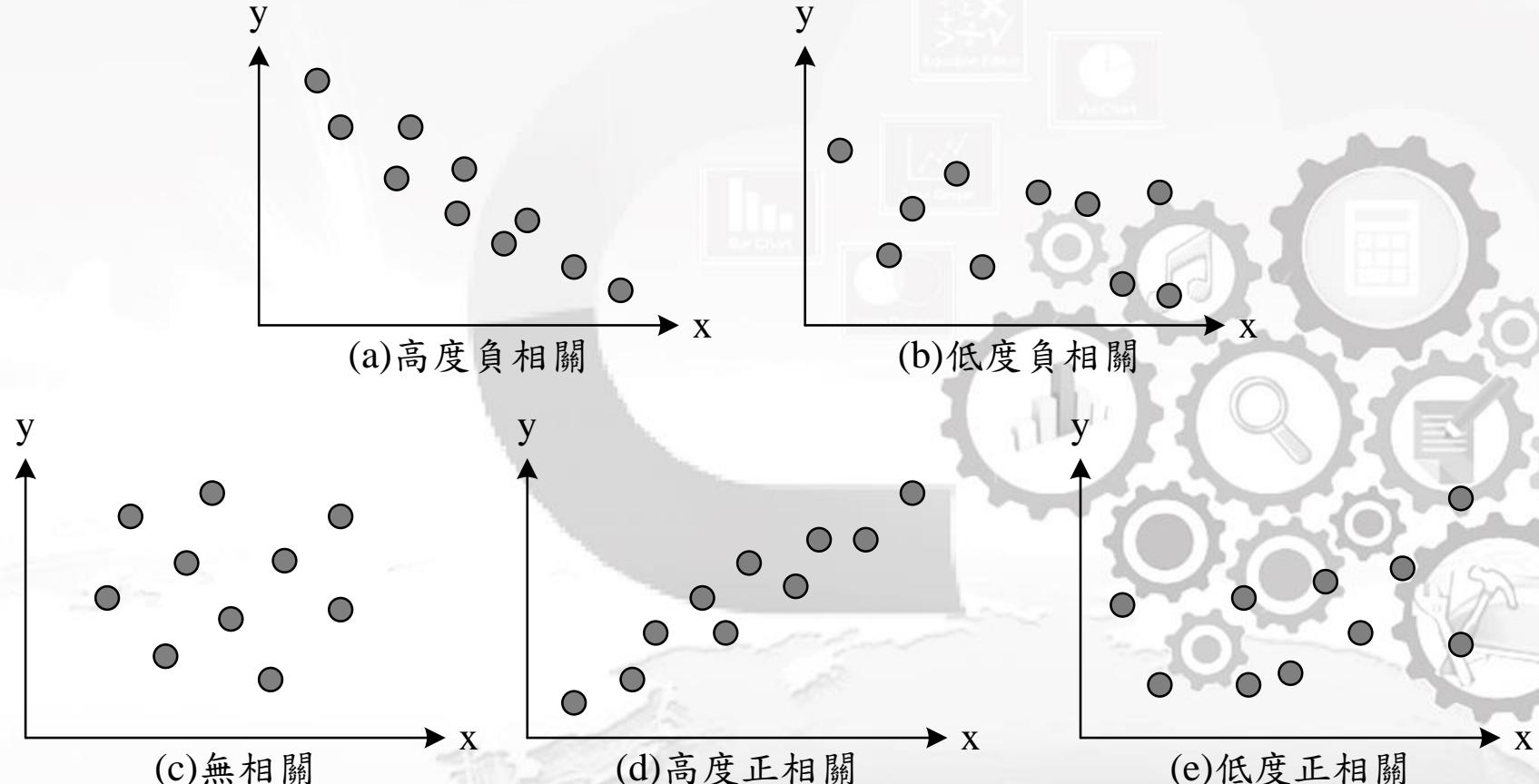
# 折線圖 (Line Chart)

- 由一條線連接數點以顯示序列，以圖表呈現資料分布的變化趨勢
- 由折線的上升或下降可清楚看出序列的變動，通常用來比較一段時間的資料變化或兩序列以上的變動情況
- 縱軸代表量測值，橫軸代表類別目錄標籤



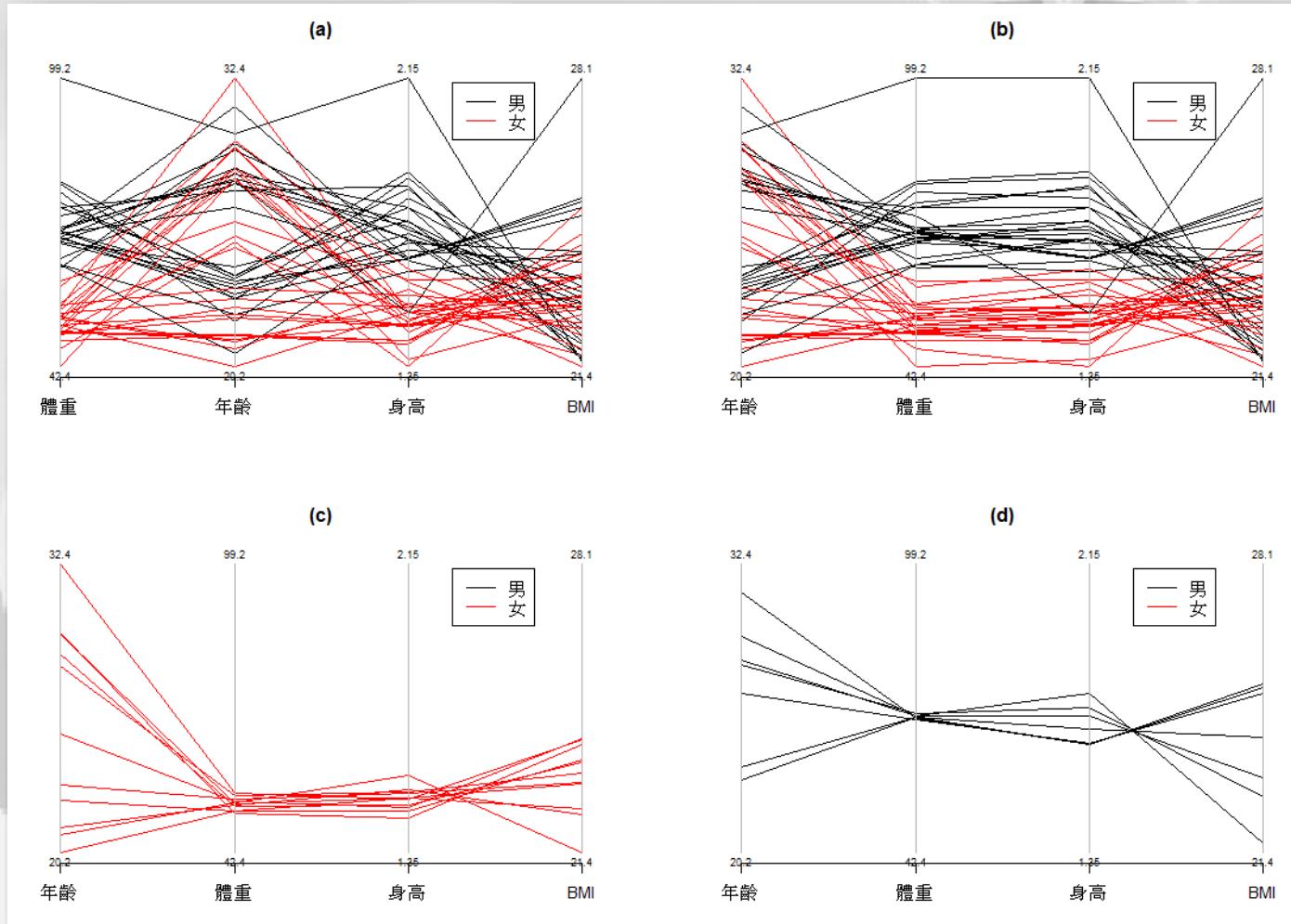
# 散佈圖 (Scatter Plot)

- 在多維空間中給出  $p$  個變數關係的點
- 由點的疏密程度和延展方向等分布特徵，初步瞭解變數的關係
- 迴歸分析常以散佈圖作為篩選獨立變數x的基本檢驗步驟

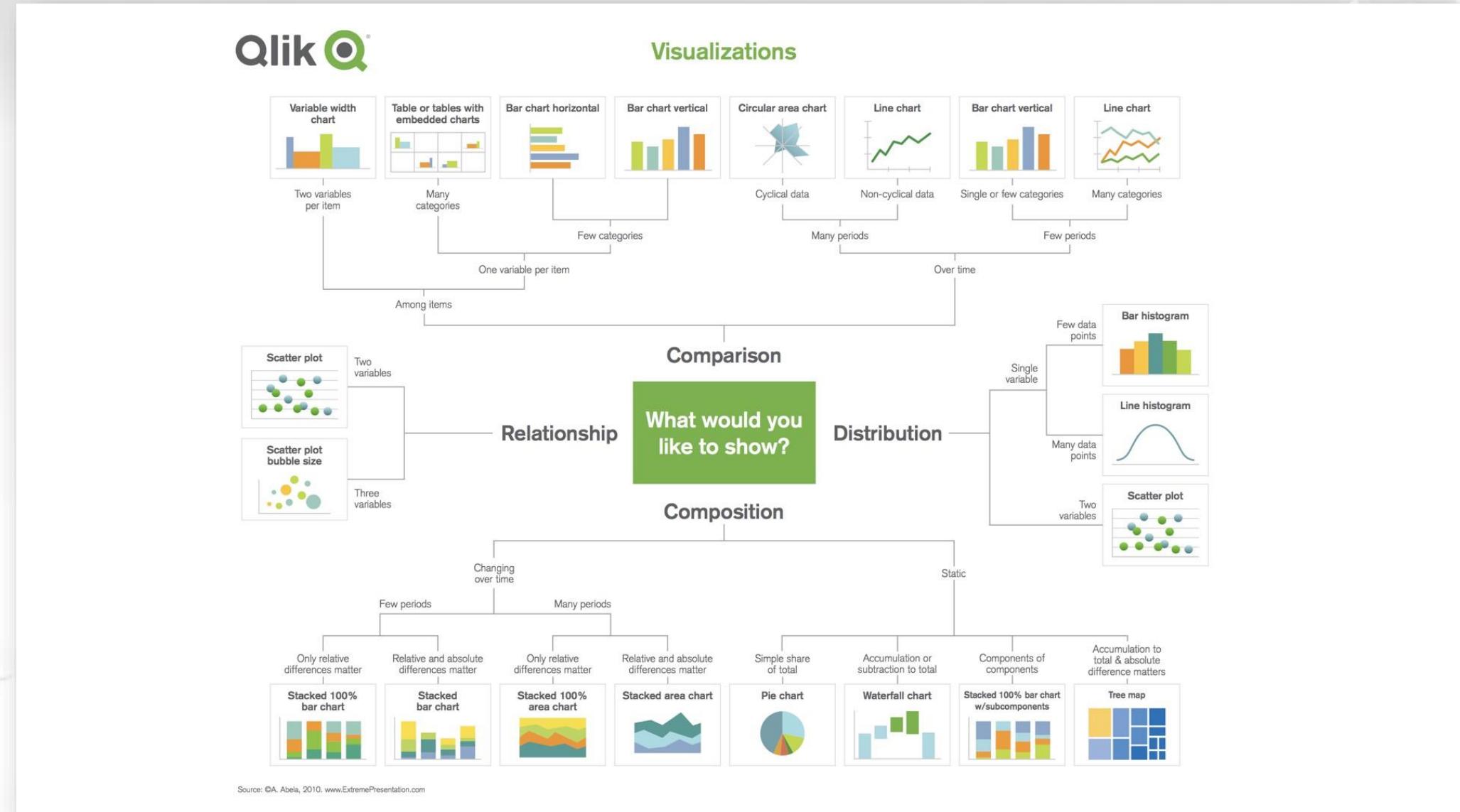


# 平行座標圖 (Parallel Coordinates Plot)

- 檢視高維度資料概況的圖形呈現方法
- 指在一份資料中，以  $p$  條垂直以及相互平行的座標軸（座標軸之間通常等距）來表示彼此之間不同的維度，每一筆資料以一條折線來呈現，折線與平行軸的相交位置為該資料於該維度變數所對應之數值



# 資料視覺化應用的分類



# R 語言的資料視覺化示範操作

The R Graphics Package 

Documentation for package 'graphics' version 3.3.0

- [DESCRIPTION file.](#)
- [Code demos](#). Use `demo()` to run them.

[graphics-package](#) [filled.contour](#) [.Pars](#) [abline](#) [arrows](#) [asp](#) [assocplot](#) [Axis](#) [axis](#) [axis.POSIXct](#) [axTicks](#) [barplot](#) [box](#) [boxplot](#) [boxplot.matrix](#)

```
> library(graphics)
> demo(graphics) #常見圖形
> demo(Hershey) #各種符號
> demo(image) #image和contours
> demo(Japanese) #日本字
> demo(persp) #曲面圖
> demo(plotmath) #數學符號
```

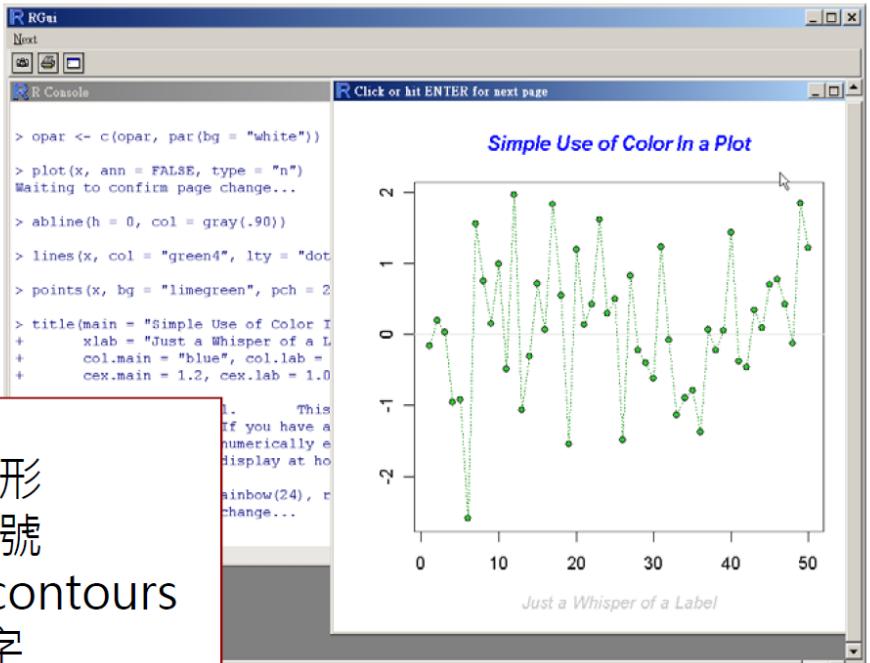
R GUI

R Console

R Click or hit ENTER for next page

Simple Use of Color in a Plot

Just a Whisper of a Label



This is a screenshot of the R graphical user interface (GUI). On the left, the R Console window shows R code for generating a plot. The code includes commands like `library(graphics)`, `demo(graphics)`, and `title(main = "Simple Use of Color In A Plot")`. The main window displays a scatter plot with green points and lines, titled "Simple Use of Color in a Plot". The x-axis ranges from 0 to 50, and the y-axis ranges from -2 to 2. The plot features a grid and some text at the bottom labeled "Just a Whisper of a Label".

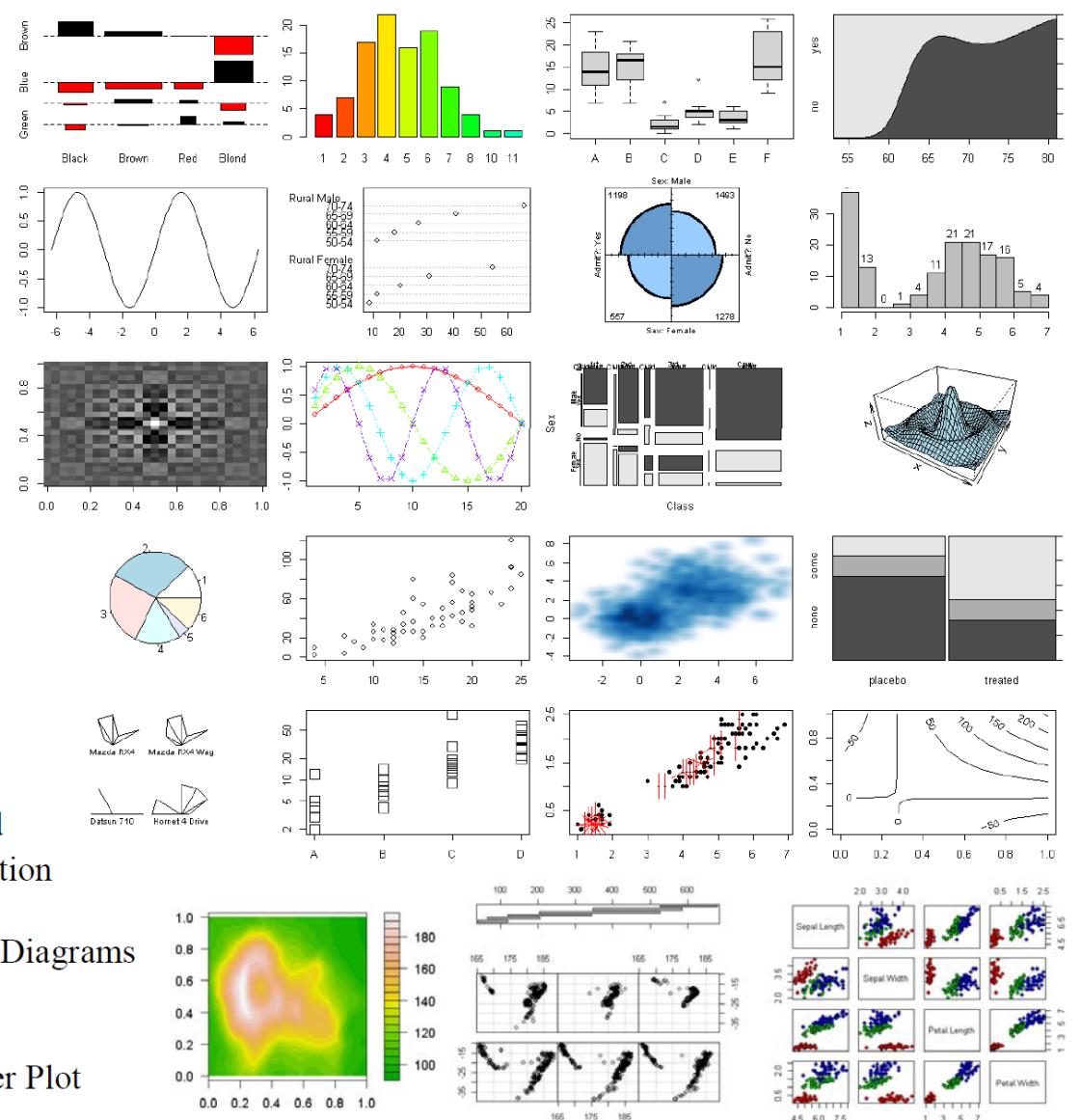
作圖準則: First calling a **high-level function** that creates a complete plot, then calling **low-level functions** to add more output if necessary.

42

# R語言相關繪圖 Packages

## Plots:

- assocplot**: Association Plots
- barplot**: Bar Plots
- boxplot**: Box Plots
- cdplot**: Conditional Density Plots
- contour**: Display Contours
- coplot**: Conditioning Plots
- curve**: Draw Function Plots
- dotchart**: Cleveland's Dot Plots
- filled.contour**: Level (Contour) Plots
- fourfoldplot**: Fourfold Plots
- hist**: Histograms
- image**: Display a Color Image
- matplot**: Plot Columns of Matrices
- mosaicplot**: Mosaic Plots
- pairs**: Scatterplot Matrices
- persp**: Perspective Plots
- pie**: Pie Charts
- plot**: Generic X-Y Plotting
- smoothScatter**: Scatterplots with Smoothed Densities Color Representation
- spineplot**: Spine Plots and Spinograms
- stars**: Star (Spider/Radar) Plots and Segment Diagrams
- stem**: Stem-and-Leaf Plots
- stripchart**: 1-D Scatter Plots
- sunflowerplot**: Produce a Sunflower Scatter Plot



# 資料清理 (Data Cleaning)

- 不正確的資料：確認資料的有效範圍及驗證資料的合理性
- 不一致的資料：數值不一致、資料內容不一致、欄位不一致
- 重複的資料：若兩項重複資料完全相同，可刪除其中一項，否則應注意哪一項紀錄為最新資料
- 冗餘的資料：針對具有相同意義或彼此間存有已知數學關係的欄位，此變數的屬性或意義可由另一變數推導而得
  - 有些冗餘資料可以經由相關分析偵測到
- 遺漏值 (Missing Value)
  - 為遺漏或錯誤的資料，可刪除該筆資料或補值
- 空白值 (Empty Value)：為無法或不需填入的資料
- 雜訊(Noise)：表示一個測量變數中的隨機錯誤或偏差
- 離群值(Outlier)：表現明顯與其他資料不一樣的資料

# 遺漏值的處理

1. 忽略變數值
2. 人工填寫遺失值
3. 使用一個全域常數填充遺漏值
4. 使用屬性平均值
5. 使用與給定變數值屬於同一類別的所有樣本之平均值
6. 使用最可能的值去填充遺漏值
  - 簡單線性迴歸
  - 多元線性迴歸
  - 類神經網路
  - Nearest-Neighbor Estimators

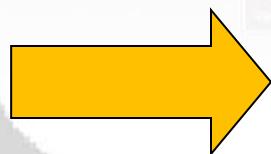
取代遺漏值可能會產生的失真或誤差，使用者必須瞭解每種取代方法的特性，避免忽略掉原本應有的資訊

# 不偏估計量 V.S. 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	0.0886	0.0886
2	0.0684	0.0684	0.0684	0.0684
3	0.3515	0.3515	0.3515	0.3515
4	0.9874	0.9874	0.9874	0.9874
5	0.4713	0.4713	0.4713	0.4713
6	0.6115	0.6115	0.6115	0.6115
7	0.2573	0.2573	0.2573	0.2573
8	0.2914	0.2914	0.2914	0.2914
9	0.1662	0.1662	0.1662	0.1662
10	0.44	0.44	0.44	0.44
11	0.6939	?	0.3731	0.6622
平均值	0.4023	0.3731	0.3731	0.3994
標準差	0.2785	0.2753	0.2612	0.2753
誤差值			0.3208	0.0317

# 利用其他變數與遺漏值之間的關係來估計遺漏值

- 補值：可利用其他變數與遺漏值之間的關係來估計遺漏值
- 例如，若「收入水準」變數發生遺漏值，或許可能用「房子坪數」這變數來做預測



# 最鄰近估計法

- 假設在現有的資料庫中發現某一顧客其購買反應的態度為一遺漏值

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	?

顧客	$d_{\text{年齡\_norm}}$	$d_{\text{性別\_norm}}$	$d_{\text{薪水\_norm}}$	加總	購買反應
A	1	0	1	2	No
B	0.33	1	0.44	1.77	Yes
C	0.38	1	0.06	1.44	Yes
D	0.66	0	0.55	1.21	Yes
E	0	1	0.67	1.67	No
F	0	0	0	0	YES

# 雜訊資料的處理方式

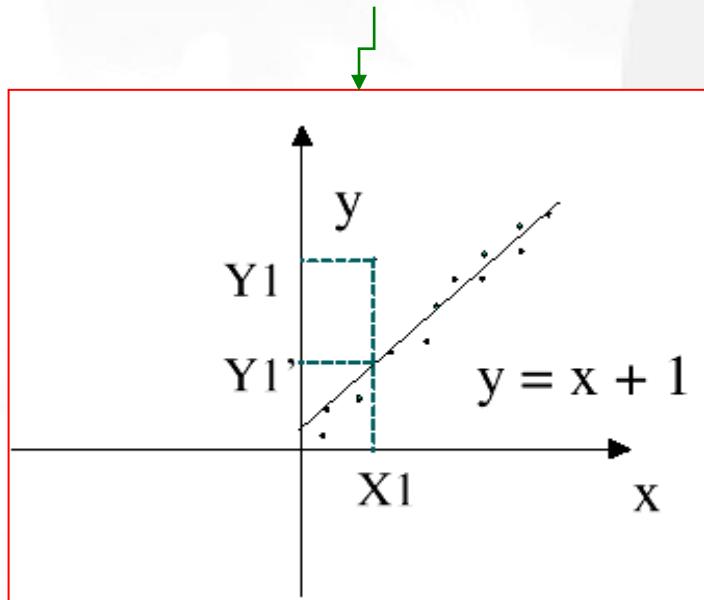
- 雜訊(**noise**)：表示一個測量變數中的隨機錯誤或偏差，可能因人為因素或機器設備產生誤差
- 處理方式：可利用資料平滑（**smooth**）技術將資料平緩化，以降低其對結果的影響
  - 分箱法(**Binning**)：利用「相鄰」值來局部平滑儲存在同一箱子的資料值。
  - 資料配適：利用資料配適為新的函數來平滑資料，例如採用簡單線性迴歸以一個解釋變數估計目標變數

# 雜訊的處理 – 資料平滑技術

Binning

Clustering

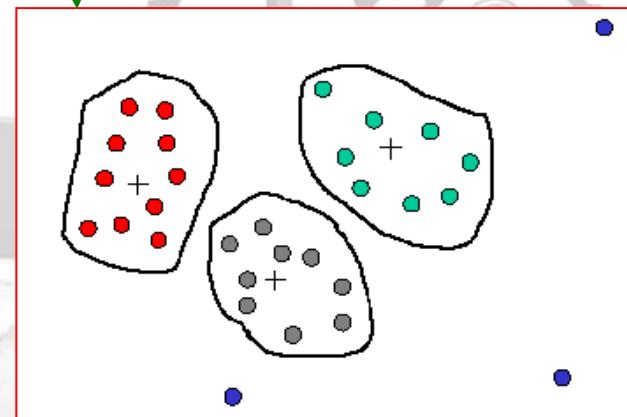
Regression



Bin1: 4, 8, 15  
Bin2: 21, 21, 24  
Bin3: 25, 28, 34

means  
Bin1: 9, 9, 9  
Bin2: 22, 22, 22  
Bin3: 29, 29, 29

boundaries  
Bin1: 4, 4, 15  
Bin2: 21, 21, 24  
Bin3: 25, 25, 34



# 離群值的處理

- **離群值**：表現明顯與其他資料不一樣的資料，會影響資料分析模式的效果，特別是預測方法或迴歸模型
- 處理方法：
  - **直接刪除**：出於儀器或工具造成的判斷錯誤，或是完全不合理的資料
  - **用其他數值替換，將資料範圍正規化**：當數值變數為空白值或是非數值資料，且具有一定的代表性時，可用其他數值來更替，將資料的範圍正規化
  - **群集分析**：藉由將類似的點結合為一個群組或族群，落在群集集合之外的值即視為離群值

# 資料轉換 (Data Transformation)

- 將資料轉換成適合資料探勘模式可處理的資料格式或為豐富化資料的內容，以轉換原始資料或重新編碼以提升資料價值
  - 資料數值的轉換(又稱特徵尺度轉換，Feature Scaling)：提升模型的收斂速度、提高模型的精準度
    - 正規化 (特徵正規化(Feature Normalization))
      - 最小值最大值正規化 (Min-Max Normalization)：將原始資料按比例縮放於  $[0, 1]$  區間，且不改變其原本分佈
      - 標準化 (Standardization)
        - Z分數標準化(Z-Score Standardization)：Z分數標準化適用於分佈大致對稱的資料，因為在非常不對稱的分佈中，標準差的意義並不明確，此時若標準化資料，可能會對結果做出錯誤的解讀
    - 資料型別的轉換
      - 離散型資料 → 連續型資料 - 資料編碼 (Encoding)
      - 連續型資料 → 離散型資料 - 資料分箱 (Binning)

# 資料正規化 (Data Normalization)

- 最常見的Normalization為0–1區間縮放，經過Normalization之後資料的範圍會介在0~1之間，原本的最大值變為1，最小值變為0
- 將屬性資料按比例縮放到一個特定的區間
  - 如類神經網路中的倒傳遞(back propagation)演算法需要對於訓練樣本輸入值範圍轉換至[0, 1]
- 可防止較大初始值域與較小初始值域屬性間互相比較的情況，以及權重過大的問題
- 方法：**極小值–極大值正規化** (min-max normalization)
  - 對原始資料進行線性轉換，假設  $\min_A$  和  $\max_A$  別為屬性A的最小值和最大值

$$v' = \frac{v - \min_A}{\max_A - \min_A} (new_{\max} - new_{\min}) + new_{\min}$$

# 標準化 (Standardization)

- 基於屬性  $A$  的平均值和屬性  $A$  的標準差  $S_A$  將資料標準化。屬性  $A$  的值  $X$  標準化後為  $Z$ ，經由下式計算而得：

$$Z = \frac{X - \bar{X}_A}{S_A}$$

- 其中  $\bar{X}_A$  與  $S_A$  分別為屬性  $A$  的平均值和標準差，當屬性  $A$  的最大值和最小值為未知，或孤立點左右極小值—極大值正規化時，可改用標準化方法
- 經過 Standardization 資料的平均值會變為 0, 標準差變為 1

# 離散型資料轉成連續型資料

- 轉換過程必須利用合適的領域知識來定義離散值的距離或相似程度
- 通常需要結合專家意見，然後以類似的矩陣定義出數值與數值之間的距離或相似程度，再利用此距離或是相似程度把離散的資料轉換為連續型的資料型態
  - 例如：學生成績的等第為A應該對應至85分，若成績為B+，則應該對應至78分

# 資料轉換-改變變數維度範例(1/2)

## Sub. 出生地

2133	台北市
2135	新竹市
2163	桃園縣
2653	彰化縣
2654	澎湖縣
3124	金門縣
3554	苗栗縣
3646	台東縣
3772	嘉義市
5449	南投縣

## Sub 北市 新北 桃縣 竹縣 竹市 苗縣

2133	1	0	0	0	0	0
2135	0	0	0	0	1	0
2163	0	0	1	0	0	0
2653	0	0	0	0	0	0
2654	0	0	0	0	0	0
3124	0	0	0	0	0	0
3554	0	0	0	0	0	1

# 資料轉換-改變變數維度範例(2/2)

## Sub. 出生地

2133	台北市
2135	新竹市
2163	桃園縣
2653	彰化縣
2654	澎湖縣
3124	金門縣
3554	苗栗縣
3646	台東縣
3772	嘉義市
5449	南投縣

## Sub 區域 都市型態

2133	北	大都會
2135	北	都市
2163	北	都市
2653	中	鄉村
2654	島	鄉村
3124	島	鄉村
3554	北	鄉村

# 類別尺度之編碼與量化

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

- Categorical scale: labeling.
- Introduce wrong patterns in calculation.
- Proxy attribute: quantify a qualitative attribute with one highly correlated with the effect
- Likert scale

# 連續型資料轉成離散型資料

- **離散化(discretization)**：將連續資料分配到數個小區間，以**類別尺度**取代原有連續資料的尺度
- 離散化後的資料在敘述上較為簡單，可使透過資料探勘或機器學習方法所得到的結果更容易瞭解與解釋
- 資料離散化可同時進行**特徵選取**與**資料維度縮簡**
- 離散化方法：
  - 分箱法、利用熵(entropy)尺度進行二維分支的ID3(Quinlan,1986)與C4.5(Quinlan,1993)等決策樹方法、群集分析
  - 二位遞迴分支演算法的D2(Catlett,1991)、使用最小敘述長度準則法(Minimum Description Length Principle, MDLP)來改善D2無限遞迴分支的缺點(Fayyad & Irani,1993)等

# 離散化過程

1. 將欲轉換的連續數值排序

2. 選擇分割或合併的準則

3. 分割或合併數值

4. 是否符合停止條件。

## ■依照不同特性可從幾個維度來看：

- 監督式 (supervised) 與非監督式 (unsupervised)
- 動態 (dynamic) 與靜態 (static)
- 全域 (global) 與局部 (local)
- 分割 (splitting) 與合併 (merging)
- 直接 (direct) 或增加 (incremental)

# 離散化的特性 (1/3)

- 監督式與非監督式
  - 監督式：考慮資訊有類別的情況，類別是已知的
    - 將連續的資料分配至領域專家根據領域知識所定義的不同類別中，使得同類別的資料有相似的特性
  - 非監督：考慮資訊無類別的情況，類別是未知的
    - 區分成等距離的區間，但會受到離散值所影響，所以我們可以採用等量的方法定義區間，讓每個區間裡面的個數都是一樣的
    - 常見的非監督式離散化方法就是等寬分箱法與等深分箱法

# 離散化的特性 (2/3)

- **動態與靜態**

- 動態離散化方法指的是在資料離散化的同時生成模式
- 靜態離散化方法指的是在模式生成前完成資料離散化

- **全域與局部**

- 局部類型的方法一次僅考慮一個變數或特徵，例如等寬分箱法與等深分箱法
- 全域型的離散化方法則是一次考慮所有的變數或特徵

# 離散化的特性 (3/3)

- 分割與合併

- 分割指的是由上而下把一個連續型的區間切割成同樣的間隔
- 合併是指反覆尋找最好的相鄰區間，然後由下而上把它們合併成一個較大的區間

- 直接的或增加的

- 直接的離散化方法是將資料範圍切為固定的 $k$ 個區間，通常 $k$ 是由使用者決定
- 增加類別的方法是一種不斷改進的簡單離散方法，雖然不必分割 $k$ 個區間，但需要有另一個停止判斷的準則

# 資料型態轉換 — 分箱法 (Binning)

- 將資料排序後，依序排入預定的箱子中，接著利用各箱子的平均值、中位數、邊界值等三種數值進行資料平滑
- 假設欲分析15件商品的庫存量，其數值依序分別是：5、10、12、12、24、32、43、55、60、65、72、77、81、90、120

## • 等寬分箱法

5 10 12 12 24	32 43	55 60 65 72	77 81 90	120
箱子一	箱子二	箱子三	箱子四	箱子五

## • 等深分箱法

5 10 12	12 24 32	43 55 60	65 72 77	81 90 120
箱子一	箱子二	箱子三	箱子四	箱子五

# 資料轉換範例

產品編號	製程A		製程B		製程良率
	加工時間(分)	機台類型	加工時間(分)	機台類型	
01	28	A01	48	B03	0.53
02	27	A01	42	B03	0.62
03	31	A03	43	B01	0.84
04	42	A02	33	B02	0.91
05	46	A02	28	B03	0.85
06	50	A01	27	B03	0.68
07	35	A02	24	B01	0.83
08	24	A03	36	B02	0.69
09	28	A02	25	B01	0.88
10	44	A03	37	B03	0.92



產品編號	製程A		製程B		產品製程 良率
	加工時間(分)	機台類型	加工時間(分)	機台類型	
01	1	A01	3	B03	低
02	1	A01	3	B03	低
03	2	A03	3	B01	高
04	3	A02	2	B02	高
05	3	A02	1	B03	高
06	3	A01	1	B03	低
07	2	A02	1	B01	高
08	1	A03	2	B02	低
09	1	A02	1	B01	高
10	3	A03	2	B03	高

假設製程加工時間 $\leq 30$ 者標示為類別1

$30 < \text{加工時間} \leq 40$ 者標示為類別2

$\text{加工時間} > 40$ 者標示為類別3

上表(連續型資料)可轉換為下表(類別  
型資料)

# 資料化約 (Data Reduction)

- 資料本身的價值因**資料解析度(Resolution)**不同而有所差別，可經由資料匯總提升資料代表的意義
- 資料蒐集階段應盡可能地蒐集所有可記錄的變數或資料，再經由資料化約，得到與原始資料具有相同資訊但卻較精簡的資料集
- 其效益為：
  - 提升資料品質
  - 縮短資料運算時間
  - 提升資料價值、知識價值的取得與增加可讀性
  - 降低資料儲存成本

# 資料維度 (Data Dimension)

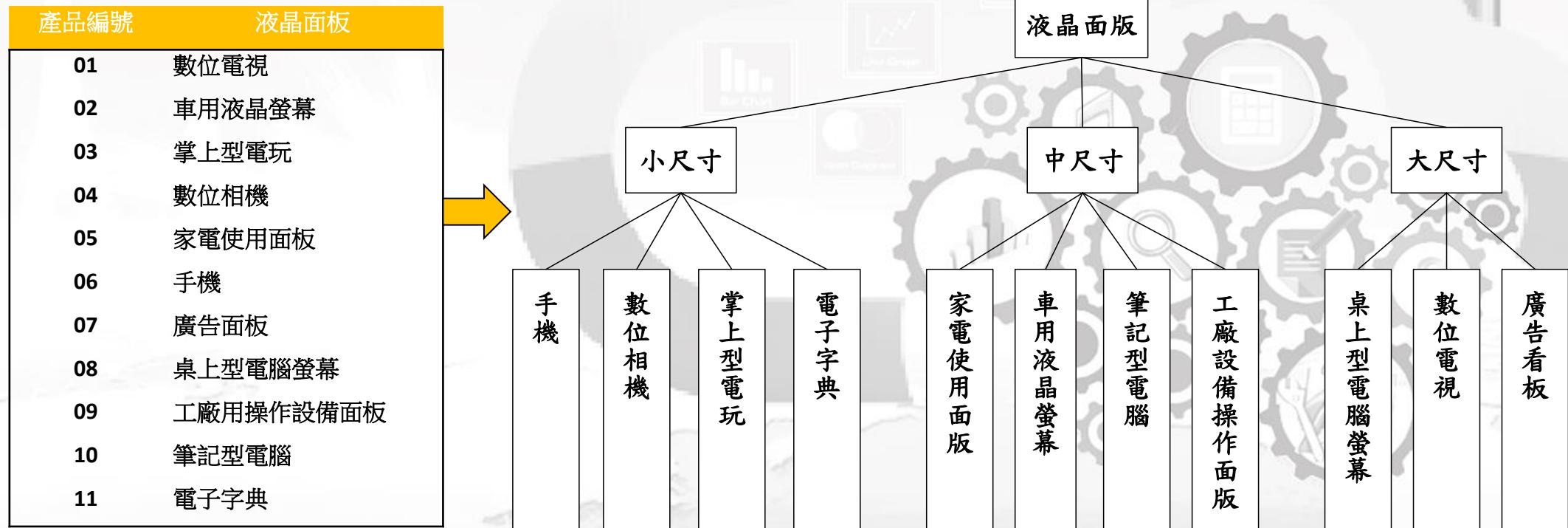
- 資料集或資料庫中的資料表
- 資料表中描述資料集合所用的特徵或屬性欄位稱為資料維度 (Data Dimension)
- 資料維度所描述的資料集合稱為資料紀錄，記錄資料集合於某一維度下的數值稱為資料數值 (value)，在某一維度下所有可能出現的數值稱為值域 (domain)
- 資料維度化約可以減少資料紀錄的長度，資料數值化約則能縮小可能的值域

# 資料維度縮減 (Data Dimension Reduction)

- 簡稱資料縮維
- 常用在分類或預測的問題
- 方式
  - 以目標變數作為比較基準，利用特徵選取法將變數維度及與目標變數不相關的屬性刪除
  - 利用主成分分析法 (PCA) 將變數作線性轉換，只留下提供較多資訊的幾個主要成分 (與目標變數高度相關的成分)，藉以縮小變數維度。
  - 此法不需要目標變數作為比較基準，目的在於找出最能解釋資料變異的線性組合

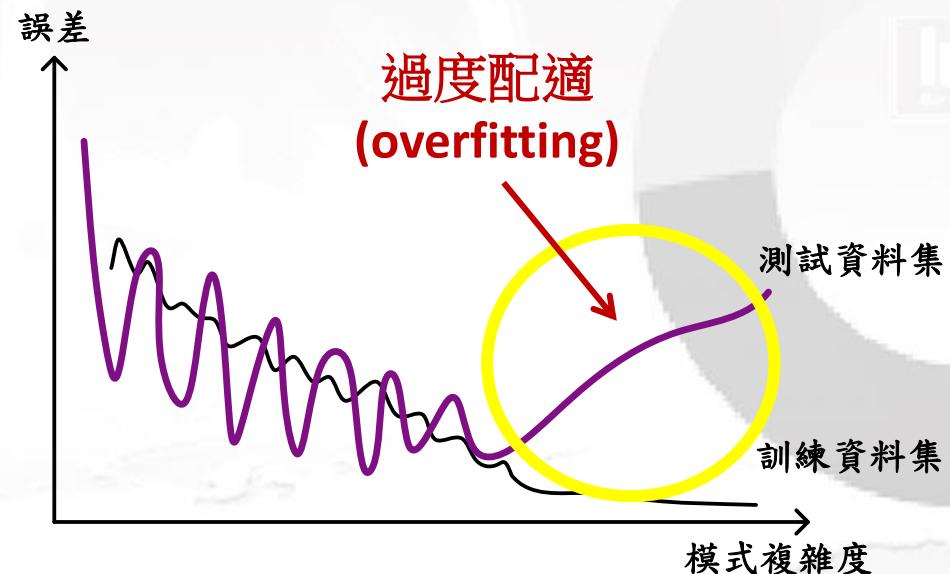
# 資料數值化約

- 資料探勘主要是找出較高層次的知識，如特殊的樣型或趨勢，因此需將原始資料中太細或較低層次的資料離散化與一般化
  - 連續型資料可使用離散化方法，將屬性值域分為若干區間
  - 離散型資料則可使用**概念階層**



# 資料分割(Data Partition)(1/2)

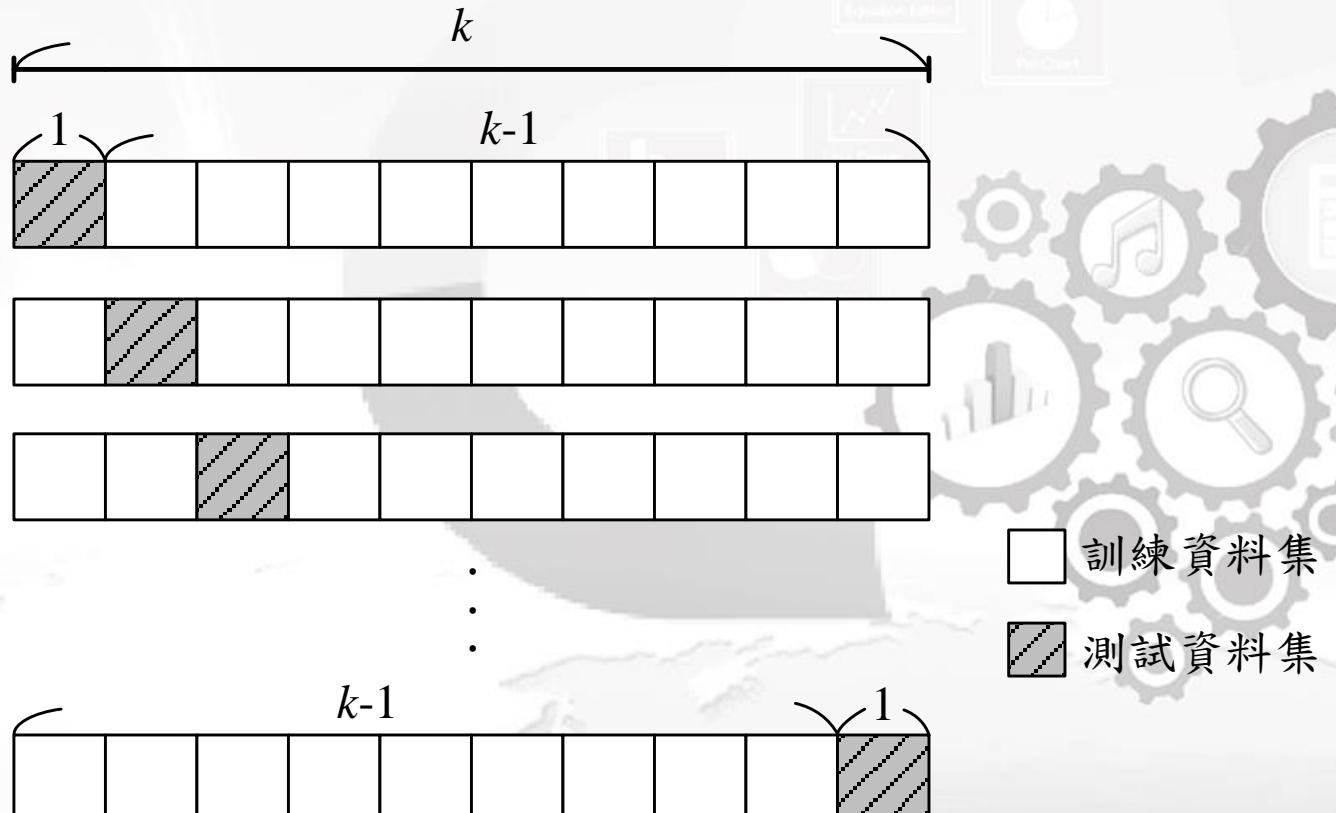
- 資料分成：
  - 訓練資料組 (training data)：建立模式
  - 測試資料組 (testing data)：評估模式
  - 驗證資料組 (validation data)：衡量模式的好壞，如分類錯誤率、均方誤差



當模式複雜度越來越高，導致訓練資料組準確率高而測試資料組的誤差卻越來越大，表示該訓練模型有過度配適情形

# 資料分割 (2/2)

- 資料分割的比例有不同的定義，均應代表原來的資料
  - 方法1：80%資料建構模式，20%用於模式效度檢驗
  - 方法2：*k-fold*交互驗證(*k fold cross-validation*) → 較佳的方法



# 資料分析之產業應用個案

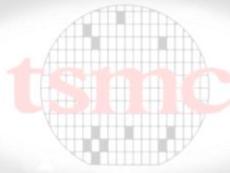
利用資料探勘提升半導體廠  
製造技術員人力資源管理品質

## Reference:

簡禎富、王興仁、陳麗妃（2005），利用資料挖礦提升半導體廠製造技術員人力資源管理品質，  
品質學報，12(1)，9-28。

# 案例背景

- 產業所屬領域：半導體產業 (新竹科學園區-某半導體公司)
- 實務專業領域：人力資源管理 (HRM)
- 資料分析對象：以個案公司人力資源管理資訊系統之資料進行**實證資料分析的探索性研究**
- 個案問題背景：
  - 該公司**製造部門的技術員來源複雜**，有外籍勞工也有本地勞工，語言、文化、學歷等背景皆不相同
  - 現場主管**抱怨新進技術員的素質無法符合公司的要求**，希望能夠招募適當的人員，以提升相關生產的績效，卻也**無法具體提出技術員遴選的條件與方式**
- 確認研究目的：建立**員工績效及人力評估最佳化模型**
- 資料彙整規劃：資料擷取、資料轉換及資料處理



# 資料擷取

- 從HR資訊系統萃取個案公司製造部門技術員的個人基本資料與績效資料，說明資料準備的實際應用過程
- 資料來源：以生產線所有的技術員為對象，共計465位
- 資料蒐集時間：90年1月1日至4月3日
- 資料屬性：
  - 員工個人資料指標，包括：姓名、工號、課別（Photo、Etch、Diff）、班別（DA、DB、NA、NB）、職等（T1~T7）、國籍（本地勞工或外籍勞工）、生日、血型、畢業學校（school）、科系別（master）以及有無其他工作經驗（experience）
  - 工作表現與績效指標，包括：提案次數（proposal）、特殊發現次數（apple）、操作錯誤次數（M.O.）、異常狀況反映（report）以及績效排名（ranking）等

# 資料轉換

- 先轉換資料格式，以減少資料變化所產生的不必要的複雜度：
  - **工作經驗**：工作經驗原有數十種描述，簡化成三種
  - **學校**：原有數十家，簡化成專科、高職、高中等三類
  - **科系**：原有37種科系，依教育部分類方式，再分成16類
  - **提案次數**：從0次到32次皆有，進一步分為4類
  - **異常狀況反映**：從0次到11次皆有，再分為4類
  - **特殊發現次數**：分布從0~4次，為增加其可讀性，亦分為3類

# 遺漏值的處理

- 採推論的補值方式

- 目的在於以其他資料提供的資訊，來推估遺漏值，並嘗試以「較合理」的方式賦予補償值意義
- 例如學校、科系資料不完整者，可檢視其工號；推論依據在於該部門技術員有許多是同校畢業生，且又一起報到並分發至同一班別者大多為相識的同學或朋友

- 採平均值的補值方式

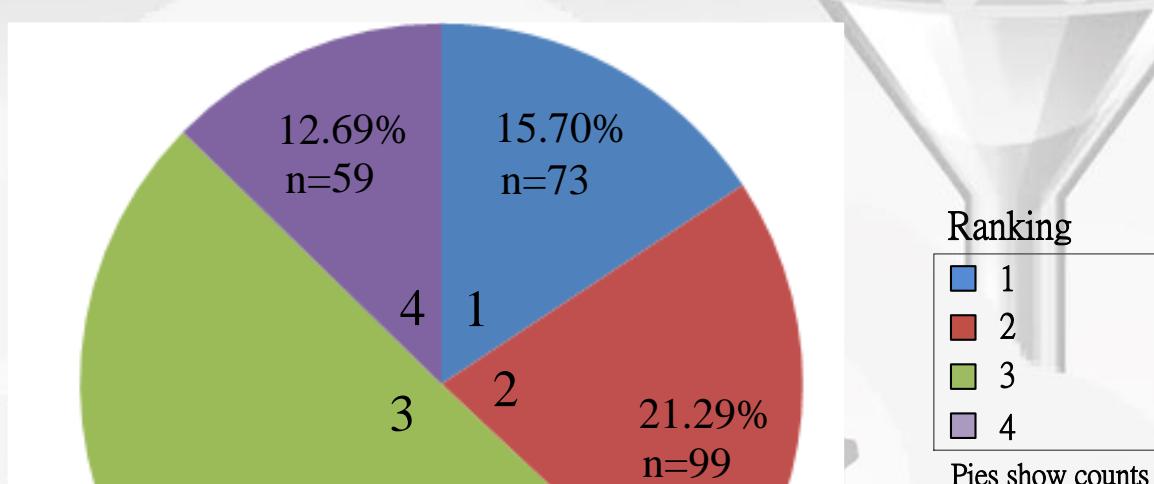
- 以平均值作為不偏估計量，讓中心群資料來取代遺漏的資料。例如，考績遺漏者以2或3代表

- 不予補值的方式

- 例如血型，在難以找到適當的處理方式時，決定不予補值

# 資料特徵強化

- 為增加潛在有用資訊，生日部分以星座來表示，共12種星座；並進一步依其年齡加以區分，共6個年齡層
- 在操作錯誤次數方面，取得前兩年的資料作補充
- 技術員績效部分，除了以年度考核為主的資訊外，另參考非直屬管理人員的意見，並分為四個等級



# 員工個人基本資料與工作績效(部分)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
工號 TA ID	課別 Section	班別 Shift	職等 T.G.	國別 Nation	年齡 age	星座 Star	血型 Blood	學歷 School	科系類別 master type	經驗 experience	提案次數 proposal	特殊發現 次數 Apple	操作錯 誤次數 M.O.	異常狀 況反映 report	考績 Ranking
(略)	ETCH	DB	T7	Local	5A	AA	O	高職	商業及管理學類	非相關	Never	Never	0	Never	4
(略)	DIFF	DA	T6	Local	5A	GG	O	高職	商業及管理學類	非相關	seldom	Never	0	Never	1
(略)	DIFF	DB	T6	Local	6A	EE	A	高職	商業及管理學類	非相關	seldom	Never	0	Seldom	1
(略)	ETCH	DB	T6	Local	6A	GG	O	高職	商業及管理學類	非相關	Never	Seldom	0	Seldom	1
(略)	PHOTO	DB	T6	Local	5B	II	AB	高職	商業及管理學類	非相關	Never	Seldom	1	Seldom	2
(略)	PHOTO	DB		Local	5B	EE	AB	高職	商業及管理學類	無	seldom	Never	0	Seldom	1
(略)	DIFF	DA	T6	Local	5A	FF	A	高中	普通科	有	Never	Never	0	Seldom	3
(略)	DIFF	DA	T7	Foreign	5B	AA	O	高中	普通科	非相關	seldom	Seldom	2	Seldom	1
(略)	DIFF	DB	T6	Local	5B	II	O	高職	商業及管理學類	非相關	seldom	Never	0	Seldom	2
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

# Sum up

- 受到資料來源的不同，資料探勘分析時需處理的資料型態也不盡相同，適當的瞭解蒐集的資料特性將有助資料探勘模式的選擇
- 有意義的資料呈現已成為資料探勘與巨量資料分析的重點，視覺化的工具將可提供資料探勘分析者更多元的整合資訊
- **資料準備**為資料探勘的重要步驟，所需耗費的時間可能遠高於其他步驟

# Workshop

- Data Analysis 基礎操作練習 - 使用R語言
  - Workshop 2 - R Data Type, Structure & Functions
  - Workshop 3 - R Programming
  - 程式練習 - 海賊王 ([1.Data Manipulation\\_One\\_Piece.R](#))