

iPAS營運智慧分析師 July 14, 2021

大數據分析

- R/Python/Julia/SQL 程式設計與應用
(R/Python/Julia/SQL Programming and Application)
- 資料視覺化 (Data Visualization)
- 機器學習 (Machine Learning)
- 統計品管 (Statistical Quality Control)
- 最佳化 (Optimization)

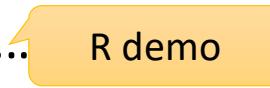


李明昌博士

alan9956@gmail.com

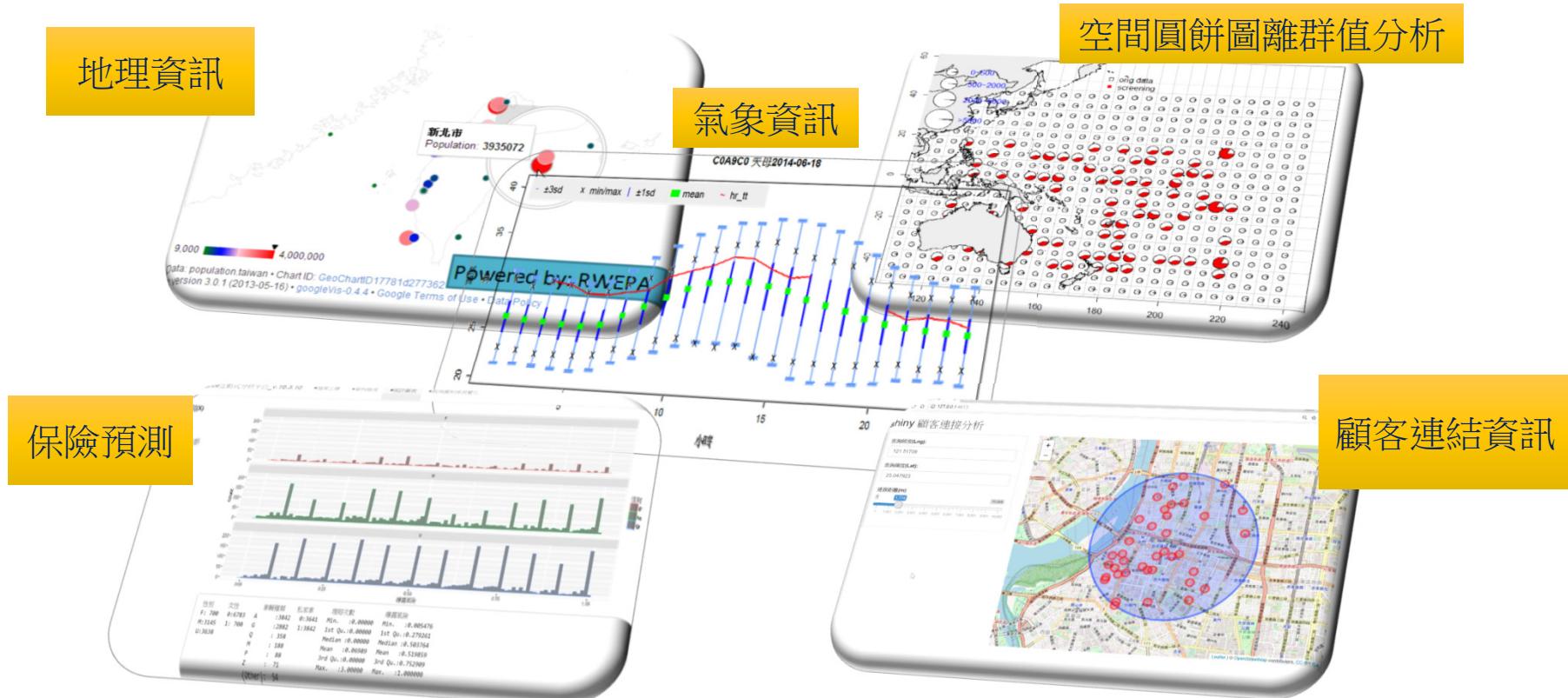
@RWEPA

大綱

• 營運智慧分析師簡介(L11營運智慧概論,L12經營管理數位化概論)	10
• (1) L111營運智慧基本知識	24
• (2) L112基礎資料分析	 R demo 82
• (3) L121經營管理基本知識	139
• (4) L122數位化企業資訊工具基本知識	214
• Recap	309
• Q & A	311

營運智慧資料分析暨視覺化應用

R + shiny → 互動式網頁



網頁呈現

中央氣象局 1,600萬筆資料



保險預測模型

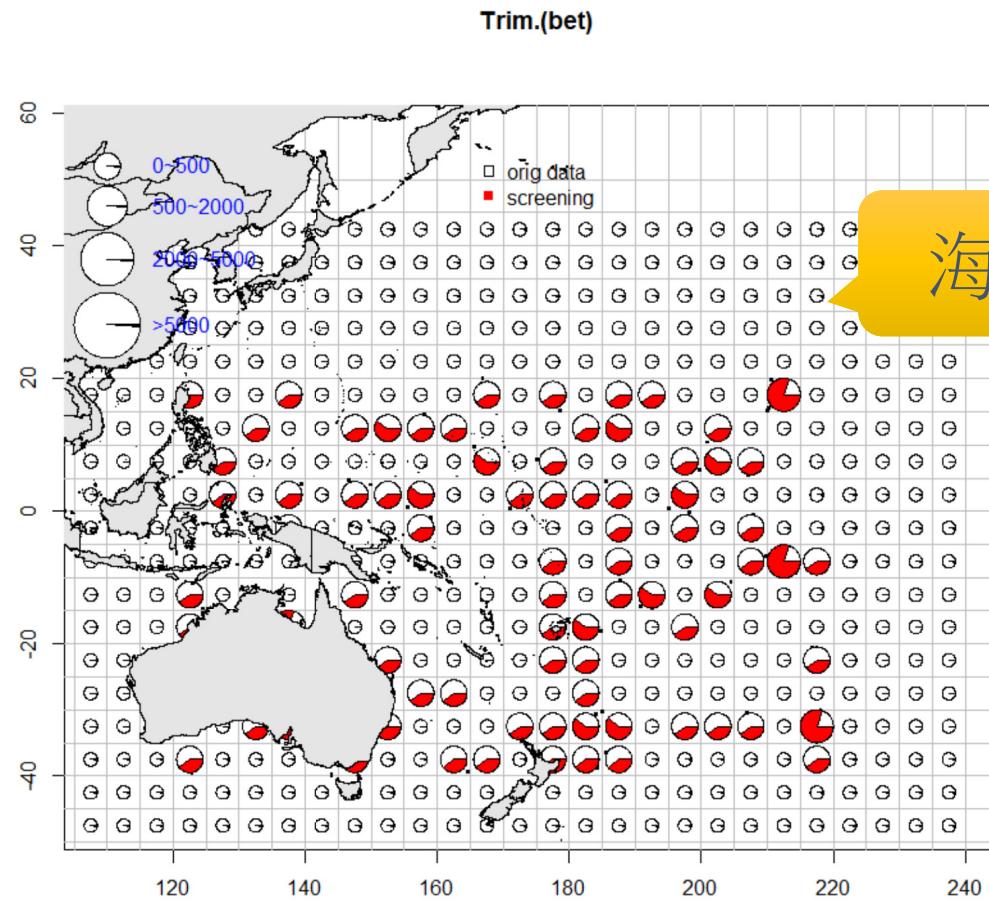
機率模型閾值調整

預測結果

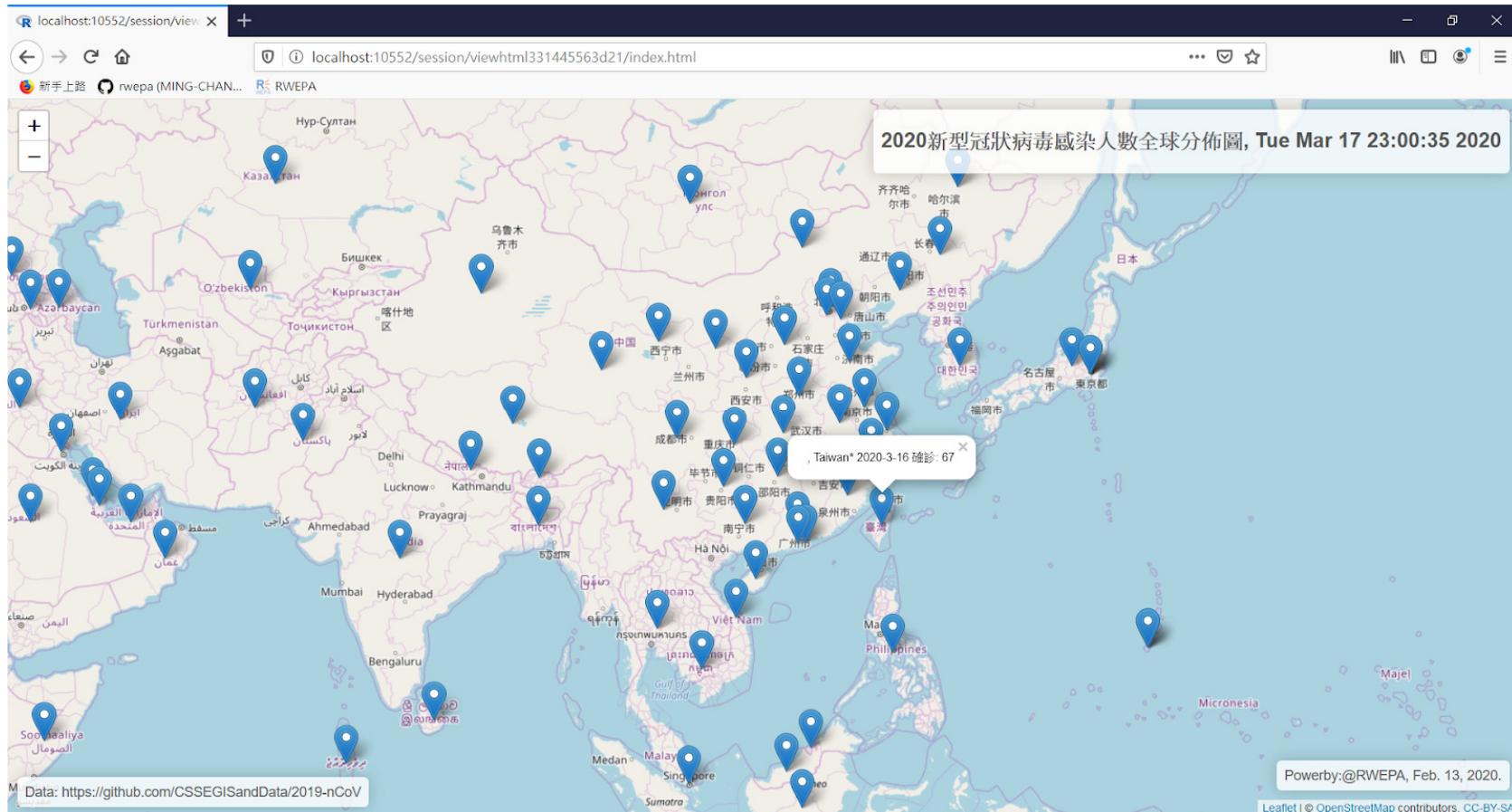
The screenshot shows the iInsurance interactive analysis platform version v.16.3.24. The top navigation bar includes links for document upload, data processing, statistical charts, model evaluation, and prediction models. A red box highlights the 'Prediction Model' dropdown menu. Below it, a modal window titled 'Prediction Data Upload' has a blue button labeled 'Review Results'. A yellow callout points to this button with the text '預測結果' (Prediction Results). Another yellow callout points to a slider labeled 'Probability Model Threshold' with the value '0.1', with the text '機率模型閾值調整' (Probability Model Threshold Adjustment) above it. The main content area displays a table of 12 entries, each with various demographic and vehicle information, along with a predicted probability and a claim status. The last two columns are highlighted with red boxes: '預測機率' (Prediction Probability) and '理賠' (Claim Status). Red arrows point from the 'Review Results' button to these columns. The table includes columns for gender, vehicle type, private car, exposure risk, exposure risk count, no claim discount, insured person age, private car age 0, private car age 1, private car age 2, private car age 0_1_2 combination, car age 0_1_2 combination, and prediction probability.

性別	女性	車輛種類	私家車	曝露風險	曝露風險對數	無索償折扣	被保險人年齡	私家車 一車齡 0	私家車 一車齡 1	私家車 一車齡 2	私家車 -車齡 0_1_2 組合	車齡 0_1_2 組合	預測機率	理賠		
M	0	A	1	0.9144422	-0.08944106	50	4	1	0	0	1	0	2	0.1069	有	
M	0	A	1	0.8158795	-0.20348856	20	4	0	0	1	1	2	2	0.1441	有	
3	M	0	A	1	0.8377823	-0.17699695	50	3	0	0	1	1	2	2	0.1866	有
4	M	0	A	1	0.4325804	-0.83798702	50	6	0	1	0	1	1	2	0.0944	無
5	M	0	A	1	0.7173169	-0.33223755	50	4	0	0	1	1	2	2	0.1218	有
6	M	0	A	1	0.8377823	-0.17699695	50	4	0	0	1	1	2	2	0.1495	有
7	M	0	A	1	0.8487337	-0.16400975	50	5	0	0	1	1	2	2	0.1422	有
8	F	1	A	1	0.8268309	-0.19015503	10	3	0	0	1	1	2	2	0.1733	有
9	M	0	A	1	0.7145791	-0.33606164	0	5	1	0	0	1	0	2	0.0694	無
10	M	0	A	1	0.3340178	-1.09656101	0	3	0	0	1	1	2	2	0.0783	無

空間圓餅圖離群值分析

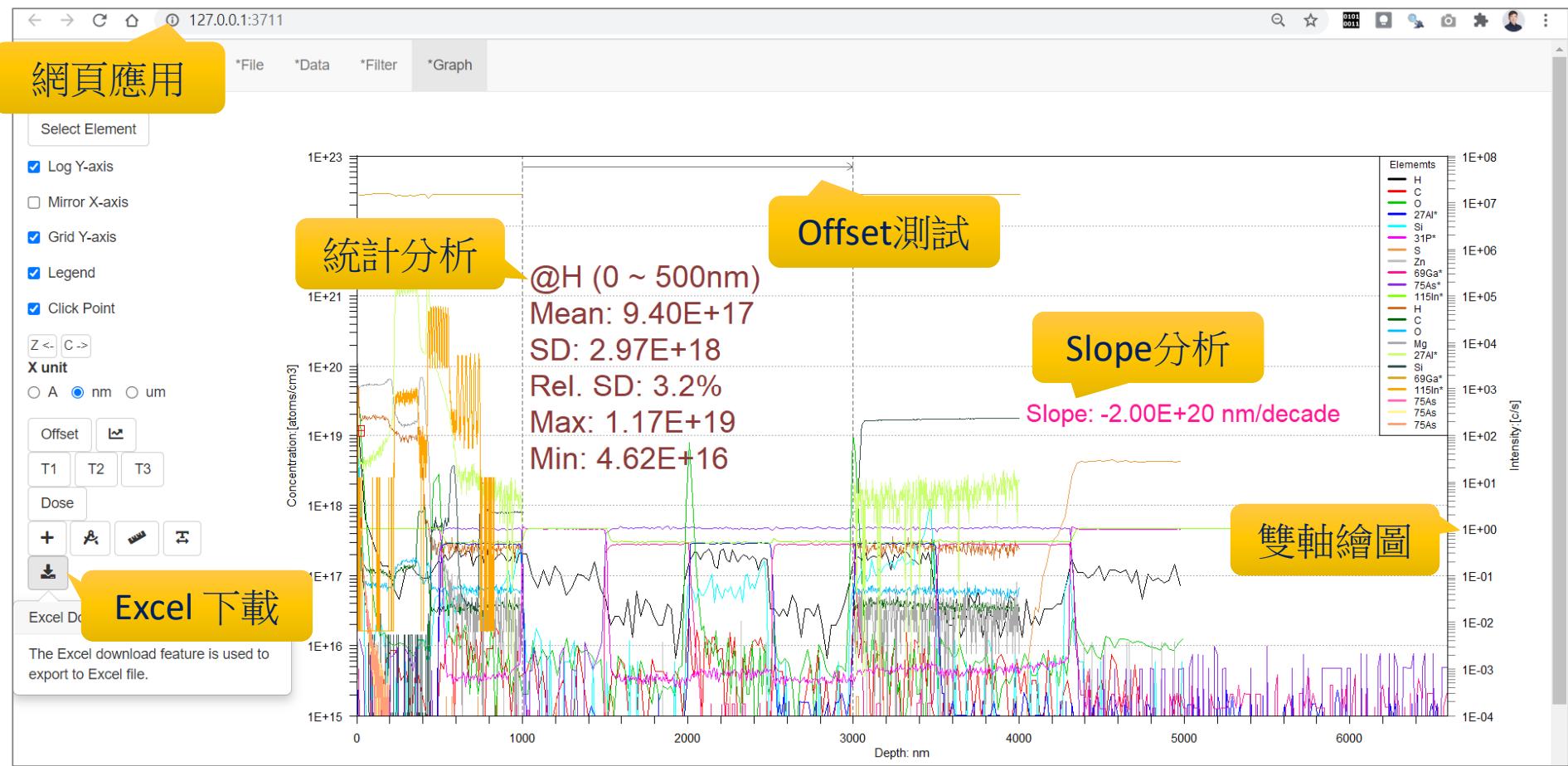


2020新型冠狀病毒視覺化應用

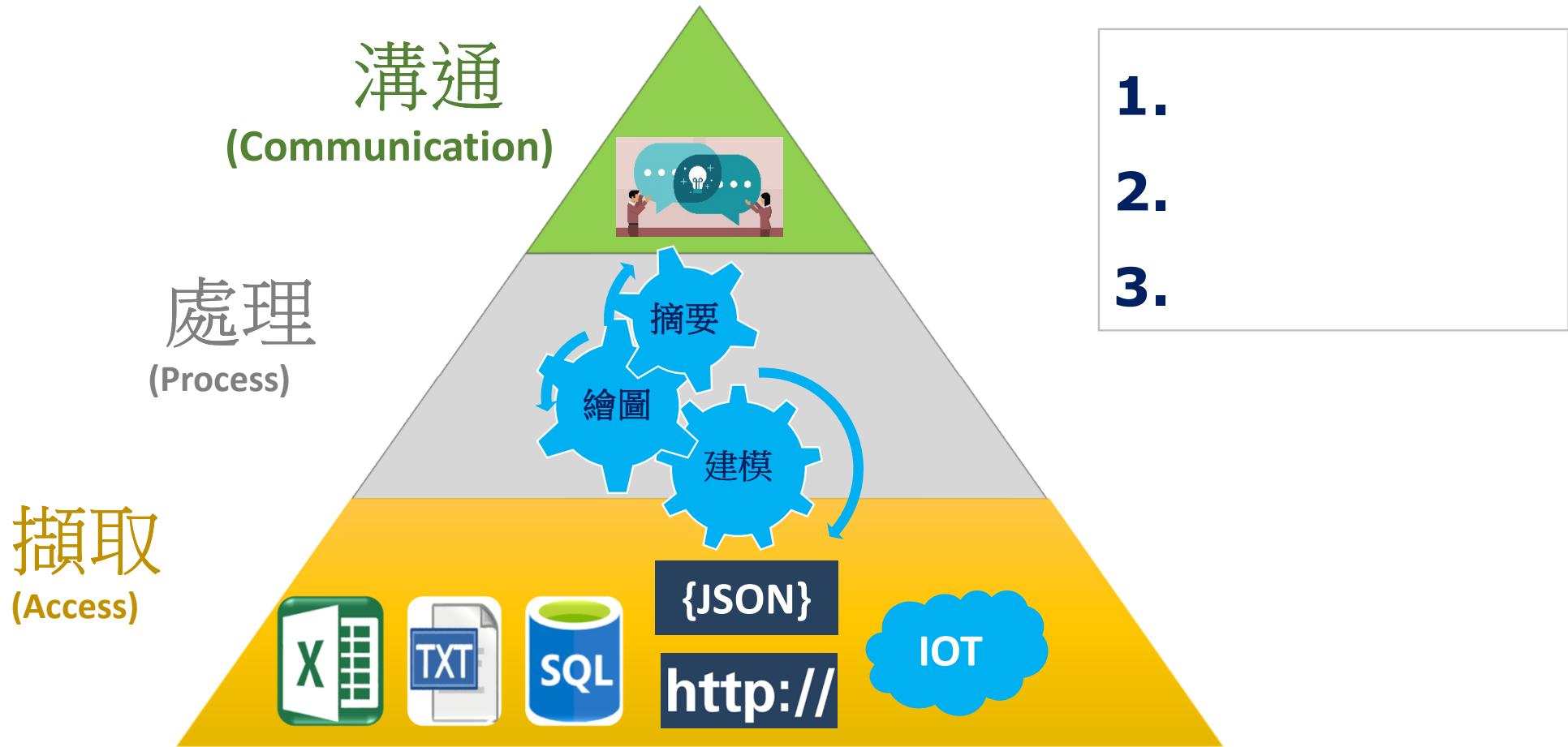


<http://rwepa.blogspot.com/2020/02/2019nCoV.html>

離子資料分析與視覺化應用



資料分析架構→APC方法



營運智慧分析師簡介

營運智慧分析師檢定

- <https://www.ipas.org.tw/OIA>

The screenshot shows the official website for the OIA examination. At the top left is the logo of the Industrial Development Bureau (IDB) and the Ministry of Economic Affairs (MEA). Next to it is the iPAS logo with the text "經濟部產業人才能力鑑定" and "Industry Professional Assessment System". A red arrow points from the text "專業工程師考試 營運智慧分析師" to the "專業工程師考試" section of the navigation bar. Below the header is a banner with a woman holding a bar chart, the text "經濟部發證・教育部認可", and a graphic for "專業 iPAS 就業 All Pass". The main content area includes sections for "考試公告" (Exam Announcements) and "最新消息" (Latest News), both with lists of links. A large callout on the right side promotes the "108年第二次營運智慧分析師-初級能力鑑定" exam, showing a "54 Days" countdown and a "立即報名" button.

最新消息、關於能力鑑定、團報優惠專區、企業、學校/培訓機構認同、檔案下載、職能基準、常見問題、

IDB
INDUSTRIAL DEVELOPMENT BUREAU,
MINISTRY OF ECONOMIC AFFAIRS
經濟部工業局

iPAS 經濟部產業人才能力鑑定
Industry Professional Assessment System

專業工程師考試 營運智慧分析師

最新消息 / 職能基準 / 職能基準檢索下載

考試公告

- 108年第二次營運智慧分析師-初級能力鑑定 - 2019/11/30 ~ 2019/11/30

more

最新消息

- 108年第一次營運智慧分析師榮譽榜
- 108年第一次營運智慧分析師考試成績
- 108年第一次初級營運智慧分析師能力鑑定考試成績於6/12(三)開放網路查詢。
- 108.03.12【考試樣題及參考書目】營運智慧分析師
- 108.02.13【公告】108年度營運智慧分析師能力鑑定簡章內文興動
- 108.01.10【公告】108年度營運智慧分析師能力鑑定簡章

108年第二次營運智慧分析師-初級能力鑑定11/30

報名截止10/15倒數

54Days

立即報名(另開新視窗)

能力指標

►1.4 能力指標：

➤ 各級等能力指標：

初級		
考科	1.營運智慧概論	2.經營管理數位化概論
能力指標	<ul style="list-style-type: none">◆ 能敏銳地理解各式資料(結構與非結構)生成的方式與商業邏輯。◆ 根據分析標的，彙集營運解決方案所需之數據資料及其來源，並將之表格或系統化。	<ul style="list-style-type: none">◆ 對企業營運之生產、行銷、研發、人資、資訊、財務與策略管理系統之流程有基本概括的了解，並能夠理解流程對應之資訊工具，以從營運業務流程中，快速掌握資料蒐集管道。

評鑑主題

L1 初級		
科目	評鑑主題	評鑑內容
1 L11 營運智慧概論	L111 營運智慧基本知識	L11101 營運智慧簡介(含營運智慧 vs. 商業智慧) L11102 營運智慧與企業管理 L11103 營運智慧與資訊管理 L11104 評估與規劃營運智慧
	L112 基礎資料分析	L11201 資料來源與資料獲取(含資安) L11202 資料性質(例如：結構性與非結構性) L11203 常用統計概念及其資料前處理
2 L12 經營管理數位化概論	L121 經營管理基本知識	L12101 企業經營環境與策略管理 L12102 企業的核心流程及其管理活動 L12103 財務會計基本知識
	L122 數位化企業資訊工具基本知識	L12201 營運智慧資訊技術(如雲端技術、無線射頻辨識技術、物聯網、大數據等) L12202 數位化企業常見資訊系統(如企業資源規劃、供應鏈管理、電子商務、雲端運算、知識管理等) L12203 數位化轉型創新與價值創造(包括商業模式、企業價值鏈、核心流程與所需之資訊科技、企業流程再造，及結合後創新與創造價值)

- 營運智慧
- 商業智慧

- 資料分析
- 統計應用

- 管理
- 產銷人發財

- 物聯網
- 大數據
- 機器學習

考試題型

專業級等	日期	時間	科目	題型	鑑定方式	考區
初級	第一次： 11/20(六)	09:00~10:15 (75分鐘)	1. 營運智慧概論	選擇題(70%)+ 非選擇題(30%)	紙筆測驗	台北. 台中. 高雄.
		10:45~12:00 (75分鐘)	2. 經營管理數位化概論	選擇題(70%)+ 非選擇題(30%)		

- 單選題 40 題 (70%) → 每題 1.75 分 , 65 分鐘寫 40 題
- 非選擇題 2 題 (30%) → 每題 15 分 , 10 分鐘寫 2 題

授證資格

專業級等	考試科目	考科及格標準/成績保留	授證資格
初級	1. 營運智慧概論 2. 經營管理數位化概論	<p>➤ 及格標準：</p> <ol style="list-style-type: none"> 1. 每科100分，該科達70分為及格（成績計算以四捨五入方式取整數）。 2. 同時報考同一級等的所有考科，平均達70分得視為及格，但單科成績不得低於50分。 <p>➤ 成績保留：</p> <p>保留及格單科成績自應考日起三年度有效。</p>	二個考科皆達及格標準。
成績保留	<p>保留及格單科成績自應考日起三年度有效。</p> <p>範例說明：</p> <p>105/05/01報考，單科及格成績可保留至108年12月31日止。</p> <p>105/12/01報考，單科及格成績可保留至108年12月31日止。</p>		

109年第一次考試

109年度第一次營運智慧分析師-初級能力鑑定各考科通過人數統計		
專業級等	初級營運智慧分析師能力鑑定(90人/179人次)	
考科	營運智慧概論	經營管理數位化概論
報考人數	90	89
到考人數	80	79
到考率	88.89%	88.76%
平均分數	62.85	66.47
及格人數	21	28
及格比例	26.25%	35.44%
本次到考總人數 計算方式:不論報考幾考科，只要有1考科到考，即算到考。報2考科，只要到考1考科也算	80 人	
本次證書共核發(當次+跨次)	24 張	獲證率:30.00%
當次報考獲證:	23 人	
跨次報考獲證:	1 人	
授證資格	1.每科100分，該科達70分為及格(成績計算以四捨五入方式取整數)。 2.同時報考同一級等的所有考科，平均達70分得視為及格，但單科成績不得低於50分。	

109年第二次考試

109年度第二次營運智慧分析師-初級能力鑑定各考科通過人數統計		
專業級等	初級營運智慧分析師能力鑑定(90人/179人次)	
考科	營運智慧概論	經營管理數位化概論
報考人數	172	173
到考人數	163	164
到考率	94.77%	94.80%
平均分數	67.79	75.51
及格人數	57	92
及格比例	34.97%	56.10%
本次到考總人數 計算方式:不論報考幾考科，只要有1 考科到考，即算到考。報2考科，只 到考1考科也算	164 人	
本次證書共核發(當次+跨次)	69 張	獲證率:42.07%
當次報考獲證:	69 人	
跨次報考獲證:	0 人	
授證資格	1.每科100分，該科達70分為及格(成績計算以四捨五入方式取整數)。 2.同時報考同一級等的所有考科，平均達70分得視為及格，但單科成績不得低於50分。	

營運智慧概論
及格比例較低?

參考書目

營運智慧分析師參考書目

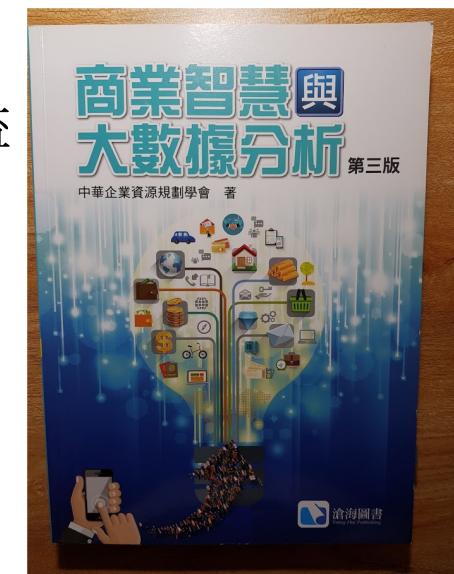
適合等級	No.	書名	作者	出版資料
初級	2	<u>認識資料科學的第一本書 Data Analytics Made Accessible</u>	Anil Maheshwari/徐瑞珠譯	碁峰出版，2017
初級	1	<u>商業智慧與大數據分析(第三版)</u>	中華企業資源規劃學會	滄海出版，2017
初級	8	<u>作業管理原理中文第一版 Principles of Operations Management 7/e</u>	Heizer、賴奕銓/審閱	雙葉書廊，2010
初級	3	<u>大數據戰略 4.0</u>	任立中總編輯	前程出版，2016
初級	7	<u>商用統計學(Lind/Statistical Techniques in Business & Economics 17e)</u>	Douglas A. Lind, William G. Marchal, Samuel A. Wathen	華泰文化，2018
中級	3	<u>Data Mining and Big Data Analytics 資料挖礦與大數據分析</u>	簡禎富、許嘉裕	前程出版，2014
中級	5	<u>資料科學的商業運用 Data Science for Business</u>	Foster Provost; Tom Fawcett; 陳亦苓譯	歐萊禮出版，2016
中級	4	<u>Google 必修的圖表簡報術</u>	Cole Nussbaumer Knaflic; 徐昊譯	商業周刊，2016

<https://www.ipas.org.tw/OIA/AbilityNewsData.aspx?nwsno=88c9122d-f8be-4dad-82bd-61fa993c5c5a>

No1商業智慧與大數據分析(第三版)

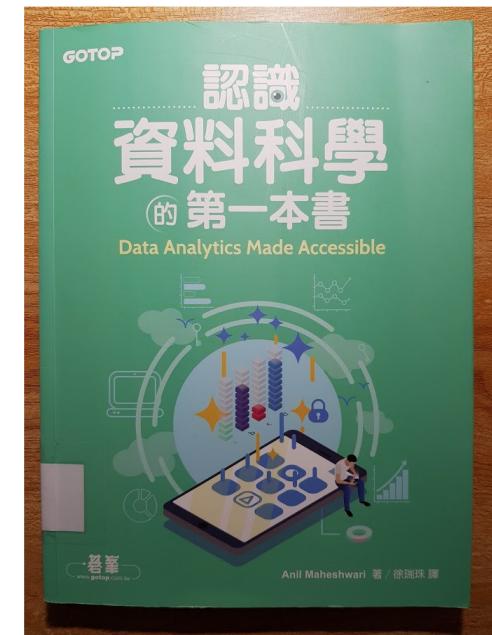
- 第1章 大數據時代的商業智慧簡介
- 第2章 BI 專案生命週期
- 第3章 維度模型化介紹
- 第4章 資料立方體與資料報表呈現
- 第5章 資料倉儲的資料建置
- 第6章 銷售與配銷分析
- 第7章 採購之關鍵績效指標
- 第8章 財務會計模組之關鍵績效指標
- 第9章 商業智慧—生產規劃與控制
- 第10章 人力資源關鍵績效指標
- 第11章 商業智慧對於企業的效益
- 第12章 大數據集群分析介紹
- 第13章 分類技術
- 第14章 關聯規則

有習題



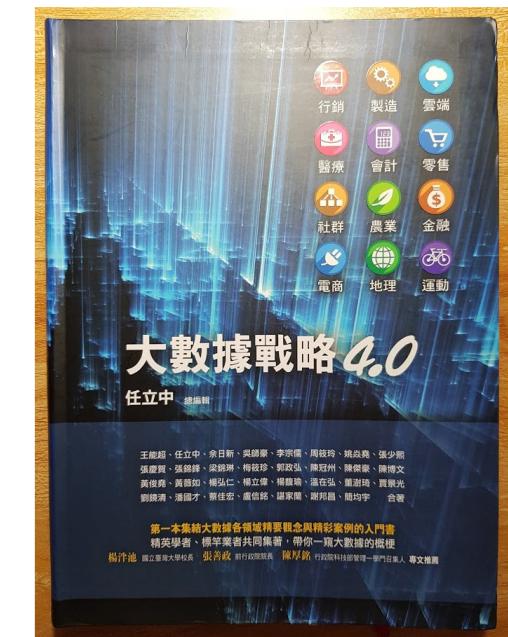
No2認識資料科學的第一本書 (Data Analytics Made Accessible)

- Ch01 資料分析概觀
- Ch02 商業智慧
- Ch03 資料倉儲
- Ch04 資料探勘
- Ch05 資料視覺化
- Ch06 決策樹
- Ch07 迴歸
- Ch08 類神經網路
- Ch09 群集分析
- Ch10 關聯規則探勘
- Ch11 文字探勘
- Ch12 單純貝式分析
- Ch13 支援向量機
- Ch14 網路探勘
- Ch15 社群網路分析
- Ch16 大數據
- Ch17 資料建模入門
- Ch18 資料科學職涯與個案研究



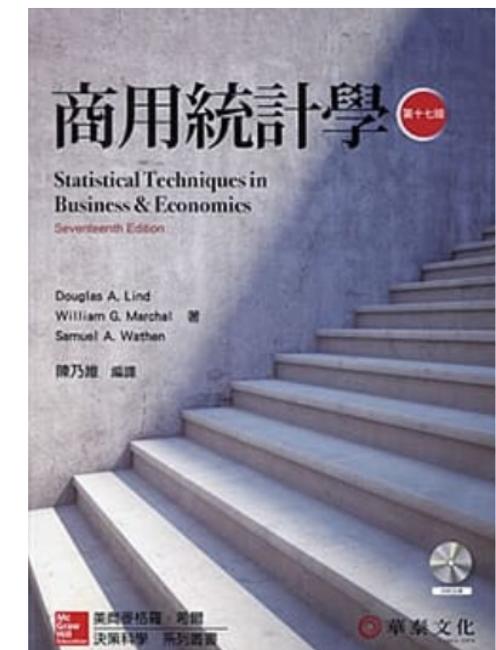
No3大數據戰略

- 第1章 行銷大數據
- 第2章 製造大數據
- 第3章 雲端大數據
- 第4章 醫療大數據
- 第5章 會計大數據
- 第6章 零售大數據
- 第7章 社群大數據
- 第8章 農業大數據
- 第9章 金融大數據
- 第10章 電商大數據
- 第11章 地理大數據
- 第12章 運動大數據



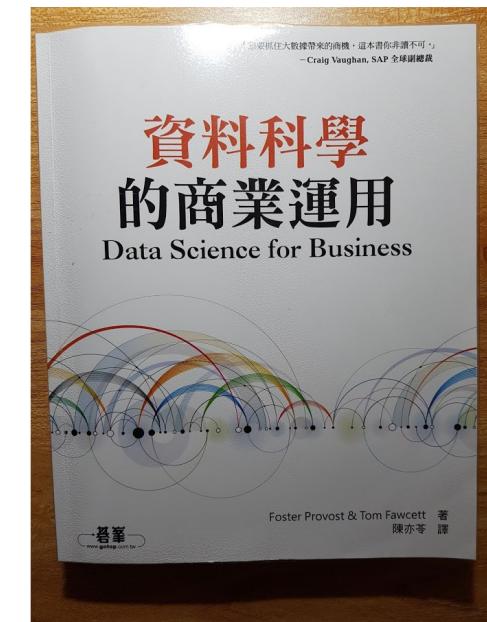
No7 商用統計學

- 第 01 章 何謂統計學？
- 第 02 章 資料的描述：次數表、次數分配與圖形的呈現
- 第 03 章 資料的描述：數值的測量
- 第 04 章 資料的描述：展現與探索資料
- 第 05 章 機率觀念概述
- 第 06 章 離散型的機率分配
- 第 07 章 連續型機率分配
- 第 08 章 抽樣方法與中央極限定理
- 第 09 章 估計與信賴區間
- 第 10 章 單組樣本的假設檢定
- 第 11 章 兩組樣本的假設檢定
- 第 12 章 變異數分析
- 第 13 章 相關分析與線性回歸
- 第 14 章 複迴歸分析
- 第 15 章 非參數方法：名目與順序尺度資料的假設檢定



No5資料科學的商業運用(中級) (Data Science for Business)

- 第一章 序論：數據分析思維
- 第二章 商業問題與資料科學解決方案
- 第三章 預測性建模入門：從關聯性到監督式區隔
- 第四章 將模型配適於數據資料
- 第五章 過適與避免過適
- 第六章 相似性、鄰近及聚類
- 第七章 決策分析思維I：怎樣的模型才是好模型？
- 第八章 將模型效果視覺化
- 第九章 證據與機率
- 第十章 文本的表述與文字採礦
- 第十一章 決策分析思維II：關於分析設計
- 第十二章 其他的資料科學任務與技術
- 第十三章 資料科學與商業策略
- 第十四章 總結



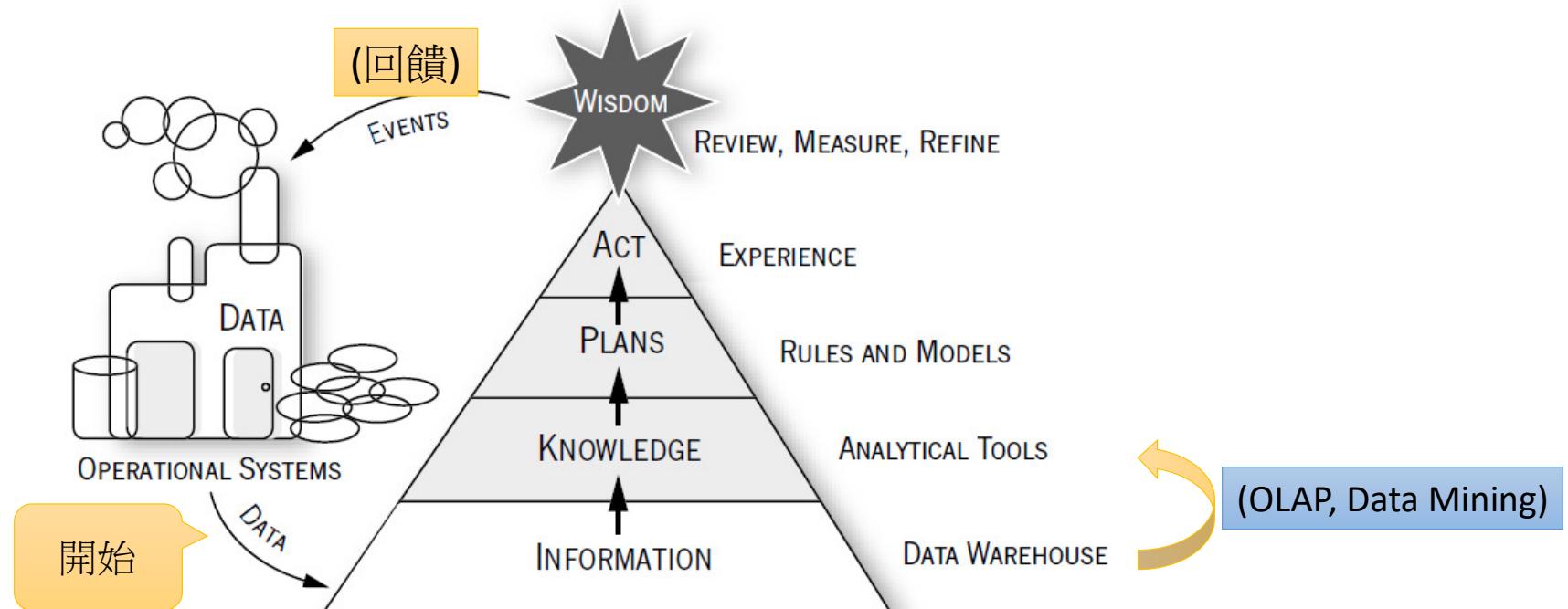
1.L111 營運智慧基本知識

- L11101 營運智慧簡介(含營運智慧vs.商業智慧)
- L11102 營運智慧與企業管理
- L11103 營運智慧與資訊管理
- L11104 評估與規劃營運智慧

商業智慧

BI As a Data Refinery

資料煉油廠



參考: Wayne Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligence Solutions, 2013. http://download.101com.com/tdwi/research_report/2003BIReport_v7.pdf

商業智慧 – 步驟1

1. 從原始資料到資料倉儲的資訊

- 第一步是從企業間交易和企業內營運系統中萃取資料，然後經過清理、轉換等處理步驟，將定義清楚且一致的細節和彙總的資料，載入資料倉儲的資料庫中，從最底層的資料轉換成資料倉儲的資訊。
- 例如：將分散在訂單、維修服務、銷售、出貨、和會員等系統中的顧客資料記錄整合成一個以顧客為主題的完整資料庫，對於瞭解顧客及其需求產生有用而完整的資訊。

2. 從資訊到知識
3. 從知識到決策
4. 從決策到行動
5. 回饋迴圈

商業智慧 – 步驟2

1. 從原始資料到資料倉儲的資訊

2. 從資訊到知識

- 使用者可以運用各種報表和分析工具，例如查詢、報表、線上分析處理（OLAP）、和資料探勘等，存取並分析資料倉儲中的資訊。
- 這些分析可以找出資料中的趨勢(Trends)、型態(Patterns, 樣式)、和例外狀況等，這些分析工具幫助使用者將資訊轉換成知識。
- 例如：零售通路從大量銷售資料中挖掘出特定類型的顧客會同時購買紙尿布和啤酒之間的關聯規則，對於賣場而言，這個發現對於商品陳列是有一定參考價值。

3. 從知識到決策

4. 從決策到行動

5. 回饋迴圈

商業智慧 – 步驟3

1. 從原始資料到資料倉儲的資訊

2. 從資訊到知識

3. 從知識到決策

- 使用者從分析所發現的趨勢和型態中可以建立業務規則，也可以將知識作為建立決策模型的依據，來規劃業務的進行並作為決策的參考。
- 例如：庫存降至10單位時，就要下單採購30個單位。
- 規則也可能是根據過去的趨勢所做的預測，或是根據假設或估計所產生的情境（what if）分析。
- 統計分析和最佳化分析也可以產生比較複雜的規則，例如機動的定價機制以回應變動的市場狀況，其規則可以用統計方法產生，也可以使用利潤最大化最佳化模型。

4. 從決策到行動

5. 回饋迴圈

商業智慧 – 步驟4

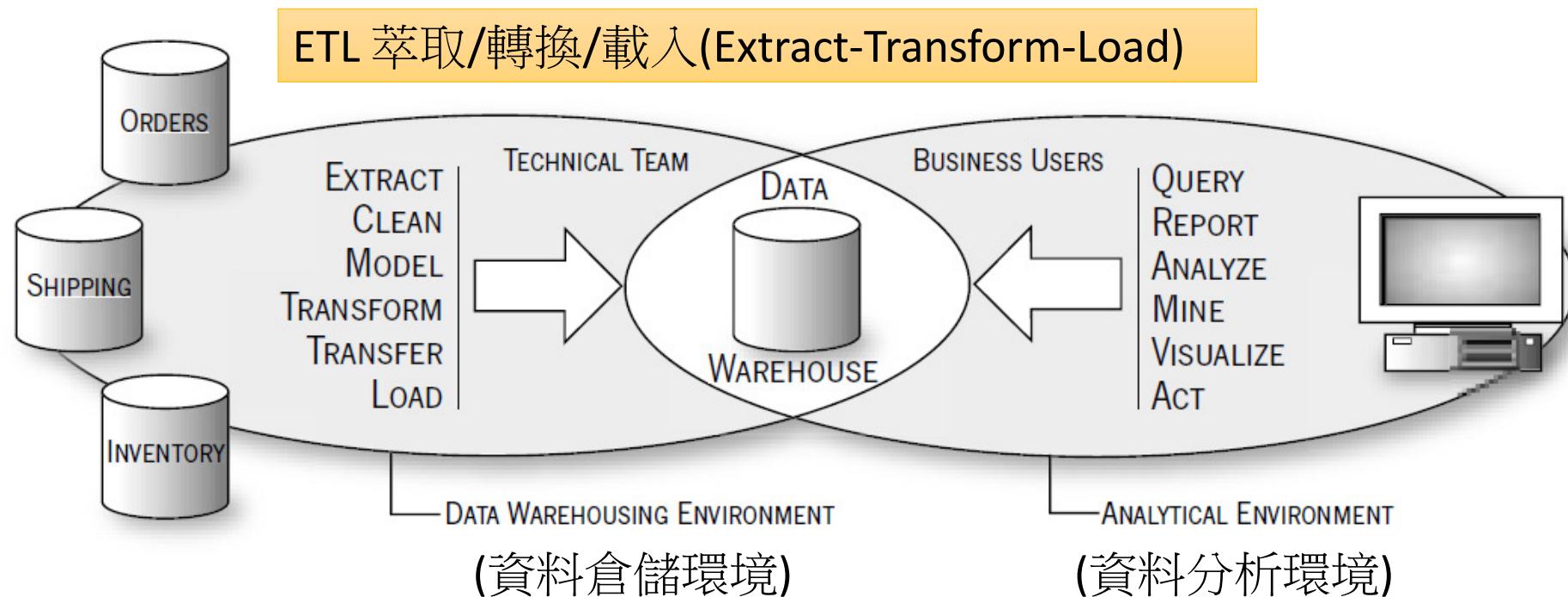
1. 從原始資料到資料倉儲的資訊
2. 從資訊到知識
3. 從知識到決策
- 4. 從決策到行動**
 - 根據前一步驟的業務規則或決策規劃，使用者要產生執行計畫。
 - 例如：行銷人員要根據顧客區隔的分析，顧客回應特定優惠的預測模型，以及過去的促銷活動經驗，來規劃各種促銷活動。
 - 針對不同的客戶透過不同的通路所提供的優惠方案為何。這些執行計畫將決策方案轉換成實際的行動。
5. 回饋迴圈

商業智慧 – 步驟5

1. 從原始資料到資料倉儲的資訊
2. 從資訊到知識
3. 從知識到決策
4. 從決策到行動
- 5. 回饋迴圈**
 1. 一旦計畫開始實施，整個循環便會重複。
 2. 營運系統中會有顧客對於優惠的反應以及後續的交易。
 3. 這些資料會被萃取出來，並和相關資料整合後載入資料倉儲。
 4. 行銷人員就可以進一步分析以評估其促銷活動的成效，並據以修正其促銷活動的規劃。

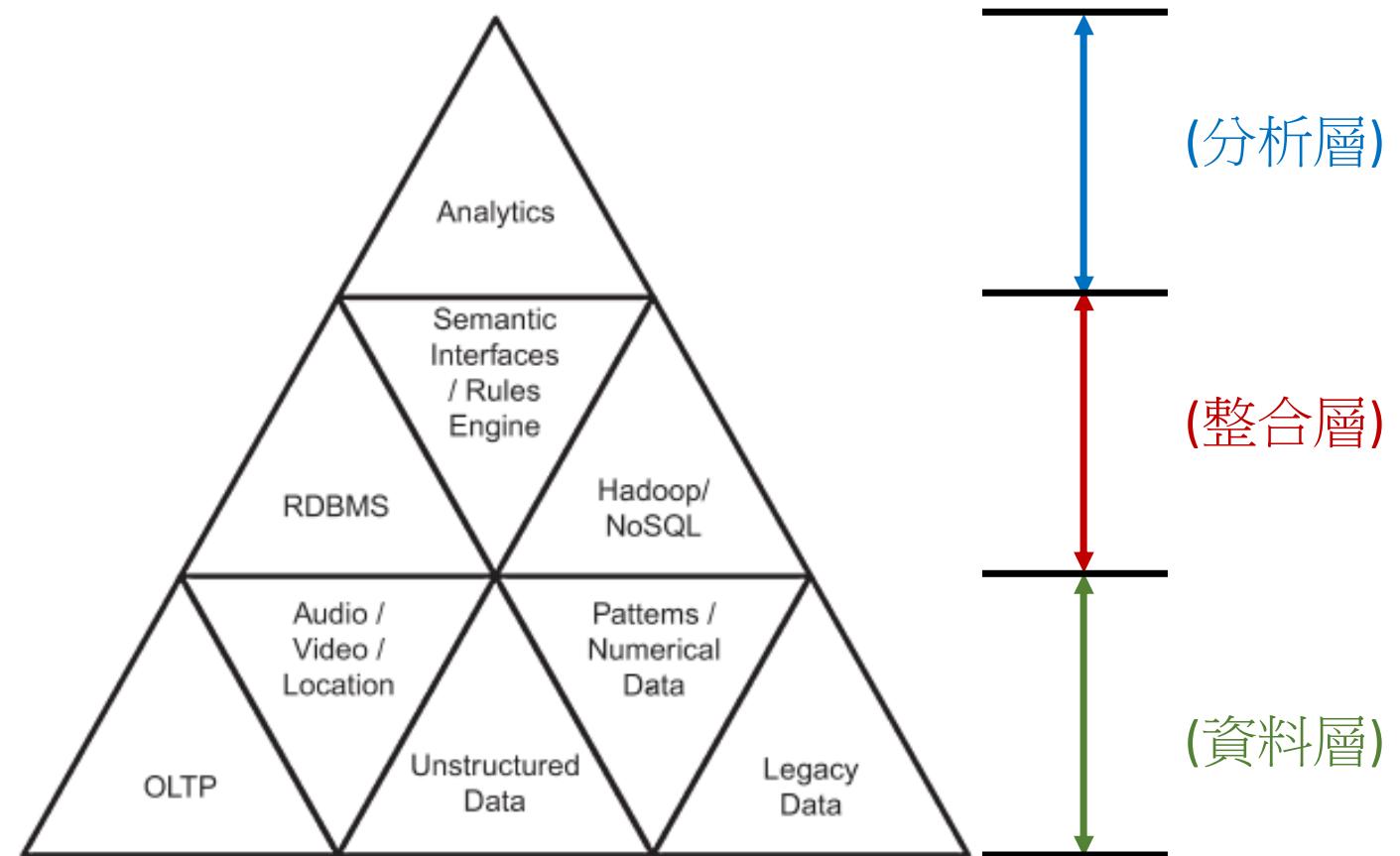
商業智慧系統的組成架構

BI Component Framework



參考: Wayne Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligence Solutions, 2013. http://download.101com.com/tdwi/research_report/2003BIReport_v7.pdf

新世代資料倉儲架構 (Components of the next-generation data warehouse)



商業智慧在企業應用

- 關鍵結果指標(key result indicators, KRI)：公司在某方面的表現，**長時間評估**，例：每月，每季報表交付董事會。
- 績效指標(performance indicators, PI)：公司該做的是什麼，例：獲利前10%的顧客，其主要產品的淨利。
- 關鍵績效指標(key performance indicators, KPI)：該做什麼可以**大幅改善**公司的績效，**短時間評估**，例：每日，每週。→中高階主管
- 顧客滿意度、稅前淨利、和員工滿意度等，是常被誤認為KPI的KRI(本質是KRI)。
- 較佳數量：10 KRI(少), 80 PI(多), 10 KPI(少)。
- 商業智慧中上列三大指標需要被檢視、衡量、和修正。

KRI vs. KPI

比較		KRI	KPI
性質差異	指標意涵	公司是否在正確的方向	公司該做些什麼才能改善這些結果
	檢視週期	每個月、每季、或每年	全時監控或每天或每週
	指標數量	至多 10項	至多 10項
	指標屬性	落後指標	領先指標/目前或未來導向
用途差異	報表類型	治理報表	管理報表
	格式工具	數位儀表板	平衡計分卡
	使用對象	董事會	各級經理人
	指標/報表	至多 10項	含PI至多 20項
	範例指標	顧客滿意度、稅前淨利、和員工滿意度	獲利率前 10% 的顧客、主要產品線的淨利、和前 10% 顧客的銷售成長百分比

BI專案生命週期



評估：

- 公司營運上的評估
- 成本效益分析
- 風險評估

規劃：

- 技術面的基礎設施-硬體平台,中介軟體平台,資料庫管理系統平台
- **非技術面**的基礎設施的評估-功能部門的運作,營運活動的作業流程,企業營運資料,企業應用系統,詮釋資料庫(Meta data repository)

專案分析：

- 需求分析
- 資料分析：
 - (1).邏輯資料模型(Logical data modeling)法—確保資料整合與一致性
 - (2).來源資料分析(Source data analysis)法—確保資料品質

資料分析流程

1. 分析外部資料
2. 建立邏輯資料模型—資料須清楚的，沒有重複，沒有不一致，與系統實作的軟硬體設備無關。
3. 分析來源資料品質
4. 規劃整體公司的邏輯資料模型
5. 解決邏輯資料及來源資料的差異
6. 設定資料清理的步驟

詮釋資料儲存庫 (Meta data repository)

- 將資料清楚且唯一地定義，可借助詮釋資料(Meta data)。
- 詮釋資料是所有的資訊系統必然存在的東西，但並不是所有的組織對這些背景資訊都詳細地定義且記錄下來。
- 詮釋資料儲存庫就是儲存這些詳細定義的背景資訊。
- 資料的背景資訊包括：
 1. 該資料的內容與資料代表的意義。
 2. 管理該資料的政策(例如：誰負責維護、更改資料的流程等等)。
 3. 該資料的技術規格(例如：資料型態、長度)。
 4. 使用該資料的資訊系統。

詮釋資料資料庫 - 分析工作

1. **分析詮釋資料資料庫需求**：針對BI專案，按重要順序列出上述詮釋資料的要項，何者為必要，何種為重要，何者為可選擇項目。
2. **分析詮釋資料資料庫的界面需求**：BI系統的詮釋資料有可能來自其他系統，所以詮釋資料資料庫的界面選擇需考慮不同詮釋資料的整合問題。
3. **分析詮釋資料資料庫的存取及呈現方式的需求**：詮釋資料資料庫中的內容，需透過何種工具存取，內容以何種檔案格式呈現。
4. **建立邏輯詮釋資料模型(Logical meta model)**：邏輯詮釋資料模型即是用來表達詮釋資料的需求，將詮釋資料的要項以實體關聯圖 (Entity-relationship diagram, ERD)呈現。
5. **建立Meta-meta data**：Meta-meta data指的就是Meta data 的背景資訊。

平衡計分卡(Balanced Scorecard, BSC)

- Robert S. Kaplan和David P. Norton在1992年哈佛商業評論發表”The Balanced Scorecard—Measures That Drive Performance”
- Kaplan and Norton在1996年發表BSC的第一本專書：The Balanced Scorecard—Translating Strategy into Action.
- BSC 最早的用意在於解決傳統的績效評核制度過於**偏重財務構面**的問題，但在實際運用後又發現平衡計分卡要與企業的**營運策略**相互結合，才能發揮企業績效衡量的真正效益與目的
- 因此平衡計分卡不僅是一個績效衡量系統，更是一個企業營運策略的管理工具。
- BSC是一種策略管理方法，透過績效管理從四個構面將組織的願景(vision)和策略化為具體的實施方案。

BSC的四大構面

- **BSC**的四個構面：
 - 財務構面—平衡財務與非財務間之指標
 - 顧客構面—平衡落後資訊與領先資訊
 - 內部流程構面—平衡企業內部與外部間的組成要素
 - 學習與成長構面—平衡短期績效與長期價值
- **BSC**的整體觀點不但可以監控目前的績效，其方法也試圖掌握反映公司未來績效的資訊。

營運智慧 (Operations Intelligence, OI)



- 營運智慧特性：
 - 是一種數據分析方法。
 - 屬於即時動態 (real-time and dynamic) 營運分析。
 - 使用企業內部或外部收集的即時資料，幫助企業的決策與執行。
 - 資料分析過程是自動化的。
 - 將分析的資訊集成到資訊系統中，提供業務主管和工作人員立即使用。
- 分析對象：
 - 串流數據(streaming data)
 - 串流事件(streaming events)
 - 當下的營運業務狀況(產銷人發財)
- 強調數據即時處理與分析後所獲得之**可見性和洞察力**的應用。

參考: <https://searchbusinessanalytics.techtarget.com/definition/operational-business-intelligence>

營運智慧 (續)

- BI應用程序主要使用對象為一線員工，輔助能夠及時掌握瞬時的變化。
- 採用即時線上分析(**on-line analysis**)的結果，而非批次分析(**batch analysis**)，並做出更明智的業務決策，或者對問題採取更好更快的行動。
- 拜**物聯網**時代來臨及資料蒐集工具的進步，製造業的智慧工廠、醫院的智慧醫療訊息網路、金融交易網路、交通訊息網、行動通訊人流網、環境(氣象、水質、空氣)監測網等，有助於相關行業快速識別並處理營運業務中的問題和機會。

營運智慧如何運作

- 營運智慧的前端元件：
 - 物聯網(Internet of Things, IoT)
 - 串流資料處理系統和大數據平台
 - 例如：Hadoop、Splunk和Spark。
- 營運智慧的後端元件：
 - 涉及**大量即時**資料處理與具備串流數據分析功能的應用程序。

營運智慧如何運作 (續)

- 各種資通訊 (Information and Communication Technology, ICT)供應商整合即時監控、資料串流和數據分析與即時商業智慧 (**Real-time business intelligence**)等工具，創造出專門的營運智慧平台。
- 企業仍可使用儀表板向員工提供即時營運指標、關鍵績效指標KPI和業務洞察等。
- 這些儀表板嵌入在工作系統中，通常包括資訊易於理解的資料視覺化，並具有發送警報以通知利害關係人的自動化程序，例如：當價格達/降到特定行情時而觸發的股票交易。

營運智慧運用案例

- Ken 與 Marlene Banwart 夫婦於 1991 年創立 Country Maid (West Bend, Iowa)，開始在地下室為當地農民市集製作糕點。



參考: <https://www.rockwellautomation.com/zh-tw/company/news/case-studies/country-maid-gains-production-intelligence-to-meet-growing-demand.html>

營運智慧運用案例 – 背景

- 2012 年，Country Maid 決定投資自動化設施以因應消費者對 Butter Braid® 糕點的需求。
 - 公司目前由員工全資金擁有。
 - 位於愛荷華州的設施內，乳粉、麵粉及糖等材料皆存放在多個儲存槽。
 - 材料轉移到工業攪拌機中與水混合，接著混合物進入旋轉的麵團饋送器並沿輸送管線傳送。
 - 再由類似擀麵棍的機器將麵團展開、分層，並填充奶油與餡料。
 - 在運送到冷凍櫃之前，由一名工作人員在 12 層的糕點上進行最後的潤色，讓頂層交織呈現，創造出 Butter Braid 糕點的招牌外觀。

營運智慧運用案例 – 現況

- 工作人員必須高度參與生產流程，包括手工批次作業、調整糕點大小，以及將每袋麵粉倒入攪拌機中。操作員也必須手動收集基本生產資料並製作報告，因此難免發生人為錯誤與不一致的情形，而且完全手動的製程也需要耗費大量時間與勞力。
- **Country Maid** 想為這項製程進行**自動化升級**，獲得進階的資料收集功能，確保每個批次皆維持相同的高品質，同時也希望提高全設施的整體營運效率。
- **Country Maid** 必須**增加一條生產線**以滿足需求，他們在設計攪拌室時有兩個選項：
 - 其中包括更換為較大的攪拌機，但這必須擴大設施空間以容納其尺寸，
 - 否則就必須安裝較小的新型攪拌機並設置全新的自動化生產線。
- **Country Maid** 系統整合專家 **Marc Banwart** 表示：「我們發現，若要自動化一台較大的新攪拌機，我們可以減少攪拌機的尺寸而無需擴大設施。這個選擇較不昂貴，而且可以實現相同目標。我們知道這項變化將有助於我們滿足不斷成長的生產需求，並可獲得精確資料，以利做出更完善的營運決策。」

營運智慧運用案例 – 解決方案

- Rockwell Automation 的 PlantPAx® 製程自動化系統是內建擴充能力的全廠控制系統，可提供 Country Maid 所需要的生產分析洞見。這項系統包括：
 - 視覺化、分析與報表入口網站，以及製程歷史記錄，如此可供操作員檢視生產趨勢，例如配料的平衡，以及調整配方的比例。
- PlantPAx 系統中的 FactoryTalk® VantagePoint EMI 軟體一
 - 追蹤與記錄資料以確認生產趨勢，這款軟體從整體生產線的可用來源收集資訊，使操作員可在各種以角色為基礎的儀表板上即時檢視資料。
 - 操作員使用這些資料做為探索與分析工具，依據溫度和麵團一致性等各種參數，對各個批次進行策略性的改善。

營運智慧運用案例 – 解決方案 (續)

- Country Maid 面臨麵團擴展階段的一致性問題，而且無法了解問題的成因。
 - 操作員使用 VantagePoint 軟體終於找出問題的根源：位於建築物外部的廠房麵粉倉庫可能會因為天氣變化而產生問題。
 - 操作員現在可根據控制箱內不斷變化的溫度情況來監控每個批次，這套系統還提供可能影響生產的參數檢視畫面。
- Interstates Control Systems, Inc. 首席控制系統開發人員 Raymond Berning 表示：「操作員可以檢視停機時間持續多久，以及在生產線上發生的時間與環節，如此精確的洞察能力提升了生產控制的成效。」

營運智慧運用案例 – 成果

- Country Maid 希望獲得高品質的資料與更緊密的製程控制，並且能根據各批次的資料做出更佳決策，如今他們的成果不僅於此：
 - 設施的產量倍增，也已實現公司的主要目標。
 - 他們的生產線速度提高 14%，攪拌時間縮短 23%。
 - 在實作 PlantPAx 製程系統之後，Country Maid 獲得豐富的資訊，不再只是每批次加入多少水或糖而已。現在操作員知道何時應調整批次的參數，而且知道為何應進行調整，無需再依靠假設或猜測
 - 實驗設計 <http://rwepa.blogspot.com/2021/06/design-of-experiments-with-r.html>
- 自動化流程讓同一名操作員變得更有效率，我們可以快速判斷是否需要對某個批次進行調整，而且我們真的辦到了。
- 透過實作高度自動化攪拌機而非採用大型攪拌機並改建設施，Country Maid 在設備上節省了 12 萬美元，並避免了相當長的生產停機時間。
- 他們也透過新的資料收集功能更有效地運用人力，每年可節省超過 45,000 美元的批次**人力成本**。

商業智慧 vs. 營運智慧 (1/4)

- 商業智慧(Business intelligence)簡稱為BI，表示處理與分析組織活動所產生之原始數據與各種應用程序。BI由幾個相關活動組成：
 - 從資料庫中萃取資料，轉換資料，並載入資料(Extraction-Transformation-Loading, ETL)到資料倉儲(Data Warehouse, DW)中。
 - 對資料倉儲中的歷史資料進行特定查詢和建立報表(ad hoc query and reporting)。
 - 多維線上分析處理(On-Line Analytical Processing, OLAP)。
 - 資料探勘 (Data Mining, DM)。
 - 機器學習 (Machine Learning, ML)。
 - 深度學習 (Deep Learning, DL)。

商業智慧 vs. 營運智慧 (2/4)

- BI系統對資料倉儲或資料市集中已清理(非分析目的的清理)和整合後的歷史數據進行分析，所獲結果與洞見多應用於策略規劃或營運業務中。
- BI應用程序之目的在讓公司高階主管，瞭解企業過去在收入、利潤和其他關鍵績效指標KPI等方面的狀況，以協助制定未來預算和戰略規劃。

商業智慧 vs. 營運智慧 (3/4)

- BI早期源自於主管資訊系統(Executive Information Systems, EISs)，負責建立方便主管閱讀的靜態營運報告，目前仍有組織以此為其主要應用，越來越多組織已經導入儀表板(dashboard)，期望藉此能夠掌握即時營運狀況。
- 當今BI工具雖然允許使用者創建自己的查詢，並提供諸多數據視覺化功能，不過其重點仍然是分析過去的歷史數據。

商業智慧 vs. 營運智慧 (4/4)

特性	商業智慧 BI	營運智慧 OI
功能	執行查詢並視覺化數據	監測和調查營運數據
資料	歷史數據分析	即時數據監控和警報
指標	收入、利潤和其它關鍵績效指標	營運業務的預測性指標
過程	被動反應的過程(reactive)	主動(proactive)預測應用
使用時機	BI涉及內部系統和外部來源的資料收集，以作為分析、特定查詢、報表和構建儀表板及資料視覺化的準備	OI的應用包括預測性維護，即時產品和促銷推薦，以及計算物流車輛的最佳運送路線

歷屆考題

- 1 下列哪些為營運智慧對企業經營管理的功效：(1)提升服務整體效率；
(2)協助公司開發新客源；(3)提升員工個人在統計方面的專業知識；
(4)降低產品的成本。
- (A) 1234
- (B) 123
- (C) 234
- (D) 124

歷屆考題

- 2 商業智慧(Business Intelligent)分析工具正逐漸變得越來越強大，用起來也更友善，其中何種運用可讓分析結果更快顯現？
- (A) 記憶體內運算(in-memory computing)
 - (B) 資料倉儲
 - (C) Hadoop NoSQL
 - (D) 資料探勘

歷屆考題

3

運用**人力資源商業智慧系統**進行決策可具有何種優勢？

- (A) 決策經常交付資淺員工執行
- (B) 可縮短決策與採取行動之時間差**
- (C) 資訊蒐集時間占很大比重
- (D) 以上皆是

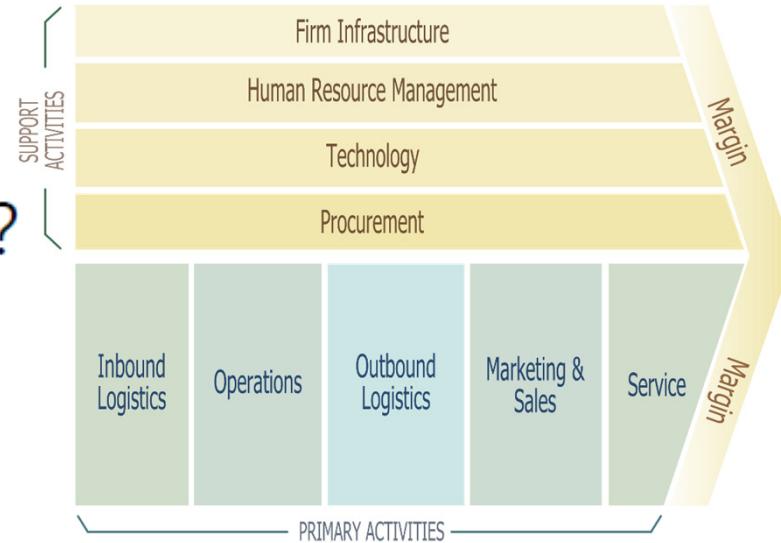
歷屆考題

4

下列何者不是價值鏈管理的重要利益？

- (A) 改善採購
- (B) 改變顧客群
- (C) 改善產品發展
- (D) 增進顧客訂單管理

參考: https://en.wikipedia.org/wiki/Value_chain



- 價值鏈（Value chain），又名價值鏈分析、價值鏈模型等。
- 麥可·波特在1985年，於《競爭優勢》一書中提出的。
- 波特指出企業要發展獨特的競爭優勢，要為其商品及服務創造更**高附加價值**，商業策略是結構企業的經營模式（流程），成為一系列的增值過程，而此一連串的**增值流程**，就是「價值鏈」。

歷屆考題

5 提高製成品的良率是商業智慧應用於生產管理的主要目的之一，請問正確的良率計算方式為何？

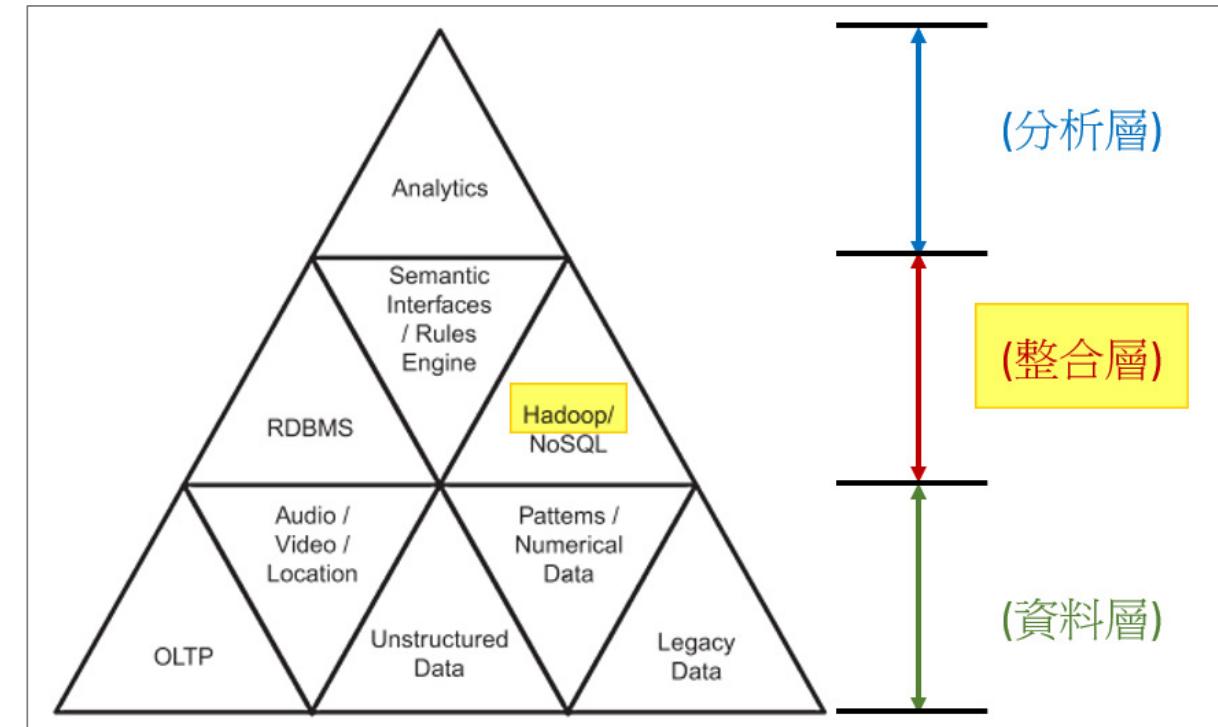
- (A) 「實際確認數量」除以「所發完工單數量」
- (B) 「實際確認數量」除以「生產數量」
- (C) 「未報廢數量」除以「實際確認數量」
- (D) 「所發完工單數量」除以「生產數量」

思考題

Hadoop在新世代資料倉儲架構中屬於哪一層級？

- (A) 分析層
- (B) 整合層
- (C) 資料層
- (D) 應用層

答案:B



★★★ 參考: <https://sites.google.com/site/shangyezhidierban/>

思考題

BI專案生命週期裡，包含了(1)系統導入階段、(2)評估階段、(3)規劃階段、(4)系統設計與建構階段、(5)專案分析階段，請問它們的先後順序為？

- (A) 12345
- (B) 23541
- (C) 23451
- (D) 23145



答案:B

思考題

關於BI的特性，下列何者為非？

- (A) BI的重點是在於以資料作為企業決策的依據
- (B) 資訊需具備即時、整合的特性
- (C) 資訊需具備多維度的特性
- (D) 透過BI系統，企業可以及時發現內部營運現況及掌握市場動態

答案：A，正確是以資訊作為企業決策的依據

思考題

在BI的規劃階段裡的企業基礎設施評估裡，可分為技術面與非技術面的基礎設施評估。請問在**技術面**的基礎設施評估裡，下列何者不包含在內？

- (A) 硬體平台
- (B) 中介軟體平台
- (C) 網路安全平台
- (D) 資料庫管理系統平台

答案：C

思考題

在BI的規劃階段裡的企業基礎設施評估裡，可分為技術面與非技術面的基礎設施評估。請問在非技術面的基礎設施評估裡，下列何者不包含在內？

- (A) 功能部門的運作
- (B) 營運活動的作業流程
- (C) 詮釋資料庫
- (D) 資源的分配

答案：D

思考題

資料分析時，應遵循以下6個步驟，請問它們的先後順序為何？(1)解決邏輯資料及來源資料的差異、(2)規劃整體公司的邏輯資料模型、(3)建立邏輯資料模型、(4)設定資料清理的步驟、(5)分析來源資料品質(6)分析外部資料。

- (A) 635214
- (B) 653241
- (C) 563142
- (D) 356124

答案：A

- 1. 分析外部資料
- 2. 建立邏輯資料模型—資料須清楚的，沒有重複，沒有不一致，與系統實作的軟硬體設備無關。
- 3. 分析來源資料品質
- 4. 規劃整體公司的邏輯資料模型
- 5. 解決邏輯資料及來源資料的差異
- 6. 設定資料清理的步驟

思考題

在規劃ETL時，首先要決定有哪些資料需要移轉，不同資料移轉的決定，攸關 ETL程式的準備，請問下列何種**移轉方式為非**？

- (A) 初始載入
- (B) 歷史載入
- (C) 隨機載入
- (D) 差異載入

答案：C

思考題

關於詮釋資料資料庫設計需包括的部份，下列何者為非？

- (A) 建構詮釋資料資料庫
- (B) 建構雛形系統
- (C) 詮釋資料資料庫工具介面
- (D) 詮釋資料搬移(Migration)流程

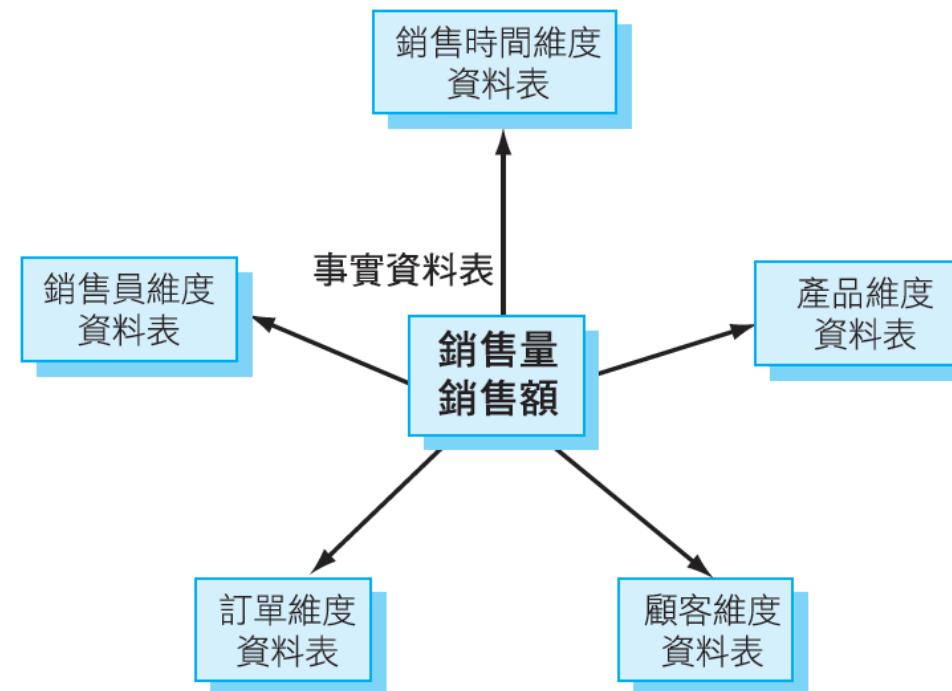
答案：B

資料倉儲-維度模型 (Dimensional Model)

- 別稱—**星狀綱要(Star Schema)**
- 模型結構
 - 中央核心：**事實資料表 (Fact table)** → 加總計算
 - 環繞核心的光芒：**維度資料表 (Dimension table)** — 群組計算 → Excel樞紐分析
 - 此模型不建議以**實體關係體圖(Entity Relationship Diagrams, ERD)**來呈現
- **事實資料表(Fact table)**
 - 只有一個事實資料表
 - 可衡量(Measurement)數值績效統稱事實(Facts)
- 環繞核心的光芒：**維度資料表(Dimension table)**
 - 允許數個維度資料表
 - 每一道光芒代表決策者觀察績效的角度
 - 因此表格內所儲存的資料就是主管查看企業營運績效時所要的觀查的角度特色

維度模型範例

- 某企業經營績效的資料有銷售量與銷售額為二個事實資料表。
- 企業主有興趣查看經營績效的角度有五個，分別為銷售時間、產品、客戶、訂單以及銷售員等五個維度資料表。



思考題

目前BI系統設計以用簡易的視覺化拖曳(Drag & Drop)方式來進行資料選取，下列有關BI系統的描述何者正確？

- (A)從**維度資料表**內挑選出來的欄位，BI系統都會事先針對這些欄位內容值相同者進行加總(Sum)的計算
- (B)從**事實資料表**中挑選出來的數值性內容的欄位，都會依據同一群組的資料值進行群組(Group By)的動作
- (C)從維度資料表內挑選出來的欄位，BI系統都會事先針對這些欄位內容值**不相同者**進行加總(Sum)的動作
- (D)從維度資料表內挑選出來的欄位，BI系統都會事先針對這些欄位內容值**相同者**進行群組(Group By)的動作

答案:D

思考題

下列有關**實體關係模型(Entity Relationship Model, ERM)**不適合BI系統設計的描述何者錯誤？

- (A) ERM內資料需正規化(Normalization)技術分析過
- (B) ERM太過細碎因此不適合非IT人員閱讀
- (C) ERM查詢時候需作很多速度慢的合併(Join) 運算
- (D) 以上皆錯

答案:D

思考題

下列哪些描述選項是維度模型化的優點 (1) 容易瞭解(Understandability)
(2) 查詢運算快(Query Performance) (3) 模型中每一個維度資料表的重要性都相同 (4) 可以容納非預期中的新資料(Accommodate Unexpected New Data) (5) 只需要進行到第三正規型關聯表 (6) 不需要反正規化技術

- (A)1345
- (B)1234
- (C)1236
- (D)1246

答案:B

思考題

維度模型中事實資料表的基本結構包含下列哪三個部份 (1) 擁有一組對應聯結到維度模型中各個維度資料表的外來鍵(Foreign Keys, FK) (2) 會有一個或者多個數值型態的欄位用來儲存事實資料 (3) 可能包含一個或者多個退化維度(Degenerate Dimensions, DD)的欄位 (4) 具有唯一性的超級鍵(Super Keys) (5) 具有一種RSA特性的加密欄位來確保資訊安全

- (A)234
- (B)145
- (C)345
- (D)123

答案:D

思考題

有關下列維度模型中事實資料表的顆粒度(Fact Table Granularity)描述何者錯誤 (1) 顆粒度是描述事實資料表中衡量欄位的精細度 (2) 同一張事實資料表中所有衡量欄位可以有不相同的精細度 (3) 如果顆粒度是指一天的營業額，則銷售量與銷售金額就是一整天訂單中的銷售量與銷售金額的加總 (4) 如果顆粒度是訂單，則銷售金額與銷售量就是加總一整張訂單上的列項目所得的銷售量與銷售金額 (5) 同一張事實資料表中所有衡量欄位都要有相同的精細度 (6) 如果顆粒度是訂單中列項目(Line Item)，則銷售數量與銷售金額都是指某個品項在某張訂單中的銷售數字 (7) 儲存顆粒的大小的設定工作通常會是使用BI系統的決策者在訪談與討論前負責設定

(A)14

(B)35

(C)27

(D)56

答案:C

(2) 同一張事實資料表中所有衡量欄位必須有相同的精細度
(7) 儲存顆粒的大小的設定工作通常會是使用BI系統的決策者在訪談與討論後負責設定

線上即時報表功能

- **Roll-up 向上匯總:** Roll-up 動作會讓資料根據所規範的屬性及其階層做匯總。
- **Drill-down 向下展開:** Drill-down 則是針對資料做細部展開，以得到更詳細的資訊。
- **Slice-and-dice 交叉剖析:** Slicing 和 Dicing 將多維度模型中的資料，做進一步的**資料限制**，以讓報表中只呈現出使用者所想要的資料。Slicing 的動作是指將某一個屬性的值規範在某一個值或某一個範圍，以用來過濾多維度模型中的資料。Dicing 的動作是指我們透過一個可能包含多個屬性的條件，只取出多維度模型中資料某一個區塊的動作。
- **Pivot 樞紐分析:** Pivot 是一個會改變報表排列的動作，原本的主要分析維度和陪襯的維度會在 Pivot 動作中被對調，整個報表的重點也可能會有所改變。
- **Drill-across:** Drill-across 動作指的是我們要將**兩個或兩個以上的多維度資料模型**建立關係，以用來比較兩個不同模型裡面的資料。

資料倉儲之資料集結 (Data Staging)

- 資料倉儲架構包括前端與後端：
 - 後端負責準備資料，亦稱為資料管理(Data management)。
 - 前端負責應用資料，亦稱為資料存取(Data access)。
- 資料集結：
 - 後端會進行 ETL 的各項步驟(抽取、清理、一致化與交付)將所產生的資料進行儲存。
 - 建議每一個 ETL 步驟都需要進行資料的集結。
- 資料集結區(Data staging area)包括：
 - 持久集結區 (Persistent staging area)：維護歷史資料而使用的集結區。
 - 臨時集結區(Temporary staging area)：資料在每次載入過程後即被刪除。

思考題

下列哪一項線上即時報表功能會有改變報表排列的動作？

- (A) Roll-up
- (B) Drill-down
- (C) Slice-and-dice
- (D) Pivot

答案:D

思考題

單一單位的「績效趨勢綜合看板」的重點在顯示每一個選取單位的營運績效的什麼內涵？

- (A) 變化
- (B) 趨勢變化
- (C) 大小
- (D) 百分比

答案:(B)

思考題

資料倉儲的資料建置規劃步驟中，一切圍繞著哪一項目？

- (A) 人員
- (B) 需求
- (C) 經費
- (D) 系統

答案:B

思考題

ETL各項步驟(抽取、清理、一致化與交付)於資料倉儲架構中何處進行？

- (A) 來源系統
- (B) 展示區
- (C) 資料集結區
- (D) 前端

答案:C

思考題

將儲存在資訊系統中龐大的生產或交易相關的原始資料，透過分析以轉成有用的資訊或知識，以提供企業內部各層級管理者，進行即時決策之參考依據，稱為：

- (A) 商業智慧
- (B) 商業知識
- (C) 資料倉儲
- (D) 資料探勘

答案: A

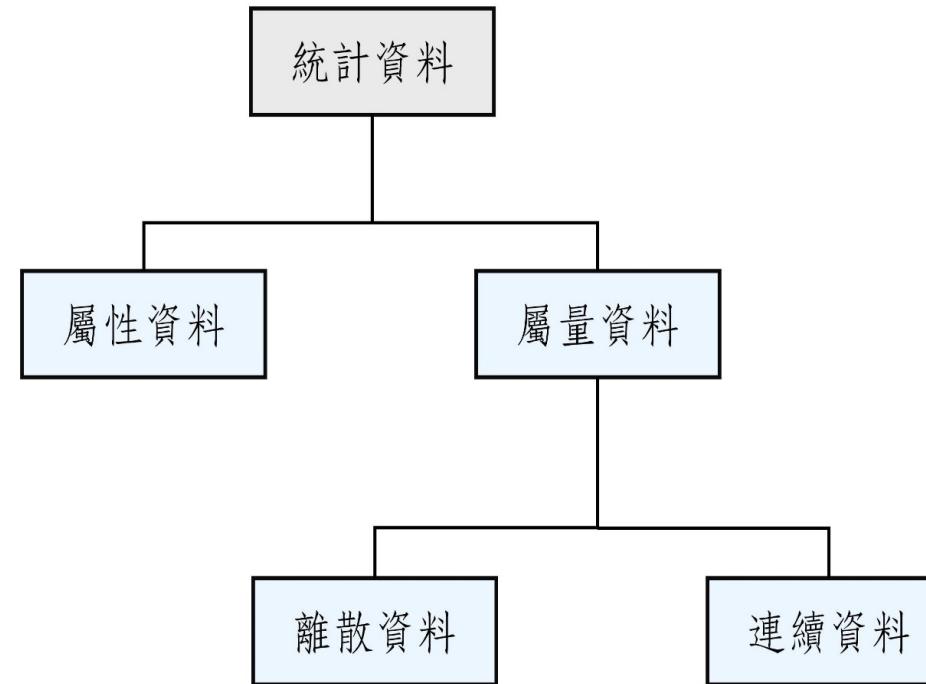
2.L112基礎資料分析

- L11201資料來源與資料獲取(含資安)
- L11202資料性質(例如：結構性與非結構性)
- L11203常用統計概念及其資料前處理

資料型態

1. 屬性資料(**qualitative data**)：不可以用數值來表示，僅以類別來區分的資料，又稱為「類別資料」(**categorical data**)。
 - 例：性別、學歷、國籍、汽車品牌為類別資料。
2. 屬量資料(**quantitative data**)：可以用數值來衡量的資料。
 - (1)離散資料(**discrete data**)：屬於可數的屬量資料，此類型資料之資料值均為離散之數值，在任兩個數值間不可能插入無限多個數值資料。
 - (2)連續資料(**continuous data**)：屬於不可數的屬量資料，此類型資料之資料值為連續之數值，在任兩個數值間可插入無限多個數值的資料。

資料型態 (續)



測量尺度 (scale of measure)

- 測量尺度是統計學和定量研究中，對不同種類的資料，依據其尺度水平區分成四大類別：
 1. 名目 (nominal)，例：{有, 沒有}, {Apple, Google, FB}
 2. 次序 (ordinal)，例：{普通, 好, 優}
 3. 等距 (interval)，例：溫度，年份
 4. 等比 (ratio)，例：絕對溫度，銷售量

參考: https://en.wikipedia.org/wiki/Level_of_measurement

屬性資料 – 表格法

- 表格法 (tabular method)：將蒐集的資料整理成表格的形式。
 1. 次數分配表(frequency distribution table)：
將資料依其類別分成若干組，並計算各組的資料個數。
 2. 相對次數分配表(relative frequency distribution table)：
將資料依其類別分成若干組，並計算各組的資料個數比例。
- 圖示法 (graphic method)：將蒐集的資料以圖形的方式呈現。
 - 長條圖 (bar graph)
 - 圓形圖 (pie chart)

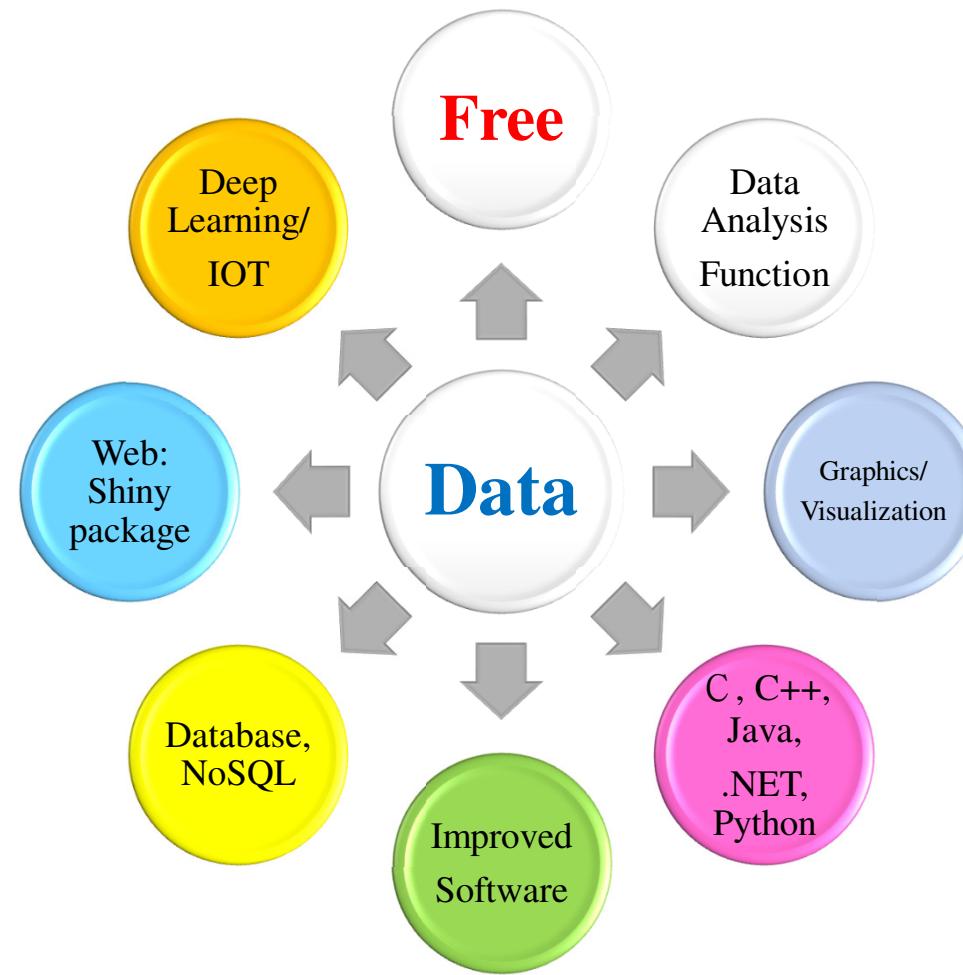
R 簡介

認識 R

- 1976 - 貝爾實驗室 John Chambers, Rick Becker, and Allan Wilks 研發 S 語言。
- 1993 - Ross Ihaka and Robert Gentleman, University of Auckland, New Zealand 研發 R 語言。
 - R 是一種基於 S 語言所發展出具備統計分析、繪圖與資料視覺化的程式語言。
- 1997年—R的核心開發團隊 (R development core team) 成立，專責R原始碼的修改與編寫。
 - 2000年2月 — R 1.0.0
 - 2013年3月 — R 2.15.3
 - 2021年5月 — R 4.1.0



R-八大功能



R-下載

- 官網: <http://www.r-project.org/>
- 選取左側 Download \ CRAN
- 選取 Taiwan CRAN

Taiwan

<https://cran.csie.ntu.edu.tw/>

- 選取 Download R for Windows

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

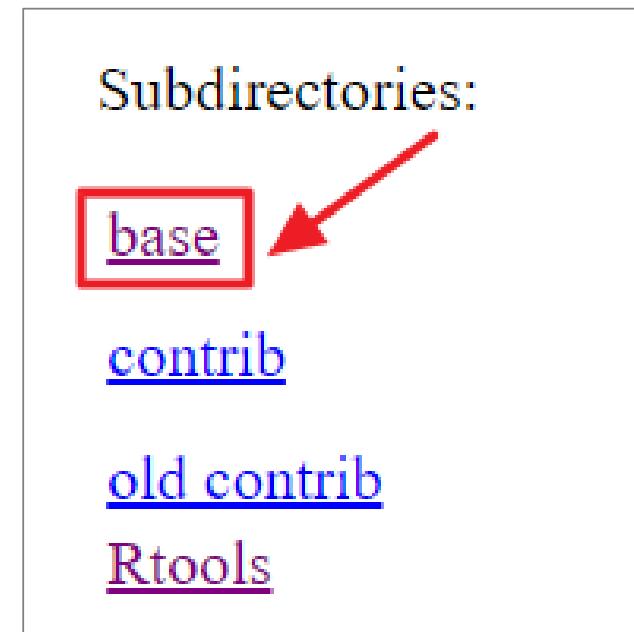
RWEPA 搜尋此網誌 (例: task)

- GitHub DataDemo
- iPAS-R-tutorial
- iPAS-Python-tutorial
- ★★★R入門資料分析與視覺化(付費,中文字幕)
- ★★★R商業預測與應用(付費,中文字幕)
- R教學-基礎篇/程式碼(免費)
- Python程式設計PDF(免費)
- ★R 4.1.0-Wndows下載
- ★RStudio-1.4.1106下載



R-下載 (續)

- 選取 base → 下載 [R-4.1.0-win.exe]

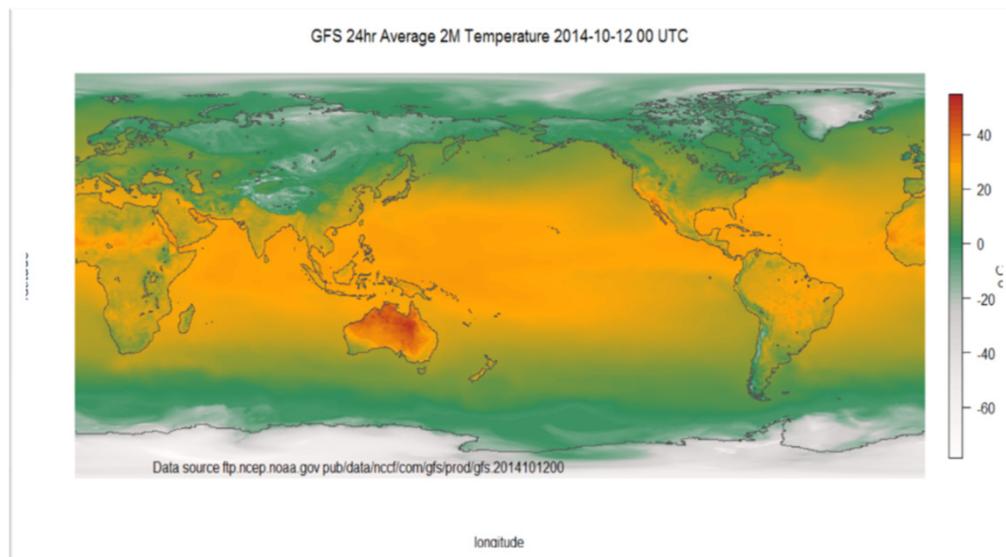


- R安裝路徑: 保留原路徑,不要修改
- https://github.com/rwepa/DataDemo/blob/master/windows_intall_R.pdf

RStudio 簡介

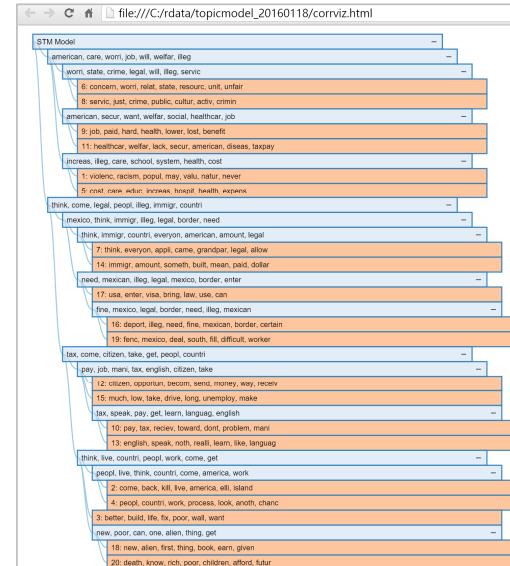
整合式開發環境 - RStudio

- <http://www.rstudio.com/>



視覺化應用

(全球2M氣溫圖)



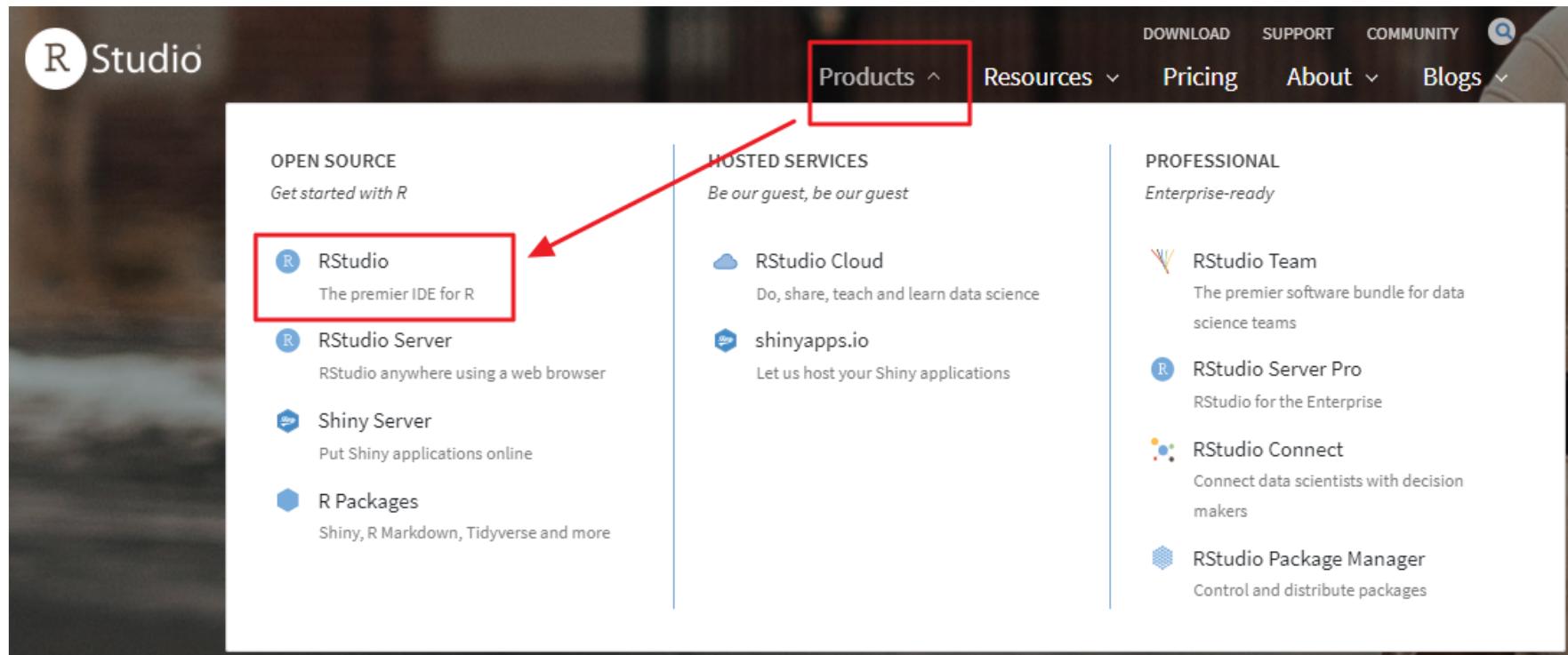
主題模型

RStudio - 特性

- 支援智慧輸入 (按Tab)
- 高亮度顯示程式碼
- 整合R程式, 控制台, 變數清單, 繪圖視窗
- 整合資料庫匯入 SQL, Spark
- 整合R套件: shiny, rmarkdown
- 安裝注意:
 - 先安裝R, 再安裝 RStudio
 - 安裝 RStudio時, 請先關閉R

RStudio 下載

- <http://www.rstudio.com/>



RStudio 下載 (續)

RStudio Desktop

Open Source License

Free

RStudio Desktop

Commercial License

\$995**單機版****DOWNLOAD****BUY**

RStudio Server

Open Source License

Free

RStudio Server Pro

Commercial License

\$4,975**伺服器版本**/year
(Named Users)**Learn more****Learn more****Learn more****BUY****免費版**

Integrated Tools for R



Priority Support



Access via Web Browser



Enterprise Security



Project Sharing

DOWNLOAD**BUY****DOWNLOAD****BUY****Learn more****Learn more****Learn more****Evaluation | Learn more****免費版****Learn more****Learn more****Learn more****Evaluation | Learn more****Learn more****Learn more****Learn more****Evaluation | Learn more**

RStudio 下載 (續)

RStudio Desktop 1.4.1717 - [Release Notes](#)

- 1.** Install R. RStudio requires [R 3.0.1+](#).

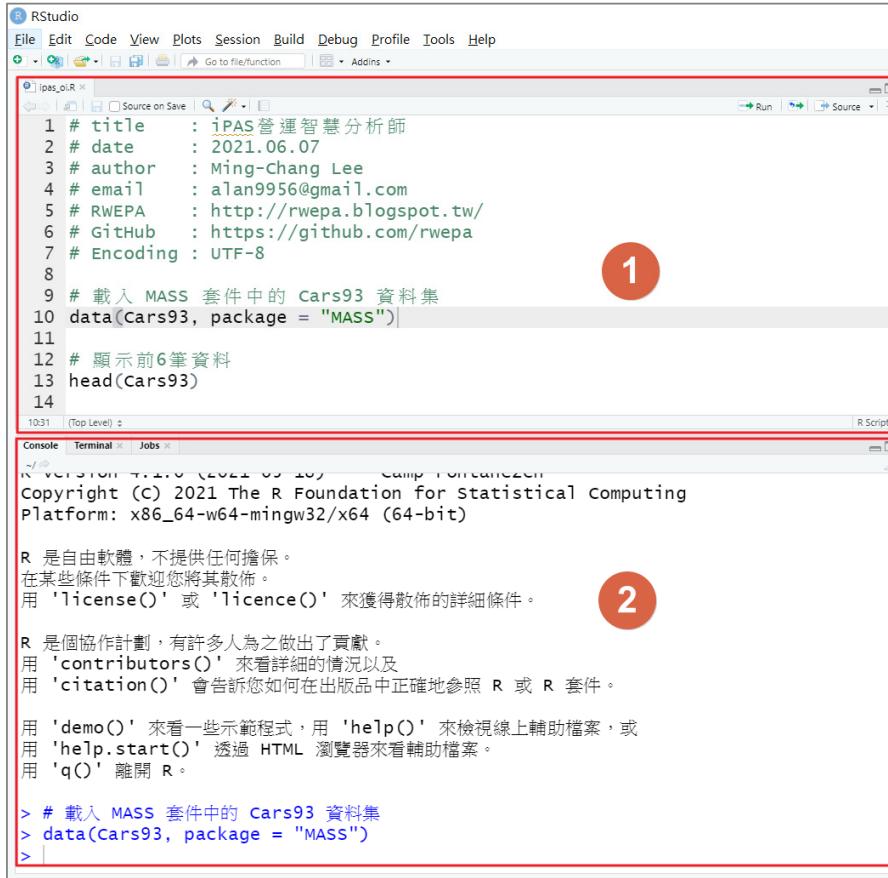
- 2.** Download RStudio Desktop. Recommended for your system:



Requires Windows 10 (64-bit)

RStudio 開啟畫面

程式碼



1 # title : iPAS營運智慧分析師
 2 # date : 2021.06.07
 3 # author : Ming-chang Lee
 4 # email : alan9956@gmail.com
 5 # RWEPA : http://rwepa.blogspot.tw/
 6 # GitHub : https://github.com/rwepa
 7 # Encoding : UTF-8
 8
 9 # 載入 MASS 套件中的 Cars93 資料集
 10 data(Cars93, package = "MASS")
 11
 12 # 顯示前6筆資料
 13 head(Cars93)
 14

1031 [Top Level] R Script

Console Terminal Jobs

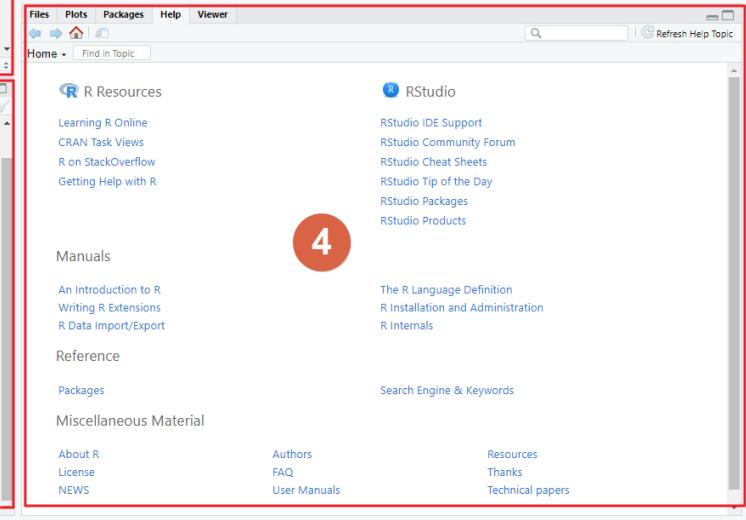
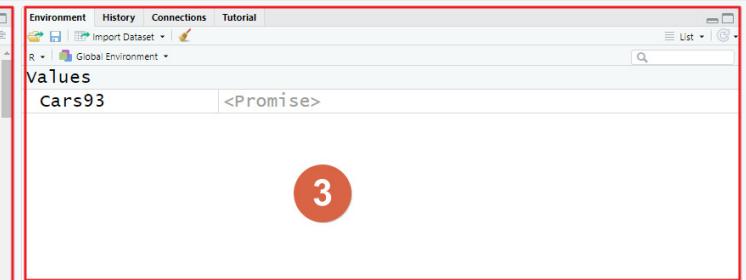
```
~/ ~ Version 4.1.0 (2021-05-10) -- "Camp Farnsworth"
Copyright (C) 2021 The R Foundation for statistical computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R 是自由軟體，不提供任何擔保。
在某些條件下歡迎您將其散佈。
用 'license()' 或 'licence()' 來獲得散佈的詳細條件。
```

2 R 是個協作計劃，有許多人為之做出了貢獻。
 用 'contributors()' 來看詳細的情況以及
 用 'citation()' 會告訴您如何在出版品中正確地參照 R 或 R 套件。

用 'demo()' 來看一些示範程式，用 'help()' 來檢視線上輔助檔案，或
 用 'help.start()' 透過 HTML 瀏覽器來看輔助檔案。
 用 'q()' 離開 R。

```
> # 載入 MASS 套件中的 cars93 資料集
> data(cars93, package = "MASS")
>
```



主控台

變數

繪圖, 說明

R - Cars93 {MASS} 範例

```
> # 載入 MASS 套件中的 Cars93 資料集
> data(cars93, package = "MASS")
>
> # 顯示前6筆資料
> head(cars93)
```

	Manufacturer	Model	Type	Min.Price	Price	Max.Price	MPG.city	MPG.highway	AirBags	
1	Acura	Integra	Small	12.9	15.9	18.8	25	31	None	
2	Acura	Legend	Midsize	29.2	33.9	38.7	18	25	Driver & Passenger	
3	Audi	90	Compact	25.9	29.1	32.3	20	26	Driver only	
4	Audi	100	Midsize	30.8	37.7	44.6	19	26	Driver & Passenger	
5	BMW	535i	Midsize	23.7	30.0	36.2	22	30	Driver only	
6	Buick	Century	Midsize	14.2	15.7	17.3	22	31	Driver only	
	DriveTrain	Cylinders	EngineSize	Horsepower	RPM	Rev.per.mile	Man.trans.avail	Fuel.tank.capacity		
1	Front	4	1.8	140	6300	2890	Yes	13.2		
2	Front	6	3.2	200	5500	2335	Yes	18.0		
3	Front	6	2.8	172	5500	2280	Yes	16.9		
4	Front	6	2.8	172	5500	2535	Yes	21.1		
5	Rear	4	3.5	208	5700	2545	Yes	21.1		
6	Front	4	2.2	110	5200	2565	No	16.4		
	Passengers	Length	Wheelbase	Width	Turn.circle	Rear.seat.room	Luggage.room	Weight	origin	Make
1	5	177	102	68	37	26.5	11	2705	non-USA	Acura Integra
2	5	195	115	71	38	30.0	15	3560	non-USA	Acura Legend
3	5	180	102	67	37	28.0	14	3375	non-USA	Audi 90
4	6	193	106	70	37	31.0	17	3405	non-USA	Audi 100
5	4	186	109	69	39	27.0	13	3640	non-USA	BMW 535i
6	6	189	105	69	41	28.0	16	2880	USA	Buick Century

R程式僅供參考學習

資料結構 - str

```
> # 資料結構  
> str(cars93)  
'data.frame': 93 obs. of 27 variables:  
 $ Manufacturer : Factor w/ 32 levels "Acura", "Audi", ... : 1 1 2 2 3 4 4 4  
 $ Model        : Factor w/ 93 levels "100", "190E", "240", ... : 49 56 9 1 6  
 $ Type          : Factor w/ 6 levels "Compact", "Large", ... : 4 3 1 3 3 3 2  
 $ Min.Price     : num 12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 .  
 $ Price         : num 15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 .  
 $ Max.Price     : num 18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3  
 $ MPG.city      : int 25 18 20 19 22 22 19 16 19 16 ...  
 $ MPG.highway   : int 31 25 26 26 30 31 28 25 27 25 ...  
 $ AirBags       : Factor w/ 3 levels "Driver & Passenger", ... : 3 1 2 1 2  
 $ DriveTrain    : Factor w/ 3 levels "4WD", "Front", ... : 2 2 2 2 3 2 2 3 2
```

93筆資料, 27個變數

資料框

- 資料 factor – 因子, 類別型變數 {名目, 次序}
- 資料 num – 數值 numeric
- 資料 int – 整數 integer

資料摘要 - summary

> # 資料摘要-敘述統計

> **summary(cars93)**

Manufacturer	Model	Type	Min.Price	Price	Max.Price	MPG.city
Chevrolet : 8	100 : 1	Compact:16	Min. : 6.70	Min. : 7.40	Min. : 7.9	Min. :15.00
Ford : 8	190E : 1	Large :11	1st Qu.:10.80	1st Qu.:12.20	1st Qu.:14.7	1st Qu.:18.00
Dodge : 6	240 : 1	Midsize:22	Median :14.70	Median :17.70	Median :19.6	Median :21.00
Mazda : 5	300E : 1	Small :21	Mean :17.13	Mean :19.51	Mean :21.9	Mean :22.37
Pontiac : 5	323 : 1	Sporty :14	3rd Qu.:20.30	3rd Qu.:23.30	3rd Qu.:25.3	3rd Qu.:25.00
Buick : 4	535i : 1	Van : 9	Max. :45.40	Max. :61.90	Max. :80.0	Max. :46.00
(Other) :57	(Other):87					
MPG.highway	AirBags	Driver & Passenger	Driver only	Luggage.room	Weight	
Min. :20.00	Driver & Passenger:16			Min. : 6.00	Min. : 1695	
1st Qu.:26.00	Driver only :43			1st Qu.:12.00	1st Qu.:2620	
Median :28.00	None :34			Median :14.00	Median :3040	
Mean :29.09				Mean :13.89	Mean :3073	
3rd Qu.:31.00				3rd Qu.:15.00	3rd Qu.:3525	
Max. :50.00				Max. :22.00	Max. :4105	

- 最小值 Min
- 25百分位數 1st Qu. Q_1
- 中位數 Median Q_2
- 75百分位數 3rd Qu. Q_3
- 最大值 Max.
- 遺漏值 NA's (Not Available)

NA's :11

次數分配表 - table

```
> # 次數分配表  
> table(cars93$type)
```

Compact	Large	Midsize	Small	Sporty	Van
16	11	22	21	14	9

個數最多: Midsize

```
>  
> # 相對次數分配表  
> prop.table(table(cars93$type))
```

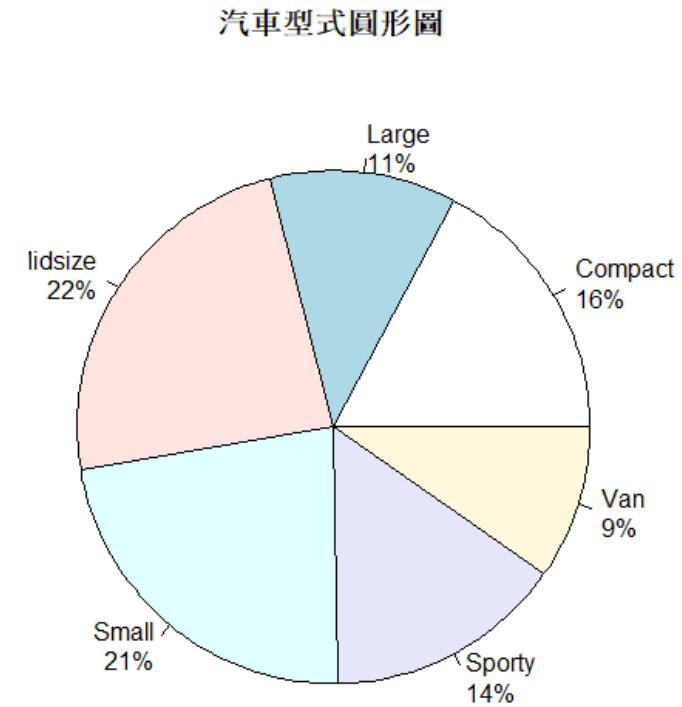
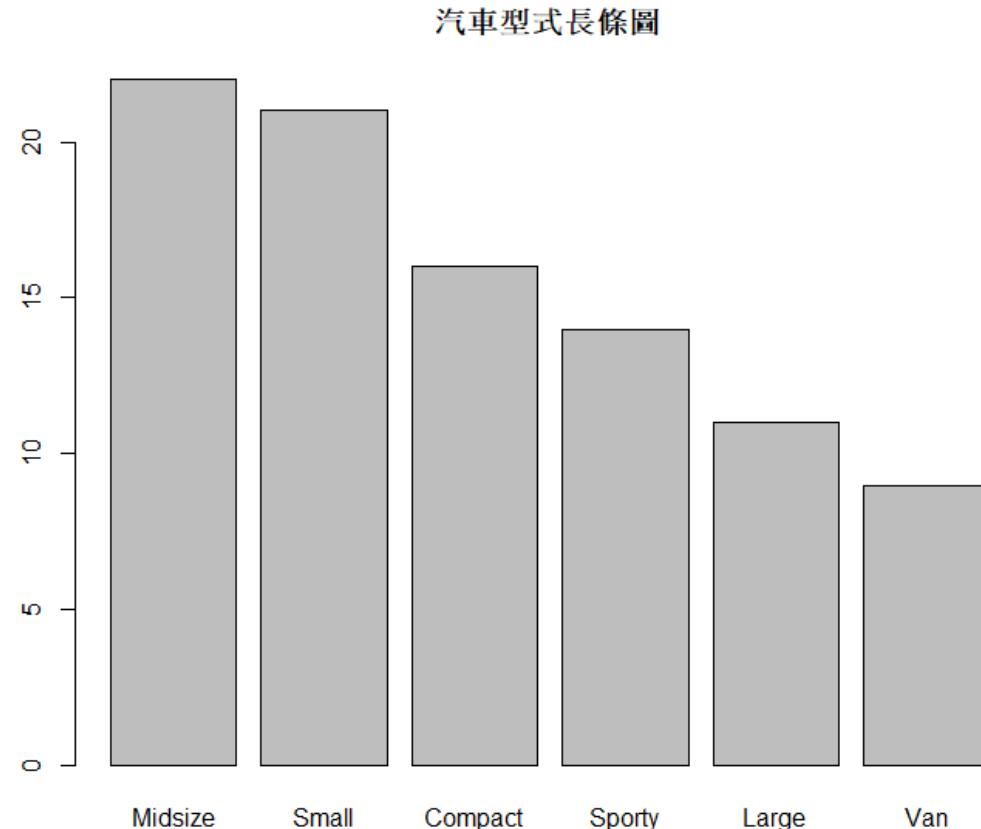
Compact	Large	Midsize	Small	Sporty	Van
0.17204301	0.11827957	0.23655914	0.22580645	0.15053763	0.09677419

```
>  
> # 相對次數分配表(%)  
> round(prop.table(table(cars93$type))*100, 2)
```

Compact	Large	Midsize	Small	Sporty	Van
17.20	11.83	23.66	22.58	15.05	9.68

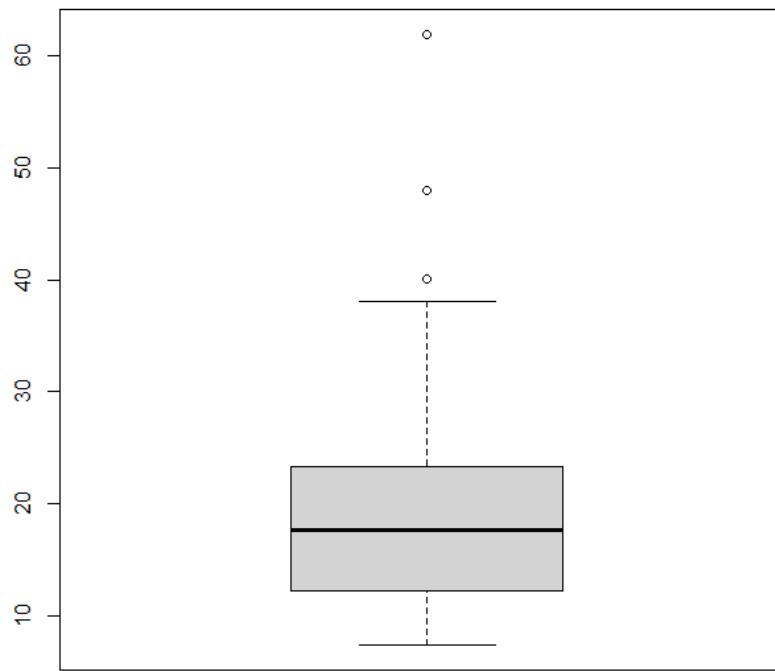
```
>
```

長條圖 - barplot, 圓形圖 - pie



R demo

盒鬚圖 boxplot



```
> # 盒鬚圖 boxplot  
> data(Cars93, package = "MASS")  
> boxplot(Cars93$Price)  
> Cars93_Price <- boxplot(Cars93$Price)  
> Cars93_Price
```

\$stats

[1,]	7.4	下限
[2,]	12.2	Q1 (25百分位數)
[3,]	17.7	- Q2 (50百分位數, 中位數)
[4,]	23.3	Q3 (75百分位數)
[5,]	38.0	上限

\$n
[1] 93

\$conf

[1,]	15.88139
[2,]	19.51861

\$out

[1] 40.1 47.9 61.9

\$group

[1] 1 1 1

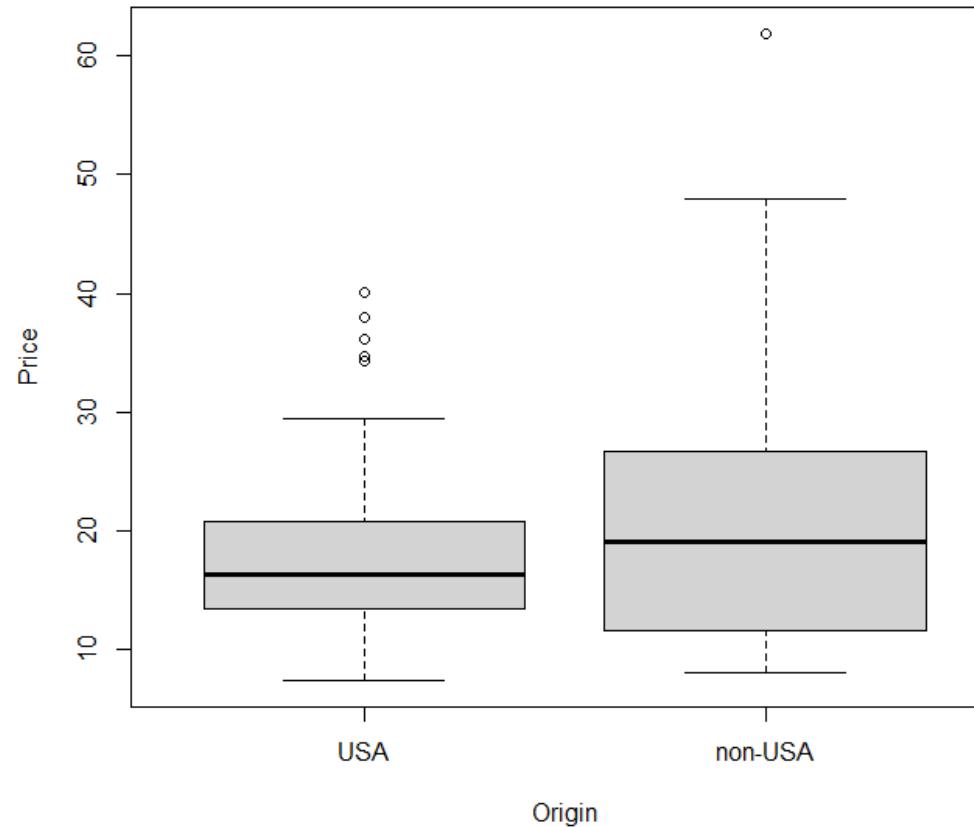
\$names

[1] "1"

上限離群值: $x > Q_3 + (Q_3 - Q_1) \times 1.5$

下限離群值: $x < Q_1 - (Q_3 - Q_1) \times 1.5$

群組盒鬚圖



- 二個群組中位數不相同
- 二個群組離群值有不同

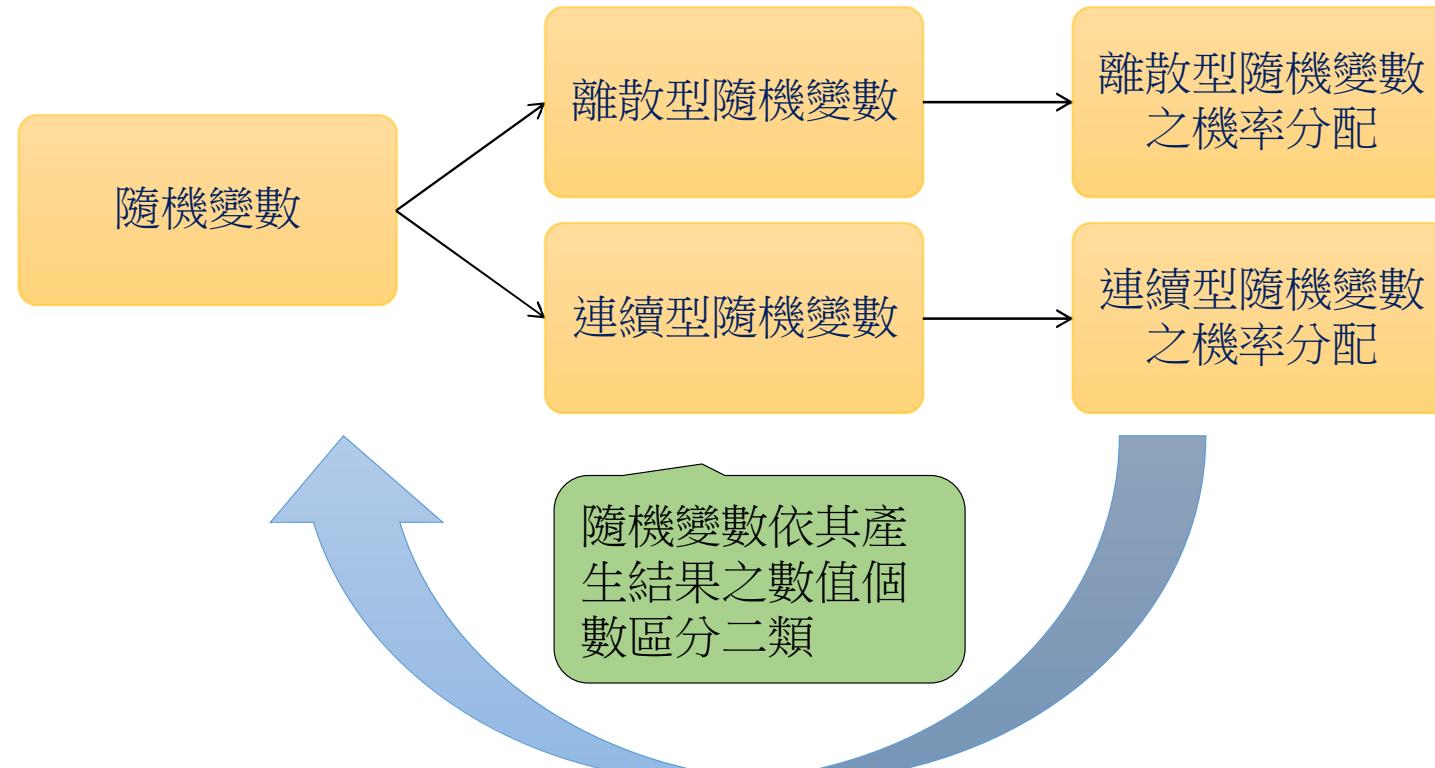
隨機變數

- 隨機變數 (random variable, 簡稱r.v.) 是一個實數值函數，定義於樣本空間的每一個元素皆可對應至一個實數。
 - 一般以大寫英文字母表示隨機變數，例： X, Y, Z 。
 - 以小寫英文字母表示對應的函數值，例： x, y, z 。
 - 例：隨機變數 X 表示公平骰子投擲一次出現的點數，則
$$P(X = 1) = f(x = 1) = \frac{1}{6}$$
。
- 隨機變數依其產生結果之數值個數可區分為兩種型態：
 - 離散型隨機變數
 - 連續型隨機變數。

機率分配

random variable

probability distribution



離散型隨機變數之機率分配

離散型隨機變數 X 之機率函數 $f(x)$ 必須滿足下列三個條件：

(1). $f(x_i) = P(X = x_i), i = 1, 2, \dots, n$ 。

(2). $0 \leq f(x_i) \leq 1, \forall x_i \in R$ 。

(3). $\sum_{i=1}^n f(x_i) = 1$ 。

連續型隨機變數之機率分配

如果函數 $f(x)$ 滿足下列三個條件，則 $f(x)$ 稱為連續型隨機變數 X 之機率密度函數：

(1). 如果 $x \in R$ ，則 $f(x) \geq 0$ ，

(2). $P(a \leq x \leq b) = \int_a^b f(x)dx$ ，

(3). $\int_{-\infty}^{\infty} f(x)dx = 1$ 。

統計量

- 集中趨勢測量

- 算數平均數
Arithmetic Mean

- 中位數
Median

- 眾數
Mode

- 資料離散程度

- 全距
Range
- 變異數
Variance
- 標準差
Standard Deviation
- 四分位距
(Interquartile range)
 $IQR = Q_3 - Q_1$

期望值 Expected Value

- 若 $f(x)$ 為隨機變數 X 之機率(密度)函數，則隨機變數 X 的平均值或期望值以 μ 或 $E(X)$ 表示，定義如下：

離散型 :
$$\mu = E(X) = \sum_{i=1}^n x_i f(x_i)$$

連續型 :
$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

變異數 Variance

- 母體變異數(Population Variance)
- 母體變異數是計算各數值與平均數差距，取平方後的算術平均數。
- 特性：
 - 在計算中使用了所有的數值。
 - 不易受極端值所影響。
 - 單位的意義不容易解釋，其為原來單位的平方。
 - 變異數不會是負數。變異數可以等於零(實務較少)。

變異數(續)

- 母體變異數(Population Variance)

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

- 樣本變異數(sample variance)

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- 標準差: 變異數的平方根

估計 (estimation)

樣本 → 母體

樣本統計量估計母體參數

- 點估計(Point estimation)
 - 利用樣本資料求得一個樣本統計量來推估母體參數的統計方法。
 - 若母體參數以 θ 表示，則以 $\hat{\theta}$ 表示用來估計母體參數之樣本統計量。
 - 此樣本統計量稱為 θ 的點估計量。
- 區間估計(Interval estimation)
 - 利用點估計量 $\hat{\theta}$ 建構一區間 $[\hat{\theta}_L, \hat{\theta}_U]$ 來推估母體參數 θ ，使得 θ 包含於 $[\hat{\theta}_L, \hat{\theta}_U]$ 為一特定之機率值。
 - $\hat{\theta}_L$ 表示估計下限 (Lower bound)， $\hat{\theta}_U$ 表示估計上限 (Upper bound)。
 - 若 $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$ ，則稱 $[\hat{\theta}_L, \hat{\theta}_U]$ 為參數 θ 的 $(1 - \alpha) \times 100\%$ 信賴區間。

參考資料：賀力行、林淑萍、蔡明春，統計學：觀念、方法、應用，前程文化，2008年。

平均數 μ 之區間估計 - 母體變異數已知

- 考慮常態母體且 σ^2 已知，樣本數 n ，樣本平均數 \bar{x} ，則母體平均數 μ 之 $(1 - \alpha) \times 100\%$ 信賴區間為

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \right]$$

- 估計誤差

$$z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

區間估計 – 範例

- 考慮某產品的重量呈現常態分配，依過去生產經驗得知產品之重量的標準差為5公克。今隨機抽取16件產品作檢查，發現其平均重量為60公克，請問：
 - 此產品平均重量之95%信賴區間為何？
 - 此產品平均重量之99%信賴區間為何？



區間估計 – 範例 (續)

- 95%之信賴區間：

- $\alpha = 0.05$ ，查詢標準常態分配之累積機率值，可知 $z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} = 1.96$ ，

- 因此95%之信賴區間為

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \right] = \left[60 - 1.96 \times \frac{5}{\sqrt{16}}, \bar{x} + 1.96 \times \frac{5}{\sqrt{16}} \right] = [57.55, 62.45]$$

- 99%之信賴區間：

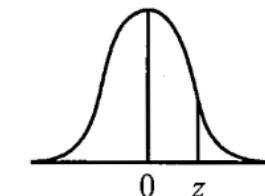
- $\alpha = 0.01$ ，查詢標準常態分配之累積機率值，可知 $z_{\frac{\alpha}{2}} = z_{\frac{0.01}{2}} = z_{0.005} = 2.575$ ，因此
99%之信賴區間為

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \right] = \left[60 - 2.575 \times \frac{5}{\sqrt{16}}, \bar{x} + 2.575 \times \frac{5}{\sqrt{16}} \right] = [56.78, 63.22]$$

參考: <http://rwepa.blogspot.com/2017/03/pnorm-qnorm.html>

附表三：標準常態分配之累積機率值

$$P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$



- 查詢 $0.05/2=0.025$
→ $1-0.025=0.975$
- 查詢 $P(Z \leq z) = 0.975$
- 查表 $z = 1.96$

- 橫列對應 1.9
- 直行對應 0.06
- p 值 → 查詢z值

標準常態分配之累積機率值(續)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

平均數 μ 之區間估計 - 母體變異數未知

- 考慮常態母體且 σ^2 未知，樣本數 n ，樣本平均數 \bar{x} ，樣本標準差 s ，則母體平均數 μ 之 $(1 - \alpha) \times 100\%$ 信賴區間為

$$\left[\bar{x} - t_{\frac{\alpha}{2}(n-1)} \times \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}(n-1)} \times \frac{s}{\sqrt{n}} \right]$$

- 估計誤差

$$t_{\frac{\alpha}{2}(n-1)} \times \frac{s}{\sqrt{n}}$$

檢定

單一樣本 T 檢定

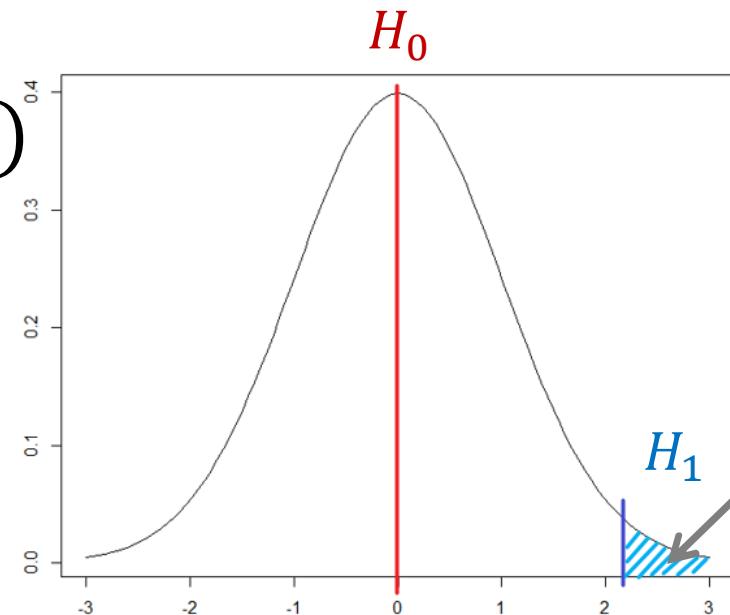
$$H_0: \mu = \mu_0$$

虛無假設 null hypothesis

$$H_1: \mu \neq \mu_0$$

對立假設 alternative hypothesis

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n - 1)$$
$$\bar{x} \rightarrow \bar{X}, s \rightarrow S$$



- p 值小, 拒絕 H_0
- t 統計值大, 拒絕 H_0

曲線下面積稱為 p 值

假設檢定

- 假設檢定的意義：首先對母體特性作一假設描述，然後利用抽樣得到的樣本資料，來驗證是否支持此假設。因此假設檢定即為利用所抽出的樣本，對母體的特性作推論。
- 建立假設：
 - 虛無假設 (null hypothesis)：對母體的未知參數所作之假設，以 H_0 表示之。
 - 對立假設 (alternative hypothesis)：與虛無假設相反之假設，以 H_1 表示之。

型 I 誤差, 型 II 誤差

- 假設檢定的兩種誤差
 - 型 I 誤差：當 H_0 為真，檢定結果為拒絕 H_0 ，所犯的錯誤，以 α 表示。
 - 型 II 誤差：當 H_0 為誤，檢定結果為接受 H_0 ，所犯的錯誤，以 β 表示。
- 發生型 I 誤差之最大機率，稱為顯著水準 (significance level)，記為 α

		真實狀況	
		H_0 真	H_0 偽
檢定結果	接受 H_0 (不拒絕)	正確的決策	(發生機率 β , 偽陰性) 犯型 II 誤差
	拒絕 H_0	犯型 I 誤差 (發生機率 α , 偽陽性)	正確的決策 (發生機率 $1 - \beta$, 檢定力 power)

假設檢定的三種類型

- 左尾檢定 : $\begin{cases} H_0: \theta \geq \theta_0 \\ H_1: \theta < \theta_0 \end{cases}$
- 右尾檢定 : $\begin{cases} H_0: \theta \leq \theta_0 \\ H_1: \theta > \theta_0 \end{cases}$
- 雙尾檢定 : $\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases}$
 - 其中 θ 表示欲檢定之母體參數， θ_0 表示特定參數值。

檢定方法

下列四種方法雖不同，但檢定的結果會一致：

1. 臨界值檢定 (critical value test), p 值推論 z 值(或 t 值)
2. Z 值檢定或 T 值檢定
3. 信賴區間檢定
4. p 值檢定

P 值檢定之步驟 (以單尾為例)

- 建立虛無假設 H_0 與對立假設 H_1 。
- 抽取一組隨機樣本並計算欲檢定之統計量(統計值)。
- 在 H_0 為真的條件下，計算 p 值。
- 將 p 值與顯著水準 α 作比較：
 - 若 p 值 $\leq \alpha$ ，則拒絕 H_0 。
 - 若 p 值 $> \alpha$ ，則不拒絕 (接受) H_0 。
- 依據檢定結果進行決策。

R - T 檢定

```
> # t-檢定
> set.seed(168)
> x <- rnorm(n = 10, mean = 5)
> x
[1] 4.476240 5.388433 5.877145 4.577346 5.834570 6.274560 5.790271
[8] 4.549820 5.140481 5.179821
>
> t.test(x, mu = 5)
 ① ②
One Sample t-test
data: x
t = 1.5402, df = 9, p-value = 0.1579
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.855214 5.762523
sample estimates:
mean of x
 5.308869
```

$$\begin{aligned}H_0: \mu &= 5 \\H_1: \mu &\neq 5\end{aligned}$$

資料轉換

資料標準化(Standardization)

- 資料標準化
 - 將資料按比例進行線性轉換
 - 將資料進行非線性轉換
 - 使資料落在某一特定的區間,例如: $[0, 1]$ 之間
 - 適用於資料有差異範圍
- 資料標準化之目的:
 - 提升模型的收斂速度
 - 提高模型的精準度
 - 適用於主成分分析, 集群法, KNN, SVM, Logistic regression

資料標準化(Standardization) (續)

- (0,1)標準化: 將資料轉換至[0, 1]區間

$$X_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- 最小最大標準化(min-max normalization)

$$X_{new} = x_{new_min} + \frac{(x - x_{min})}{(x_{max} - x_{min})} \times (x_{new_max} - x_{new_min})$$

- Z-score標準化: 將資料轉換為平均值為0, 標準差為1的分配.

$$X_{new} = \frac{x - \bar{x}}{\sigma}, \bar{x} \text{ 平均值, } \sigma \text{ 標準差}$$

資料編碼轉換

- 類別型資料 → 數值型標籤資料，表示標籤編碼 (Label encoding)
 - 例: 有 → 1, 沒有 → 0
- 類別型資料 → 獨熱編碼 (One-hot encoding, 單熱編碼)

學歷	V1	V2	V3	V4
國中以下	1	0	0	0
高中	0	1	0	0
大專院校	0	0	1	0
研究所以上	0	0	0	1

- 數值型資料 → 類別型資料
 - 例: 將年收入轉換為 {高, 中, 低} 類別型資料

歷屆考題

- 6 資料處理流程應遵循以下四個步驟，請問它們的順序為何？(1)一致化；(2)萃取；(3)交付；(4)清理。
- (A)4213
 - (B)4132
 - (C)2413
 - (D)2143

歷屆考題

7 某家製造商想瞭解對該公司主要產品進行促銷是否有助於增加銷量，假設隨機取樣 10 家經銷商在降價前後兩個月的銷售資訊，並進行樣本平均數是否相等的假設檢定，假設所得出之 **t 檢定統計值為 0.9**，試問根據檢定結果，促銷的效果為何？

- (A) 促銷前後銷量差距顯著
- (B) 促銷前後銷量差距不顯著**
- (C) 資訊不夠無法分析
- (D) 只有在 10% 的顯著水準才有差異

- p 值小, 拒絕 H_0
- t 統計值大, 拒絕 H_0

歷屆考題

8 假設有一組資料為某襯衫尺寸，分為特小、小、中、大、特大，請問此資料的特性屬於何種資料？

(A) 名目資料

(B) 次序資料

(C) 等距資料

(D) 比值資料

測量尺度

歷屆考題

9 變數分為質變數和量變數，若某資料為學生體重，請問此變數為量變數的哪一類？

(A)連續變數

(B)離散變數

(C)以上皆是

(D)以上皆非

歷屆考題

- 10 群集分析可以運用在各種具有大量數據的領域中，以下那種不是群集分析的應用？
- (A) 產品組合
(B) 市場區隔
(C) 行銷策略
(D) 信用評估
- 集群分析是非監督式學習
 - 信用評估比較偏向監督式學習

歷屆考題(非選擇題)

11

某房仲業者打算建構一**房價預測模型**，以提供業務更精準的定價參考。試問

- 除了內政部實價登錄資料庫外，請列舉二個其他對房價建模有關的資料來源？
- 下表為實價登錄之部分欄位，若要用於資料分析，「交易筆棟數」欄位將無法直接使用，請寫下無法使用之原因及你的解決方式。

鄉鎮市區	交易標的	土地區段位置/建物區段門牌	土地移轉統使用分區或交易年月	交易筆棟數	
大安區	房地(土地+建物)	臺北市大安區和平東路三段1	19.39 住	10106	土地1建物2車位0
松山區	房地(土地+建物)	臺北市松山區三民路68巷1~3	35.53 住	10107	土地1建物1車位0
中正區	房地(土地+建物)	臺北市中正區忠孝東路二段1	8.46 商	10107	土地3建物1車位0
大同區	土地	橋北段二小段601~630地號	5.5 其他	10107	土地1建物0車位0

解答：

- 房仲業本身自己的資料庫及其他重要環境及經濟資料，如人口、收入、治安、公共設施位置(捷運、醫院、市場、宮廟、殯儀館)....).
- 原因：交易筆棟數之欄位每一筆記錄包含了超過一個變數的訊息。解決方式：將此欄位拆成「土地」、「建物」、「車位」三個欄位。

3.L121經營管理基本知識

- L12101企業經營環境與策略管理
- L12102企業的核心流程及其管理活動
- L12103財務會計基本知識

歷屆考題

12

下列何者能衡量企業的營業績效？

- (A) 流動比率
- (B) 存貨週轉
- (C) 速動比率
- (D) 利息保障倍數

歷屆考題

13

管理具備四大功能，各功能各司其職並前後相關。請從以下敘述選擇正確的功能與依序步驟？

- (A) 控制—規劃—領導—組織
- (B) 規劃—組織—領導—控制
- (C) 組織—控制—領導—規劃
- (D) 領導—控制—規劃—組織



歷屆考題

14

要獲得即時化生產(Just in time)的效益，其中關鍵在於製造廠商與供應廠商必須培養何種關係？

- (A) 合夥關係
- (B) 競爭關係
- (C) 對立關係
- (D) 無關係

歷屆考題

- 15 負責計算企業生產與配銷過程中相關的各項成本是指何種人員？
- (A) 管理會計人員
 - (B) 成本會計人員
 - (C) 財務人員
 - (D) 行銷專案人員

歷屆考題

- 16 以下何者不是常用的獲利率指標？
- (A)邊際利潤率
 - (B)投資報酬率
 - (C)每股盈餘
 - (D)流動比率

銷售與配銷分析

顧客區隔分析 (Customer Segmentation Analysis)

- 根據各種歷史銷售交易，顧客區隔分析能讓企業組織將其顧客清楚地定義成各種不同的群組，然後據此研究在不同群組顧客的行為模式，稱之為「顧客區隔分析」。
- 這種分析能夠幫助行銷經理對於該企業的顧客群有更清楚地瞭解，因而定位其目標市場。
- 針對不同的顧客群，提供符合他們喜好以及需求的產品及服務。

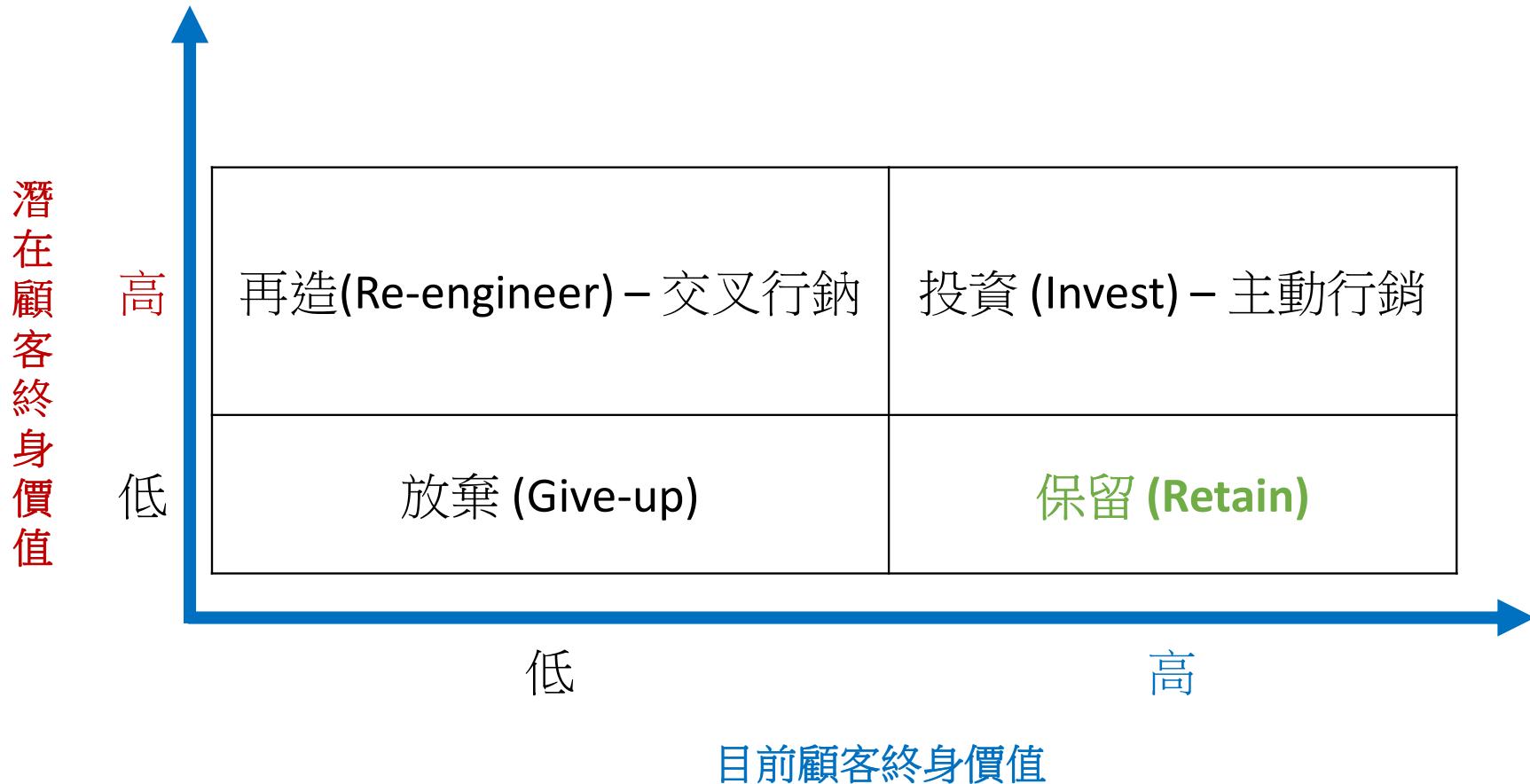
顧客區隔分析常用指標

指標名稱	指標定義與公式
顧客獲利能力 (Customer profitability)	<ul style="list-style-type: none">定義：某一特定市場中，由某些特定顧客所帶來的收益占該企業總收益之比例。公式：$\text{某些特定顧客所帶來的收益} \div \text{銷售總收益} \times 100\%$
特定市場占有率 (Specific market share index)	<ul style="list-style-type: none">定義：企業在某特定市場的市場占有率與該企業總市場占有率的比例。公式：$\text{特定區域的市場占有率} \div \text{所有市場的占有率} \times 100\%$
行銷活動之成本收益比例 (Campaign/Event cost revenue ratio)	<ul style="list-style-type: none">定義：某一特定市場，由行銷活動與事件所為企業所帶來收益與行銷活動與事件所花費的成本之比例。公式：$\text{行銷活動與事件所為企業所帶來的收益} \div \text{行銷活動與事件所花費的成本} \times 100\%$

消費時間頻率與金額分析 (Recency, Frequency, Monetary analysis, RFM)

- 在動態顧客行為方面，RFM 呈現忠誠度和滿意度相關的行為變數。
- 使用資料挖礦的**集群 (Clustering)** 技術：運用動態顧客資料庫做為分析，其結果比靜態的人口統計變數（例：僅考慮男性，女性）來區隔市場更具有價值性。
- 消費時間 (**Recency**)：該顧客最後下訂單的時間為何？分析最近購買的顧客是否傾向再度購買？
- 消費頻率 (**Frequency**)：該顧客下了多少次的訂單？分析較常購買的顧客是否會較易回應？
- 消費金額 (**Monetary**)：該顧客總共花費金額？分析金額較多的顧客未來更會消費的可能性？

顧客終身價值分析



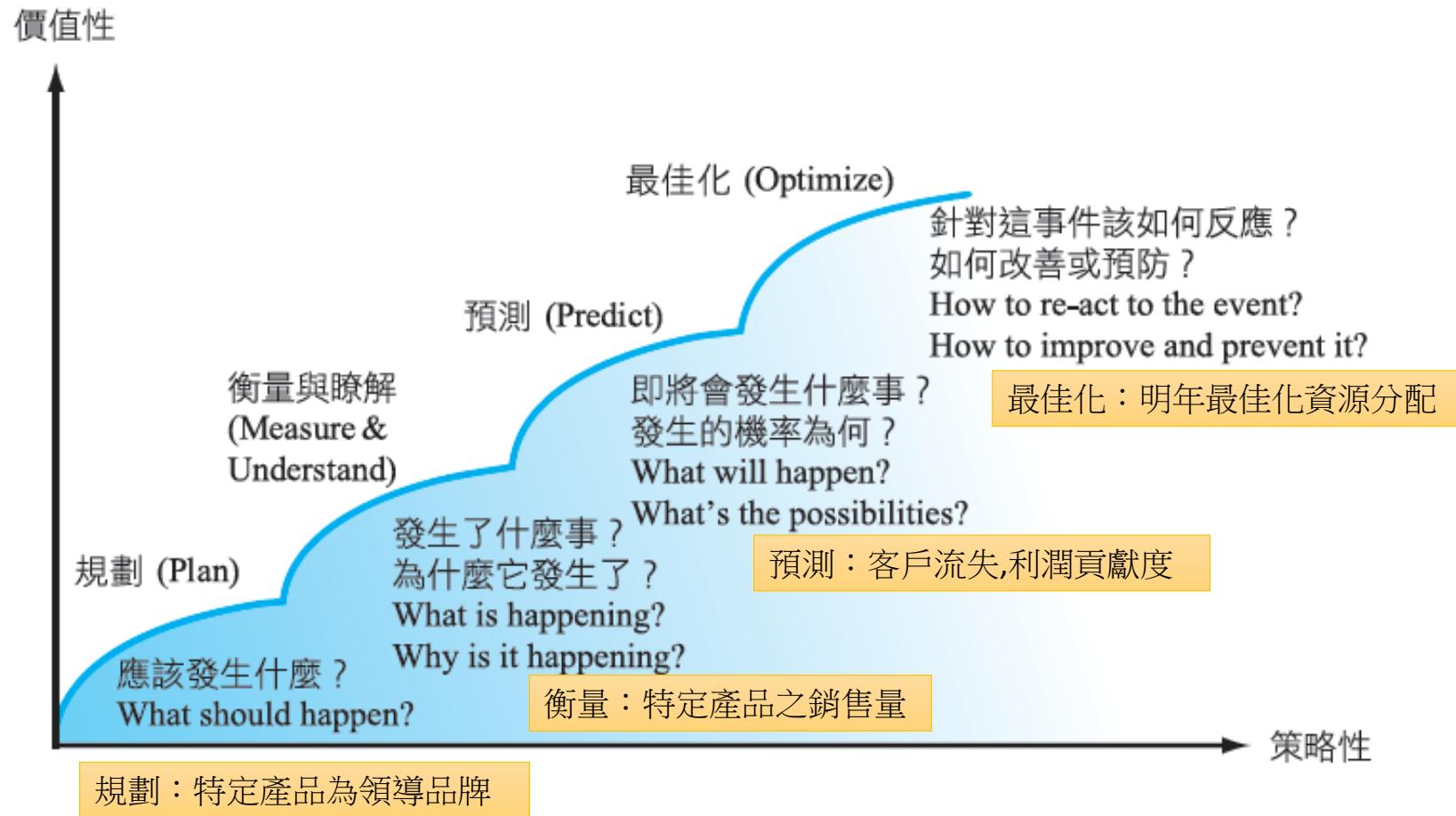
思考題

在顧客終身價值分析中，如果現有的價值高，但是潛在的價值低時，則必須採取哪一種行銷方法？

- (A) 投資(Investment)
- (B) 保留 (Retain)
- (C) 放棄 (Give-up)
- (D) 再造 (Re-engineering)

答案:B

商業智慧分析應用之層級



思考題

顧客利潤貢獻度分析是屬於商業智慧應用層級中的哪一種？

- (A) 規劃(Plan)
- (B) 衡量與瞭解 (Measure & Understand)
- (C) 預測 (Predict)
- (D) 最佳化 (Optimize)

答案: C

思考題

在銷售漏斗分析中，下列何者為常用的分析指標? (1) 詢價單/報價單比例 (2) 顧客抱怨比例 (3) 報價單/銷售訂單比例 (4) 既有顧客的收益成長比例率 (5) 銷售成功/失敗比例

- (A) 123
- (B) 134
- (C) 145
- (D) 135

答案: D

思考題

在銷售人力分析的常用分析指標中，用來衡量銷售人力的銷售週期時間通常是詢價單至下列哪些文件為止? (1)契約 (2)交貨文件 (3)訂單 (4) 請款 (5) 報價

- (A) 145
- (B) 135
- (C) 345
- (D) 234

答案: B

思考題

產品可供應天數是下列哪一種財務指標為計算之基準？

- (A) 淨銷售額
- (B) 邊際貢獻度
- (C) 股東權益
- (D) **銷貨成本**

答案: D

思考題

在商業智慧的銷售與配銷分析中，銷售成功/失敗比例常被使用於下列哪兩種常用分析中？(1) 產品存貨分析 (2) 顧客獲取與保留分析
(3) 銷售漏斗分析 (4) 銷售人力分析 (5) 訂單成交分析

- (A) 13
- (B) 25
- (C) 34
- (D) 45

答案: C

思考題

在銷售通路分析的整合性報表中，下列哪些項目常被用來衡量銷售通路的權重排名？(1) 邊際利潤貢獻度 (2) 顧客終身價值 (3) 顧客滿意度 (4) 顧客獲利能力 (5) 成功接單比例

- (A) 135
- (B) 245
- (C) 234
- (D) 124

答案: A

思考題

考慮某特定銷售通路的呆滯存貨量並且監控是否有例外事件發生，是屬於商業智慧應用層級中的哪一種？

- (A) 規劃(Plan)
- (B) 衡量與瞭解 (Measure & Understand)
- (C) 預測 (Predict)
- (D) 最佳化 (Optimize)

答案: B

採購之關鍵績效指標

思考題

收集分析並研究預測現有供應市場的短、中、長期供應趨勢，針對易受國際局勢與技術發展影響的採購品，從區域性、供應商類型、採購品組合與分類等不同構面進行分析，稱為：

- (A) 採購績效研究
- (B) 供應商績效研究
- (C) 供應商管理
- (D) **採購研究**

答案:D

思考題

在顧客導向與消費意識抬頭的現代環境中，採購-供應關係中，還有另外二項採購績效構面亦被多位學者提出，包含「保持良好供應商關係」以及

- (A) 降低採購成本
- (B) 準時交貨
- (C) 滿足顧客需求
- (D) 正確迅速的處理訂單

答案:C

思考題

以下哪項並非是在考慮供應商總數目時之重要因素？

- (A) 採購成本
- (B) 交貨準時性
- (C) 採購風險
- (D) 正確迅速的處理訂單

答案:A

思考題

以下哪種情形發生時，採購部門應積極了解原因，儘快做出決定是否尋找替代供應商，或是採取措施要求供應商進行改善？

- (A) 採購金額高、供應績效不佳
- (B) 採購金額低、供應績效不佳
- (C) 次要供應商
- (D) 以上皆是

答案:A



思考題

以下關於供應商績效評估之敘述，何者正確？

- (A) 量化評估資料可由電腦取得運算
- (B) 質化評估資料可由訪談資料得到
- (C) 供應商績效評估之量化質化資料均應收集
- (D) 以上皆是

答案:D

思考題

下列哪項是1990年代以後採購部門的主要任務？

- (A) 尋找最具競爭力的供應商
- (B) 強化供應商與企業組織內部的研發設計連結
- (C) 發展與管理供應商群
- (D) 以上皆是

答案:D

思考題

Monczka 等人主張以採購總成本比較供應商之間的績效，以下有關採購總成本之論述，何者為是？

- (A) 最佳評估結果是2.0
- (B) 指標結果值越高則代表採購績效愈佳
- (C) 指標結果值越高則代表採購績效愈差
- (D) 指標值介於正負1之間

答案:C

參考: Purchasing And Supply Chain Management 4th edition, 2009.

<http://www.mim.ac.mw/books/Purchasing%20And%20Supply%20Chain%20Management%204th%20edition.pdf>

思考題

供應商的供應績效評估，可達到以下哪個目的？

- (A) 比較某單一供應商歷史績效
- (B) 比較數個競爭供應商的供應績效
- (C) 透過不同時期的分析，可看出每家供應商在各次考核期的績效狀況
- (D) 以上皆是

答案:D

思考題

有關績效指標「採購品平均運送的時間」的敘述，何者正確？

- (A)指某採購品從下達採購單開始，到採購品送達為止，其經過的時間總值
- (B)可用來評估供應商的供貨效率
- (C)指標數值越大，代表平均運送的時間越長
- (D)以上皆是

答案:D

$$\text{採購品平均運送的時間} = \frac{\text{從下採購單至收貨之間的天數加總}}{\text{運送次數}}$$

財務會計模組之關鍵績效指標

模擬題

下列哪些是說明財務報表的横向分析 (1) 以多期的資料進行分析，以瞭解企業之未來可能發展趨勢，亦可據以訂定來年之成長目標 (2) 可比較前後期各金額項目的增減情況 (3) 分析同一財務報表的各金額項目對於全體金額的關係 (4) 為一種趨勢分析 (5) 其分析方法有指數法與成長率法

- (A) 34
- (B) 345
- (C) 1245
- (D) 2345。

答案:C

- **横向分析**：分析財務報表金額項目之多個期間的變化情況，或比較財務報表金額項目之前後期的增減情況。若針對多年期連續財務報表間進行比較分析，就是一般所謂的**趨勢分析**。
- **縱向分析**：分析同一財務報表之各金額項目相對於整體金額的關係，或各金額項目間的關係而言，而一般所採用的分析方式是比率分析。

參考: <https://yamol.tw/exam-105+%E5%B9%B4BI%E8%A6%8F%E5%8A%83%E5%B8%AB20130051983-51983.htm>

模擬題

下列哪一種報表是在表達公司**某一時間點的財務狀況**，是一種靜態的財務報表

- (A) 股東權益變動表
- (B) 資產負債表
- (C) 現金流量表
- (D) 損益表

答案:B

思考題

下列哪些是 ERP 系統對於財務分析的助益 (1) ERP 系統能簡化會計處理而即時提供整合性資料 (2) ERP 系統即時提供財務分析資訊 (3) ERP 系統能自動提供財務問題的解決方法 (4) ERP 系統可以作更精細的財務分析

- (A) 234
- (B) 134
- (C) 124
- (D) 1234

答案:C

財務報表 (Financial statement)

- 財務會計係將整個企業體所發生的交易事項記錄下來，而最終目的為編製對外發布之財務報表。
- 企業主要財務報表：
 - 資產負債表：報導企業**在一特定時間**之資產、負債、業主權益等財務狀況之報表。
 - 股東權益變動表：報導一特定期間內股東權益的變動情況
 - 損益表：報導企業在一特定期間的**經營損益情況**。
 - 現金流量表：報導在一特定期間內，有關企業之營業活動、投資活動及融資活動(又稱為理財活動)之現金流入及流出情形。

思考題

下列哪些敘述是正確的 (1) 損益表係用以表達某一時點的財務狀況 (2) 現金流量表主要係報導在一特定期間內，有關企業之營業活動、投資活動及融資活動之現金流入及流出情形 (3) 期末應作調整分錄而將收入與費用的帳戶調整成可以真正代表當期之收入與費用，以便正確地計算當期損益 (4) 雖然 ERP 系統可以將前端的交易資料即時、自動過帳，但仍須以單據層層傳送而再由財務會計人員將交易資料輸入會計系統

- (A) 24
- (B) 23
- (C) 134
- (D) 1234

答案:B

槓桿度分析之 KPI

損益表

銷貨淨額	(營收淨額)
- 銷貨成本	(營業成本)
銷貨毛利	(營業毛利)
- 銷管費用	(營業費用)
營業利益	
+ 營業外收入	(含利息收入)
- 營業外費用	(含利息費用)
± 非常損益	
税前利益	
- 所得稅	
税後純益	
每股盈餘	

營運槓桿度

$$= (\text{銷貨淨額} - \text{變動營業成本及費用}) \div \text{營業利益}$$

(每月/內部經理人)

財務槓桿度

$$= \text{營業利益} \div (\text{營業利益} - \text{利息費用})$$

(每月,季/內外人士)

槓桿度分析

- 營運槓桿度

- 衡量企業風險及資本密集的程度，即它代表的是企業對固定成本的使用程度。
- 營運槓桿程度越大，風險也越高。
- 營運槓桿可以迅速知道不同百分比的銷貨變動對利潤的影響。

- 財務槓桿度

- 用來衡量公司舉債經營是否對股東有利的指標。
- 財務槓桿度越高，即財務彈性越大，表示固定財務成本越高，故使得所得稅及利息費用前純益變動對每股盈餘之變動影響效果越大。

思考題

下列哪一種財務分析只牽涉到**損益表**上的數字

- (A) 經營能力分析
- (B) 財務結構分析
- (C) 償債能力分析
- (D) 槓桿度分析

答案:D

思考題

下列哪些財務分析可能同時牽涉到資產負債表與損益表上的數字 (1)
經營能力分析 (2) 財務結構分析 (3) 槓桿度分析 (4) 債債能力分析 (5)
獲利能力分析

- (A) 13
- (B) 234
- (C) 145
- (D) 1234

答案:C

思考題

下列何者不是財務結構分析的KPI?

- (A) 負債比例
- (B) 負債對權益比率
- (C) 流動比例
- (D) 長期資金占不動產、廠房及設備比例

- 流動比率 = 流動資產 ÷ 流動負債
- 流動比率愈大代表短期償債能力愈強，流動比率太小容易造成資金週轉不靈的危險。

答案:C

思考題

下列是哪一種財務分析的KPI?

$$\frac{\text{稅後純益} + \text{所得稅} + \text{利息費用}}{\text{利息費用}}$$

- (A) 速動比例
- (B) 流動比例
- (C) 利息保障倍數
- (D) 殖利率

答案:C

- **利息保障倍數**：企業對於長期借款的債權人有定期支付利息與到期還本的義務，而債權人可以用利息保障倍數來瞭解其利息的保障程度。
- 比率的意義為：企業賺取的利潤相當於當期利息費用的倍數。

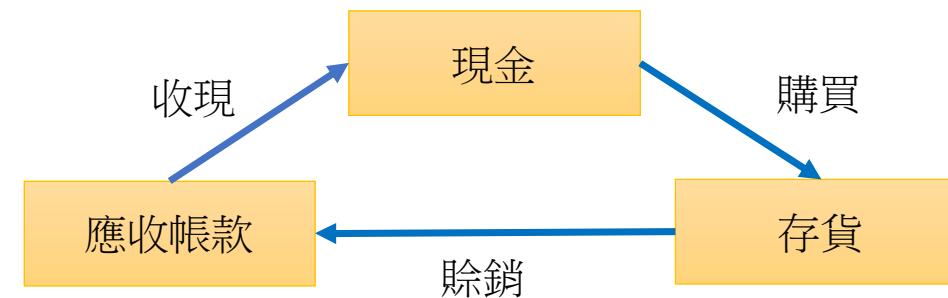
思考題

營業循環日數是哪些日數合計而來？(1)平均付款日數 (2)平均收現日數 (3)平均銷貨日數 (4)現金變現日數。

- (A) 14
- (B) 23
- (C) 234
- (D) 1234

答案:B

- 營業循環日數 = 平均銷貨日數 + 平均收現日數
- 一個營業循環所需日數，稱為營業循環日數。



思考題

下列是哪一種財務分析的KPI?

$$\frac{\text{每股股利}}{\text{每股盈餘}}$$

- (A) 本益比
- (B) 價格與股利比
- (C) 股利分配率
- (D) 殖利率

答案:C

- 本益比 = 每股市價 ÷ 每股盈餘
- 價格與股利比 = 每股市價 ÷ 每股股利
- 股利分配率 = 每股股利 ÷ 每股盈餘
- 殖利率 = 每股股利 ÷ 每股市價

思考題

下列是哪一種財務分析的KPI?

$$\frac{\text{營業活動現金流量} - \text{現金股利}}{\text{固定資產毛額} + \text{長期投資} + \text{其他資產} + \text{營運資金}}$$

- (A) 現金流量比率
- (B) 現金利用比率
- (C) 現金再投資比率
- (D) 現金流量允當比率

答案:C

- 現金流量比率 = 營業活動現金流量 ÷ 流動負債
- 現金流量允當比率 = 最近五年度營業活動現金流量 ÷ [最近五
年 (資本支出 + 存貨增加額 + 現金股利)]
 - 衡量企業由營業活動產生之現金流量能否支應其購買存貨、資本支出及現金股利的資金需求。
 - 若現金流量允當比率大於1，表示企業營業活動之現金流量能夠充分支應三項支出(資本支出、購置存貨，以及支付現金股利)，不需向外界借款或現金增資來籌資，且還有剩餘資金。
 - 若小於1，則需對外借款或籌資或變賣資產等措施來支應。

生產規劃與控制

思考題

目標與實際報廢率偏差一指標之計算，有牽涉到以下哪些值？(1) 目標報廢數目；(2) 實際報廢數目；(3) 目標生產數目；(4) 實際生產數目；(5) 目標交貨數目。

- (A) 1234
- (B) 1245
- (C) 2345
- (D) 1345

- 目標報廢率=目標報廢數目/(目標生產數目-目標報廢數目) × 100%
- 而實際報廢率=實際報廢數目/實際生產數目 ×100%

答案:A

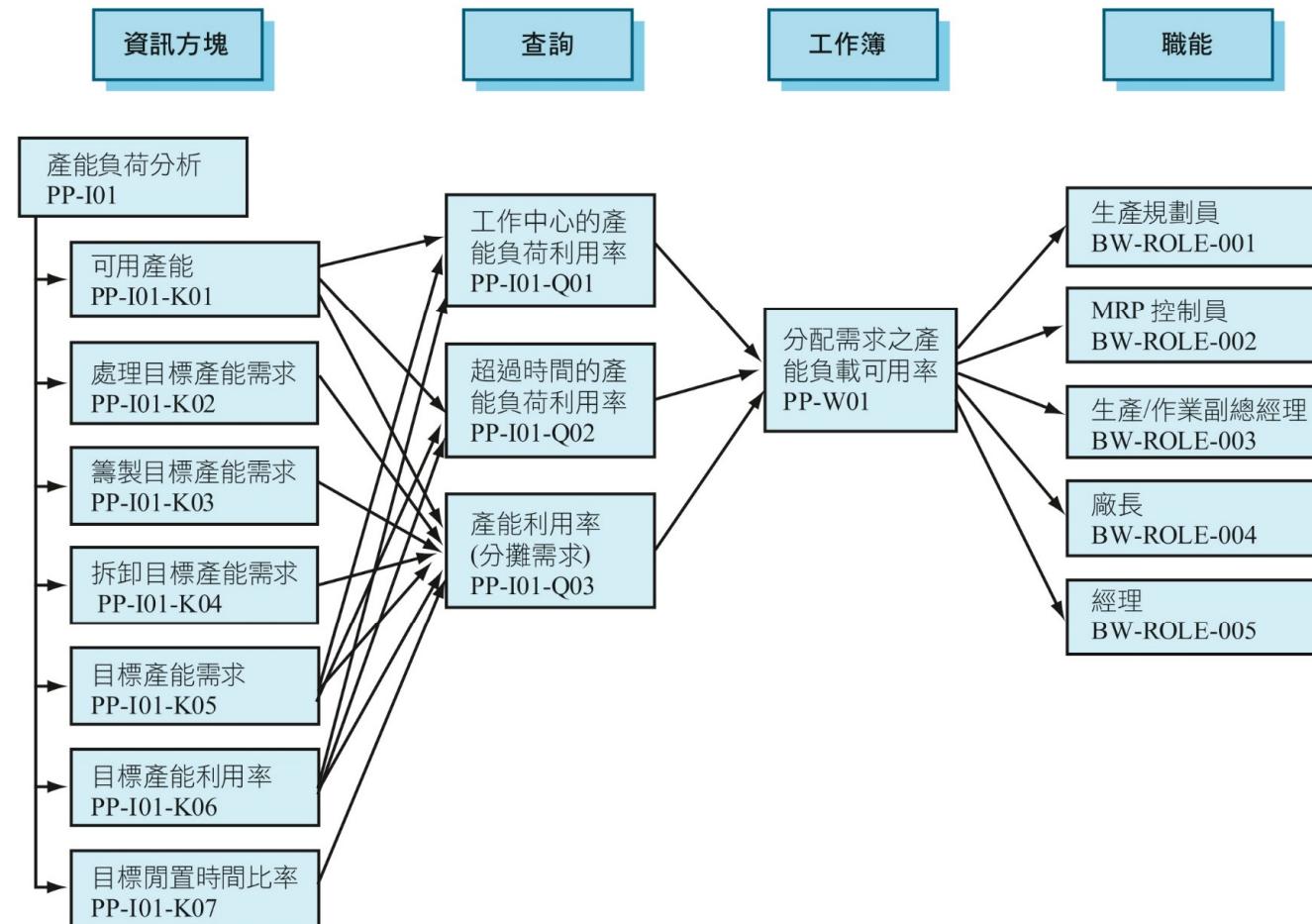
思考題

請問下列何圖表中，詳列多維度分析中維度資訊？

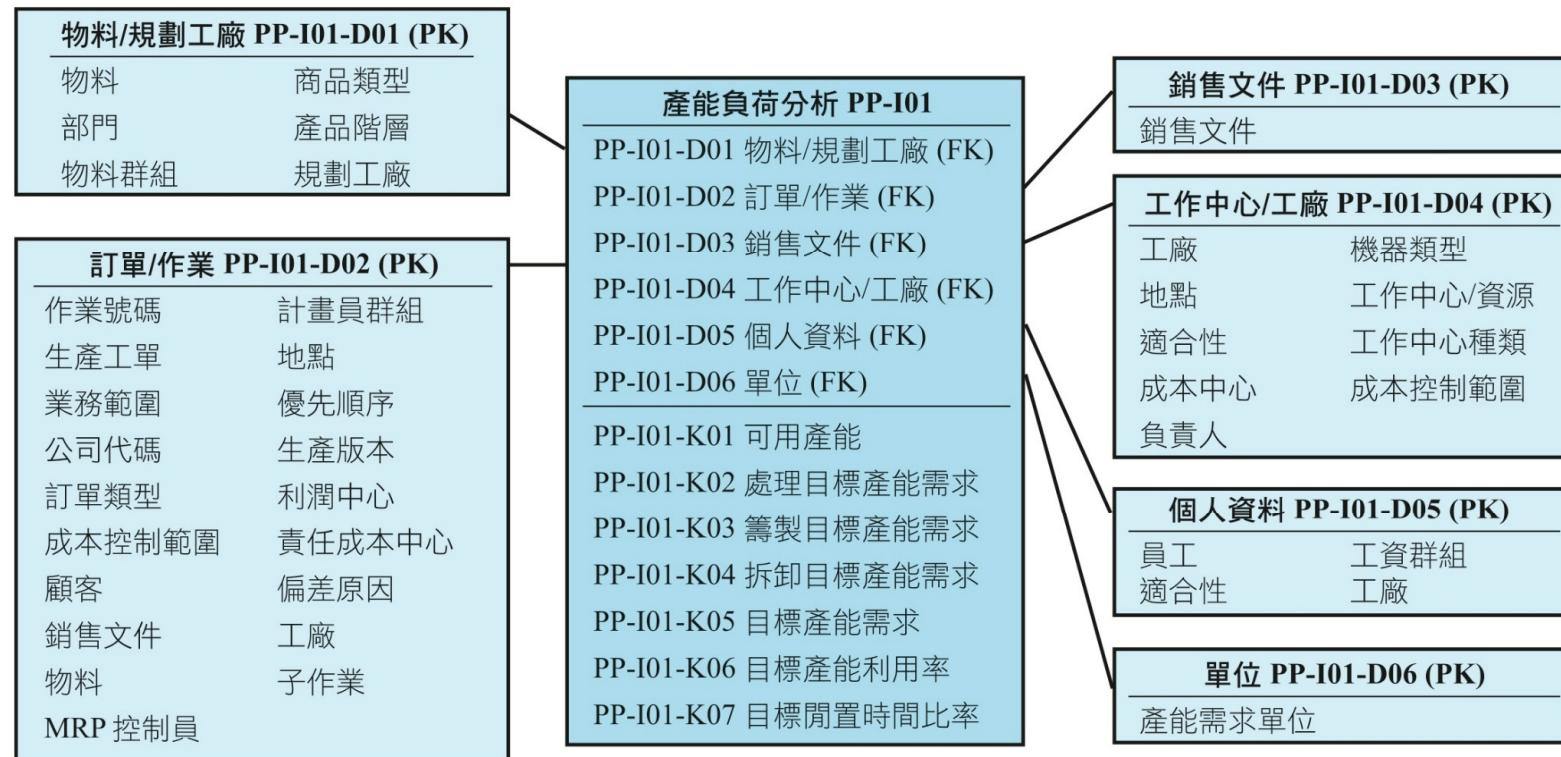
- (A) 資料流 Data flow
- (B) 星型綱要 Star schema
- (C) 要因圖 (特性要因圖,魚骨圖) Case and Effect Diagram
- (D) 途程

答案:B

Data flow - 產能負荷分析



Star schema - 產能負荷分析

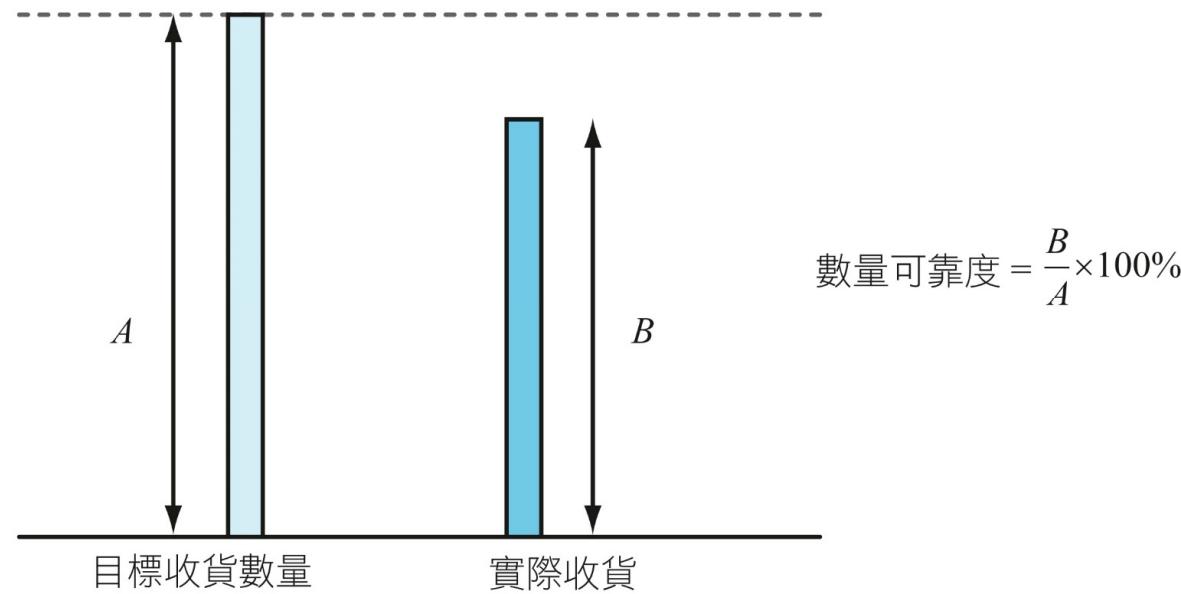


思考題

數量可靠度是由下列哪兩者所計算而得？(1) 目標收貨數量 (2) 所有生產工單數目 (3) 所發放工單數量 (4) 實際收貨 (5) 未報廢數量。

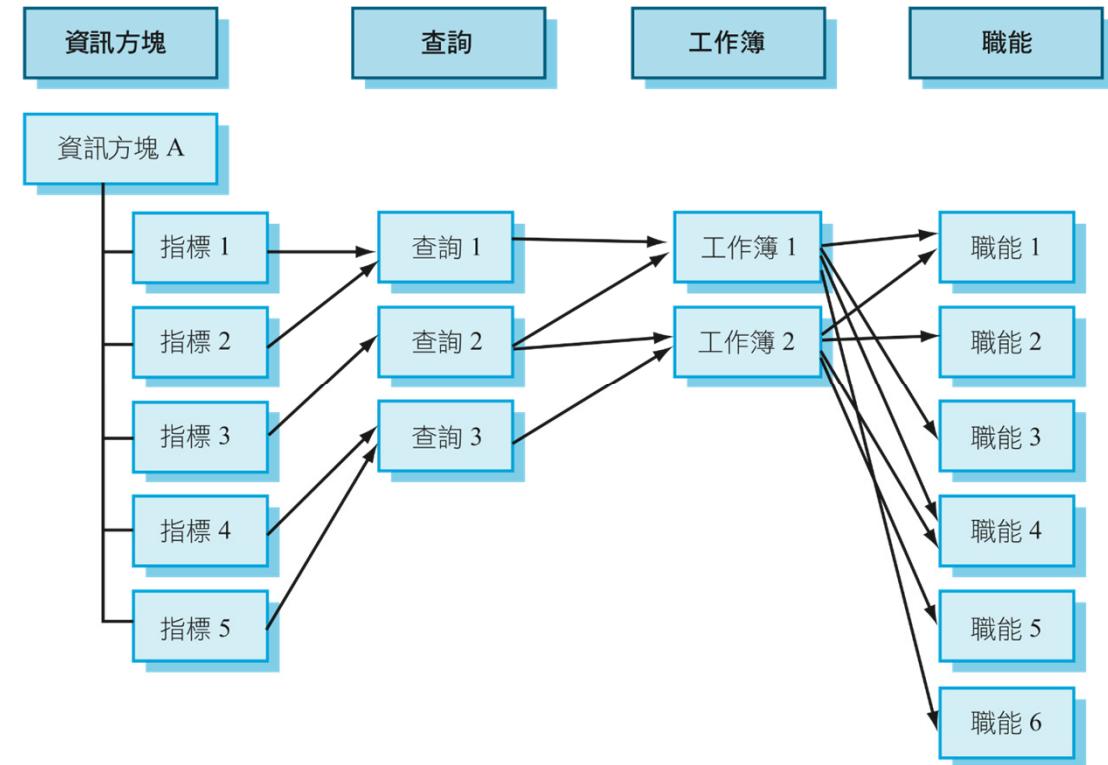
- (A) 23
- (B) 25
- (C) 34
- (D) 14

答案:D



Data flow – 以生產工單角度分析

- 資訊方塊(InfoCube): 定義的KPI與查詢(Query)將被儲存於相關之資訊方塊中。
- **查詢(Queries)**: 集群(Grouping)了某一群KPI，以期能更方便簡潔的呈現有興趣的指標。
- 工作簿(Workbook)：使用者介面，方便使用者以不同形式呈現指標。
- 職能/角色(Roles)：透過職能的界定，分類不同職能將使用不同之查詢。



思考題

下列何者集群(grouping)了某一群KPI，以期能更方便簡潔的呈現有興趣的指標？

- (A) 職能
- (B) 工作簿
- (C) 資訊方塊
- (D) **查詢**

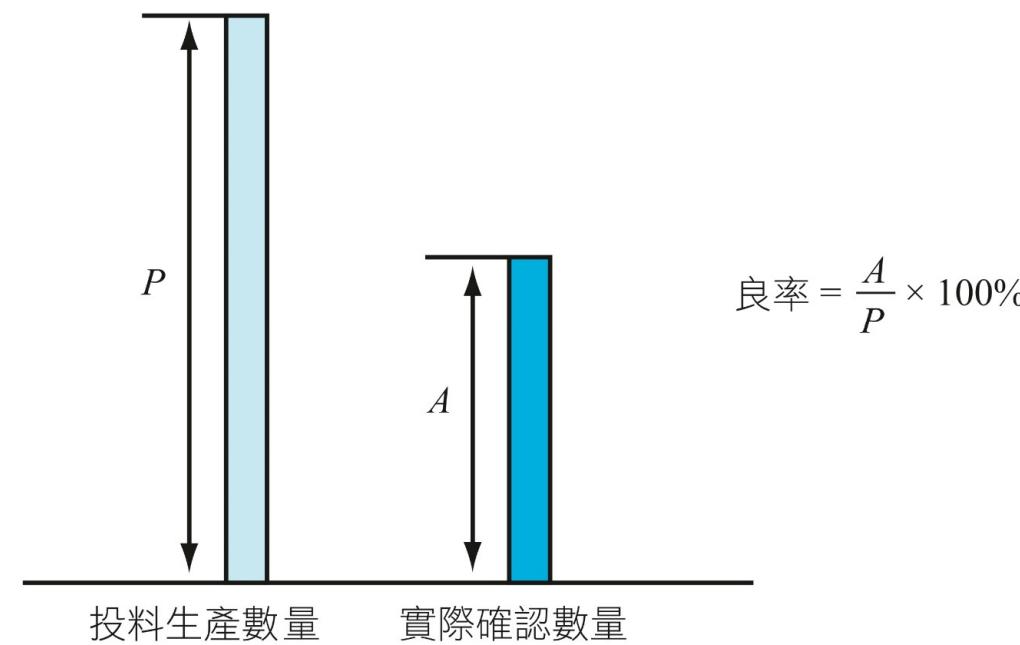
答案:D

思考題

良率是由下列哪兩者所計算而得？(1) 實際確認數量 (2) 所有生產工單數目(3) 所發完工單數量 (4) 生產數量 (5) 未報廢數量。

- (A) 23
- (B) 25
- (C) 34
- (D) 14

答案:D



思考題

發料數量偏差百分比-調整後之指標計算，其所謂的調整是根據何值？

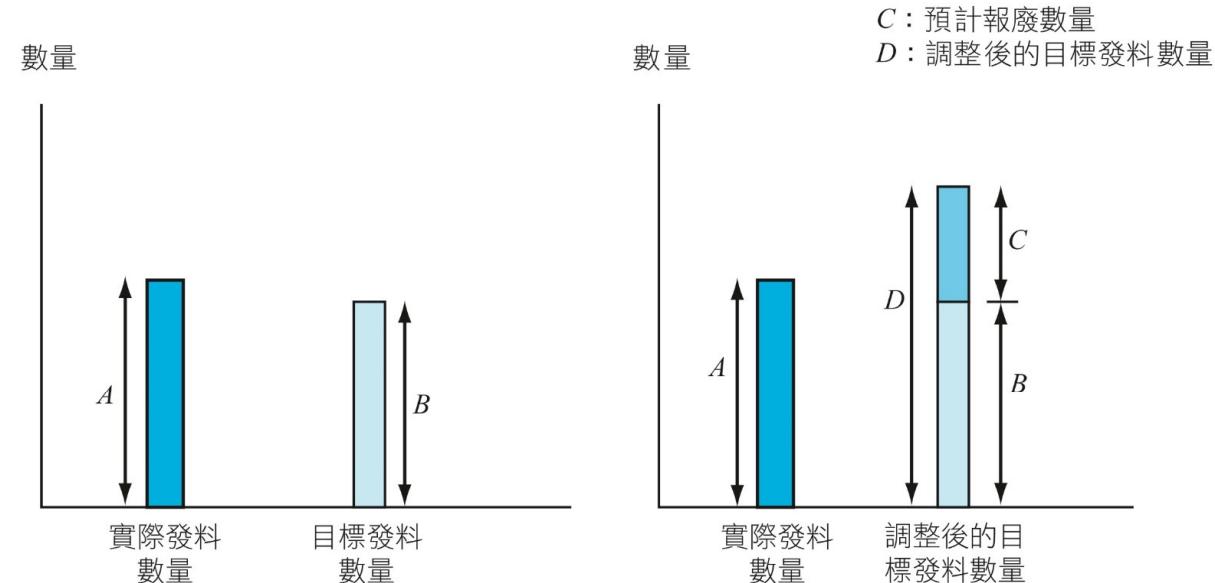
- (A) 良率
- (B) 允收機率
- (C) 報廢率
- (D) 量產率

答案:C

- 發料數量偏差百分比-調整後指標：考慮報廢率下備料之正確性。
- 指標小於100%時，表示多備料的情形發生，不利成本競爭。
- 本指標大於100%時，較易有不足料影響生產的情形，進而影響交期。

發料數量偏差百分比 - 調整前/後

- 實際發料數量與調整後目標發料數量之差額占調整後目標發料數量之百分比。
- 調整後目標發料數量為依**報廢率**調整後之目標發料數量，假若公司之備料是有考慮報廢率調整，則依此指標來反應物料耗用之情形。
- 若為**負百分比**，則表示多備料的百分比，即目標定太高。
- 若為**正百分比**，則表示少備料的百分比，即目標定太少。



$$\text{發料數量偏差百分比——調整前} = \frac{(A - B)}{B} \times 100\%$$

$$\text{發料數量偏差百分比——調整後} = \frac{(A - D)}{D} \times 100\%$$

思考題

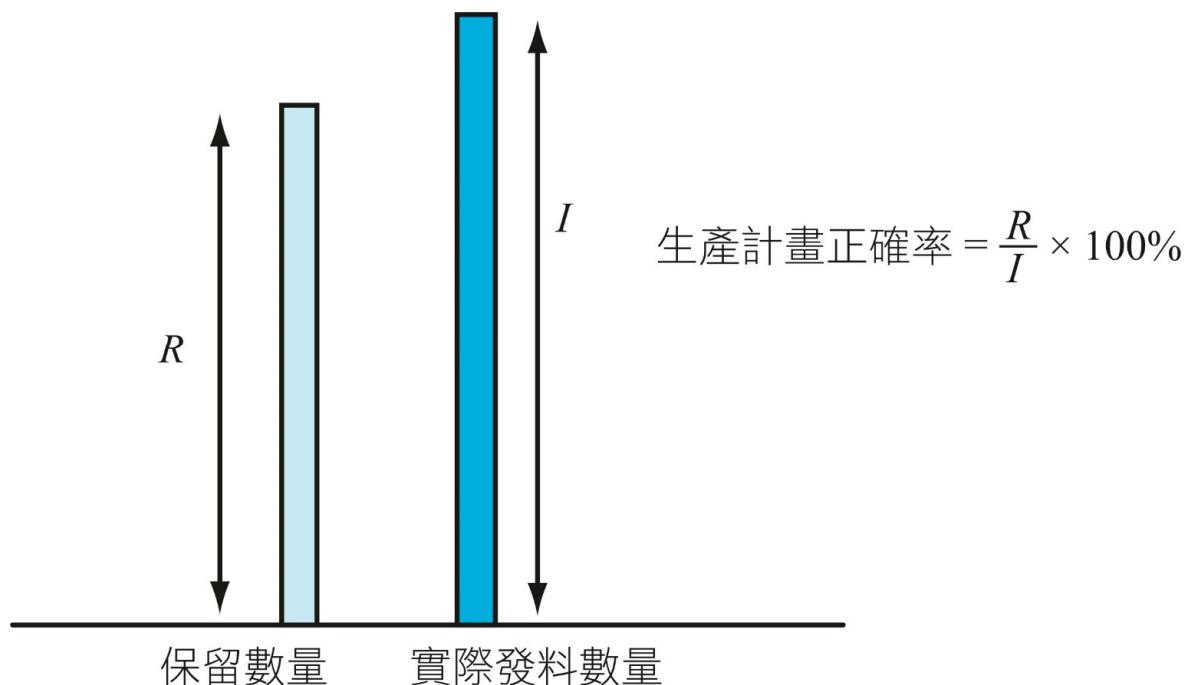
生產計畫正確率是由下列哪兩者所計算而得？(1) 實際確認數量；
(2) 保留數量；(3) 實際發料數量；(4) 生產數量；(5) 未報廢數量。

- (A) 23
- (B) 25
- (C) 34
- (D) 14

答案:A

生產計畫正確率

- 本指標以**保留數量**為基礎來計算生產計畫正確率；
- 生產工單開立時，可保留生產時所需之相依物料，**保留之數目根據工單而定**，發料時則依保留文件進行計畫性發料。
- 假如實際發料數目與保留數目相同，則表示實際生產數量完全依當初規劃的數量，反映出計畫正確率百分之百。



思考題

下列何者在生產上無附加價值？(1) 署置；(2) 銷售；(3) 處理；(4) 採購；(5) 拆卸。

- (A) 23
- (B) 15
- (C) 34
- (D) 14

答案:B

思考題

產能利用率是由下列哪兩者所計算而得？(1) 目標產能需求；(2) 工單產能需求；(3) 生產線所有產能；(4) 三班產能；(5) 可用產能。

- (A) 23
- (B) 15
- (C) 34
- (D) 14

答案:B

- 產能利用率(Capacity utilization)= $\frac{\text{實際產出}}{\text{設計產能}}$
- 考慮以下資訊，計算汽車維修部門的效率與產能利用率：
設計產能 = 每天 50 輛卡車，有效產能 = 每天 40 輛卡車，實際產出 = 每天 36 輛卡車
- 效率 = 實際產出 / 有效產能 × 100%
 $= \text{每天 } 36 \text{ 輛卡車} / \text{每天 } 40 \text{ 輛卡車} \times 100\% = 90\%$
- 產能利用率 = 實際產出 / 設計產能 × 100%
 $= \text{每天 } 36 \text{ 輛卡車} / \text{每天 } 50 \text{ 輛卡車} \times 100\% = 72\%$

思考題

請依先後順序排列下列工單相關時間：(1) 交期；(2) 發放日；(3) 生產開始；(4) 生產完工；(5) 工單完工。

- (A) 23451
- (B) 15234
- (C) 12345
- (D) 23541

答案:A



人力資源關鍵績效指標

資訊蒐集與進行決策在整體時間中所占的比例

傳統人力資源決策行為

用於資訊蒐
集的時間



- 資訊蒐集時間占 80%
- 經常交付給資淺員工執行
- 決策時間短，占 20%
- 由較低層管理人員親自決策，
拉長決策與採取行動之時間差
- 無法釐清分割決策與資訊蒐集活動

運用人力資源商業智慧 系統協助進行決策

用於進行決
策的時間

- 決策時間占 80%
- 由高層管理人員親自決策
- 可縮短決策與採取行動之時間差

思考題

運用人力資源商業智慧(BI)進行決策可以

- (A) 經常交由資淺員工執行
- (B) 資訊蒐集時間占很大比重
- (C) 決策時間比重增加
- (D) 無法釐清與分割資訊蒐集與決策間的時間

答案:C

思考題

平衡計分卡(Balanced Scorecard, BSC)績效管理系統中，屬於內部流程績效指標的是

- (A) 財務面向
- (B) 客戶面向
- (C) 學習與成長
- (D) 以上皆非

答案:C

思考題

將人力資源績效與組織**策略**銜接，是將人力資源績效指標建立在

- (A) 作業層次
- (B) **策略**層次
- (C) 溝通層次
- (D) 以上皆非

答案:B

思考題

下列哪一個是人力資源績效指標可呈現的方式

- (A) 數字
- (B) 符號
- (C) 文字
- (D) 以上皆是

答案:D

思考題

在平衡計分卡(Balanced Scorecard, BSC)績效管理系統中，人力資源績效指標通常歸屬於哪一個面向

- (A) 財務面向
- (B) 客戶面向
- (C) 內部流程面向
- (D) 學習與成長面向

答案:D

思考題

在平衡計分卡(Balanced Scorecard, BSC)系統中，學習與成長面向通常不包含下列哪一個子面向

- (A) 資訊基礎設施
- (B) 組織與法令
- (C) 客戶滿意度
- (D) 人力資源

答案:C

思考題

運用**策略**地圖來設計規劃人力資源績效指標的用途在於

- (A) 能夠**策略**聚焦，讓指標能夠銜接組織策略
- (B) 人力資源可以有效分散配置
- (C) 開發組織有限人力
- (D) 以上皆是

答案:A

思考題

人力資源績效指標「員工工作士氣」的特性**不屬於**？

- (A) 量化指標
- (B) 整體人力資源運用效能
- (C) 結果導向指標
- (D) 質化指標

答案:A

思考題

人力資源績效指標「流動率」的特性不屬於？

- (A) 量化指標
- (B) 整體人力資源運用效能
- (C) 質化指標
- (D) 過程導向指標

答案:C

思考題

人力資源績效指標「雇用一位新進人員的成本」的特性屬於人力資源體系中的？

- (A) 人力確保
- (B) 人力開發
- (C) 人力報償
- (D) 人力維持

答案:A

思考題

人力資源績效指標「人員補缺平均工作天數」可以用人力資源哪一項的效率來表現

- (A) 招募與任用
- (B) 績效管理
- (C) 訓練發展
- (D) 薪資福利。

答案:A

思考題

人力資源績效指標在運用時，應該如何？

- (A) 兼顧評估之目的與影響目的的結果
- (B) 兼顧整體與個別人力資源效能
- (C) 兼顧量化與質化指標
- (D) 以上皆是

答案:D

4.L122數位化企業資訊工具基本知識

- L12201營運智慧資訊技術(如雲端技術、無線射頻辨識技術、**物聯網**、**大數據**等)
- L12202數位化企業常見資訊系統(如企業資源規劃、供應鏈管理、電子商務、雲端運算、知識管理等)
- L12203數位化轉型創新與價值創造(包括商業模式、企業價值鏈、核心流程與所需之資訊科技、企業流程再造，及結合後創新與創造價值)

歷屆考題

- 17 在電子商務時代，客服中心(Call Center)的建置，通常已與企業的何種資訊系統運用作緊密結合，以有效支援企業與客戶互動的相關活動？
- (A) 製造執行系統
 - (B) 顧客關係管理系統
 - (C) 供應鏈管理系統
 - (D) 企業資源規劃系統

歷屆考題

18

下列何者是一種利用文字探勘(Text Mining)的技術來監視消費者在社群網路上的對話，衡量消費者的行為並檢測企業促銷活動績效的一種技術與工具？

- (A) P2P
- (B) Crowd Sourcing
- (C) Social Monitor Service (SMS)
- (D) Crowdfunding

- A. 對等網路 (peer-to-peer, P2P)
- B. 群眾外包
- C. 社媒監控
- D. 群眾募資

社媒監控, <http://chinetekstrategy.com/blog/2017/11/23/social-monitoring1/>

群眾外包 (crowdsourcing)

- 群眾外包是一種特定的取得資源的模式。
- 在該模式下，個人或組織可以利用大量的**網路使用者**來取得需要的服務和想法。
- 「眾包」（crowdsourcing）是在2006年混合群眾（crowd）和外包（outsourcing）詞義而產生的混成詞。
- 這種通過將工作先分配給很多參與者再合成為最終結果的模式，在電子時代來臨前就已經取得了成功。
- 嘉和外包的區別在於，眾包的物件可以是一群沒有被特別定義的群體（而非被指派的，特定的群體）並且眾包包括了混合的由下而上和自頂向下的過程。眾包的優勢包括：最佳化的價格，速度，品質，靈活性和多樣性。

參考: <https://en.wikipedia.org/wiki/Crowdsourcing>

歷屆考題

- 19 Davenport and Short 提出企業流程再造實施的五個步驟，包括 1. 了解現行流程 2. 確定需要重新設計的流程 3. 建立新流程的雛型 4. 確定資訊科技的功用 5. 建立企業願景及活動目的。請問正確的順序為何？
- (A) 12345
(B) 52143
(C) 13524
(D) 53241
-
- ```
graph LR; A[建立企業願景及活動目的] --> B[確定需要重新設計的流程]; B --> C[了解現行流程]; C --> D[確定資訊科技的功用]; D --> E[建立新流程的雛型]
```

## 歷屆考題

20

物聯網可以說是一種結合分散式運算、行動運算、感測網路、人工智慧與人機界面的一個科技平台，主要結構可分為三個層級，分別為感測器層、網路層以及分析層。在物聯網中常使用的射頻識別(RFID)晶片應該被歸類為哪一層？

- (A) 感測器層
- (B) 網路層
- (C) 分析層
- (D) 以上皆非

## 歷屆考題

21

下列何者指的是供應鏈的訂單量因為消費者的需求改變而向上游擴散的擾動變化，零售端的小小波動觸發了庫存水平的大幅擺動。

- (A) 需求預測準確性
- (B) 能見度
- (C) 長鞭效應 Bullwhip effect
- (D) 供應鏈成本

## 歷屆考題

- 22 某運輸公司利用儲存於資料庫中關於客戶需求、可用於駕駛員、車輛、貨物重量的資料找出數以千計規則，並以此推理規劃出最佳路線以及駕駛員、車輛和貨物指派，這樣的系統技術屬於：  
(A) 作業系統  
(B) ERP 系統  
(C) 資料庫系統  
(D) 專家系統

# 物聯網 (Internet of Things, IoT)

- 物聯網，或稱智慧聯網，就是將智慧化的感測器加上網際網路，再結合雲端資料儲存、分析能力，以實現各種智慧應用。
- 更簡單的說，物聯網就是「萬物相聯的網際網路」。

# 物聯網的架構

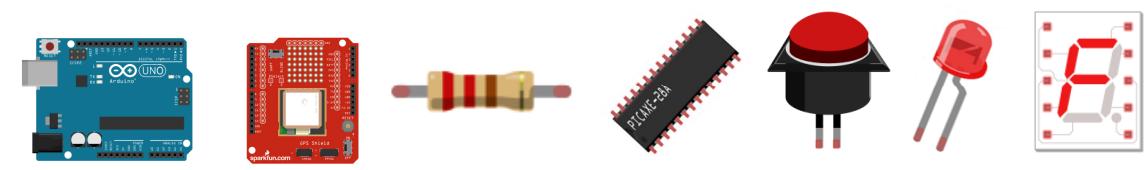
- 應用層：資訊一旦進到了應用層，可建立企業服務應用，包括智慧家庭、智慧電網、智慧醫療、智慧工廠、智慧農業與智慧城市等多種應用領域。



- 網路層：主要任務是處理下層傳來的資訊，判斷是要送往上層，也就是雲端的主機上，或者直接採取適當的動作。

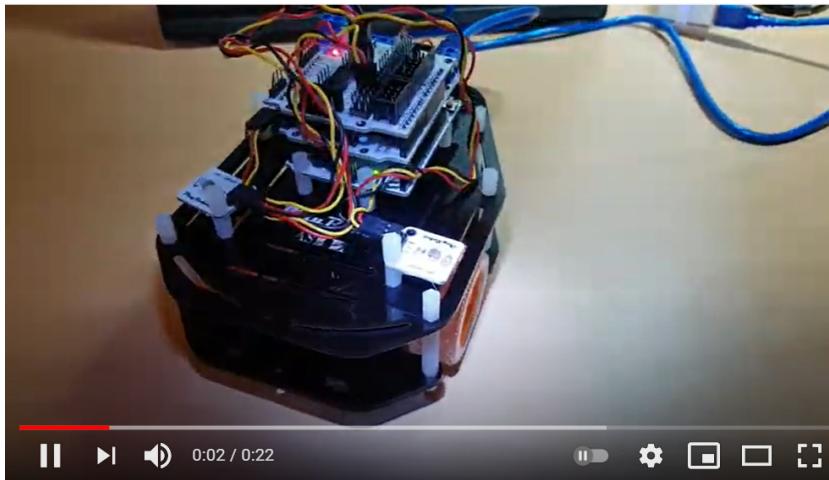


- 感測層：藉由各種感測器來擷取所需的資訊，對訊號作正確的解讀與數位化後送到網路層，供進一步分析應用。

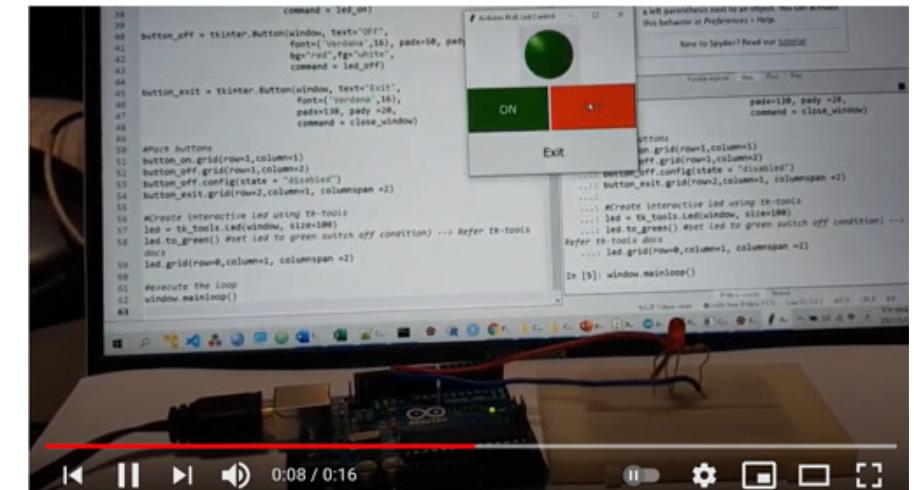


# Arduino demo

- Arduino car 光敏電阻示範
- <https://youtu.be/SXyq6urITQo>
- <http://rwepa.blogspot.com/2021/05/arduino-car.html>



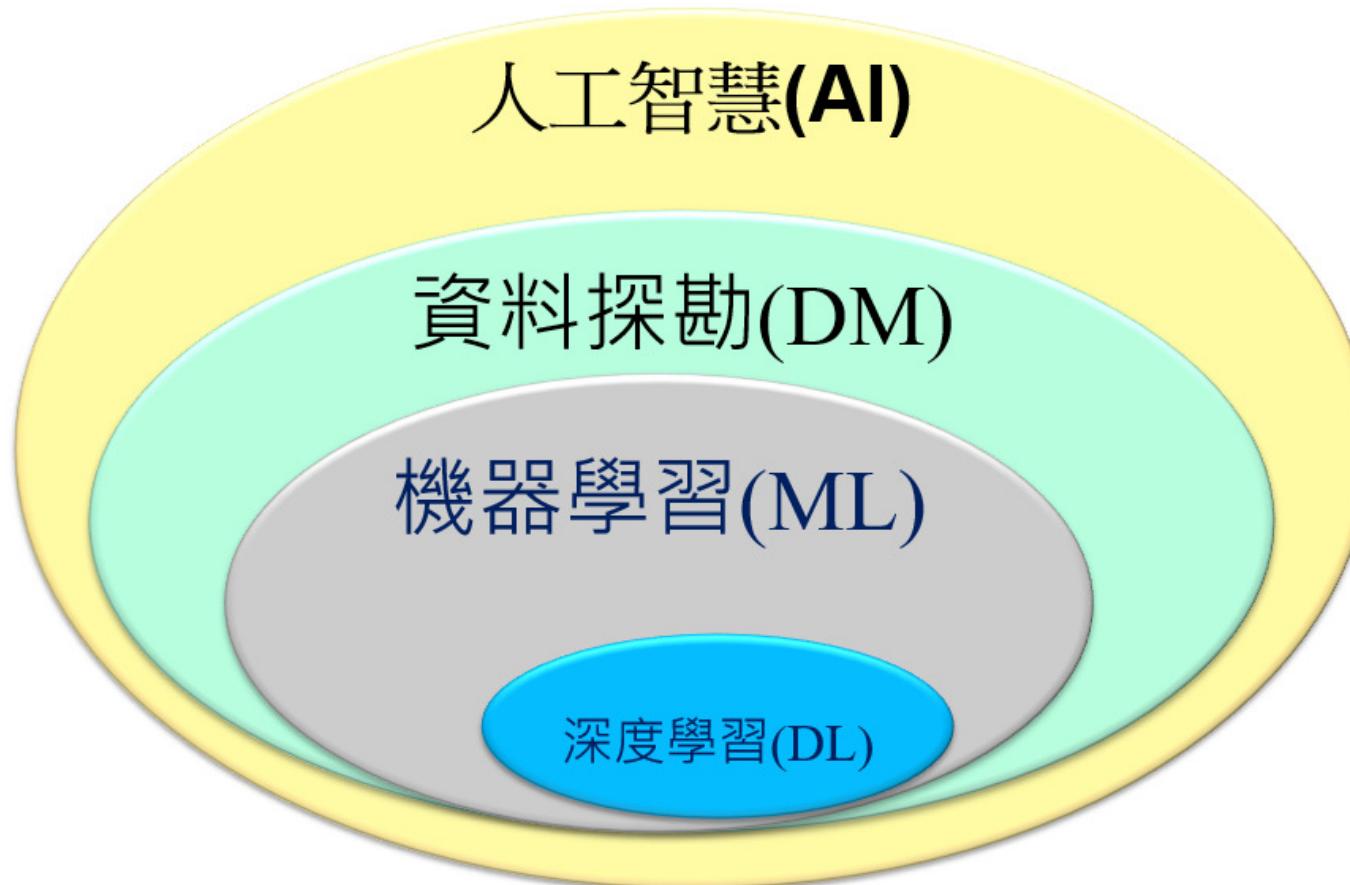
- Arduino + Python tkinter LED應用
- <https://youtu.be/LjgFIm1S7tw>
- <http://rwepa.blogspot.com/2021/05/arduino-python-tkinter-led.html>



# 大數據分析與應用

---

# 深度學習的發展



# 深度學習發展史



- 1943年：美國數學家 Walter Pitts 和心理學家 Warren McCulloch 提出人工神經元。
- 1957年：美國心理學家 Frank Rosenblatt 提出了感知器(Perceptron)。
- 1980年：多層類神經網路失敗，淺層機器學習方法(SVM等)興起。
- 2006年：Geoffrey Hinton 成功訓練多層神經網路(限制玻爾茲曼機, RBM)，命名為深度學習。
- 2012年：ImageNet 比賽讓深度學習重回學界視野，開啟 NVIDIA GPU 為重要運算硬體。

# 機器學習 Machine learning

- 非監督式學習 (**Unsupervised learning**)
  - No label or target value given for the data
- 監督式學習 (**Supervised learning**)
  - Telling the algorithm what to predict
- 半監督學習 (**Semi-supervised learning**)
  - 具有少量標記資料
- 強化學習 (**Reinforcement learning**)
  - 為了達成目標，隨著環境的變動，而逐步調整其行為，並評估每一個行動之後所到的回饋是正向的或負向的。
- 深度學習 (**Deep learning**)



# 監督式學習 vs. 非監督式學習

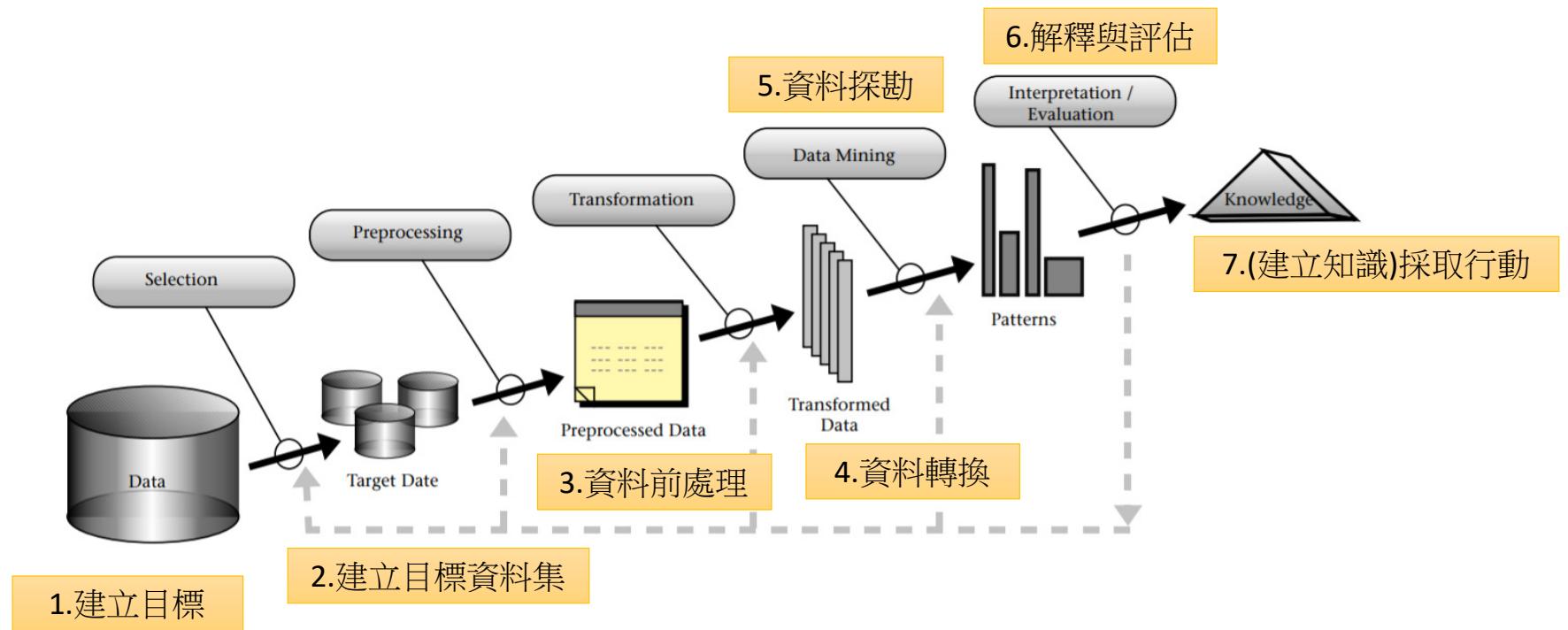
- 非監督式學習 **Unsupervised learning**
  - 集群法 Clustering
  - 關聯規則 Association rule
  - 主成分分析 Principal Component Analysis
- 監督式學習 **Supervised learning** (執行  $X \rightarrow$  預測  $\rightarrow Y$ ): 分類與數值預測
  - 迴歸分析 Regression analysis
  - 廣義線性模型 General linear model (GLM)
  - 天真貝氏法 Naïve-Bayes
  - K近鄰法 k-nearest neighbors (KNN)
  - 決策樹 Decision tree
  - 支持向量機 Support vector machine (SVM)
  - 類神經網路 Neural network (NN)
  - 集成學習 Ensemble learning: 使用多種學習算法來獲得比單獨使用演算法更好預測結果

# KDD 簡介

---

# 資料庫中的知識發掘 (Knowledge Discovery in Database, KDD)

- Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth , *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, August 1996 Pages 82–88. <https://www.kdnuggets.com/gspubs/aimag-kdd-overview-1996-Fayyad.pdf>



# 思考題

以下為KDD的個程序：1.解釋與評估，2.資料前處理，3.建立目標資料集，4.資料轉換，5.採取行動，6.訂定目標，7.資料探勘，請排序之？

- (A) 1234567
- (B) 6324715
- (C) 6342517
- (D) 6342175

答案:B

KDD 七大步驟：

- 1. 訂定目標
- 2. 建立目標資料集
- 3. 資料前處理
- 4. 資料轉換
- 5. 資料探勘
- 6. 解釋與評估
- 7. 採取行動

# CRISP-DM 簡介

---

# 資料探勘生命週期—CRISP-DM

- 跨產業資料探勘標準作業流程  
(CRoss Industry Standard Process for Data Mining)
- CRISP-DM是於1990年起，由SPSS以及NCR兩大廠商在合作戴姆克萊斯勒-賓士(Daimler Benz)的資料倉儲以及資料探勘過程中發展出來的。

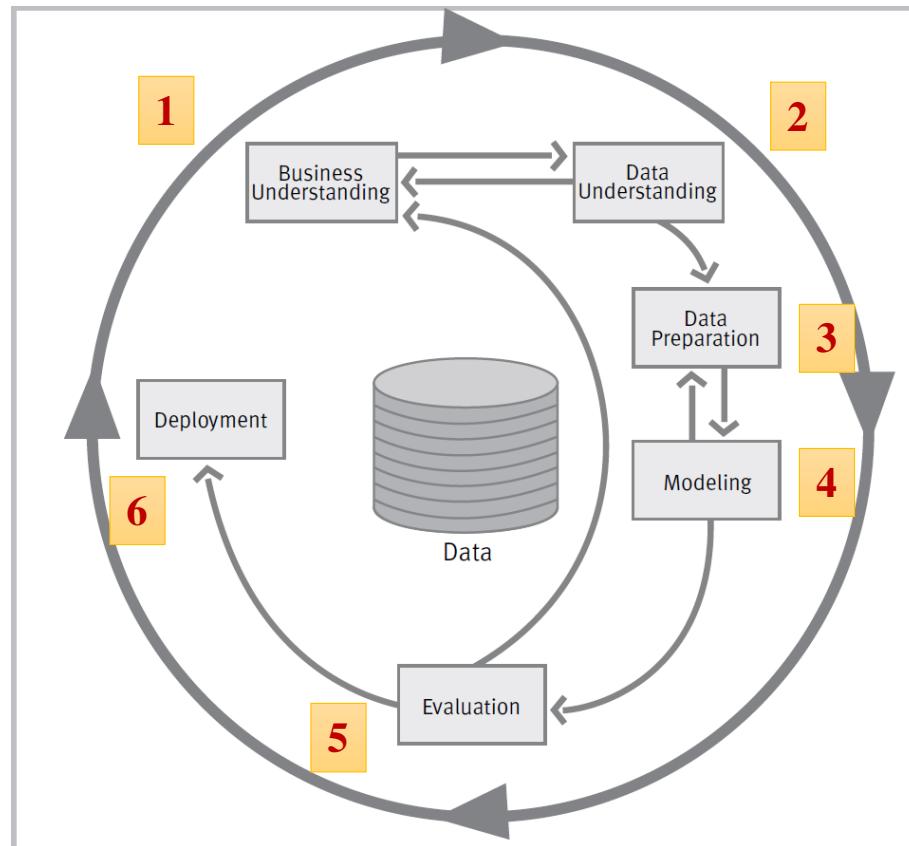
# CRISP-DM 資料探勘流程(續)

- 步驟 1：商業理解
- 步驟 2：資料理解
- 步驟 3：資料準備
- 步驟 4：模式建立
- 步驟 5：評估與測試
- 步驟 6：佈署應用

佔整專案時間的~80%

- 訓練資料70%
- 測試資料30%

# CRISP-DM 資料探勘流程(續)



參考 [https://en.wikipedia.org/wiki/Cross\\_Industry\\_Standard\\_Process\\_for\\_Data\\_Mining](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)

# 數值模型績效指標

- 不可直接使用誤差的算術平均!

$$\cancel{\text{Total error}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$

- 均方誤差 (Mean Squared Error, MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- 均方根誤差 (Root Mean Squared Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- 平均絕對誤差 (Mean Absolute Error, MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

# 類別模型績效指標 – 混淆矩陣

- <http://rwepa.blogspot.com/2013/01/rocr-roc-curve.html>

```
| 真實P類別 真實N類別

預測P類別 | TP真陽數 FP假陽數
預測N類別 | FN假陰數 TN真陰數

| P N

1.TPR(True positive rate) 真陽性率, 愈大愈好 -----
=TP/ (TP+FN)
=TP/ P
=Sensitivity 瞩敏度
=Recall 召回率
=Probability of detection
=Power
實際為陽性的樣本中，判斷為陽性的比例。
例如真正有生病的人中，被醫院判斷為有生病者的比例。
```

# 集群法

---

# 集群法 (Clustering)

- 集群法或稱為聚類分析,集群分析(Cluster analysis),叢集分析:是一種物以類聚方法.
- 每個集群的相似性是以資料間的距離來判斷.
- 分組後在同一集群組內的樣本點具有高度的相似性.
- 不同群組間的樣本點則具有高度的異質性.
- 集群法屬於非監督式學習法(Unsupervised learning),即資料沒有標籤(unlabeled data).
- 無法藉由的反應變數(Response variable, Y)來做分類之訓練.
- 因為資料沒有標籤,與監督式學習法不同,非監督式學習法較無法衡量演算法的正確率.

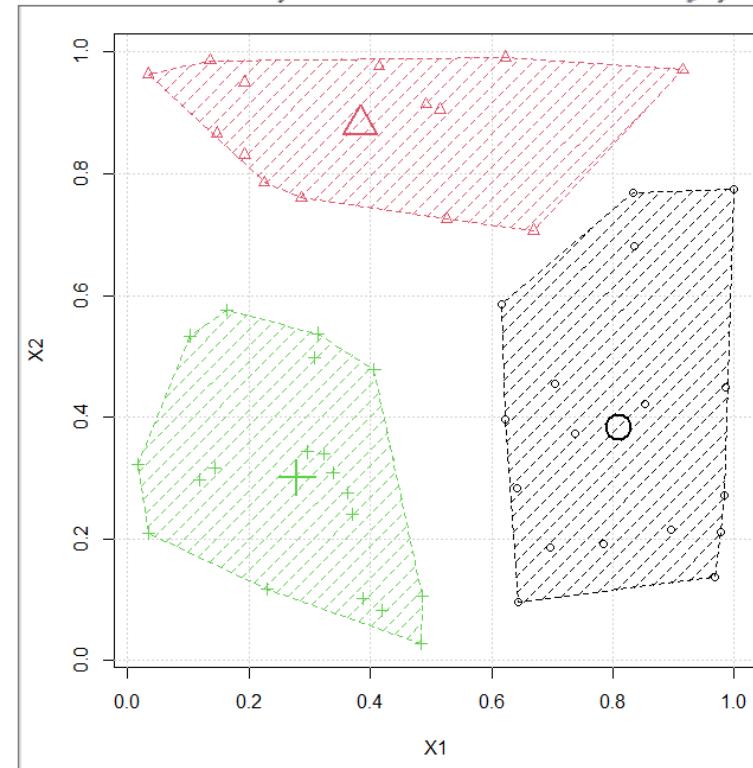
# 集群法 (續)

- 集群法的二大類別
  - 非階層式集群法(或切割式集群法 **Partitional clustering**)
    - K平均集群法(K-means集群法)，K表示集群個數，須提供給演算法。
    - K-medoid集群法,使用分割環繞物件法 (**Partition Around Medoid, PAM**)
  - 階層式集群法(**Hierarchical clustering**)
    - 凝聚階層法(**agglomerative hierarchical**)-由下往上
    - 分割階層法(**divisive hierarchical**)-由上往下

# animation 套件

```
library(animation)
```

```
kmeans.ani(x = cbind(x1 = runif(50), x2 = runif(50)),
 centers = 3,
 hints = c("Move centers!", "Find cluster?"),
 pch = 1:3,
 col = 1:3)
```



# K-means集群法實施步驟

1. 先選定**集群個數(k)**,或依過去實務經驗,可以選定 $k=2,3,4,5\dots$
2. 隨機給定k個資料點作為k個集群的中心(簡稱群心).
3. 將所有資料點指派至**距離最近**的群心所在的集群.
4. 重新更新k個集群的群心.
5. 重複3-4步驟,直到所有群心沒有太大的變動(或是收斂至事先約定條件),則結束整個演算法.

## 思考題

以 K-means 進行集群分析時，K值表示什麼？

- (A) 樣本數
- (B) 集群距離
- (C) 集群數
- (D) 整體平均值

答案:C

# 思考題

以下為KDD的5個步驟：1.選取中心點，2.計算距離與集群分配，3.確認集群收斂情況，4.設定K值，5.中心點調整，請排序之？

- (A) 12345
- (B) 42315
- (C) 14325
- (D) 41235 (最適當)

答案：

K-means步驟：

- 1. 設定K值
- 2. 選取中心點
- 3. 計算距離與集群分配
- 4. 中心點調整
- 5. 確有集群收斂情況

# K-近鄰法 (K最鄰近法)

---

# K-近鄰法 (K-Nearest Neighbors)

- 在圖型識別領域中，最近鄰居法（**KNN**演算法，**K**-近鄰演算法）是一種用於分類和迴歸的無母數統計方法。
- 在這兩種情況下，輸入包含特徵空間（**Feature Space**）中的**k**個最接近的訓練樣本。
  - 在**k-NN**分類中，輸出是一個分類族群。一個物件的分類是由其鄰居的「**多數表決**」確定的，**k**個最近鄰居（**k**為正整數，通常較小）中最常見的分類決定了賦予該物件的類別。若**k = 1**，則該物件的類別直接由最近的一個節點賦予。
  - 在**k-NN**迴歸中，輸出是該物件的屬性值。該值是其**k**個最近鄰居的值的**平均值**。

# K-近鄰法

- 步驟1 決定參數K(最鄰近物件的個數)。
- 步驟2 使用屬性資料，計算未知物件與K個訓練資料的距離總和。
- 步驟3 依距離排序，最小距離的K個最鄰近物件。
- 步驟4 依據K, 決定分類問題的多數類別，或K個數值的平均值。

# KNN demo

```
library(animation)

設定動畫參數
ani.options(interval = 1, nmax = 10)

建立訓練集, 測試集
set.seed(168)
df <- iris[iris$Species != "setosa",]
df$Species <- factor(df$Species)
ind <- sample(2, nrow(df), replace = TRUE, prob = c(0.8, 0.2))
traindata <- df[ind == 1, 3:4]
testdata <- df[ind == 2, 3:4]

KNN示範
knn.ani(train = traindata, test = testdata, cl = df$Species[ind == 1], k = 20)
```

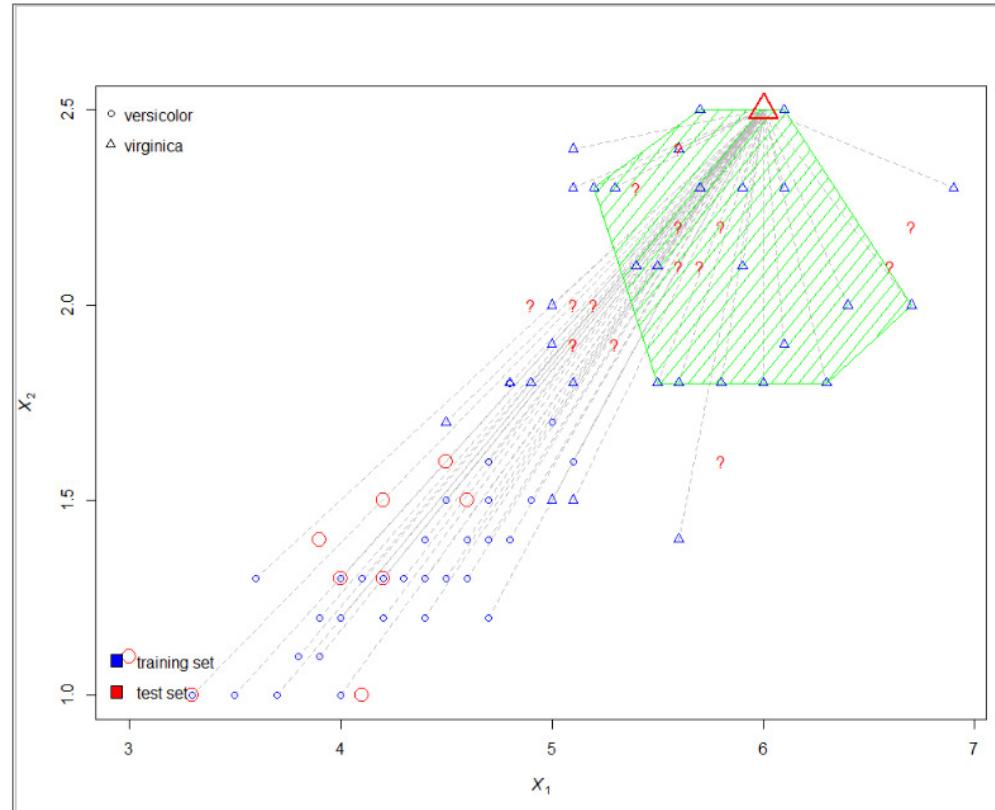
1

2

3

4

# KNN demo (續)



# 決策樹

---

# 決策樹

- 決策樹是一種類似流程圖的樹狀結構，包含：
  - 最上層：根節點(**Root node**)
  - 中間層：節點(**Node**)
  - 最底層：葉節點(**Leaf node**)，顯示分類／預測結果
- 每一節點表示一個屬性分類的測試條件，如同「**IF-THEN**」的控制結構，每個分支表示測試結果，並依此決定資料將分類於此節點的那一棵子樹(**Branch**)，並繼續作為分類的條件和最後的決策。

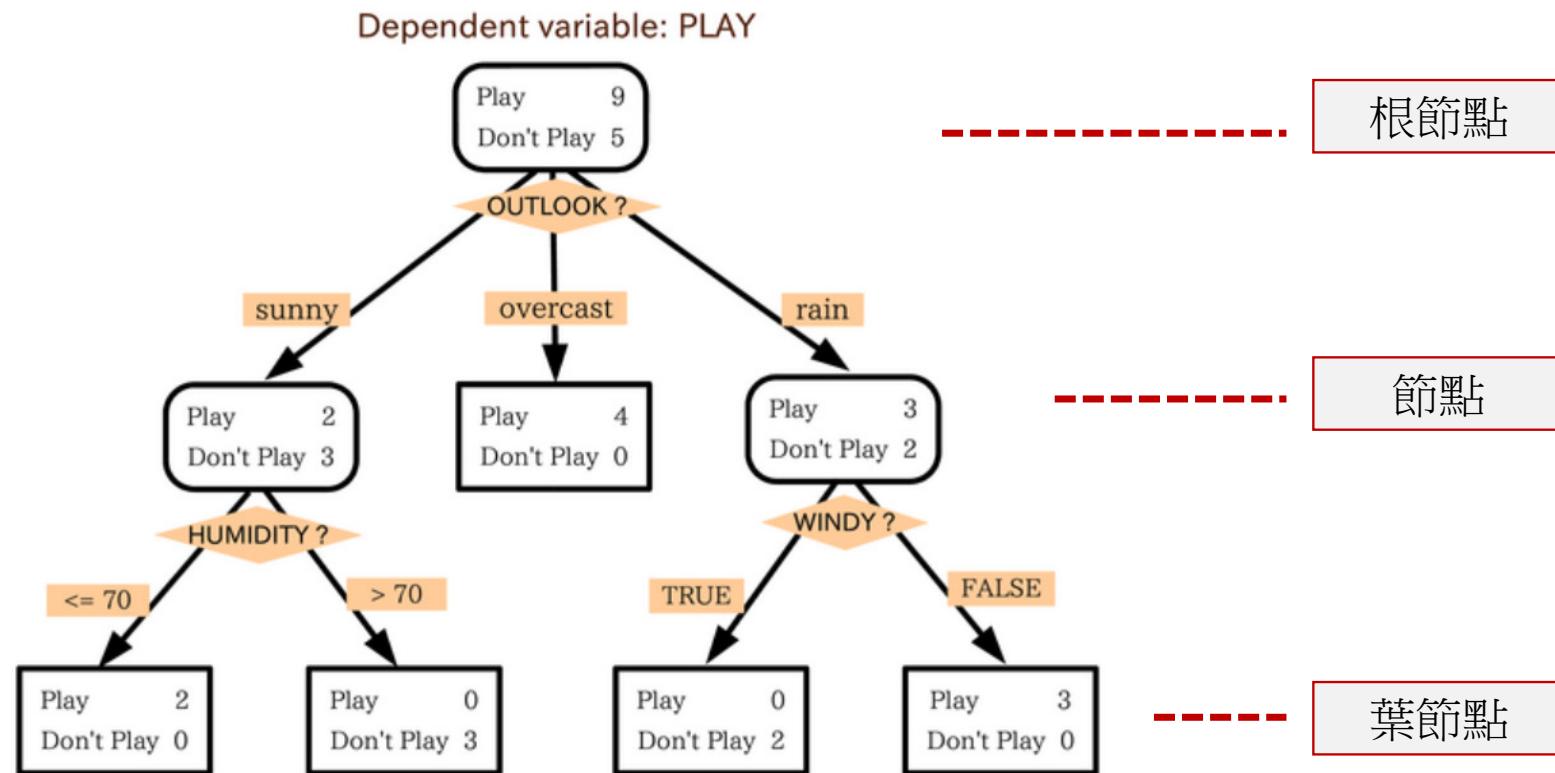
# 預測明天是否會玩高爾夫

| Independent variables |             |          |       |            | Dep. var |
|-----------------------|-------------|----------|-------|------------|----------|
| OUTLOOK               | TEMPERATURE | HUMIDITY | WINDY | PLAY       |          |
| sunny                 | 85          | 85       | FALSE | Don't Play |          |
| sunny                 | 80          | 90       | TRUE  | Don't Play |          |
| overcast              | 83          | 78       | FALSE | Play       |          |
| rain                  | 70          | 96       | FALSE | Play       |          |
| rain                  | 68          | 80       | FALSE | Play       |          |
| rain                  | 65          | 70       | TRUE  | Don't Play |          |
| overcast              | 64          | 65       | TRUE  | Play       |          |
| sunny                 | 72          | 95       | FALSE | Don't Play |          |
| sunny                 | 69          | 70       | FALSE | Play       |          |
| rain                  | 75          | 80       | FALSE | Play       |          |
| sunny                 | 75          | 70       | TRUE  | Play       |          |
| overcast              | 72          | 90       | TRUE  | Play       |          |
| overcast              | 81          | 75       | FALSE | Play       |          |
| rain                  | 71          | 80       | TRUE  | Don't Play |          |

- X: 4個
- Y: 1個
- 14列, 5行
- Y: Play 9  
Y: Don't 5

Source: <https://zh.wikipedia.org/wiki/決策樹>

# 決策樹結構圖



# 決策樹建立流程

- 資料準備
- 建立決策樹
- 選取決策樹演算法
- 決策樹修剪
- 萃取分類規則

# 建立決策樹

- 步驟1: 將原始資料分成兩組：訓練集、測試集。
- 步驟2: 將訓練集放入決策樹的樹根。
- 步驟3: 使用訓練集來建立決策樹，而在每一個內部節點，則依據資訊理論(Information Theory)來評估選擇哪個屬性繼續做分支的依據。
- 步驟4: 進行決策樹修剪(事前/事後)，以提升預測能力與速度。
- 將以上(1)-(4)步驟不斷遞迴進行，直到所有的新內部節點都是樹葉節點為止。

# 決策樹停止條件

- 決策樹停止情形
  - 該群資料中，每一筆資料都已經歸類在同一類別下。
  - 該群資料中，已經沒有辦法再找到新的屬性來進行節點分割。
  - 該群資料中，已經沒有任何尚未處理的資料。
  - 滿足使用者定義的停上條件

# 常用的屬性選擇指標

- 資訊獲利 (Information Gain) [熵愈小，獲利愈大]
  - ID3
  - C4.5
  - C5.0
- 吉尼係數 (Gini Index) – CART [吉尼係數愈小者]
- $\chi^2$ 獨立性檢定 – CHAID [卡方統計量愈大者]

# 決策樹演算法

- ID3 (Iterative Dichotomizer 3, 叠代二元樹第3代, Quinlan, 1979)
  - 可處理離散型資料。
  - 兼顧高分類正確率以及降低決策樹的複雜度。
  - 必須將連續型資料作離散化的程序。
- C4.5 (Quinlan, 1993)
  - 改良自ID3演算法。
  - 先建構一顆完整的決策樹，再針對每一個內部節點，依使用者定義的預估錯誤率(Predicted Error Rate)來作決策樹修剪的動作。
  - 不同的節點，特徵值離散化結果是不相同的。

# 決策樹演算法 (續)

- CART (Classification and Regression Trees, Breiman, 1984)
  - 是以每個節點的動態臨界值作為條件判斷式。
  - CART藉由單一輸入的變數函數，在每個節點分隔資料，並建立一個二元決策樹。
  - CART是使用 Gini Ratio來衡量指標，如果分散的指標程度很高，表示資料中分佈許多類別，相反的，如果指標程度越低，則代表單一類別的成員居多。

# 決策樹演算法 (續)

- CHAID (Chi-Square Automatic Interaction Detector, Gordon, 1980)
  - 利用卡方分析(Chi-Square Test)預測二個變數是否需要合併，如能夠產生最大的類別差異的預測變數，將成為節點的分隔變數。
  - 計算節點中類別的 P 值 (P-Value)，以 P 值大小來決定決策樹是否繼續生長，所以不需像 C4.5 或 CART 要再做決策樹修剪的動作。

# 思考題

下列何者屬於監督式學習（supervised learning）演算法？

- (A) 主成分分析（Principal Component Analysis, PCA）
- (B) K平均法（k-means）
- (C) 集群分析（clustering analysis）
- (D) 決策樹（decision tree）

- A. 主成分分析：維度縮減技術。
- B. k 平均法：集群分析之一，非監督式學習。
- C. 集群分析：非監督式學習。
- D. 決策樹：監督式學習。

# 思考題

下列哪一個是分類技術的學習演算法？

- (A) 決策樹
- (B) Apriori 演算法
- (C) k平均法
- (D) 線性迴歸

答案:A

- A. 決策樹：監督式學習一分類法
- B. Apriori 演算法 - 關聯規則：非監督式學習
- C. K平均法：非監督式學習
- D. 線性迴歸：監督式學習一迴歸法

# 思考題

下列哪一個資料探勘的應用不屬於分類技術？

- (A) 判別電子郵件是否為垃圾郵件
- (B) 判別信用卡客戶是否為如期還款
- (C) 判別哪些商品常常被一起購買
- (D) 判別 X 光片的細胞是否異常

答案:C

- A. 分類法 {有，沒有}
- B. 分類法 {有，沒有}
- C. 關聯規則：非監督式學習
- D. 分類法 {有，沒有}

## 思考題

在分類技術中，若需要建立分類模型通常會把資料分割成訓練資料和測試資料，請問目的為何？

- (A) 可以利用訓練資料和測試資料建立不同的分類模型
- (B) 使用訓練資料建立的多個分類模型可以利用測試資料評估優劣
- (C) 分類學習演算法需要兩份資料才能建立分類模型
- (D) 測試資料可以驗證訓練資料是否有問題

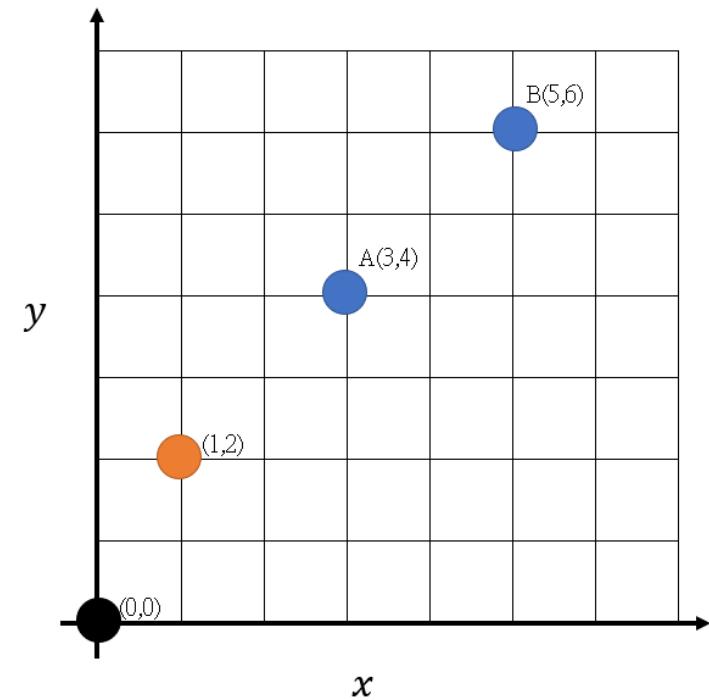
答案:D

# 思考題

已知類別 A 的屬性為(3, 4)，類別 B 的屬性為(5, 6,)，根據 K最鄰近法的類別判別方式並利用歐式距離當作距離衡量指標，請問屬性 (1, 2) 的資料應該屬於哪一個類別？

- (A) 類別 A
- (B) 類別 B
- (C) 到類別 A 和類別 B 距離一樣，所以兩個類別皆對
- (D) 到類別 A 和類別 B 距離一樣，所以無法判斷

答案:A



## 思考題

K最鄰近法的特點不包含下列哪一項？

- (A) 不需要事先建立分類模
- (B) 學習演算法的運算成本過高
- (C) 可以避免類別分布不平衡的問題
- (D) K值的決定不會影響結果

答案:D

K值的決定會影響結果

## 思考題

建立決策樹的分類模型需要衡量亂度，請問下列哪一個為亂度的衡量指標？

- (A) 歐氏距離
- (B) 熵 (Entropy)
- (C) 支持度 (Support)
- (D) 精確度 (Precision)

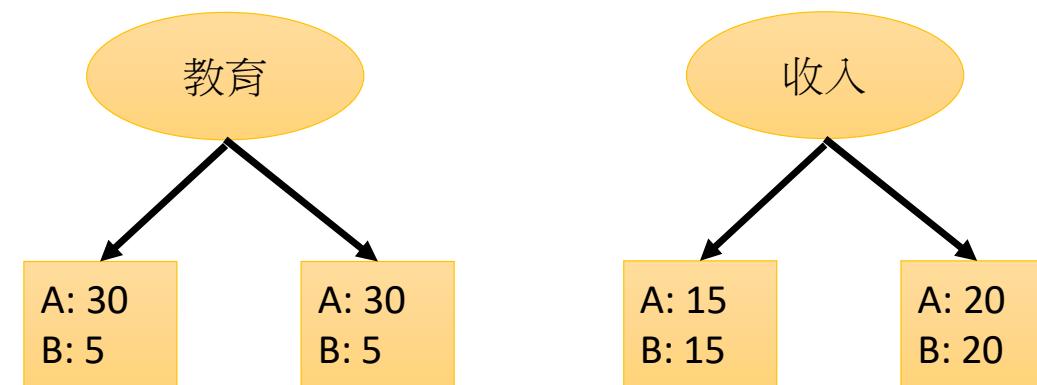
答案:B

# 思考題

使用教育與收入兩個屬性建構的決策樹如下，依據亂度衡量的觀點，試問哪一個屬性建構的結果較佳？

- (A) 教育
- (B) 收入
- (C) 兩個屬性一樣好
- (D) 無法判斷

答案:A



## 思考題

下列哪一個不是決策樹停止繼續分割的條件？

- (A) 分割內都是同一個類別
- (B) 分割內的屬性值皆一樣
- (C) 分割內的每個類別個數都一樣
- (D) 分割內有子樹是空的

答案:C

- 決策樹停止繼續分割的條件：
  - 1. 分割內都是同一個類別
  - 2. 分割內的屬性值皆一樣
  - 3. 分割內有子樹是空的

# 決策樹亂度的衡量

- 亂度表示決策樹分割後，同一分割中類別的不一致程度。
- 亂度越高，表示分割結果類別越不一致。亂度越低，表示分割結果類別越一致。
- 選取屬性可以使資料的亂度降低最多者為分割屬性。
- 分割前後亂度的差異稱為增益量(Gain),  $Gain(A) = I(D) - I_A(D)$ ,  $A$ 表示屬性， $I(D)$ 與 $I_A(D)$ 表示分割前後之亂度，選增益量較大的分割。
- 亂度的測量指標對決策機分類結果影響不大。
- 考慮亂度的衡量中， $k$ 為類別數， $p_i$ 第*i*個類別所佔比例
  1. Entropy 熵 =  $-\sum_{i=1}^k p_i \log_2 P_i$
  2. Gini係數 =  $1 - \sum_{i=1}^k p_i^2$

## 思考題

下列哪一個不是決策樹的特點？

- (A) 決策樹的建置對資料的分配沒有要求
- (B) 決策樹的建置成本低、建置速度快
- (C) 小型的決策樹容易對模型進行解釋
- (D) 亂度的測量指標影響很大，需依據資料選擇

答案:D

# 分類範例-銀行貸款是否會如期還款

|    | A  | B   | C 1  | D 2 | E 3 | F 4 | G      |
|----|----|-----|------|-----|-----|-----|--------|
| 1  | 編號 | 資料集 | 是否負債 | 性別  | 婚姻  | 收入  | 是否如期還款 |
| 2  | 1  | 訓練集 | 是    | 男   | 未婚  | 低   | 否      |
| 3  | 2  | 訓練集 | 否    | 女   | 未婚  | 低   | 否      |
| 4  | 3  | 訓練集 | 是    | 男   | 未婚  | 高   | 是      |
| 5  | 4  | 訓練集 | 否    | 女   | 結婚  | 低   | 是      |
| 6  | 5  | 訓練集 | 否    | 男   | 未婚  | 高   | 是      |
| 7  | 6  | 訓練集 | 是    | 女   | 未婚  | 高   | 否      |
| 8  | 7  | 訓練集 | 否    | 女   | 結婚  | 低   | 是      |
| 9  | 8  | 訓練集 | 是    | 男   | 結婚  | 高   | 否      |
| 10 | 9  | 訓練集 | 否    | 男   | 未婚  | 低   | 是      |
| 11 | 10 | 測試集 | 是    | 女   | 結婚  | 低   | 否      |
| 12 | 11 | 測試集 | 否    | 女   | 結婚  | 高   | 是      |
| 13 | 12 | 測試集 | 是    | 男   | 結婚  | 高   | 否      |

訓練集 9個  
測試集 3個  
屬性 4個

# 決策樹-使用 Gini 系數

訓練集 9個  
是: 5  
否: 4

- 步驟1 計算訓練集的亂度  $I(D) = 1 - \sum_{i=1}^k p_i^2 = 1 - \left(\frac{5}{9}\right)^2 - \left(\frac{4}{9}\right)^2 = 0.494$
- 步驟2 計算“是否負債”增益量

$$Gini_{\text{負債}=\text{是}} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini_{\text{負債}=否} = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.32$$

$$I_{\text{負債}}(D) = \frac{4}{9} \times 0.375 + \frac{5}{9} \times 0.32 = 0.344$$

$$Gain_{\text{負債}} = 0.494 - 0.344 = 0.15$$

是否負債  
是: 4(還款Y:1,N:3)  
否: 5(還款Y:4,N:1)

# 決策樹-使用 Gini 系數

- 步驟3 計算”性別”增益量

$$Gini_{\text{性別}=\text{男}} = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$Gini_{\text{性別}=\text{女}} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$I_{\text{性別}}(D) = \frac{5}{9} \times 0.48 + \frac{4}{9} \times 0.5 = 0.489$$

$$Gain_{\text{性別}} = 0.494 - 0.489 = 0.005$$

性別  
男: 5(還款Y:3,N:2)  
女: 4(還款Y:2,N:2)

# 決策樹-使用 Gini 系數

- 步驟4 計算”婚姻”增益量

$$Gini_{\text{婚姻}=\text{結婚}} = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$$

$$Gini_{\text{婚姻}=\text{未婚}} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$I_{\text{婚姻}}(D) = \frac{3}{9} \times 0.444 + \frac{6}{9} \times 0.5 = 0.481$$

$$Gain_{\text{婚姻}} = 0.494 - 0.481 = 0.011$$

婚姻  
結婚: 3(還款Y:2,N:1)  
未婚: 6(還款Y:3,N:3)

# 決策樹-使用 Gini 系數

- 步驟5 計算”收入”增益量

$$Gini_{\text{收入=高}} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$Gini_{\text{收入=低}} = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

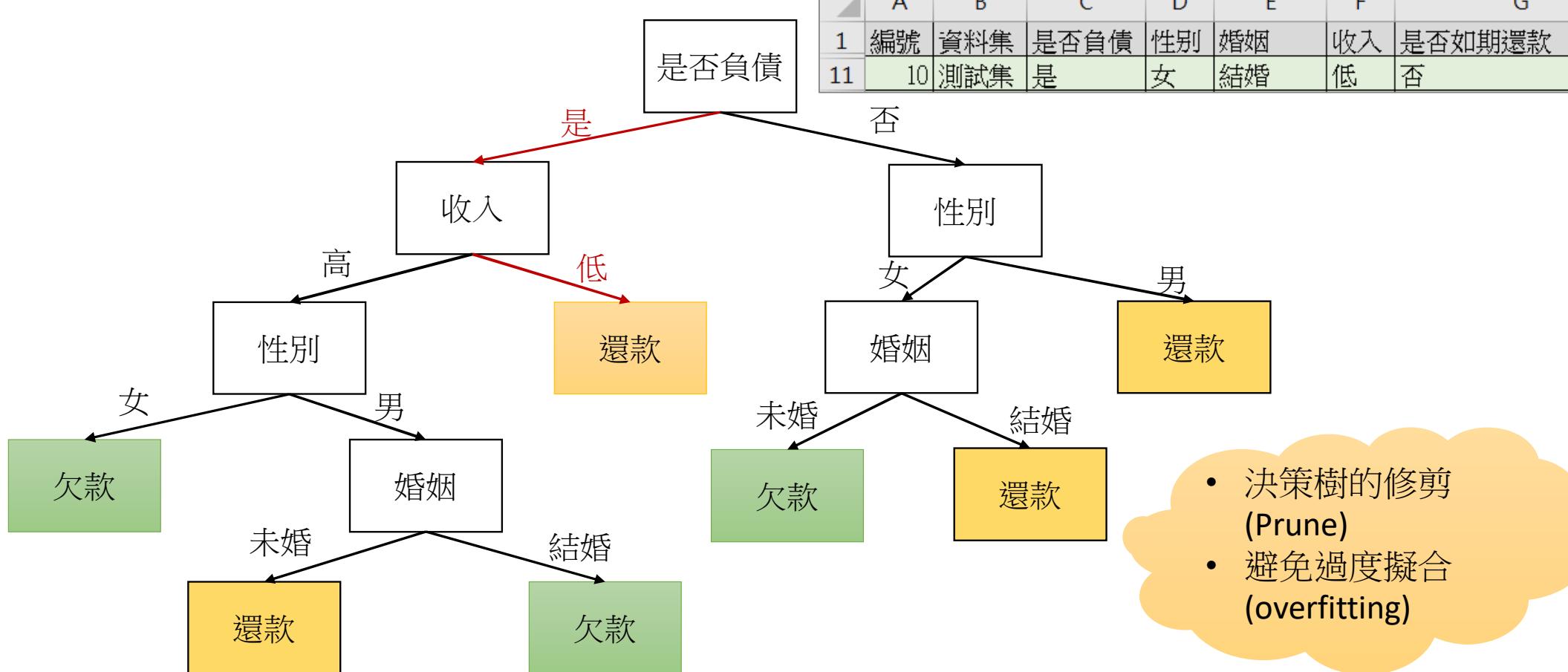
$$I_{\text{收入}}(D) = \frac{4}{9} \times 0.5 + \frac{5}{9} \times 0.48 = 0.489$$

$$Gain_{\text{收入}} = 0.494 - 0.489 = 0.005$$

收入  
高 : 4(還款Y:2,N:2)  
低 : 5(還款Y:3,N:2)

- 比較四個屬性的增益量，選擇”是否負債”作為第 1 層凡分割條件。

# 銀行貸款-決策樹

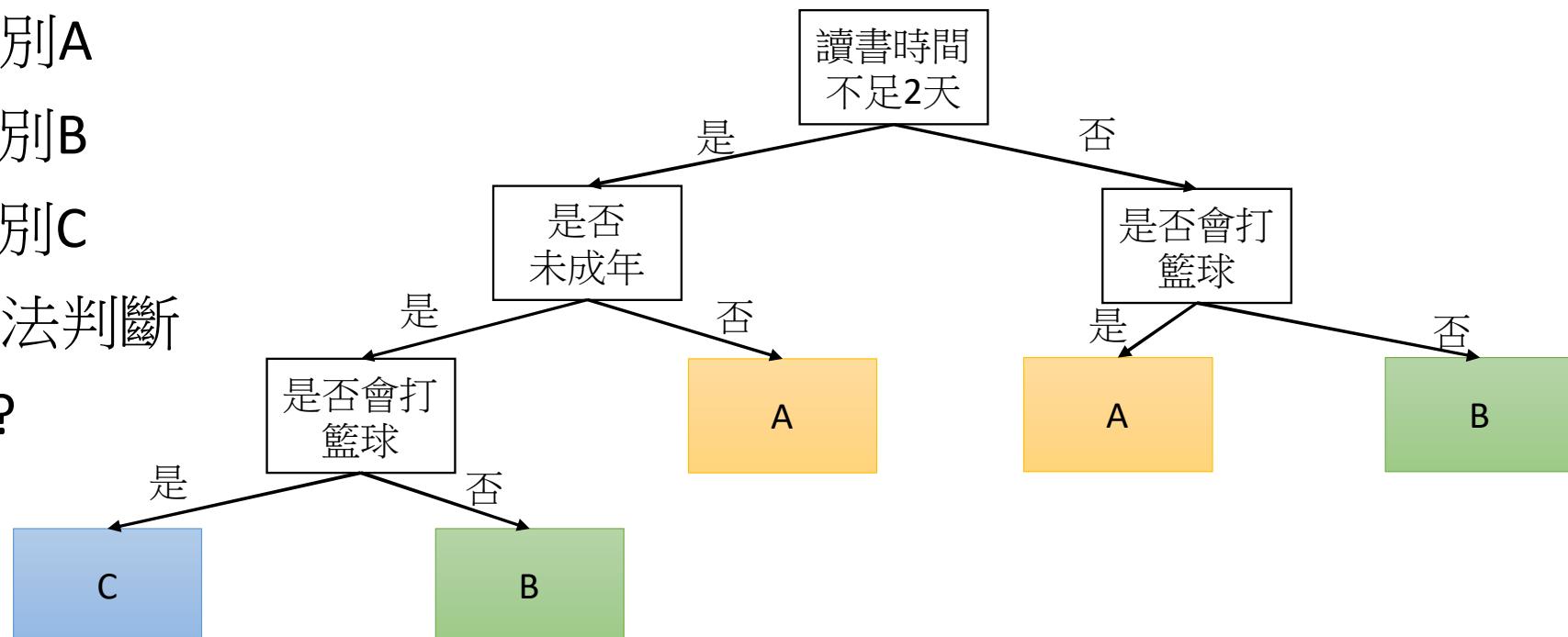


# 思考題

考慮下列決策樹，已知小明年齡是12歲，讀書時間為1天，不會打籃球，則小明應該是屬於A,B,C類別？

- (A) 類別A
- (B) 類別B
- (C) 類別C
- (D) 無法判斷

答案:?



# 思考題

考慮分類模型的混淆矩陣如下表所示，此模型精確度為何？

- (A) 66.67%
- (B) 30%
- (C) 80%
- (D) 70%

|      |     | 預測類別 |     |
|------|-----|------|-----|
|      |     | 類別1  | 類別0 |
| 實際類別 | 類別1 | 40   | 20  |
|      | 類別0 | 10   | 30  |

答案:D

$$\text{精確度} = \frac{40+30}{40+10+20+30} = 70\%$$

# 思考題

考慮分類模型的混淆矩陣如下表所示，以類別1為評估精確度的類別，則此模型的精確度為何？

- (A) 66.67%
- (B) 30%
- (C) 80%
- (D) 70%

|      |     | 預測類別 |     |
|------|-----|------|-----|
|      |     | 類別1  | 類別0 |
| 實際類別 | 類別1 | 40   | 20  |
|      | 類別0 | 10   | 30  |

答案:A

$$\text{精確度} = \frac{40}{40+20} = 66.67\%$$

# 關聯分析

---

# 項目集 Itemsets

- Itemsets (I)項目集
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- Support count ( $\sigma$ )支持數量
  - Frequency of occurrence of an itemset
  - E.g.  $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$
- Support 支持度
  - Fraction of transactions that contain an itemset
  - E.g.  $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$
- Frequent Itemset 頻繁項目集
  - An itemset whose support is greater than or equal to a minimum support threshold  
(大於或等於最小門檻值)

| TID | Items                     |
|-----|---------------------------|
| 1   | Bread, Milk               |
| 2   | Bread, Diaper, Beer, Eggs |
| 3   | Milk, Diaper, Beer, Coke  |
| 4   | Bread, Milk, Diaper, Beer |
| 5   | Bread, Milk, Diaper, Coke |

# 關聯規則及評估指標

$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

| TID | Items                     |
|-----|---------------------------|
| 1   | Bread, Milk               |
| 2   | Bread, Diaper, Beer, Eggs |
| 3   | Milk, Diaper, Beer, Coke  |
| 4   | Bread, Milk, Diaper, Beer |
| 5   | Bread, Milk, Diaper, Coke |

(信賴度)  $confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}$

# 關聯規則評估

- 關聯規則的評估指標
  - 支持度（Support）
    - 同時購買項目集X和Y的機率， $P(X \cup Y)$
  - 信賴度（Confidence）
    - 購買項目集X的情況下，也會購買項目集Y的機率， $P(Y|X)$
  - 增益值（lift）
    - 交易資料中，觀察到項目集X，再觀察到項目集Y的增益值為： $\frac{confidence(x \rightarrow Y)}{support(Y)}$

# 挑出有用的關聯規則

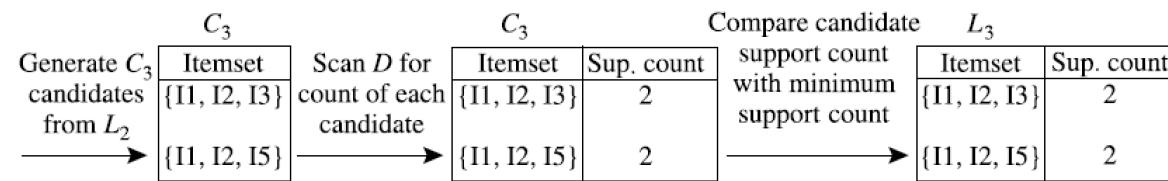
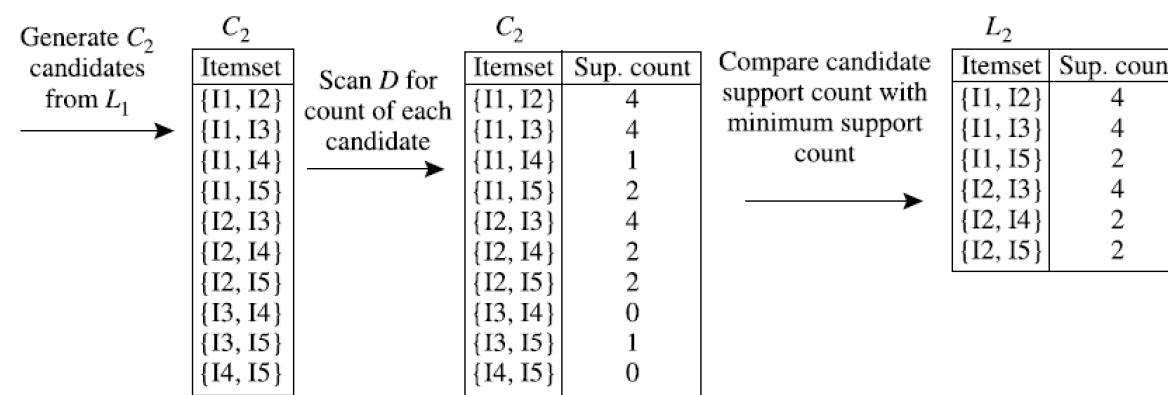
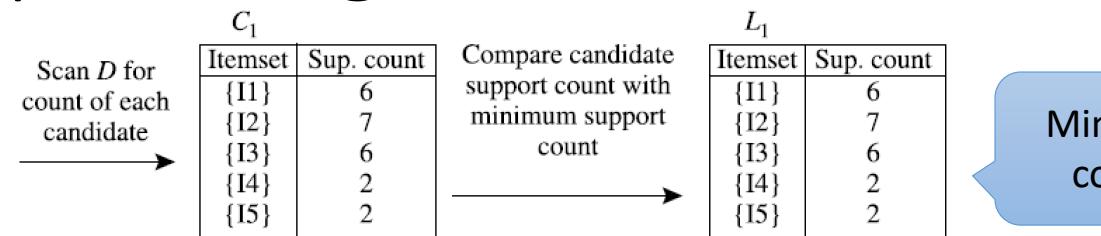
- 增益值  $> 1$ ，表示X與Y呈現正相關，規則才具有行銷實用性。
- 增益值 = 1，表示X與Y呈現正相關，結果與亂數取得方式相似。
- 增益值  $< 1$ ，表示X與Y呈現負相關，比亂數取得之結果更差。

# Apriori Algorithm

關聯規則中找出頻繁項目集的方法：Apriori演算法。

Transactional Data for an *AllElectronics* Branch

| TID  | List of item IDs |
|------|------------------|
| T100 | I1, I2, I5       |
| T200 | I2, I4           |
| T300 | I2, I3           |
| T400 | I1, I2, I4       |
| T500 | I1, I3           |
| T600 | I2, I3           |
| T700 | I1, I3           |
| T800 | I1, I2, I3, I5   |
| T900 | I1, I2, I3       |



Min support count = 2

## 思考題

平衡計分卡(Balanced Score Card)常用來衡量 BI 對企業的整體效益，但其可能因下列何種原因而失效

- (A) 度量方式可能因某種程度的量化本色而定義不正確
- (B) 過度依賴工具操作，使得員工喪失計算能力
- (C) 改善的結果往往落於操作層次
- (D) 企業主對於 BI 的不支持。

答案:C

- A. 質化本色
- B. 與題目無關
- C. 正確答案
- D. 與題目無關

## 思考題

「雖然企業有 BI，但我們公司主管們還是習慣以直覺來作決策」以上是有關影響 BI 成效因素中的

- (A) 資料因素
- (B) 流程因素
- (C) 組織文化因素
- (D) 制度化因素。

答案:C

## 思考題

根據調查資料顯示，企業 BI 的使用者族群人口最多是

- (A) 商業及財務分析人員
- (B) 總經理
- (C) 現場作業人員
- (D) 客戶

答案:A

## 思考題

下列哪一個選項是經由自動或半自動的方法尋找及分析大量的資料，以建立有效模式(Pattern) 及規則(Rule) 的方法。

- (A) 資料探勘
- (B) 資料庫
- (C) 系統分析
- (D) 程式設計

答案:A

## 思考題

下列哪一個選項是資料探勘領域中的重要議題，其基本概念為從大量交易中挖掘出項目(Item)之間價值的相關性，探索資料項目之間彼此蘊含的關係？

- (A) 分類法
- (B) 關聯規則
- (C) 集群法
- (D) 關聯規則

答案:D

## 思考題

「尿布(Diaper)」與「啤酒(Beer)」的故事是行銷界的經典神話，是屬於下列哪一種資料探勘領域技術的應用。

- (A) 集群法
- (B) 類神經網路
- (C) 關聯規則
- (D) OLAP & Cube查詢

答案:C

## 思考題

在使用關聯規則方法時頻繁項目集(Frequent Itemset)觀念甚為重要，  
Frequent又可稱為

- (A) Small
- (B) Large
- (C) Middle
- (D) Big

答案:B, Large表示出現機率很高

## 思考題

在零售產業應用中比較常聽到的購物籃分析(Market Basket Analysis, MBA), 是屬於下列哪一種技術。

- (A) 集群法
- (B) 類神經網路
- (C) 關聯規則
- (D) 貝氏分類法

答案:C

## 思考題

7. LOWS 零售店所有曾經賣過的商品項目的數量有3 種(X 、Y 、Z) , 這3種項目可能組合所產生的Itemset 的數量等於7, 包含有X 、Y 、Z 、XY 、XZ 、YZ 、XYZ 七種組合 , 假設某一零售店所有曾經賣過的商品項目的數量為N, 則這 N 個項目可能組合所產生的Itemset 的數量為

- (A)  $2^N - 1$
- (B)  $3 \times N$
- (C)  $e^N - 1$
- (D)  $N^2 - 5$

答案:A

## 思考題

在關聯規則分析時，如果場景是一個零售超市，某項目集 W 出現在幾台購物車中的觀念是指

- (A) 支持度 (Support)
- (B) 信賴度 (Confidence)
- (C) 支持數量 (Support count)
- (D) 提升度 (Lift)

答案:C

## 思考題

在關聯規則分析時，如果想了解某項目集Z 出現的次數占總交易資料庫的比例有多少，則可以透過下列哪一選項求出？

- (A) 提升度
- (B) 信賴度
- (C) 支持數量
- (D) 支持度

答案:D

## 思考題

當進行關聯規則分析時，交易資料庫中所有可能的Itemset 都可以透過查詢比對交易資料庫而求算出其支持數量 $\sigma$ 與支持度 $s$  兩項資訊，並提供作為後續產生關聯規則的基礎，下列有關於支持數量 $\sigma$ 與支持度 $s$ 的觀念何者錯誤？

- (A) 支持數量 $\sigma$  與支持度 $s$  的管理意涵是相同的
- (B) 支持數量 $\sigma$  與支持度 $s$  的管理意涵是不同的
- (C) 支持數量 $\sigma$  與支持度 $s$  的差別就在支持數量 $\sigma$  是以數量表示
- (D) 支持數量 $\sigma$  與支持度 $s$  的差別就在支持度 $s$  是以比例來表示

答案:B

# 思考題

有一個關聯規則  $R1: X \Rightarrow Y[s, c, Lift]$ ，下列哪一觀念正確？

- (A)  $Y$  稱為關聯規則  $R1$  的後繼項目集
- (B)  $Y$  稱為關聯規則  $R1$  的前導項目集
- (C)  $s$  稱為關聯規則  $R1$  的後繼項目集
- (D)  $Lift$  稱為關聯規則  $R1$  的前導項目集

答案:A

$R1: X \Rightarrow Y[s, c, Lift]$

X 前導項目集

Y 後繼項目集

## 思考題

有一個關聯規則  $R2: W \Rightarrow Z[s, c, Lift]$  已經被認定為有趣的強關聯規則，下列哪一觀念錯誤？

- (A)  $s$  必須大於等於最小支持度(minsup) 值
- (B)  $c$  必須大於等於最小信賴度(minconf) 值
- (C) Lift 值必須等於0
- (D) Lift 值必須大於1

答案:C

## 思考題

下列選項中哪一個是關聯規則探勘中找出頻繁項目集(Frequent Itemset)的方法？

- (A) Normalization 方法
- (B) ER Model 方法
- (C) SQL 查詢
- (D) Apriori 方法

答案:D

## 思考題

在關聯規則分析中使用 Apriori 方法時必須產生各個階段的候選項目集 (Itemset) , 而「產生」兩個字是指結合(Join) 以及哪一個動作 ?

- (A) 選擇 (Select)
- (B) 修剪 (Prune)
- (C) 授權 (Grant)
- (D) 插入 (Insert)

答案:B

## 思考題

有一個關聯規則  $R3: X \Rightarrow Y[s, c, Lift]$ ，下列哪一選項的意涵為某一條關聯規則在預測結果時研究是否能比隨機發生的機會好多少倍？

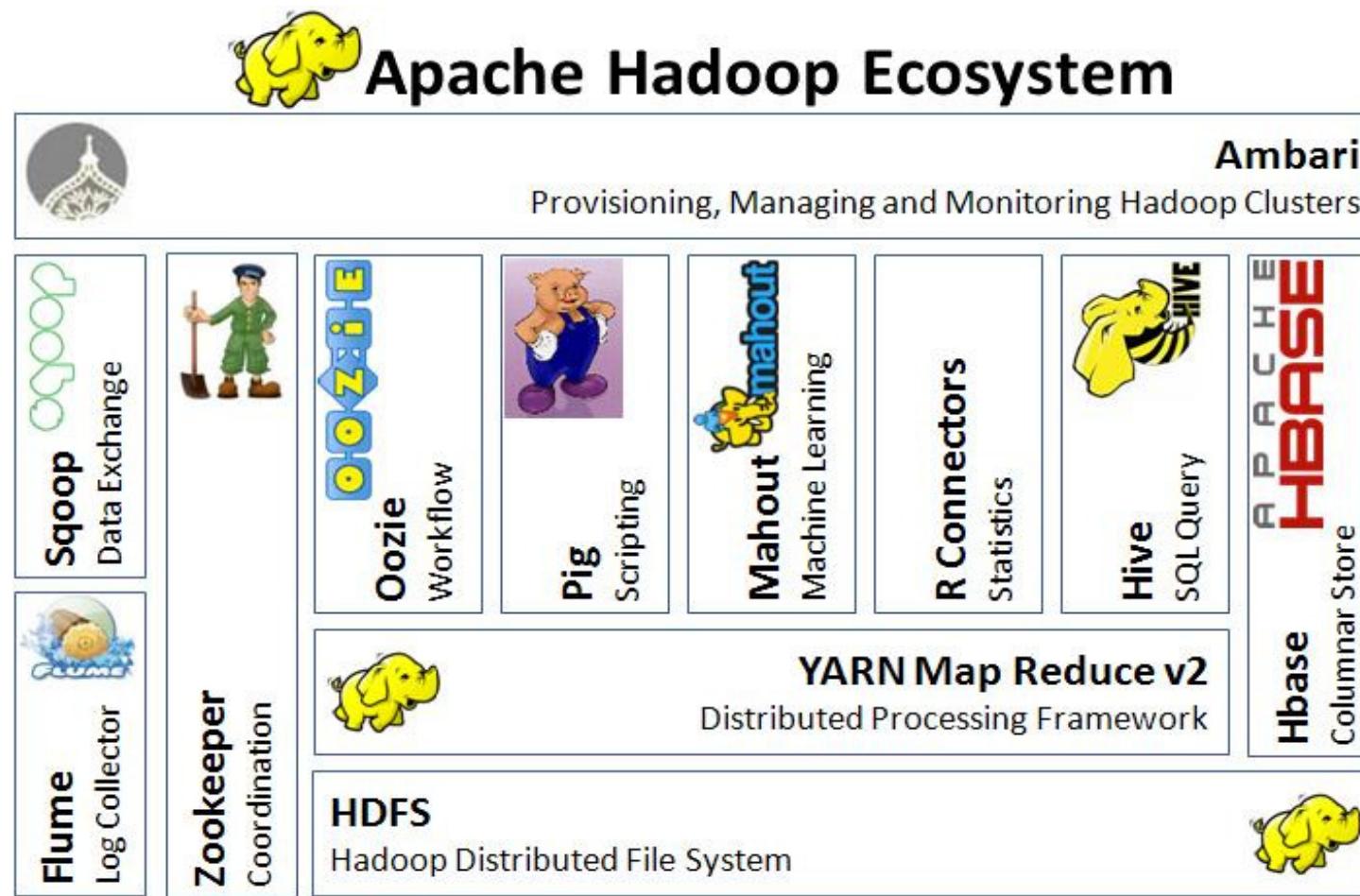
- (A) 提升度 Lift 值
- (B) 支持度 s 值
- (C) 信賴度 c 值
- (D) 支持數量 σ 值

答案:A

# Apache Hadoop

- **Apache Hadoop**: 開放源始碼軟體的集合(框架)，該集合使用許多計算機組成的網路來解決涉及大量數據和計算的問題。
- **HDFS**: Hadoop 的核心包括 Hadoop 分散式檔案系統（Hadoop Distributed File System, HDFS）
- **MapReduce** : Hadoop 以 MapReduce 做為大數據的分散式資料儲存和計算的軟體框架。HDFS 可作為MapReduce 編譯模型的部分組成。
- **HBase**: 運作在 HDFS 之上的分散式資料庫(NoSQL)

# Hadoop 生態圖譜系統

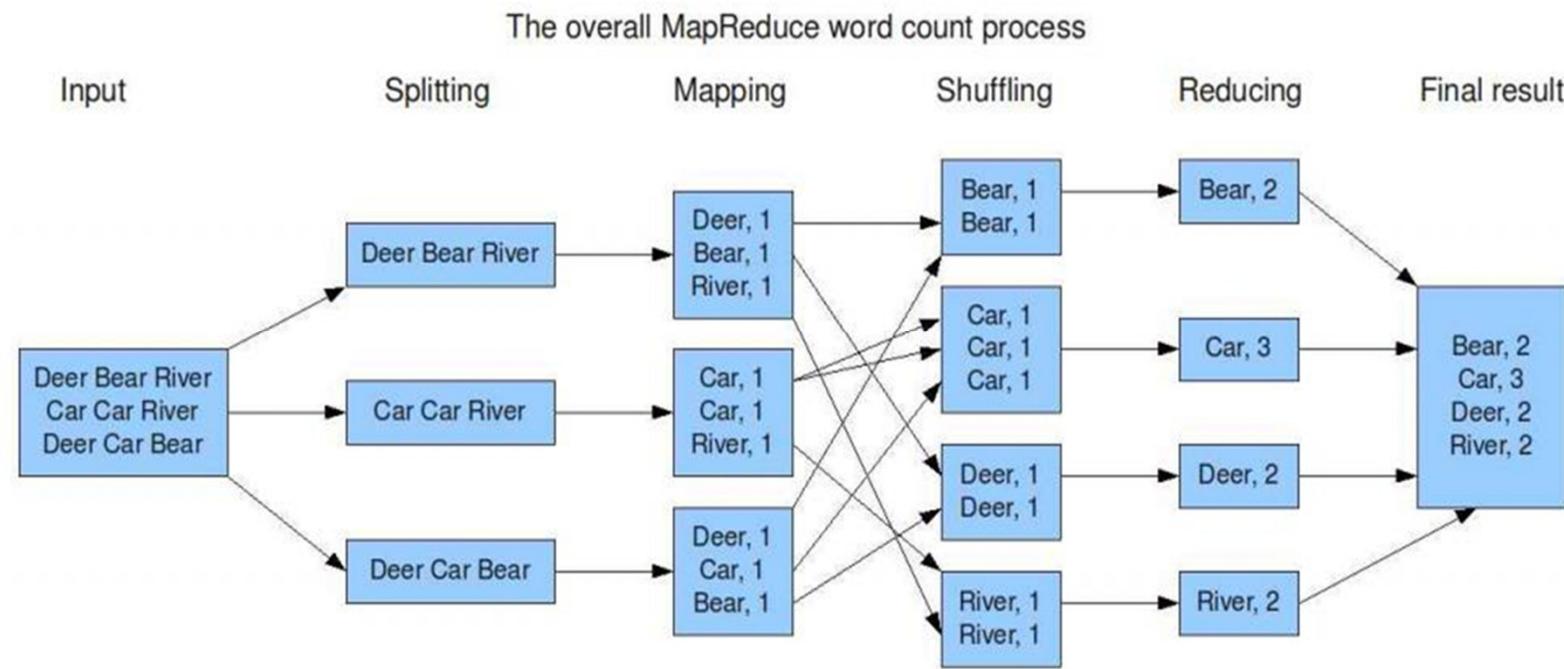


# Hadoop 生態圖譜系統

- HDFS
  - 分散式檔案儲存系統
- MapReduce
  - 資料分散式計算框架
- Hbase
  - 資料存儲系統
- Hive
  - SQL 查詢
- R connectors
  - 和 R 統計軟體連接, 如 Rhadoop 套件
- Mahout
  - 機器學習程式庫
- Pig
  - 大量資料分析的語言
- Oozie
  - Workflow 管理
- Zookeeper
  - 資料序列化處理與任務調度
- Flume
  - 非結構化資料(log)收集處理
- Sqoop
  - 與關聯式資料庫資料交換
- Ambari
  - 監控管理
- And more

# 簡單 MapReduce 範例

## word count - (key, value)



# Recap

---

# 學習重點

- 營運智慧資料分析暨視覺化應用
- 營運智慧分析師檢定四大主題：營運智慧基本識，基礎資料分析，經營管理基本知識，數位化企業資訊工具基本知識。
- 營運智慧
- 商業智慧
- 資料分析
- 統計應用(估計,檢定)
- 物聯網
- 大數據
- 機器學習(集群法,關聯規則,KNN,決策樹), Hadoop
- R, RStudio

# Q & A

李明昌

Email: [alan9956@gmail.com](mailto:alan9956@gmail.com)

URL: <http://rwepa.blogspot.tw/>

WEB: @RWPEA