

# 進階程式設計-第1章 進階程式設計簡介

## 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](mailto:alan9956@gmail.com)

<http://rwepa.blogspot.com/>

# 大綱

- 1.1 教師簡介
- 1.2 程式設計簡介
- 1.3 進階程式設計應用-商業智慧 (Business Intelligence, BI)
- 1.4 進階程式設計應用-機器學習 (Machine Learning)

# 1.1 教師簡介



# 教師簡介 <http://rwepa.blogspot.com/>

- 姓名：李明昌 (ALAN LEE)
- 現職：中華R軟體學會 常務理事  
臺灣資料科學與商業應用協會 常務理事
- 學歷：中原大學 工業與系統工程所 博士
- 經歷：
  - 淡江大學 兼任教師
  - 佛光大學 兼任教師
  - 國立台北商業大學 兼任教師
  - 育達科技大學 資訊管理系(所) 專任助理教授
  - 東吳大學 兼任教師
  - 崇友實業 行銷企劃專員
  - 國航船務代理股份有限公司 海運市場運籌管理員
- 大專院校、資策會、工業技術研究院、國家發展委員會、中央氣象局、公平交易委員會、各縣市政府與日本名古屋產業大學等公民營單位演講達300餘場, 2800小時以上.
- 連絡資訊：[alan9956@gmail.com](mailto:alan9956@gmail.com)



- iPAS 巨量資料分析師 證照推廣
- iPAS 營運智慧分析師 證照推廣

## 1.2 程式設計簡介

# 程式設計

- 電腦程式設計（英語：**Computer programming**），或稱程式設計（**programming**），是使用程式解決出特定問題的**過程**，也是軟體開發過程中的重要步驟。
- 程式設計方法往往以某種程式設計語言為工具，給出這種語言下的程式。
- 程式設計過程一般包括分析、設計、編碼、測試、除錯等不同階段。

[https://en.wikipedia.org/wiki/Computer\\_programming](https://en.wikipedia.org/wiki/Computer_programming)

# 程式設計發展

- 第一個電腦程式通常可以追溯到 1843 年，當時數學家Ada Lovelace發布了一種計算伯努利數序列的演算法，該演算法旨在由查爾斯·巴貝奇的分析機執行。
- 程式設計發展
  - 機器語言
  - 組合語言
  - 高階計算機語言：
    - 程序導向：Fortran(1957)，Cobol(1960)，Pascal(1971)，C(1972)
    - 物件導向：C++(1983)，Python(1991)，**R(1993)**，Java(1995)，C#(2000)



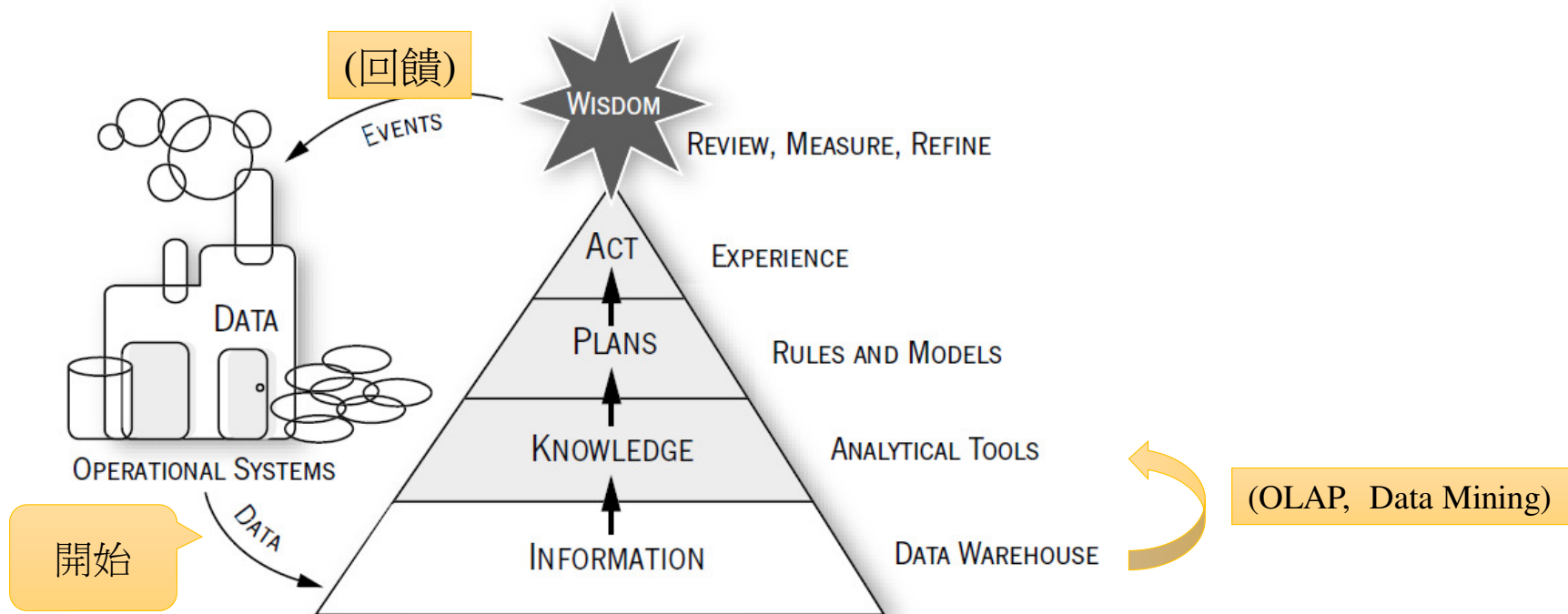
## 1.3 進階程式設計應用- 商業智慧 (Business Intelligence, BI)



# 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](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

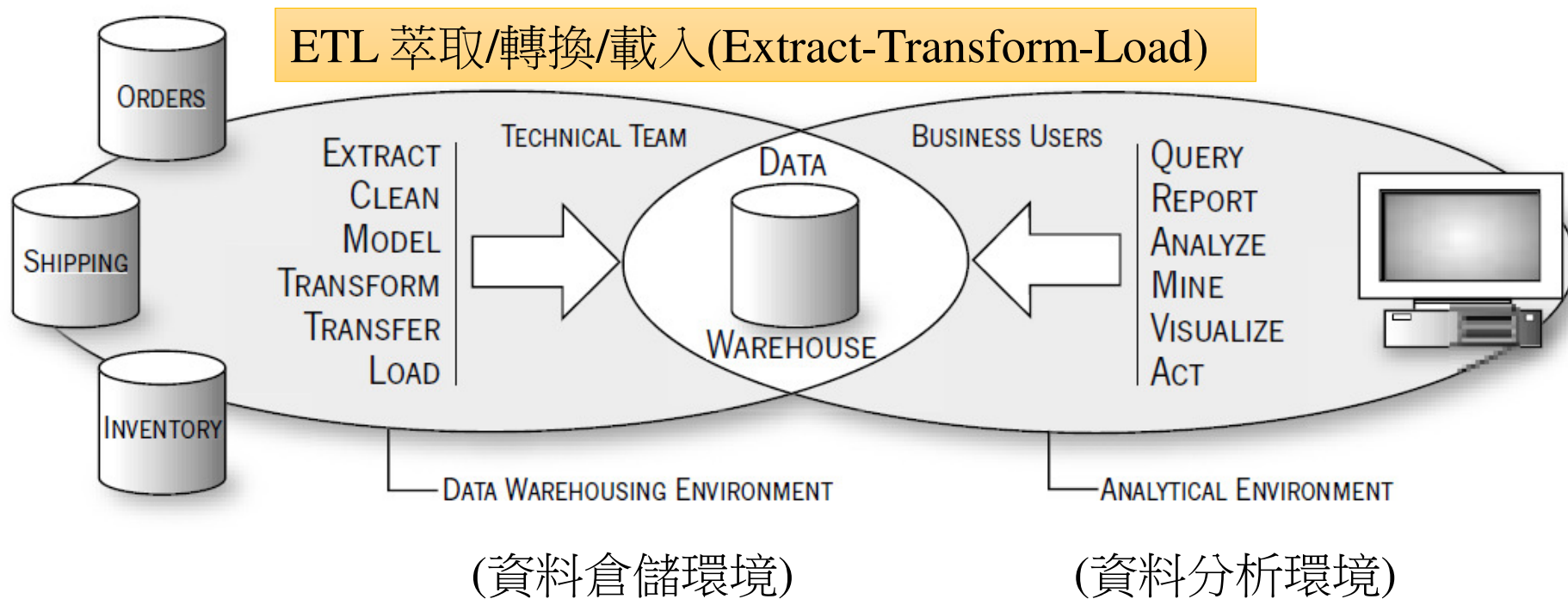
## 5. 回饋迴圈

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.



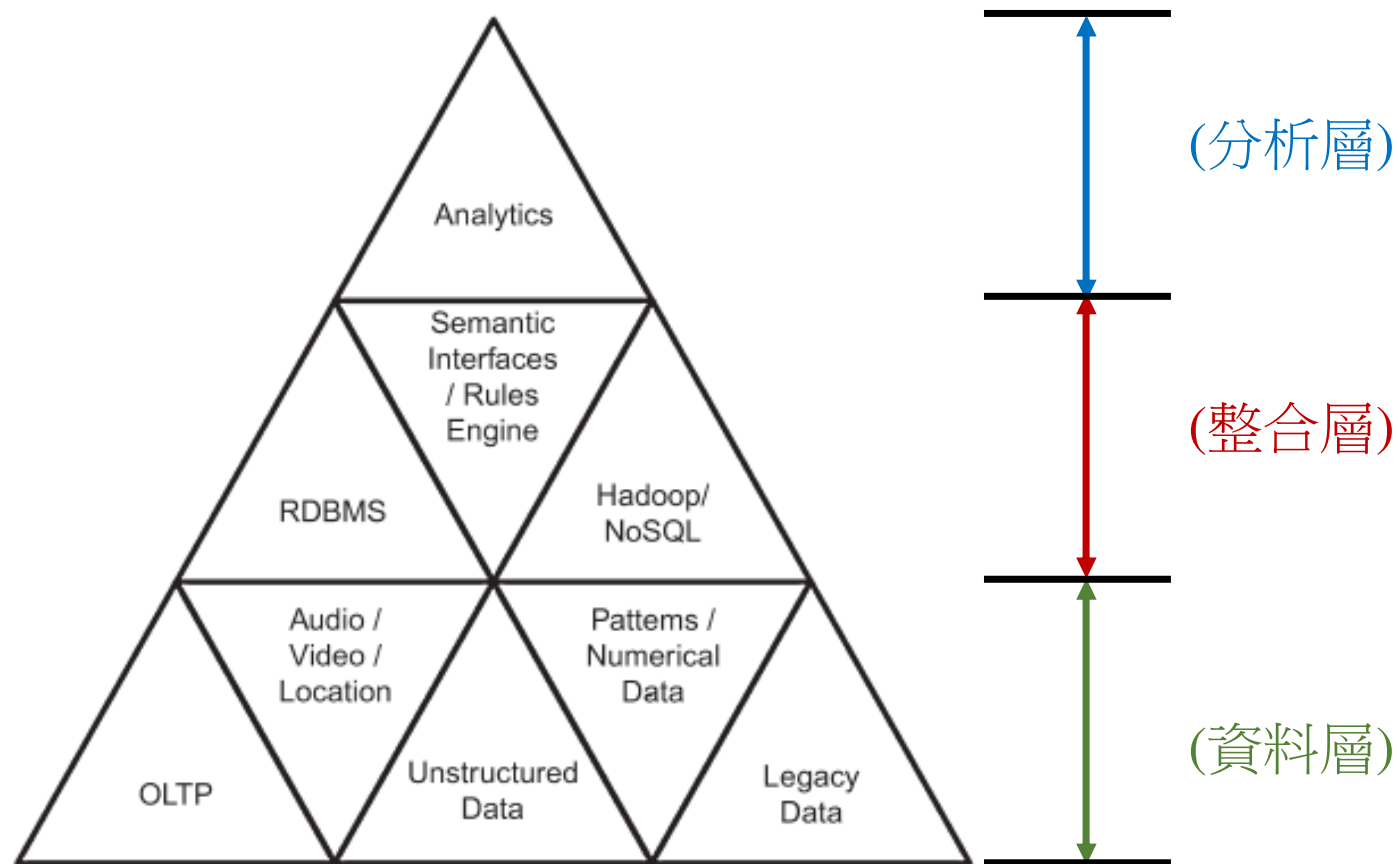
# 商業智慧系統的組成架構

## 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](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.



# 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

# 1.4 進階程式設計應用- 機器學習 (Machine Learning)

# 深度學習發展史




資料探勘 (Data Mining)

- 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(Support Vector Machine, 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

- 監督式學習 (Supervised learning)
    - Telling the algorithm what to predict
  - 非監督式學習 (Unsupervised learning)
    - No label or target value given for the data
  - 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

- 監督式學習 Supervised learning - 執行  $X \rightarrow Y$ 
  - 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
- 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

# CRISP-DM標準流程

---

# 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)
- 資料探勘方法論
- 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)

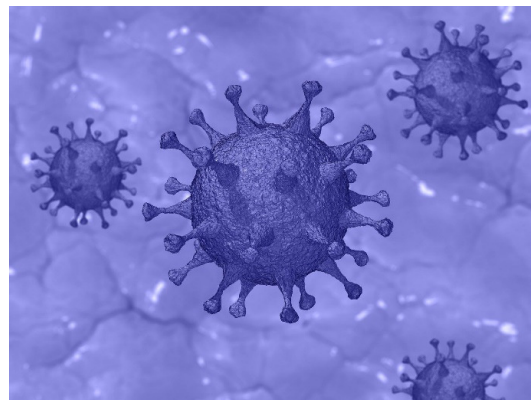
- 步驟 1：商業理解
- 步驟 2：資料理解
- 步驟 3：資料準備
- 步驟 4：模式建立
- 步驟 5：評估與測試
- 步驟 6：佈署應用

佔整專案時間的  
~80%

- 訓練資料70%
- 測試資料30%

## 步驟1.商業理解

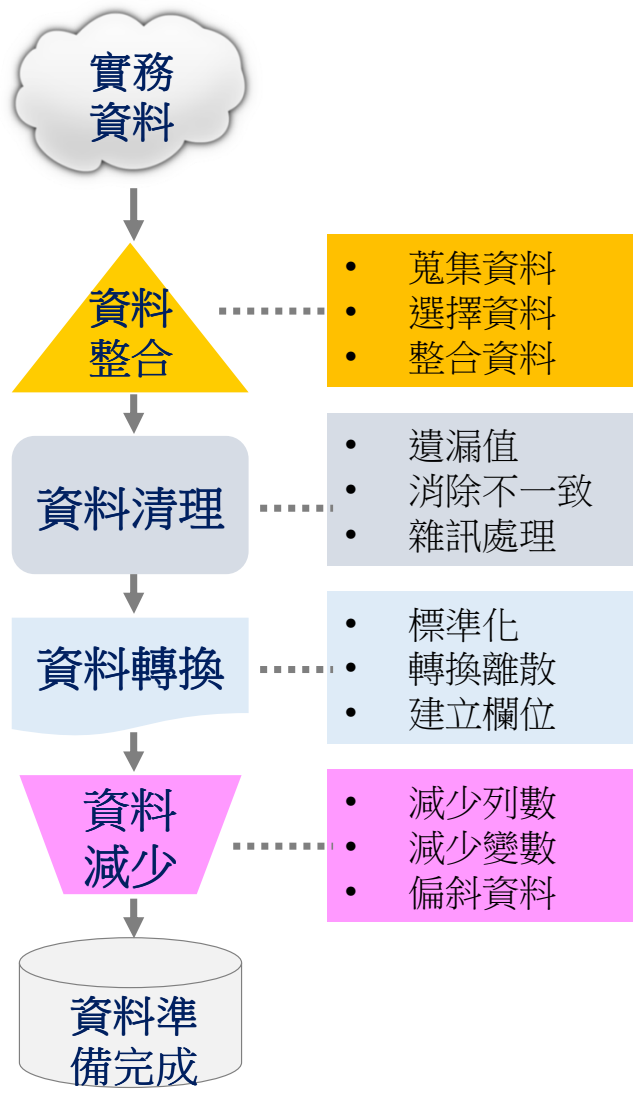
- 終極目標是要解決具體的產業問題，諸如提高購買率、找出詐欺交易、銷售預測與異常偵測等，因此以專業知識 (domain knowledge)進行商業理解是重要的第一步，處理重點：
  - 擬定商業目標
  - 進行當前處境評估
  - 決定資料探勘目標/成本
  - 產生專案計劃
  - 解決顧客問題



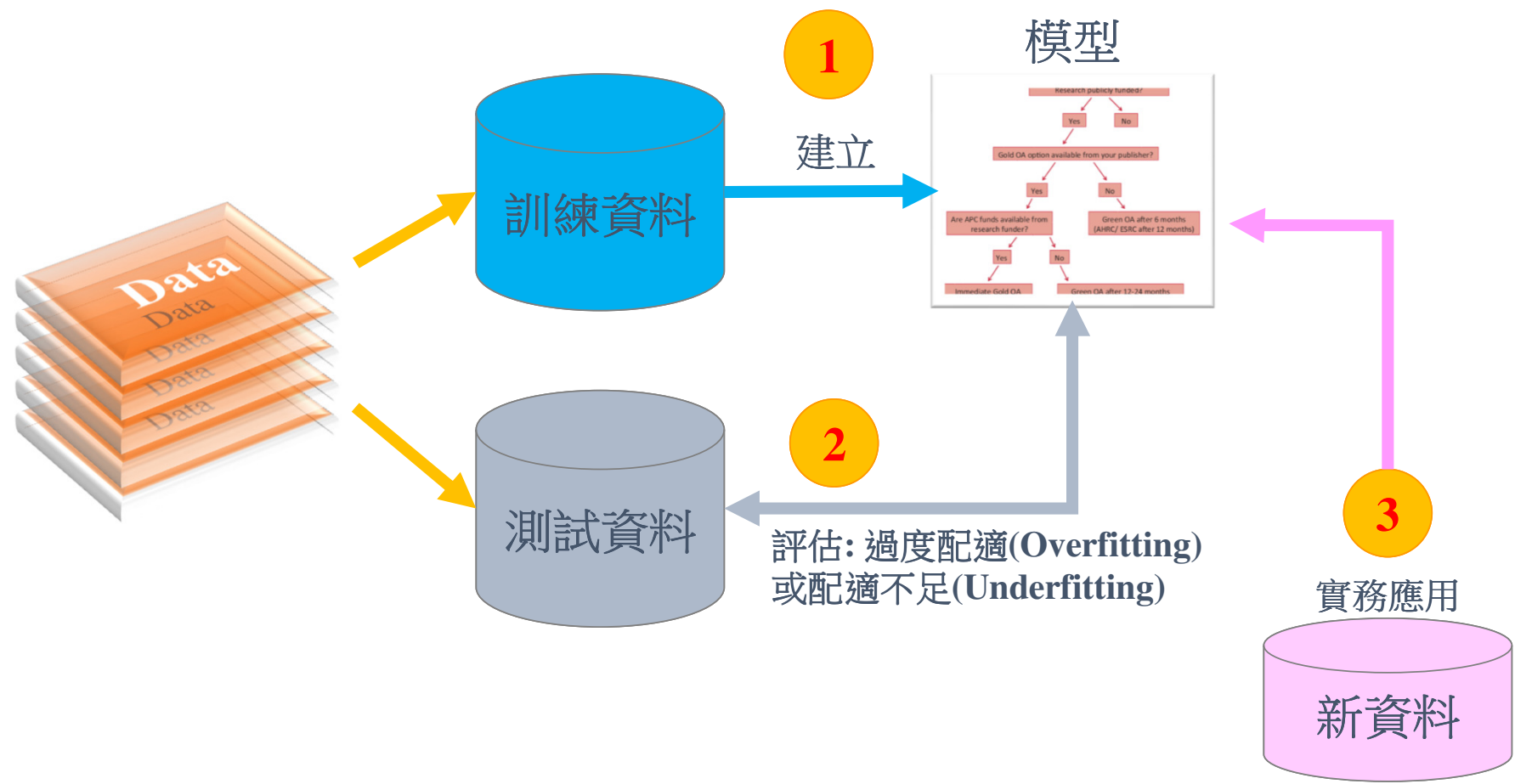
## 步驟2. 資料理解

- 包括描述資料、探索資料、核驗資料品質
- 敘述統計分析
  - 六力分析(summary函數)
- 繪圖
  - 依群組特性
  - 依時間特性
  - 新增評估欄位
  - 趨勢
  - 離群值 (outlier)
  - 散佈圖、散佈圖矩陣
  - 盒鬚圖

# 步驟3.資料準備

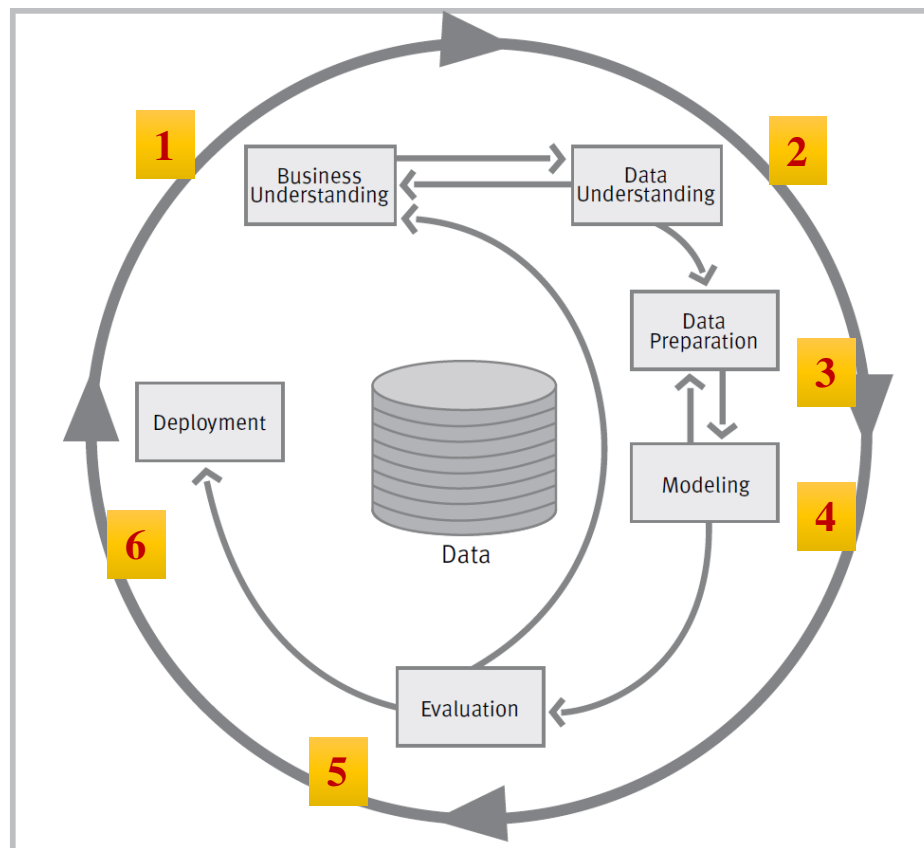


# 步驟4.模型建立、步驟5.評估與測試





# 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!

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



- 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

- <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
```

混淆矩陣  
(Confusion Matrix)

```
# 1.TPR(True positive rate) 真陽性率, 愈大愈好 -----
# =TP/ (TP+FN)
# =TP/P
# =Sensitivity 靈敏度
# =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。
```

## 參考資料

- OLAP, [https://en.wikipedia.org/wiki/Online\\_analytical\\_processing](https://en.wikipedia.org/wiki/Online_analytical_processing)
- OLTP, [https://en.wikipedia.org/wiki/Online\\_transaction\\_processing](https://en.wikipedia.org/wiki/Online_transaction_processing)
- RWEPA, <http://rwepa.blogspot.com/>

# 謝謝您的聆聽

## Q & A

李明昌

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

YouTube: <https://www.youtube.com/@alan9956>

WEB: <http://rwepa.blogspot.com/>

