2024 TWDS 台灣資料科學領航計畫

第 9 組 資料科學/機器學習

導師 : 吳彥霖 Tom

交流主題

六月份　資料科學與機器學習的職涯發展路徑

七月份　資料科學與機器學習的實務流程

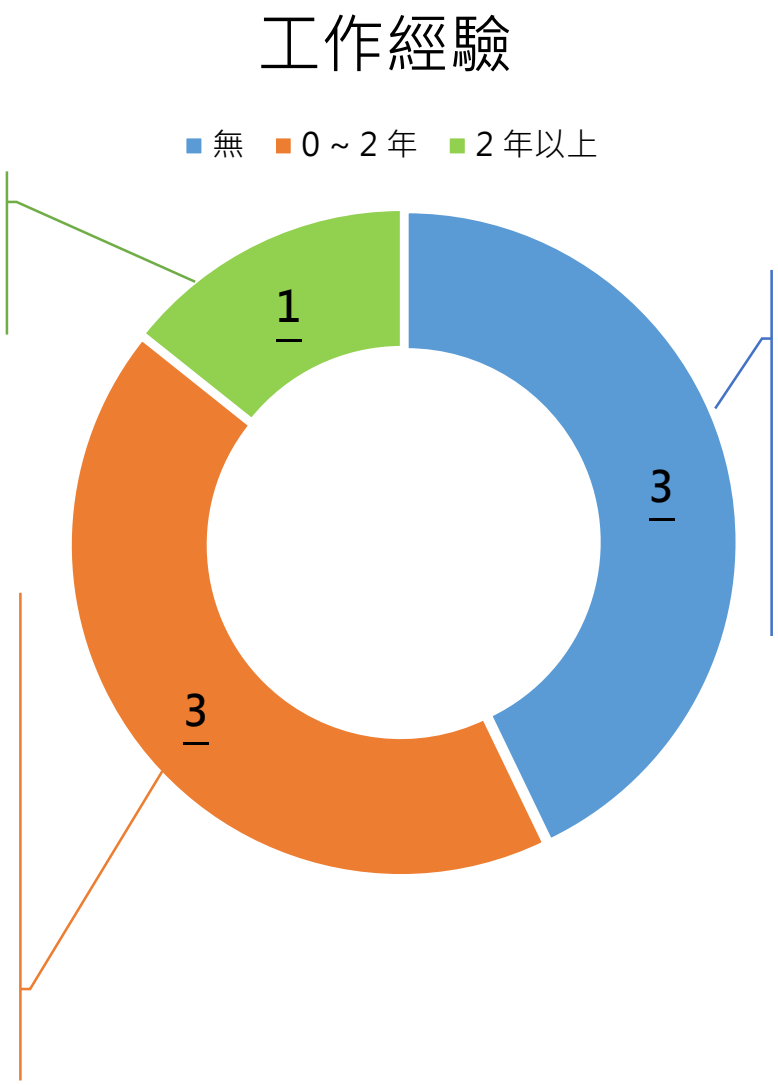
八月份　資料科學與機器學習的實務應用

資料科學與機器學習的職涯發展路徑

Tom

2024/6/18

小組學員的期待收穫

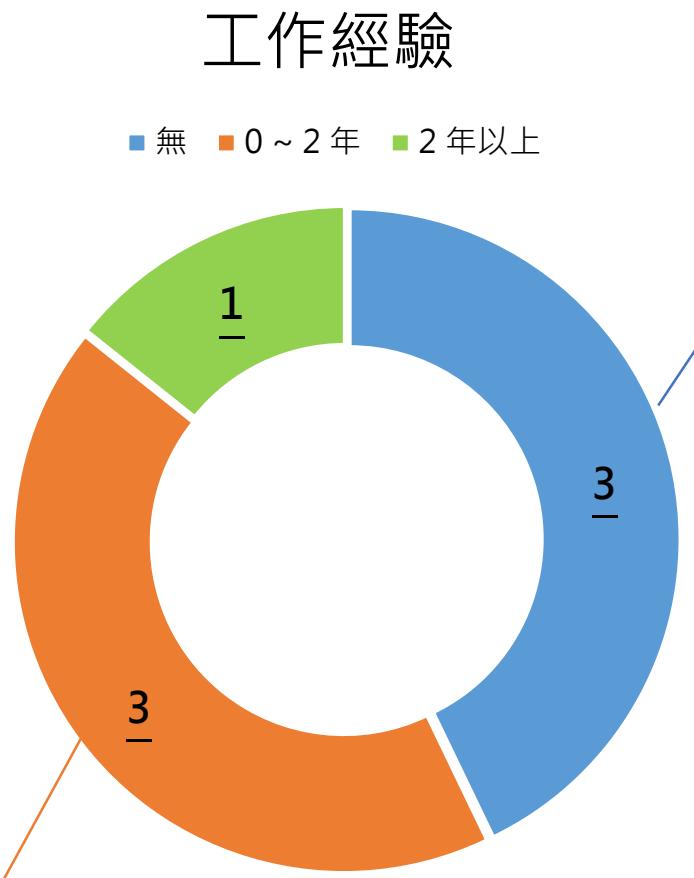


- 獲得大家在數據上的經驗分享
- 職涯的規劃藍圖

- 瞭解踏入這領域需具有的能力
- 瞭解能力養成的學習方向
- 準備履歷中吸引人的專案與作品集
- 瞭解職缺面試會問那些深度的問題，與回答技巧
- 瞭解工作中的實際應用

- 獲得新的知識與見解
- 瞭解最新的技術趨勢、行業發展動態、解決問題的新方法
- 獲得解決現實分析問題的能力
- 職涯的發展方向
- 瞭解各產業的應用與差別
- 結交志同道合的夥伴，彼此交流經驗與想法

希望 Tom 分享的主題與建議



- 實務經驗的分享
- 對新鮮人的建議
- 對於目前的職位，在學生時期與工作後，分別做了哪些準備？
- 擔心 Side Project 太普遍或太沒料，可以怎麼做？

- 如何透過資料科學提升製程能力？
- 如何實現 MLOps？
- 若未來工作想往不同產業發展，建議可以如何準備？
- 當資料品質很好時，但模型表現不佳，下一步可以如何做？
- 若研發中的模型，還尚未取得顧客的資料，只有內部模擬的資料，該如何解決績效問題？

資料科學（英語：data science）又稱數據科學，

是一門利用資料（數據）學習知識的學科，其目標是透過**從資料中提取出有價值**的部分來生產資料產品[1]，學科範圍涵蓋了：資料取得、資料處理、資料分析等過程，**舉凡與數據有關的科學均屬資料科學**。

資料科學結合了諸多領域中的理論和技術，包括應用**數學**、**統計**、**圖型識別**、**機器學習**、**資料視覺化**、**資料倉儲**以及高效能計算。

資料科學透過運用各種相關的資料來幫助非專業人士理解問題。

Reference :
■ https://zh.wikipedia.org/zh-tw/%E6%95%B0%E6%8D%AE%E7%A7%91%E5%AD%A6

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

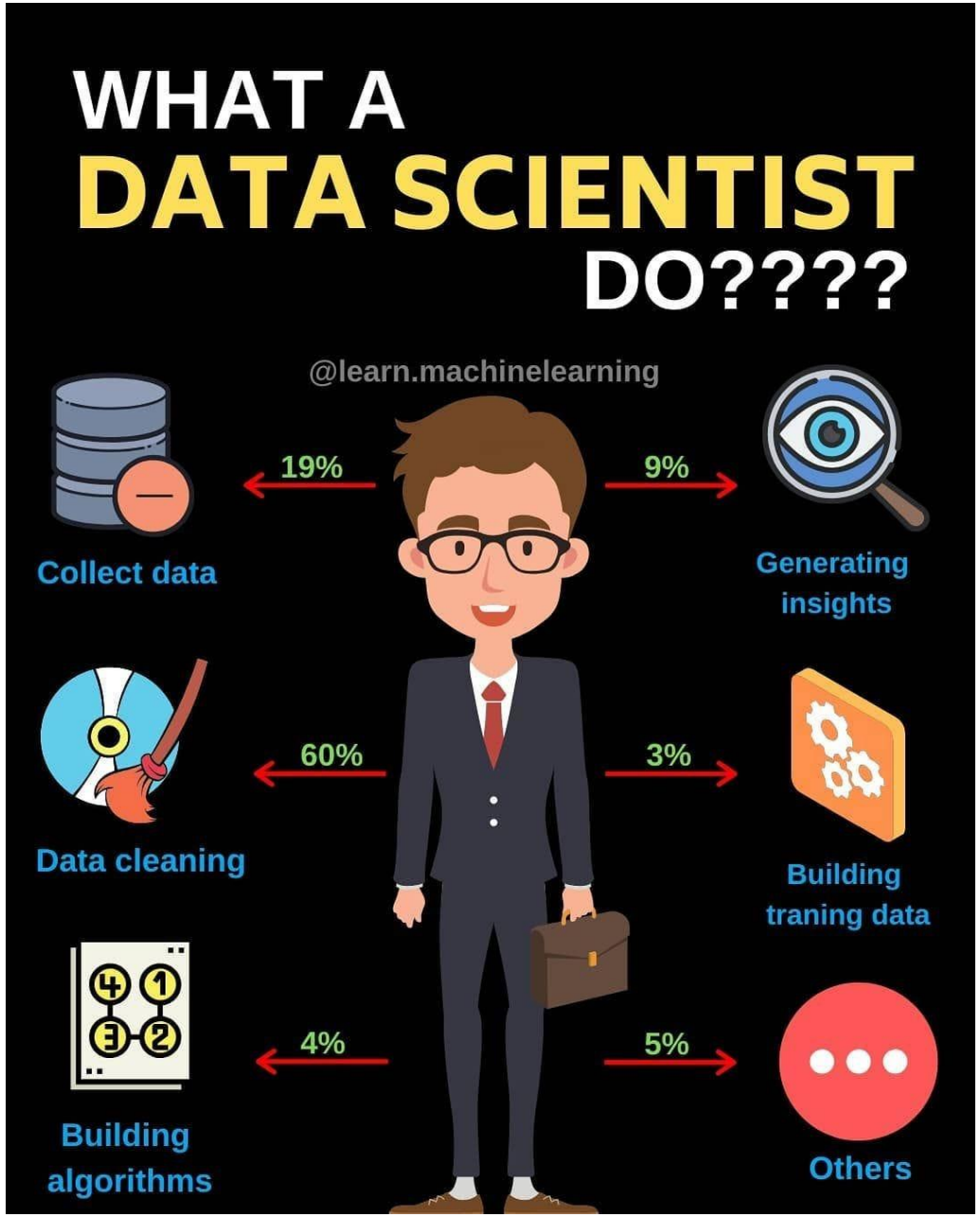
Andrew J Buboltz, silk screen on a page from a high school yearbook, 8.5" x 12", 2011 Tamar Cohen

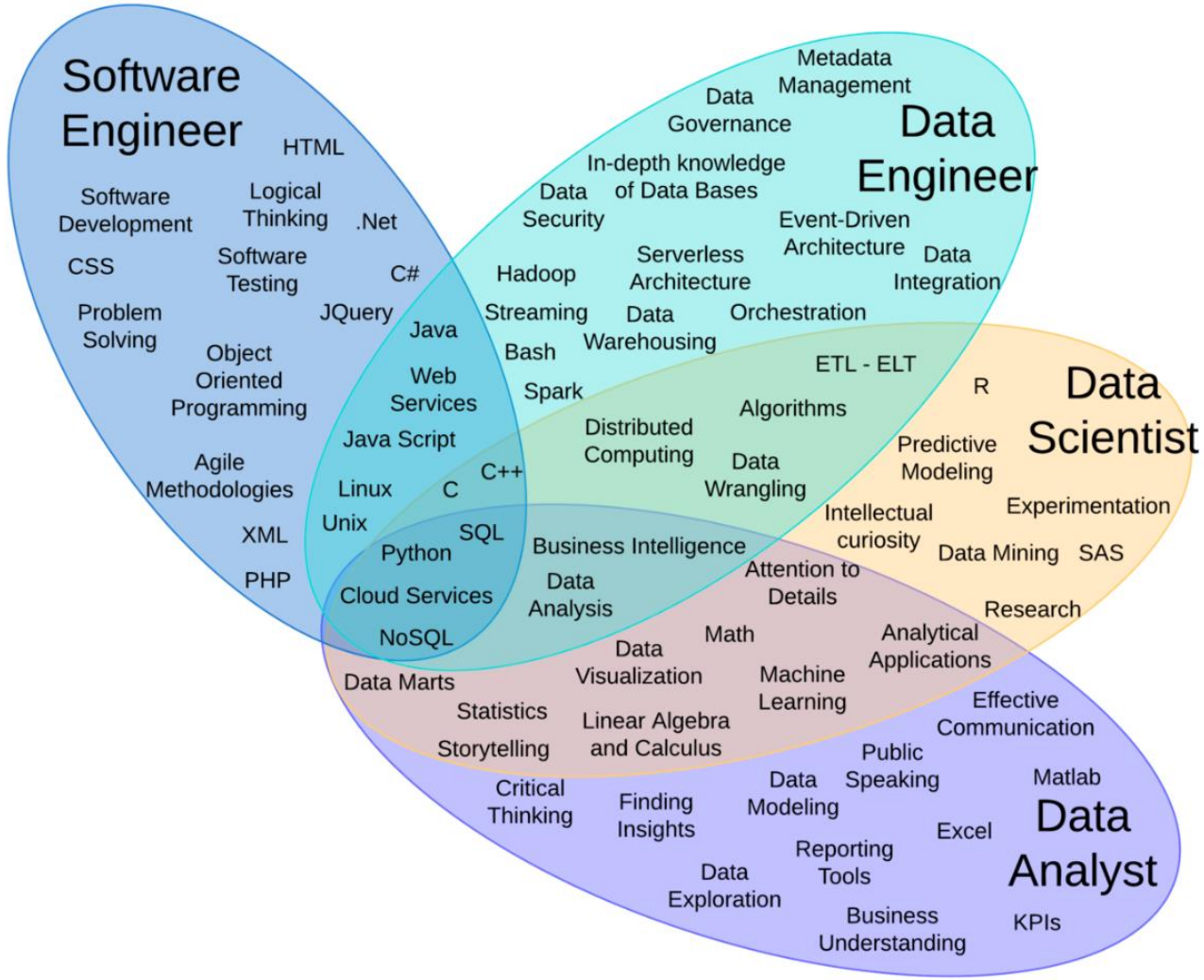
Reference :

■ Data Scientist: The Sexiest Job of the 21st Century, Thomas H. Davenport and DJ Patil, Harvard Business, October 2012. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

為什麼自己想踏入

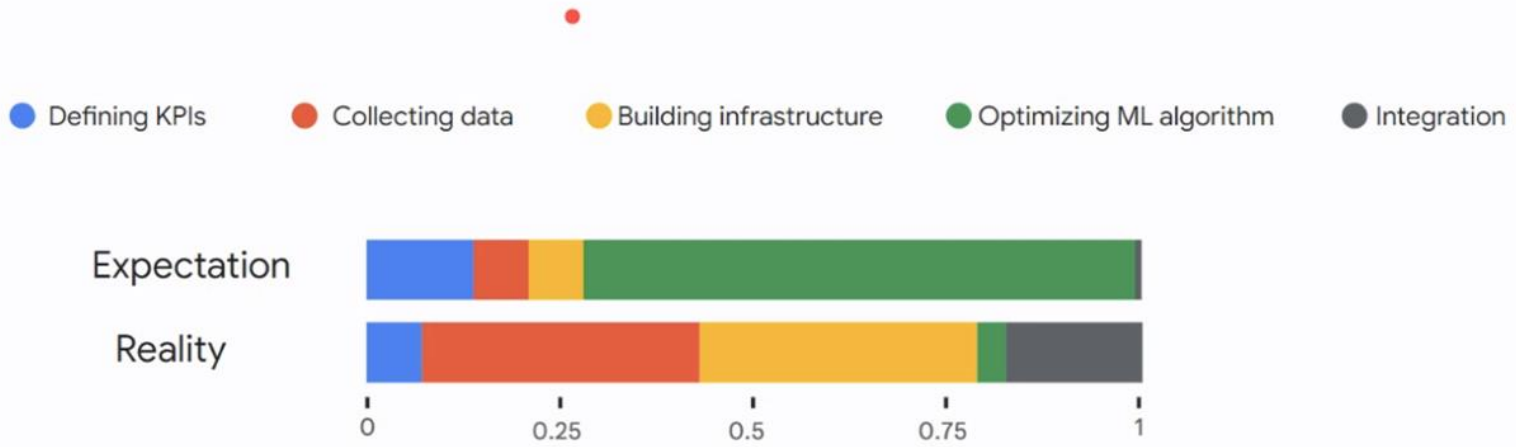
資料科學與**機器學習**的領域？


WHAT A
DATA SCIENTIST
DO????
@learn.machinelearning
19%
Collect data
9%
Generating insights
60%
Data cleaning
3%
Building traning data
4%
Building algorithms
5%
Others

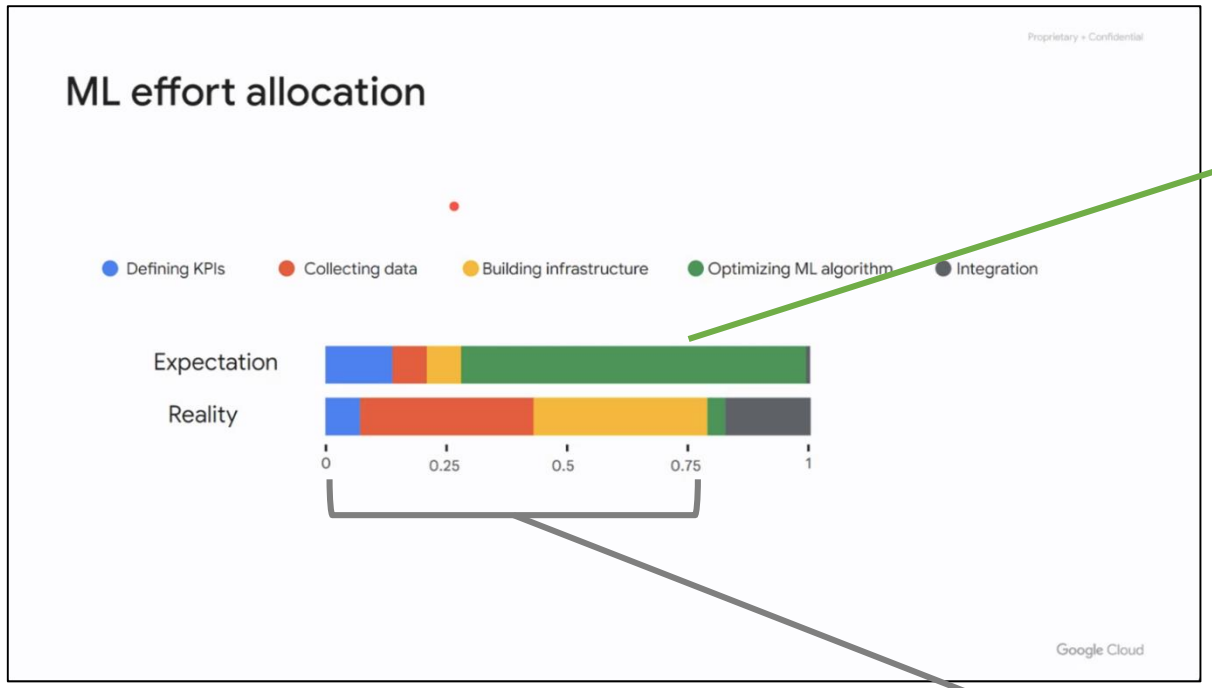

Software Engineer
HTML
Software Development
Logical Thinking
.Net
CSS
Software Testing
C#
Problem Solving
JQuery
Object Oriented Programming
Java
Web Services
Java Script
Agile Methodologies
Linux
C
C++
Unix
XML
SQL
PHP
Python
Cloud Services
NoSQL
Data Engineer
Metadata Management
Data Governance
In-depth knowledge of Data Bases
Data Security
Event-Driven Architecture
Data Integration
Hadoop
Serverless Architecture
Streaming
Data Warehousing
Orchestration
Bash
Spark
ETL - ELT
Distributed Computing
Algorithms
Data Wrangling
Data Scientist
R
Predictive Modeling
Intellectual curiosity
Experimentation
Data Mining
SAS
Business Intelligence
Data Analysis
Attention to Details
Research
Math
Analytical Applications
Data Visualization
Machine Learning
Data Marts
Statistics
Storytelling
Linear Algebra and Calculus
Effective Communication
Public Speaking
Matlab
Critical Thinking
Finding Insights
Data Modeling
Excel
Data Analyst
Reporting Tools
Data Exploration
Business Understanding
KPIs

資料科學的日常工作

ML effort allocation



資料科學的日常工作



理想上 ...

- ML 模型的期望 -> 準度高
- ML 模型的執著 -> 反覆調參

在實際中，我們常會花一半以上的功夫 ...

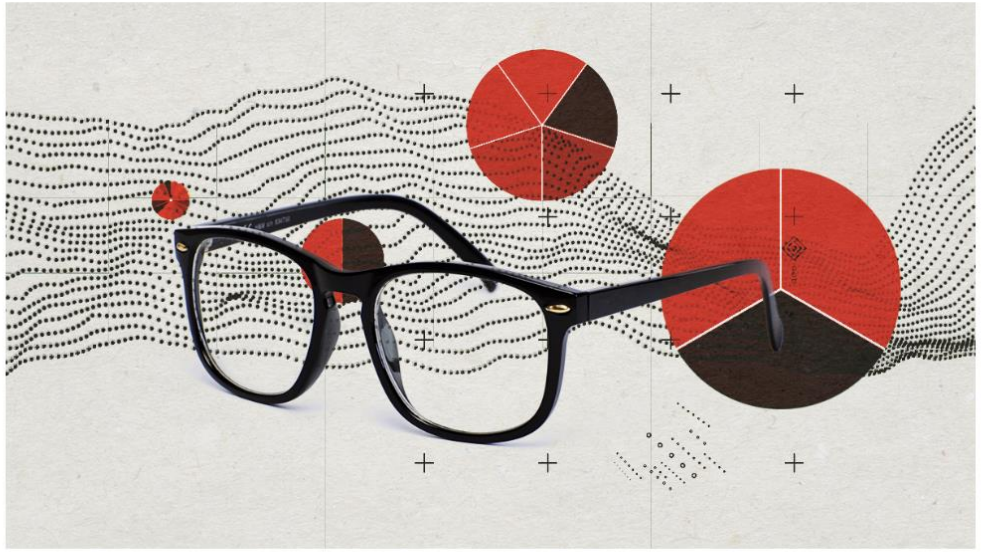
- 專案的對焦：理解業務痛點與解決目標
- 資料的蒐集：
  資料庫、欄位定義、取得途徑 ... 等
- 資料品質的檢視
- 使用的方式：
  流程機制面、UI 需求、軟硬體規格 ... 等

Analytics And Data Science

Is Data Scientist Still the Sexiest Job of the 21st Century?

by Thomas H. Davenport and DJ Patil

July 15, 2022



HBR Staff/StudioM1/Moritz Otto/Getty Images

Reference :

■ Is Data Scientist Still the Sexiest Job of the 21st Century?, Thomas H. Davenport and DJ Patil, Harvard Business, July 2022. https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century

資料科學的 10 年轉變 ...

- 在資料科學的職缺中，相較與以往，

  程式能力的重要性將有所**降低**，

  因某些功能正越來越自動化。

(然而，**數據清理**是這趨勢中的一個**明顯例外**。)

需借助 Domain Know-how

Reference :

■ Is Data Scientist Still the Sexiest Job of the 21st Century?, Thomas H. Davenport and DJ Patil, Harvard Business, July 2022.
https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century

資料科學的技能養成

硬性能力的培養

- 統計 : 敘述統計、假設檢定、機率 ...
- 資料視覺化
- 理解能力
- 表達能力

  用**白話的方式**介紹演算法/技術
- 程式語言 : Debug 的能力
- 學習新知的熱情 : 技術與 Domain 知識
- 專案規劃與管理

  - 專案其中(末)報告
  - 參與論壇/分享/展覽活動
  - 工作/小組會議

有目的性得做 !!

- 深耕專業技能
- 提升其他技能
- 「需求」導向

Side Project

- 學校課堂或線上課程的專案
- 實驗室的研究/產學計劃
- 競賽、黑客松等活動

  Kaggle https://www.kaggle.com/

  獎金獵人 https://bhuntr.com/tw

  ……
- 兼職接案
- 學習筆記文章
- 工作專案整理

職涯與面試的建議

- **面試的目的**：尋找 **最適合** 的人才　<- 能力強不代表最佳
  職缺　<- 待遇好不代表最佳
- **硬性技能可以培養**

  職涯的長期、中期、短期的目標是什麼？
  思考在職涯的下一個階段，自己想累積什麼經驗與能力？　<- 逐步提升自己的勞務價值

軟性技能的準備

- 為什麼想加入 ____？
- 如何解決與同事或外部利益相關單位的衝突？
- 過往的求學與工作中，是否有失敗的經驗？

面試最後，進一步可多問的問題

- 團隊成員的背景？　<- 同儕環境
- 團隊未來的發展與目標是什麼？　<- 主管對團隊的規劃
- 團隊期待自己的貢獻是什麼？　<- 主管對我們的定位

若想往不同產業發展，建議可 ...

- 選擇一個有需求且有錢景的產業
- 主動建立人脈，詢問業內資訊與尋求機會
- 掌握常用分析方法與其應用場景
- 參加相關的培訓課程/專案　<- 累積作品與經驗
- 定期更新履歷內容　<- 專案價值量化

Q & A

資料科學與機器學習的實務流程

Tom

2024/7/29

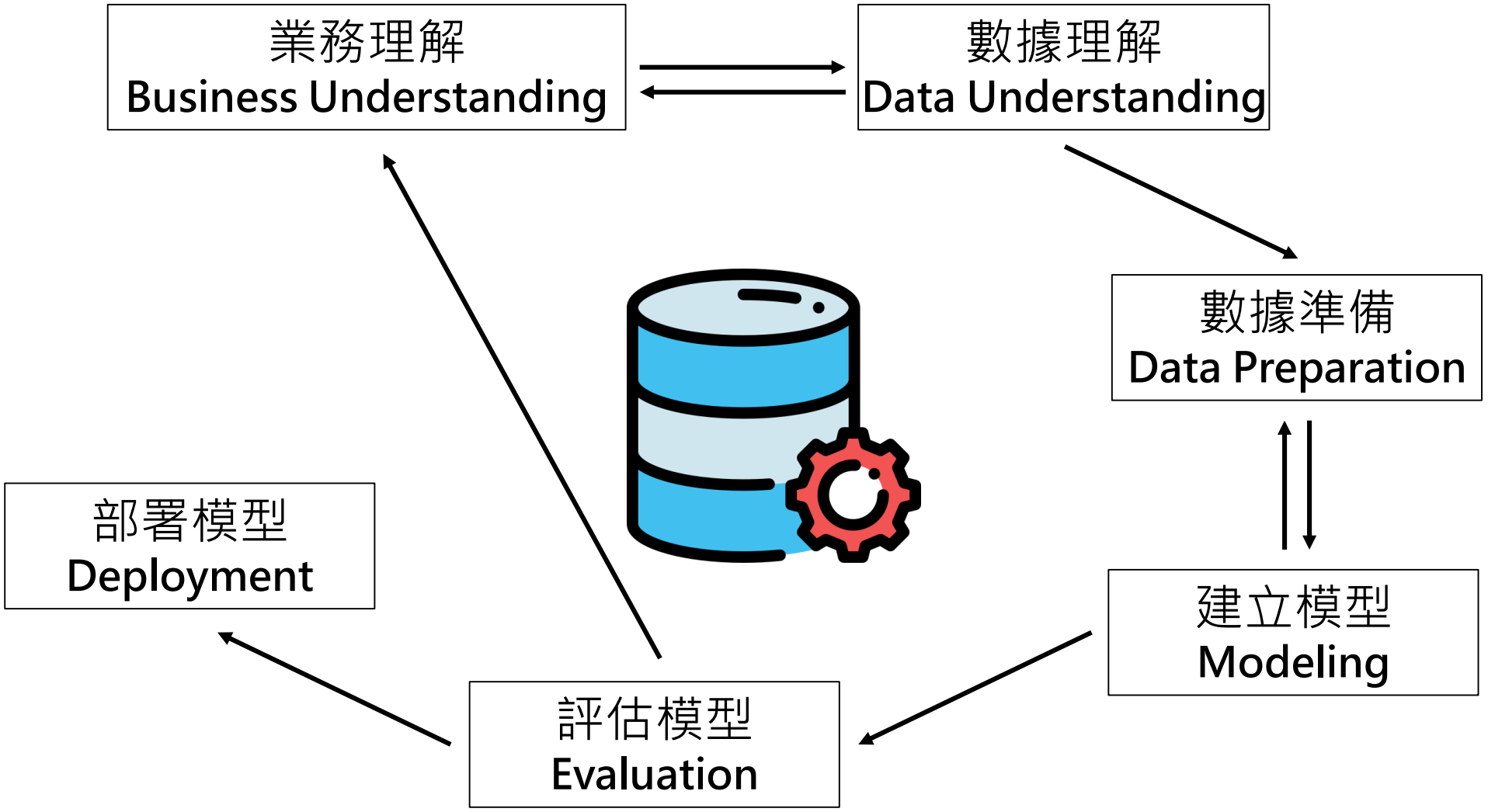
CRISP - DM

CRoss-Industry Standard Process for Data Mining

- 跨領域資料探勘標準程序
- 一個廣泛應用在資料探勘的標準流程方法論
- 目的 :
  - 提供一個結構化的框架，指導資料探勘專案的實施。
  - 提高專案的成功率，使最終產出的分析結果，更具價值與實際應用性。

CRISP - DM

CRoss-Industry Standard Process for Data Mining



CRISP - DM

CRoss-Industry Standard Process for Data Mining

業務理解
Business Understanding

- 業務流程
- 業務使用的技術/工具
- 業務術語
- 業務痛點/問題
- 專案的目標
- 專案的效益

……

評估模型
Eva uation

CRISP - DM

CRoss-Industry Standard Process for Data Mining

業務理解
Business Understanding

數據理解
Data Understanding

- 資料來源
- 資料的描述/定義
- 資料的品質

(探索性資料分析)

……

評估模型
Evaluation

CRISP - DM

CRoss-Industry Standard Process for Data Mining

數據準備
Data Preparation

- 資料前處理
  (填補、刪除、轉換...)
- 特徵轉換、特徵選擇

……

評估模型
Evaluation

CRISP - DM

CRoss-Industry Standard Process for Data Mining

業務理解
Business Understanding

數據理解
Data Understanding

數據準備
Data Preparation

建立模型
Modeling

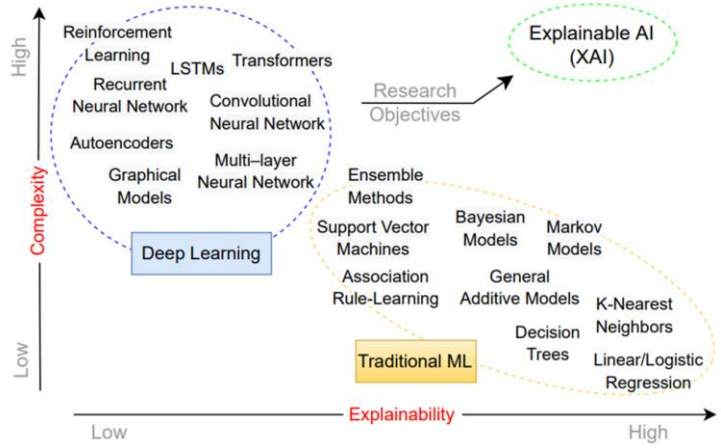


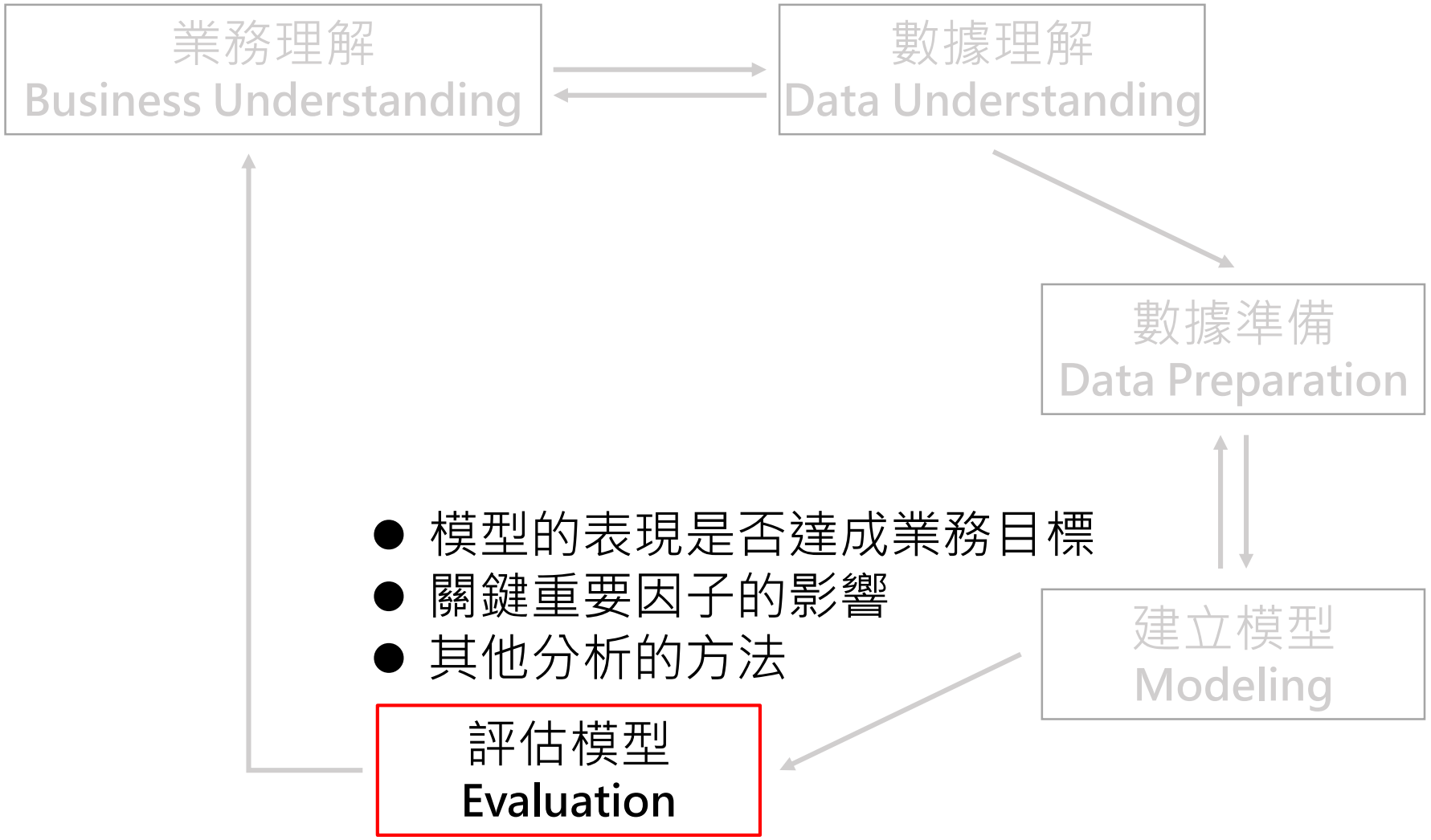
Figure 1. Classification of AI models according to their level of complexity, explainability, and their potential in modern AI applications.

- 模型的描述
- 模型的假設
- 模型的驗證方式(指標)
- 模型的驗證結果

Reference :
Clement T, Kemmerzell N, Abdelaal M, Amberg M. (2023). XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process, Machine Learning and Knowledge Extraction 5, no. 1: 78-108. https://doi.org/10.3390/make5010006

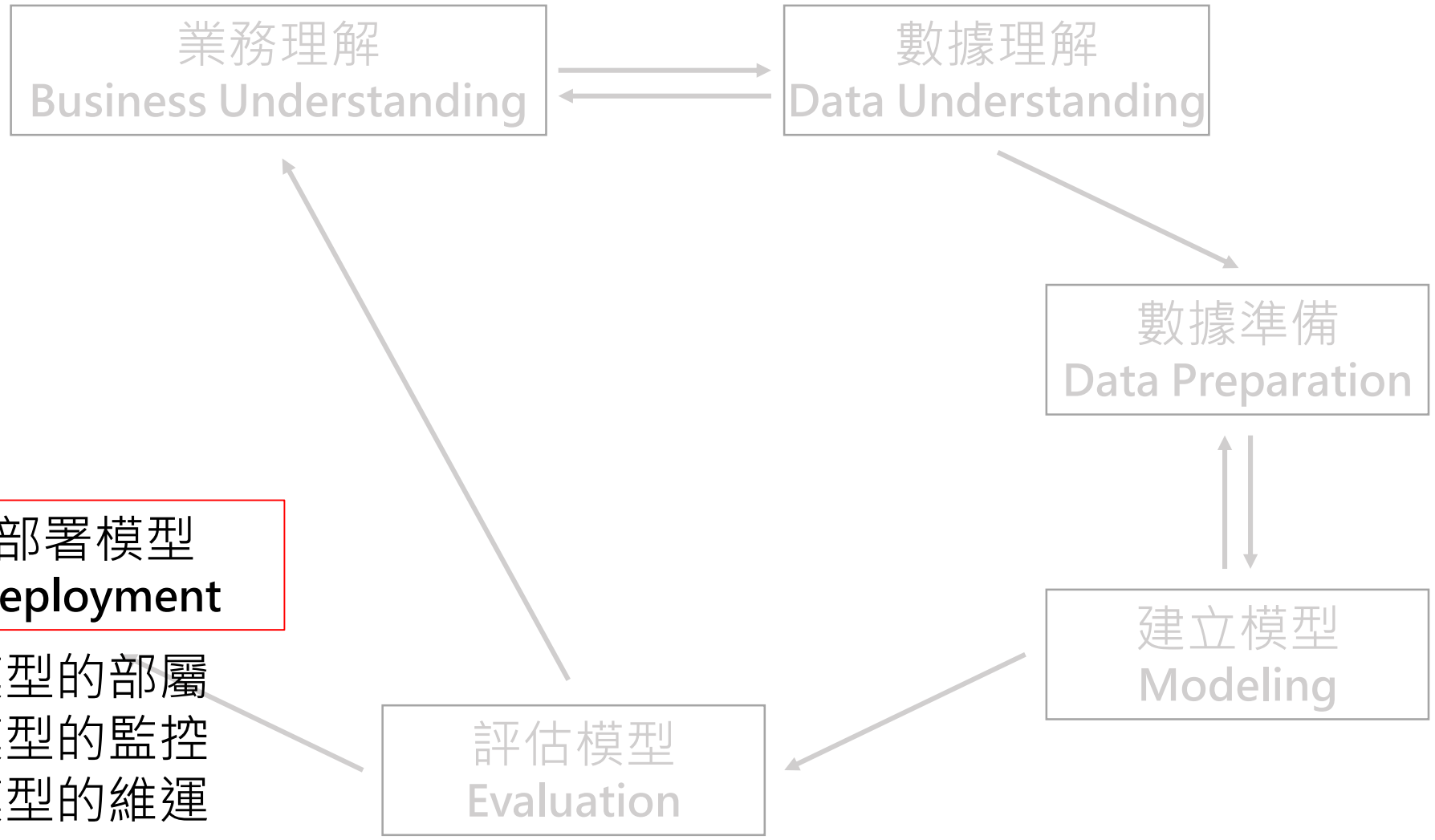
CRISP - DM

CRoss-Industry Standard Process for Data Mining



CRISP - DM

CRoss-Industry Standard Process for Data Mining



- 模型的部屬
- 模型的監控
- 模型的維運

……

CRISP – ML(Q)

CRoss-**I**ndustry **S**tandard **P**rocess for **M**achine **L**earning with **Q**uality assurance

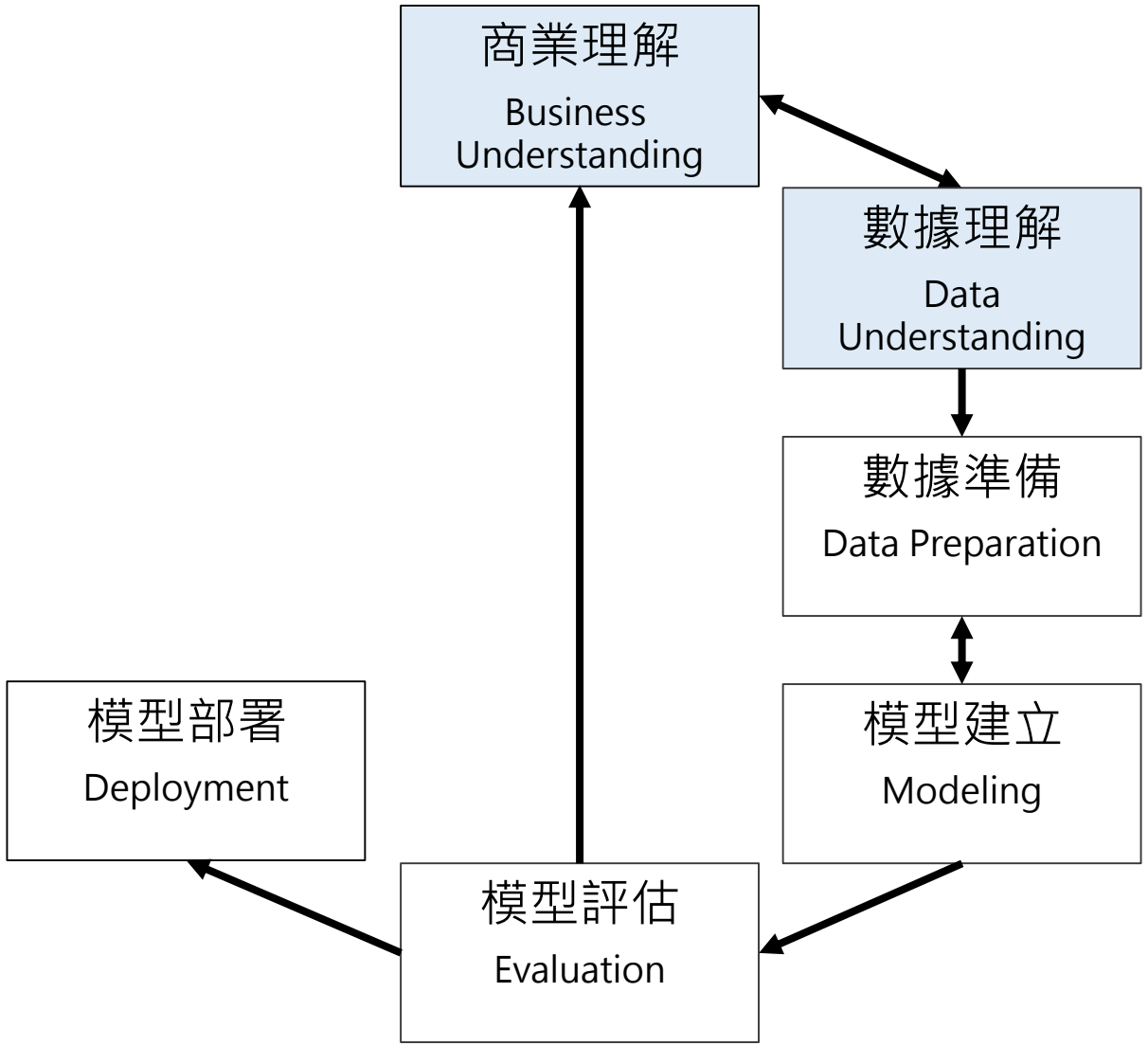
- 一種機器學習專案管理的方法論，基於 CRISP-DM 的基礎，且擴充了機器學習與模型品質保證的要素。
- 目的：
  - 幫助企業與資料科學家，系統性地進行機器學習專案。
  - 確保專案的每個階段都有高質量和**可重複性**，最後產出高效且可信賴的模型。

Reference:

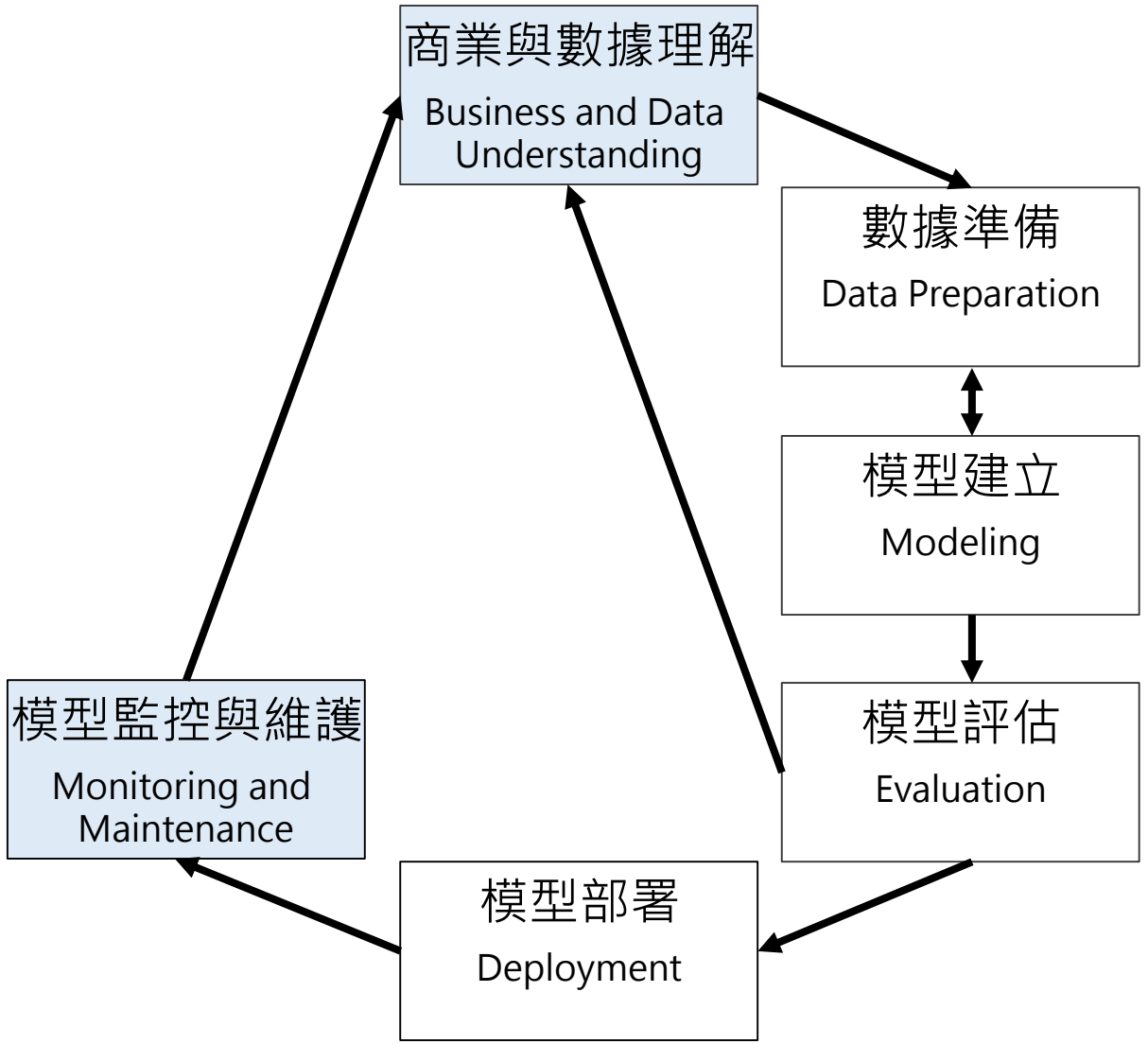
[1] Studer, S., Bui, T.B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. (2020). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. ArXiv, abs/2003.05155.

CRISP – DM v.s. CRISP – ML(Q)

CRISP-DM



CRISP-ML(Q)



CRISP – ML(Q)：模型監控與維護

表現監控

Performance Monitoring

- 評估指標的追蹤
  指標數值變差，不能直接表示模型表現變差。
- 異常資料分析

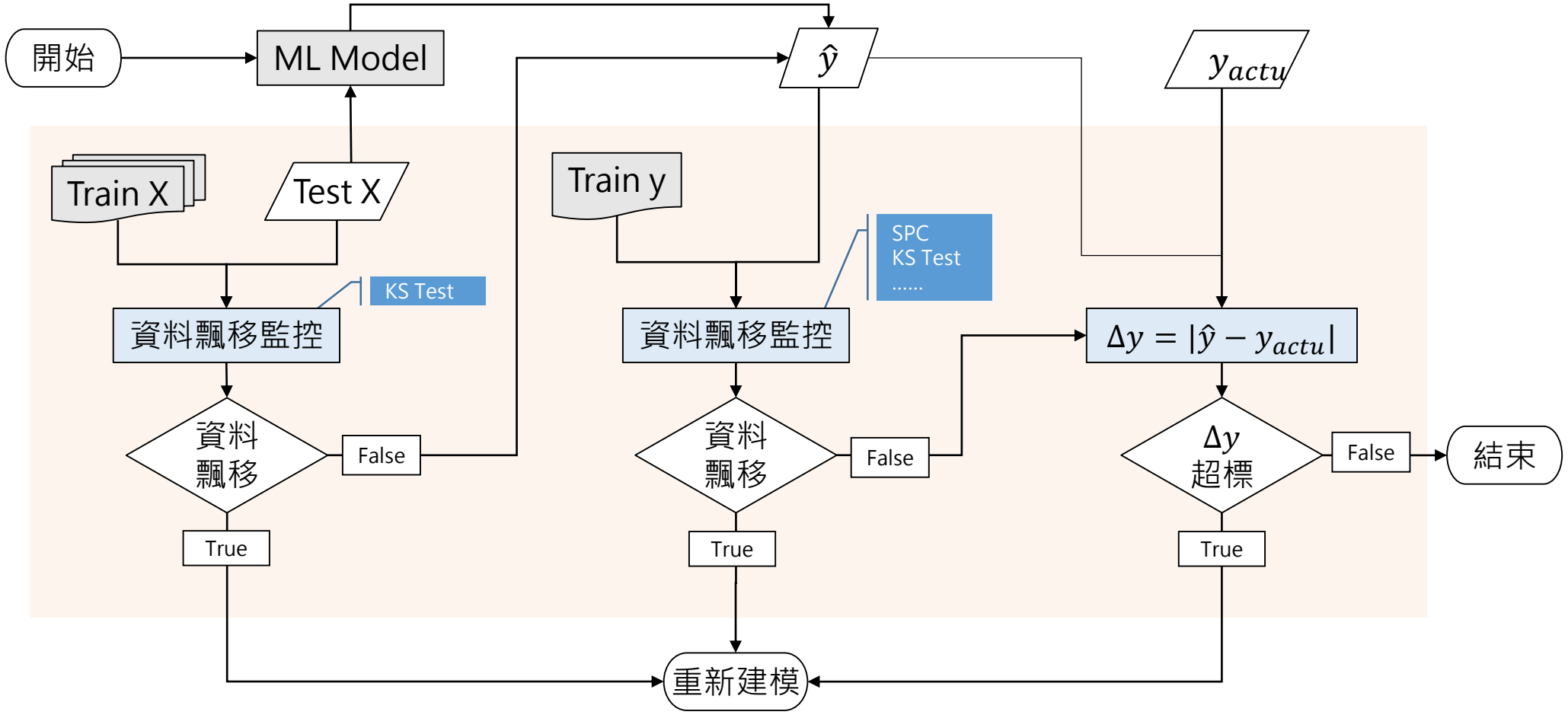
模型飄移

Model Drift

- 概念飄移 Concept Drift
- 資料飄移 Data Drift
  數據分佈不平穩、硬體老化、軟/硬體更新

CRISP – ML(Q)：模型監控與維護

【案例分享】



學員問題回饋

希望有 **Junior to Senior 的過程**，

或是 **Senior 怎麼樣更進階**一點的過程。

從 Junior 到 Senior 的建議

	Junior	to	Senior
Hard Skill	◆ 分析方法 - 統計、機率、... - 鼓勵創新 ◆ 軟體工程 - Git、Coding Style、Docker ...		◆ 深根專業 (技術、領域經驗、管理...) ◆ 掌握前沿技術與潮流 ◆ 理解領域知識
Soft Skill	◆ 任務 Task 的管理與執行 ◆ 與團隊夥伴的對焦/溝通 ◆ 發現問題 -> 釐清問題		◆ 專案的規劃與管理 ◆ 與需求者的訪談/對焦/溝通/談判 ◆ 發現問題 -> 釐清問題 -> **提出建議** ◆ 鼓勵擔任講師/導師 ◆ 心態轉換，**化被動成主動**!!

Q & A

資料科學與機器學習的實務應用

Tom

2024/8/29

實務應用的目的

目標變數　　模型　　特徵變數

$$y = f(x_1, x_2, \ldots, x_p)$$

Goal

- 尋找影響 y 的關鍵因子 $x_{k_0}, x_{k_1}, \ldots, x_{k_M}$

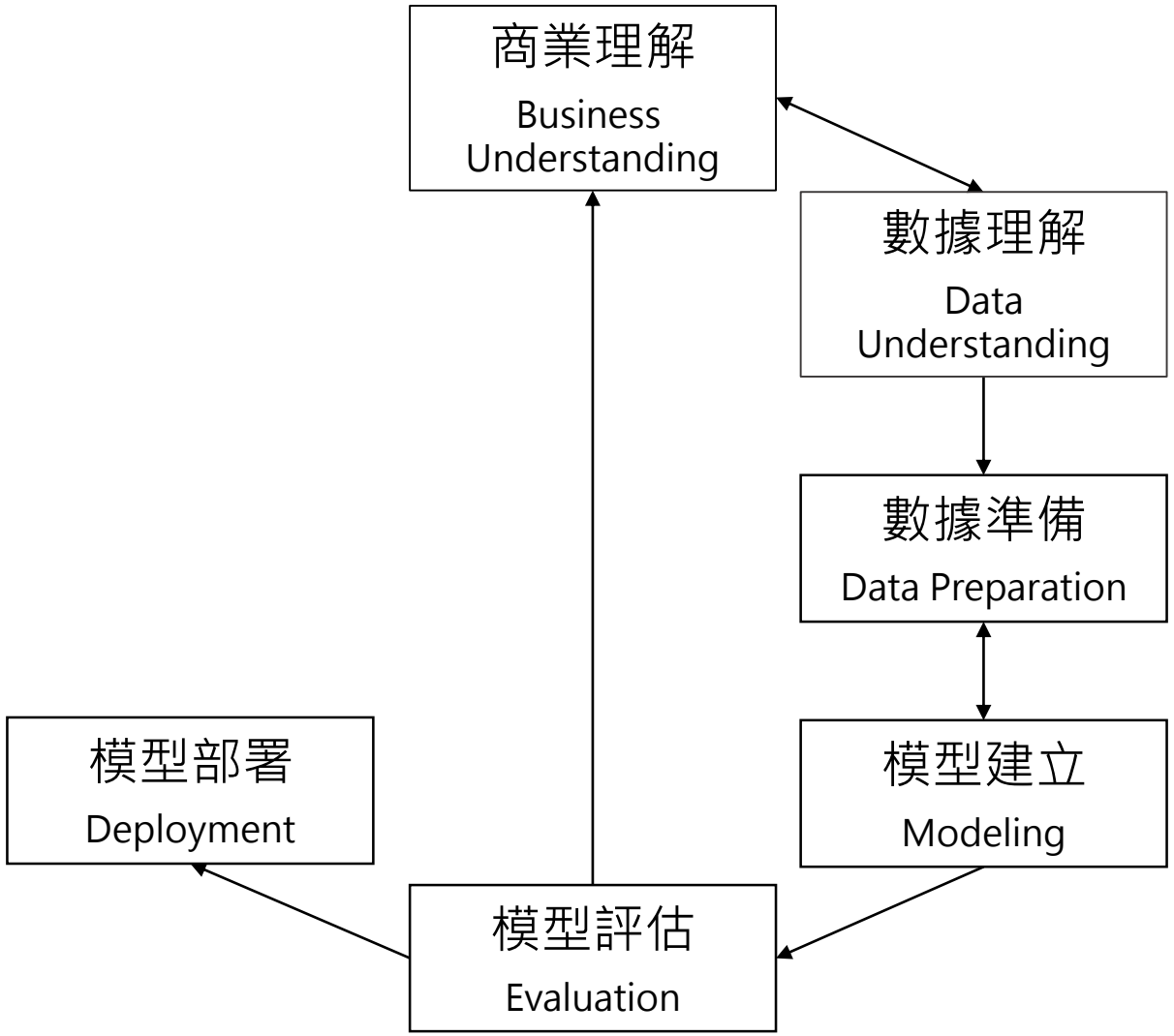
  統計、實驗設計、模擬、...

- 建立預測模型 f

  迴歸建模、機器學習、深度學習、...

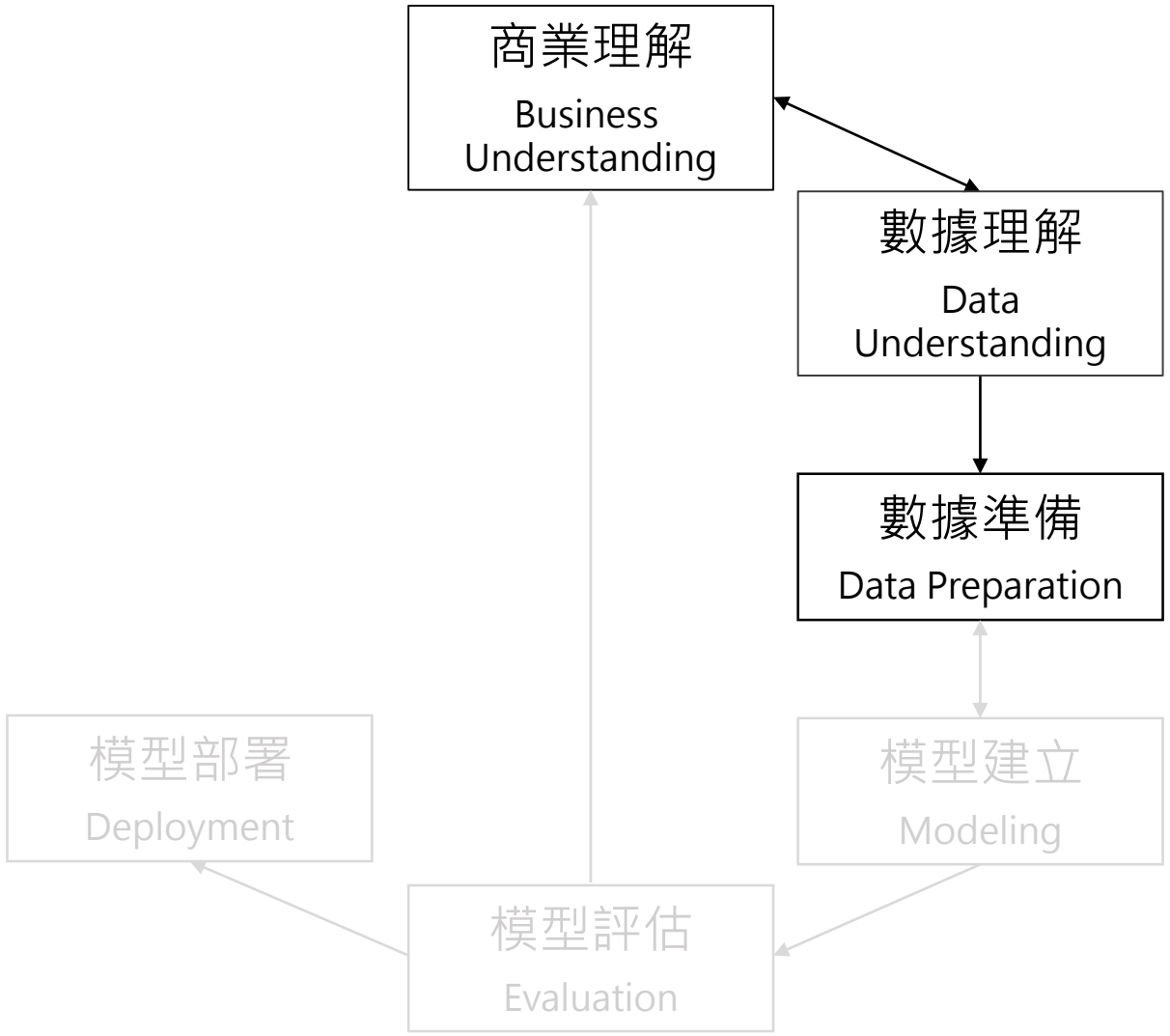
實務應用 – 案例

CRISP-DM



實務應用 – 案例

CRISP-DM

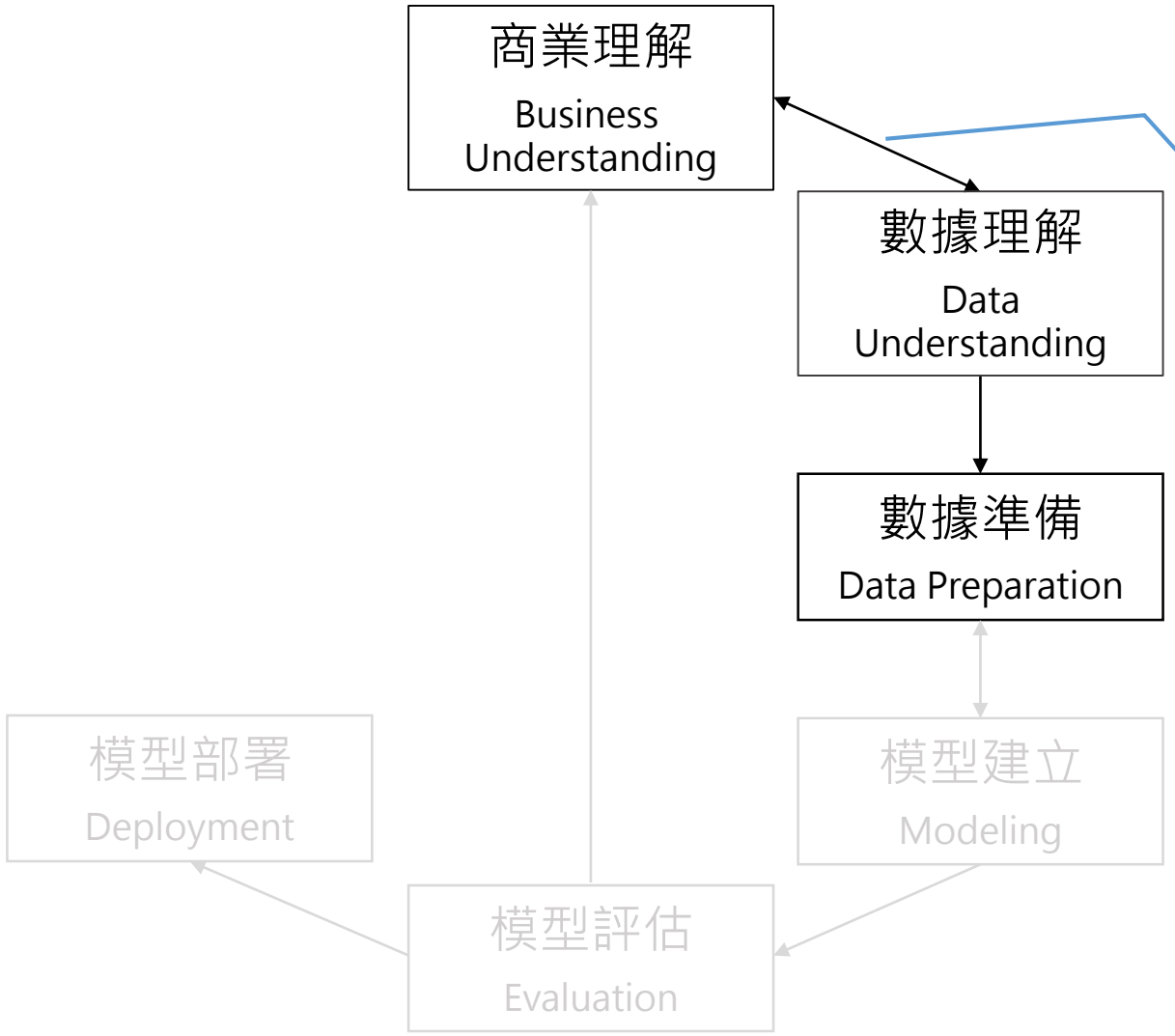


Goal :

尋找影響製程良率 y 的關鍵因子 x 有哪些？

實務應用 – 案例

CRISP-DM



Goal :

尋找影響製程良率 y 的關鍵因子 x 有哪些？

- XXX 製程
- 兩條生產線 : Line 1、Line 2
- 三道製程 :
  Process 1-1 ~ 1-20
  Process 2-1 ~ 2-70
  Process 3-1 ~ 3-50
- 一爐製作 48 個產品
- 11 種 Defect Code (報廢種類)

實務應用 – 案例

Line 1

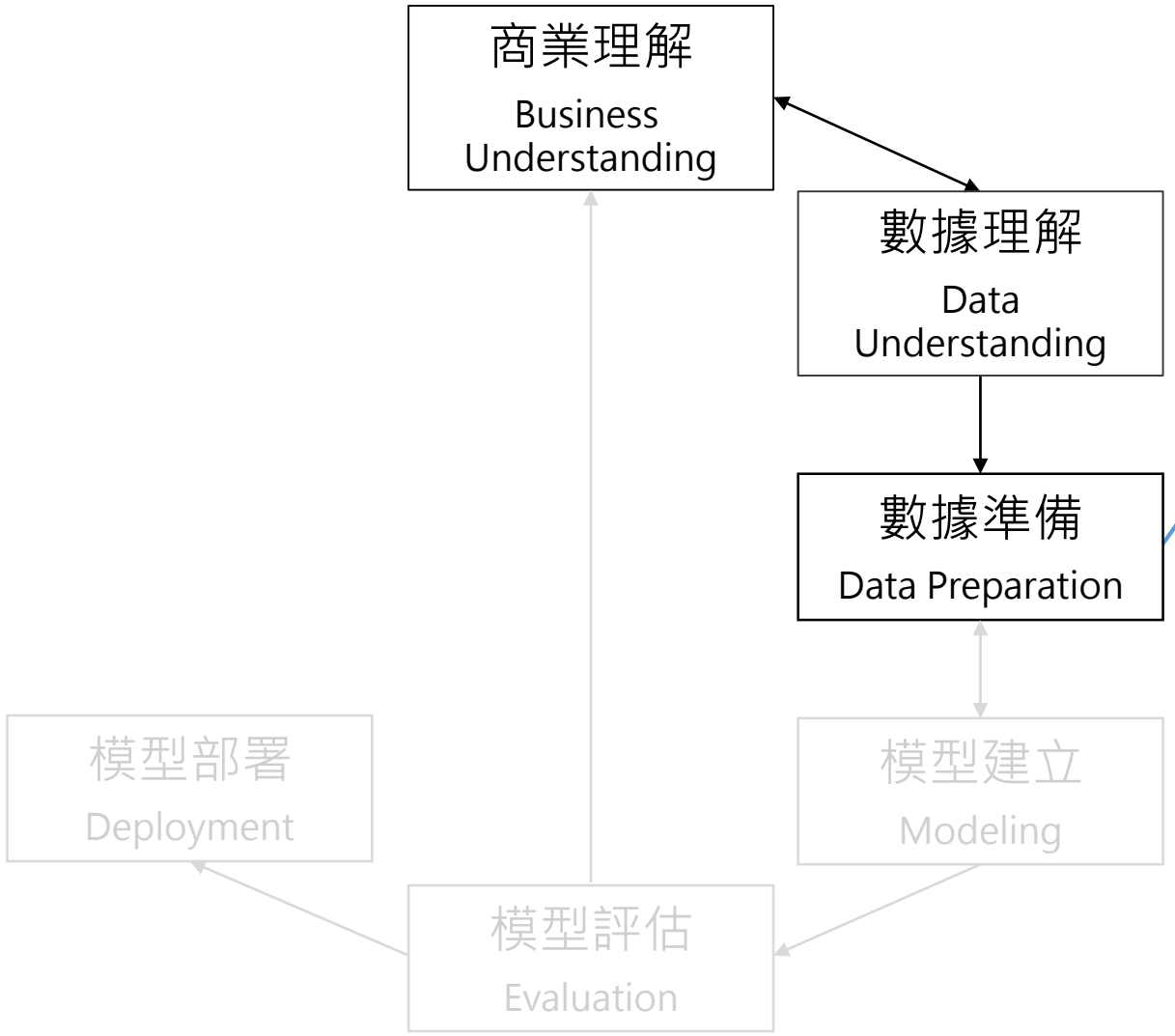
ID	1-1	…	1-20	2-1	…	2-70	3-1	…	3-50	OK	NG
0001	23	…	0.4	2.6	…	140	1000	…	2050	47	1
0002	30	…	0.6	3.1	…	145	1200	…	2100	48	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5221	18	…	0.55	3.5	…	138	1150	…	2150	46	2

Line 2

ID	1-1	…	1-20	2-1	…	2-70	3-1	…	3-50	OK	NG
0001	27	…	0.45	2.9	…	280	2000	…	2100	48	0
0002	36	…	0.46	4.1	…	255	3200	…	2200	47	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8309	28	…	0.56	6.1	…	240	2150	…	2350	45	3

實務應用 – 案例

CRISP-DM



Goal :

尋找影響製程良率 y 的關鍵因子 x 有哪些？

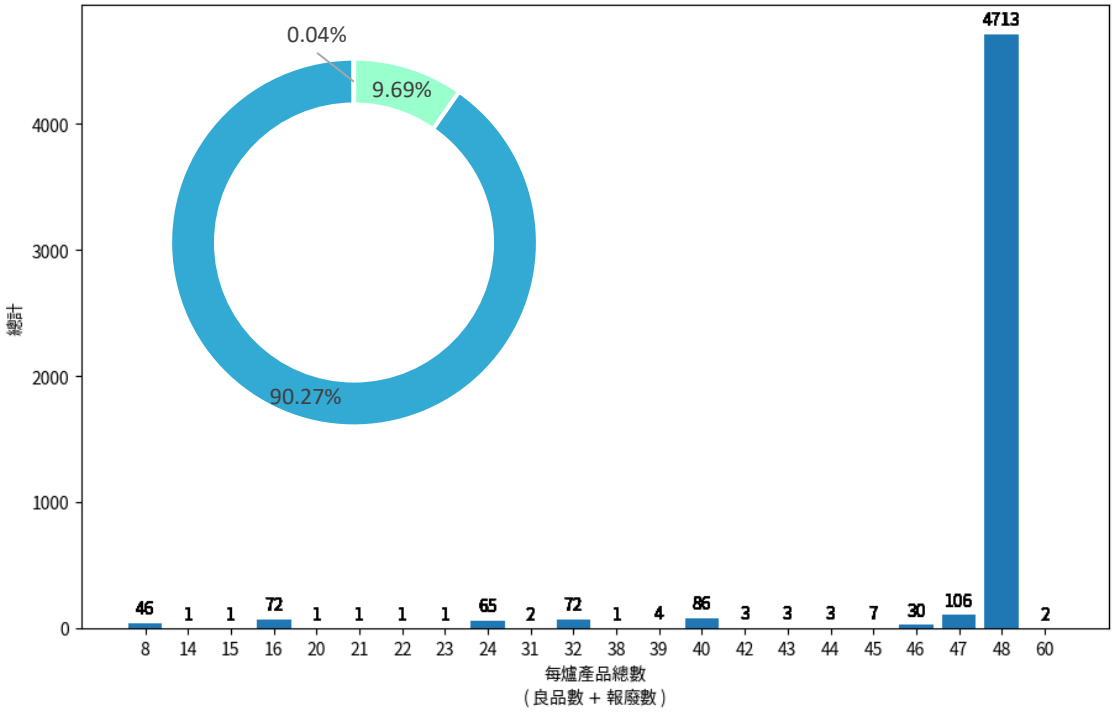
- 探索性資料分析 (Exploratory Data Analysis, EDA)
  - 資料品質： x 因子前處理問題：缺失值
    x 因子落在管制規格外
    x 因子共線性
    y 總數差異問題

實務應用 – 案例

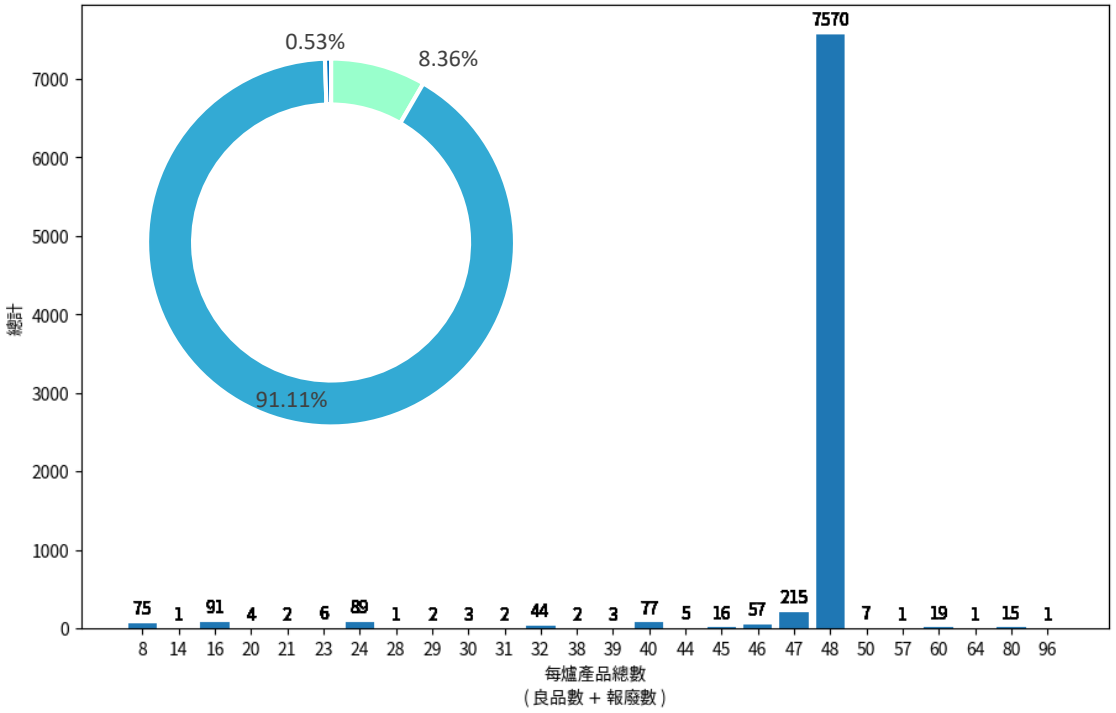
資料品質問題

y 總數差異

Line 1



Line 2

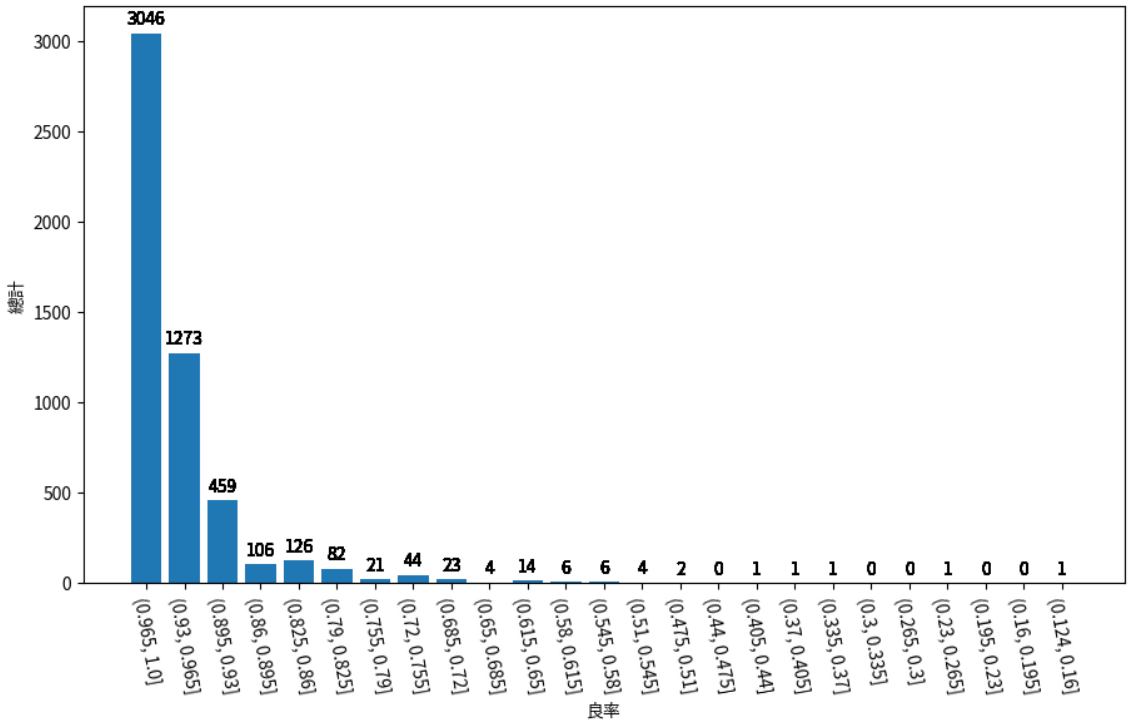


實務應用 – 案例

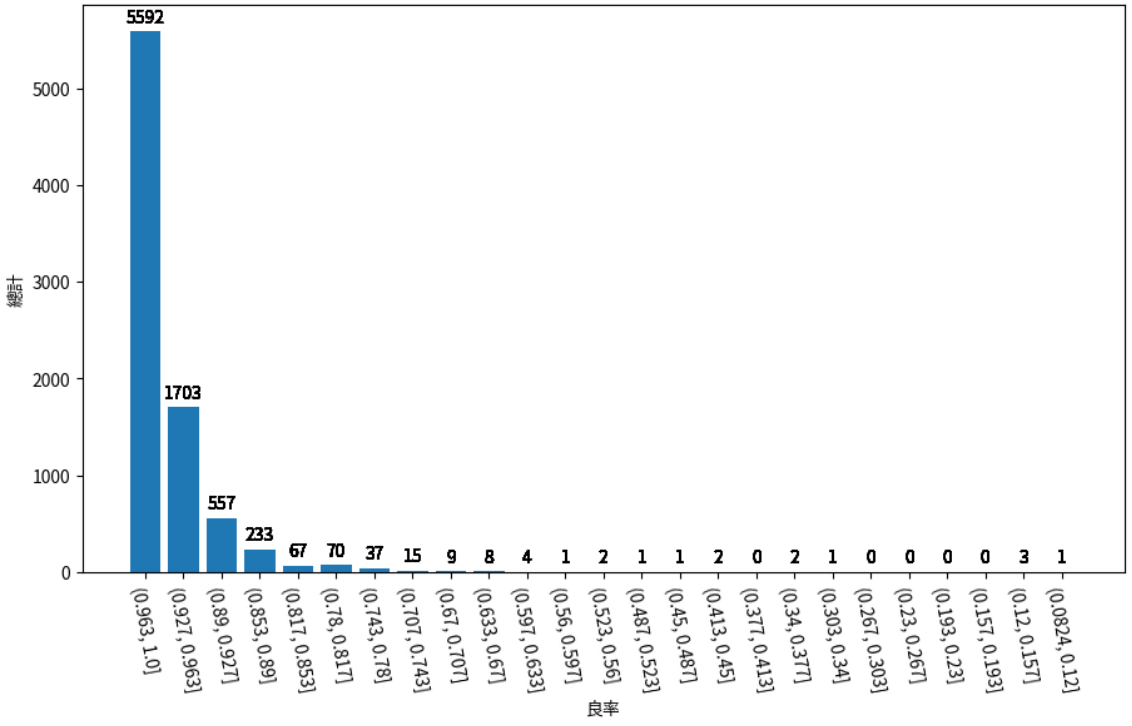
資料品質問題

y 總數差異 --> 影響良率

Line 1



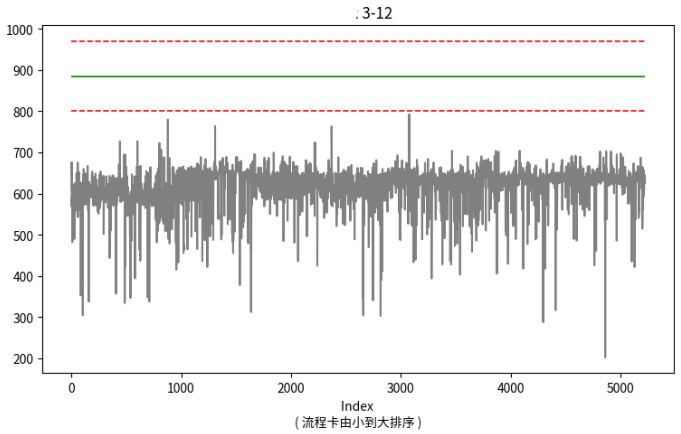
Line 2

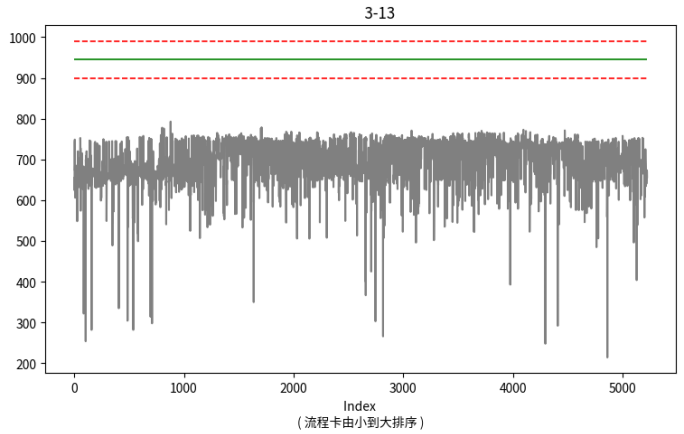


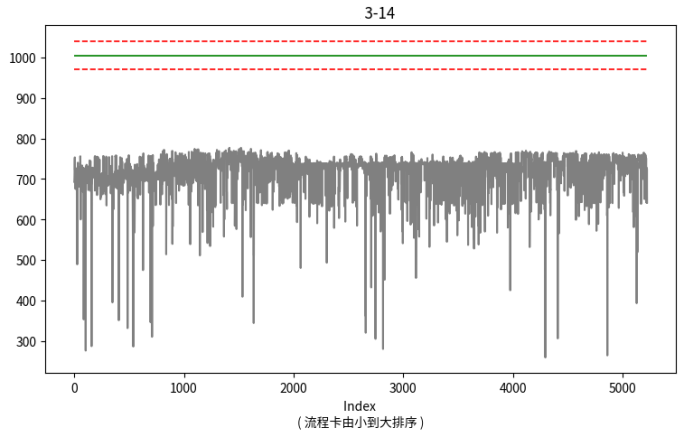
實務應用 – 案例

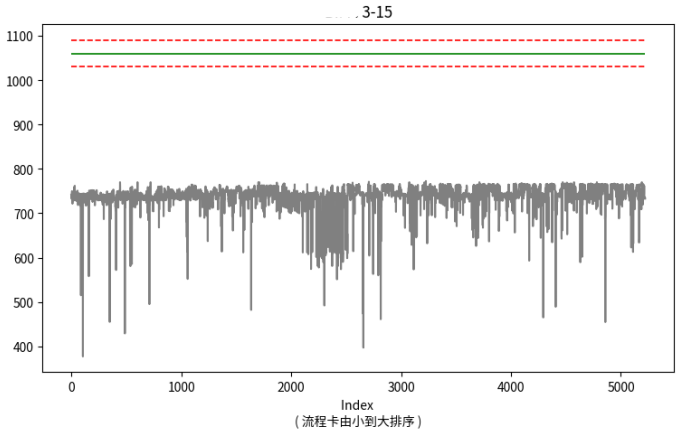
資料品質問題

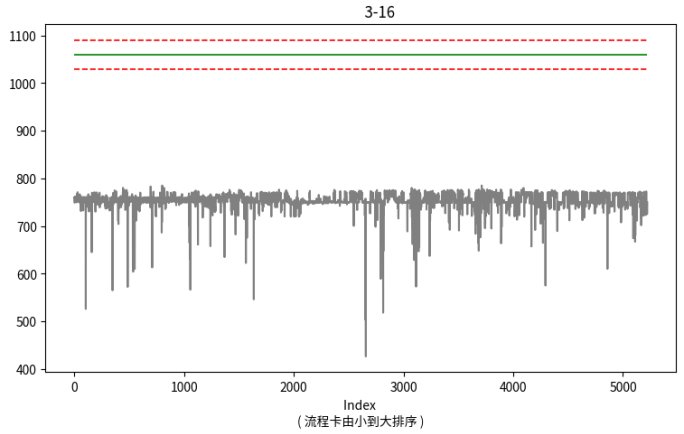
x 因子規格問題

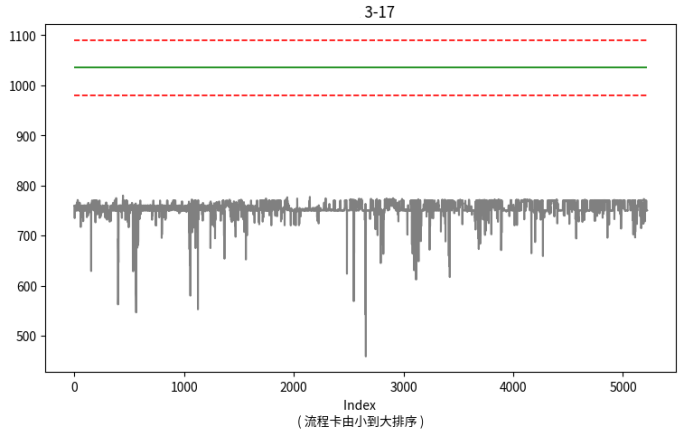








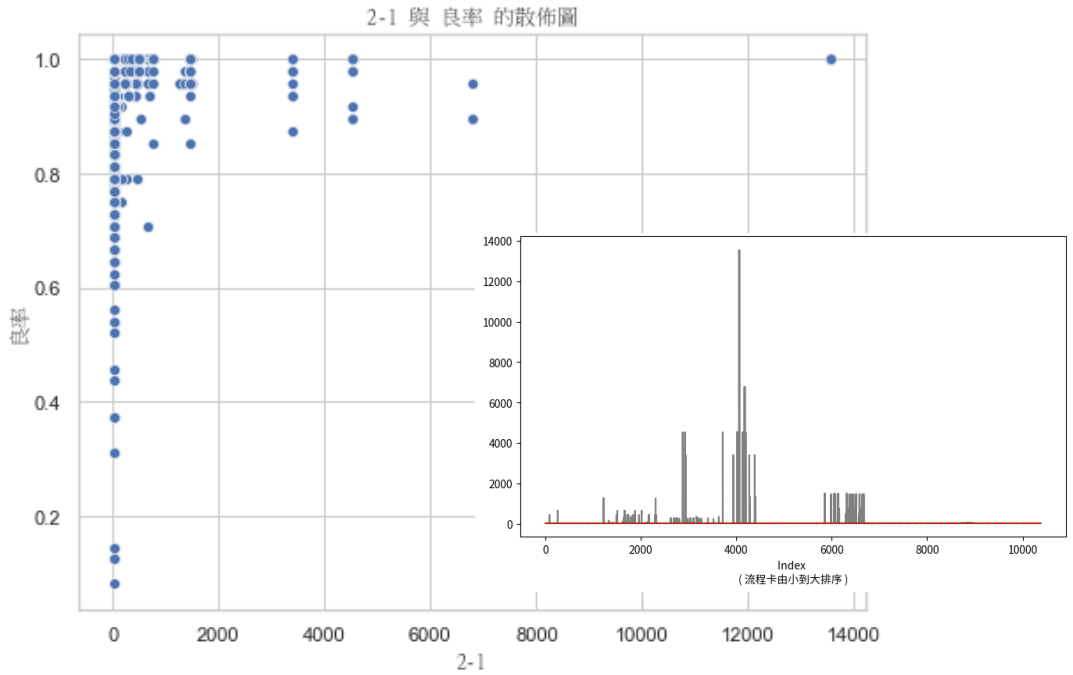


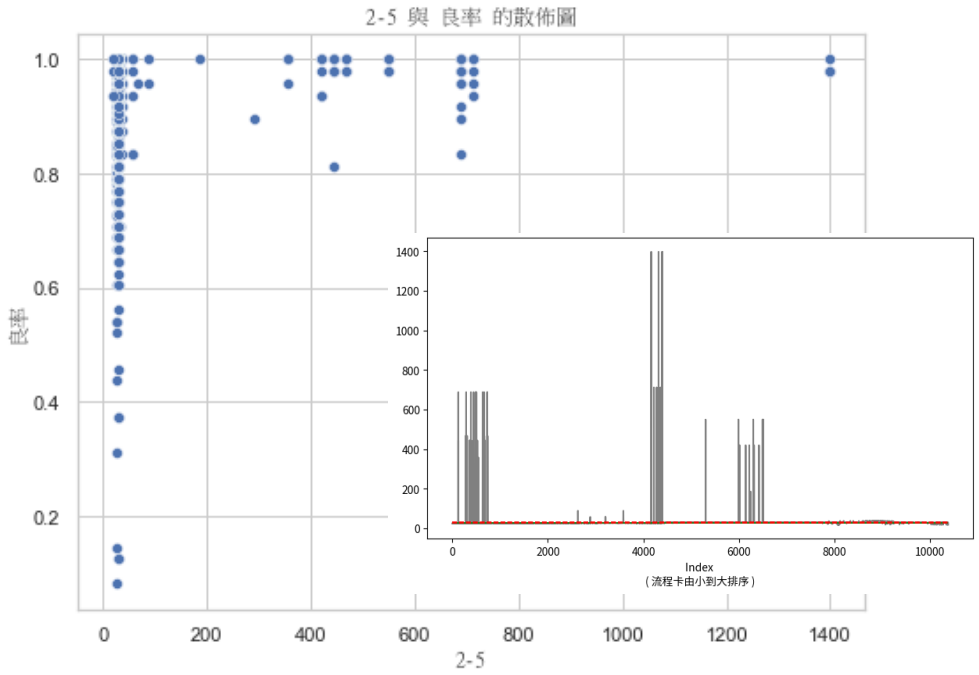


實務應用 – 案例

資料品質問題

x 因子規格問題





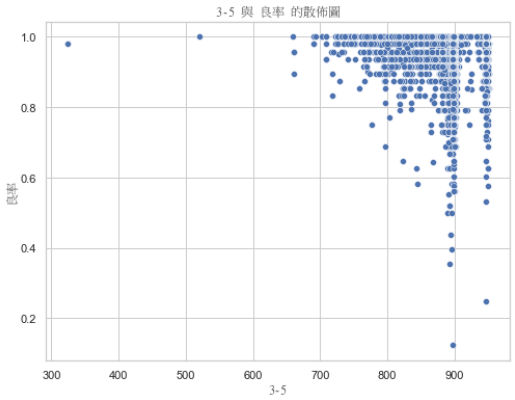
實務應用 – 案例

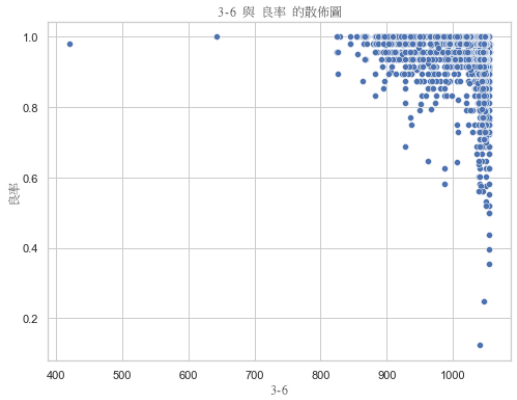
資料品質問題

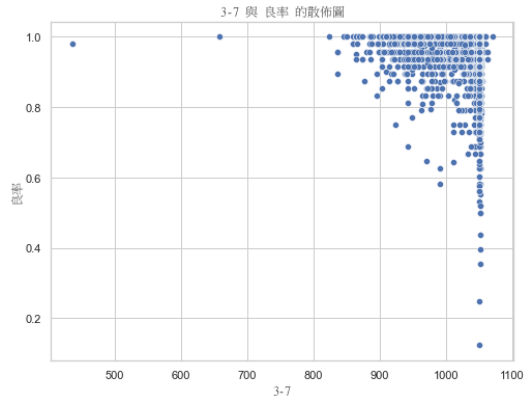
x 因子共線性

Line 1

	3-5	3-6	3-7
3-5	1		
3-6	0.893223	1	
3-7	0.883601	0.988765	1

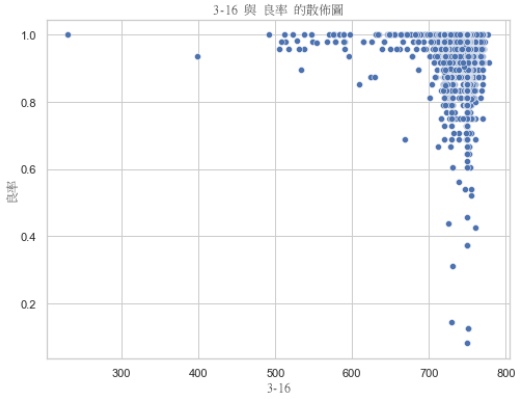


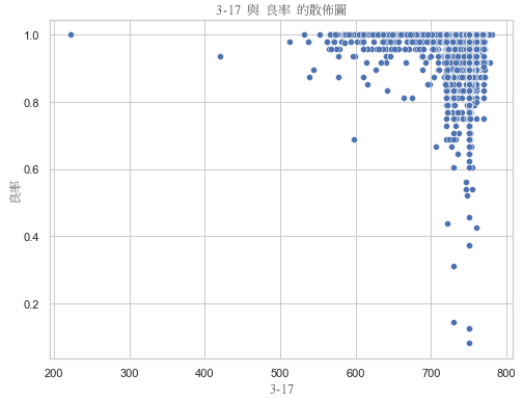


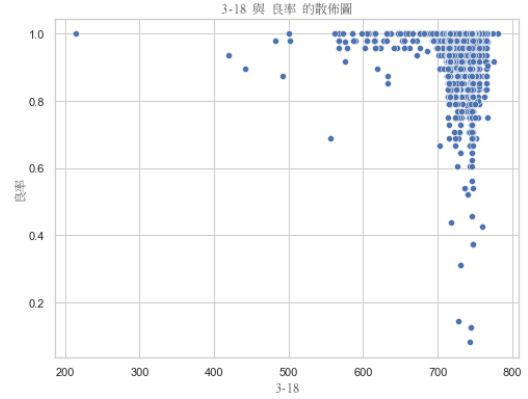


Line 2

	3/16	3/17	3/18
3/16	1	0.763482	0.842846
3/17	0.763482	1	0.805194
3/18	0.842846	0.805194	1

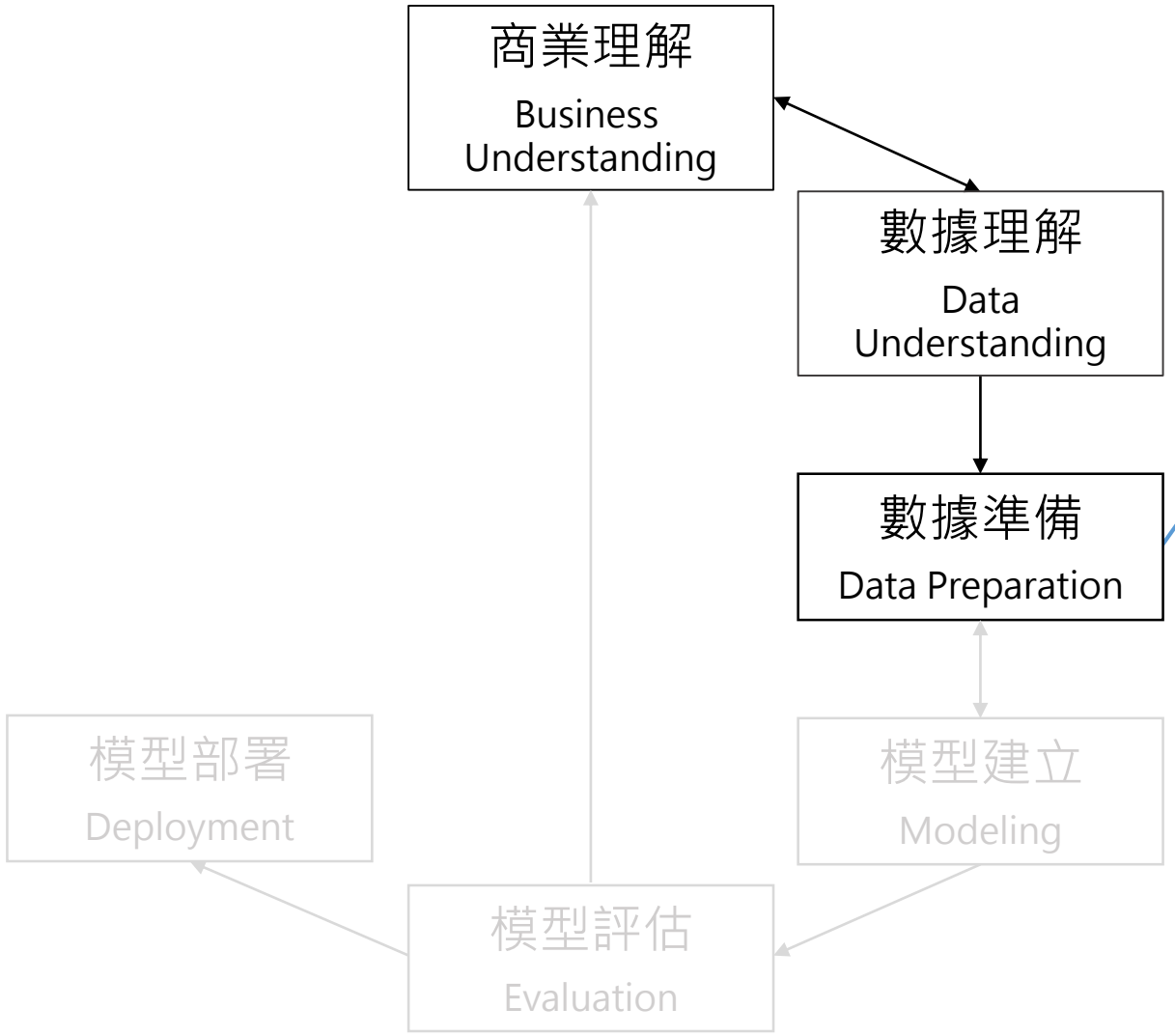






實務應用 – 案例

CRISP-DM



Goal :

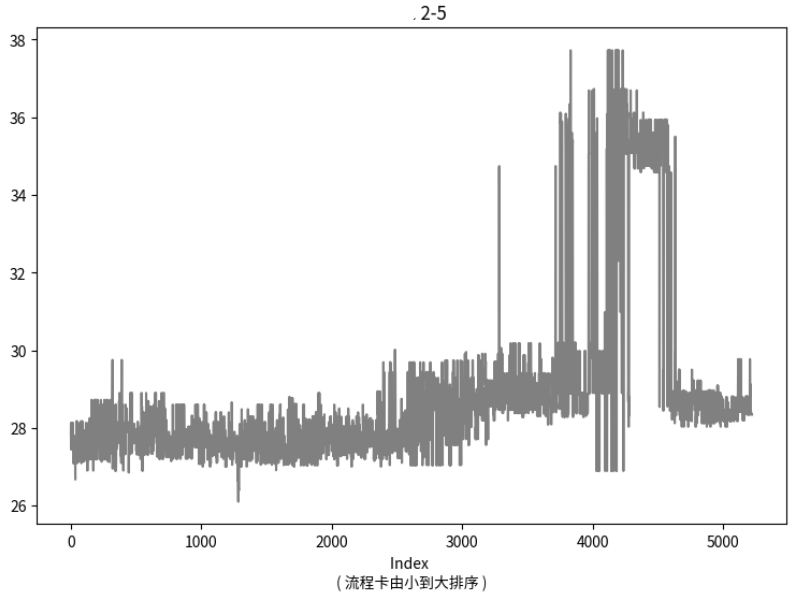
尋找影響製程良率 y 的關鍵因子 x 有哪些？

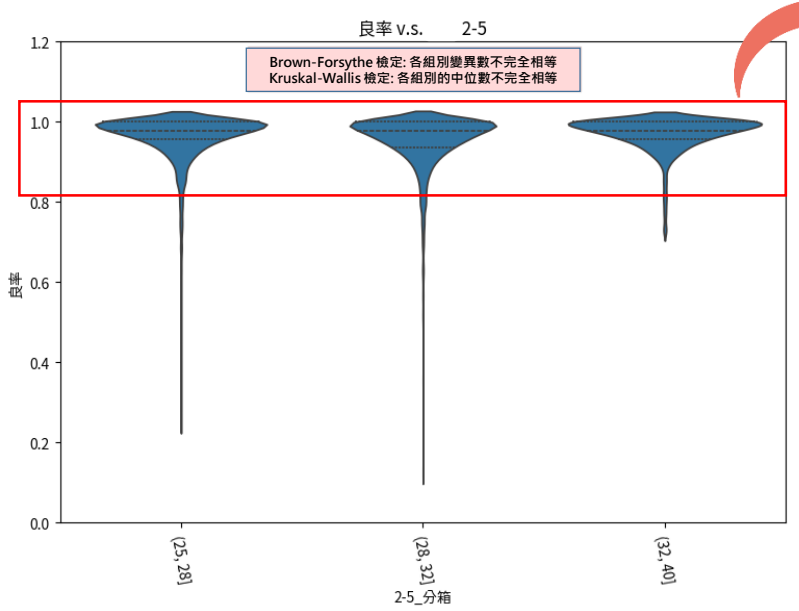
- 探索性資料分析 (Exploratory Data Analysis, EDA)
  - 資料品質： x因子前處理問題
    x因子落在管制規格外
    x因子共線性
    y總數差異問題
  - 相關性分析
    相關係數：
    Pearson、Spearman、Chatterjee、Mutual Info.
    假設檢定：
    中位數/變異數同質性檢定

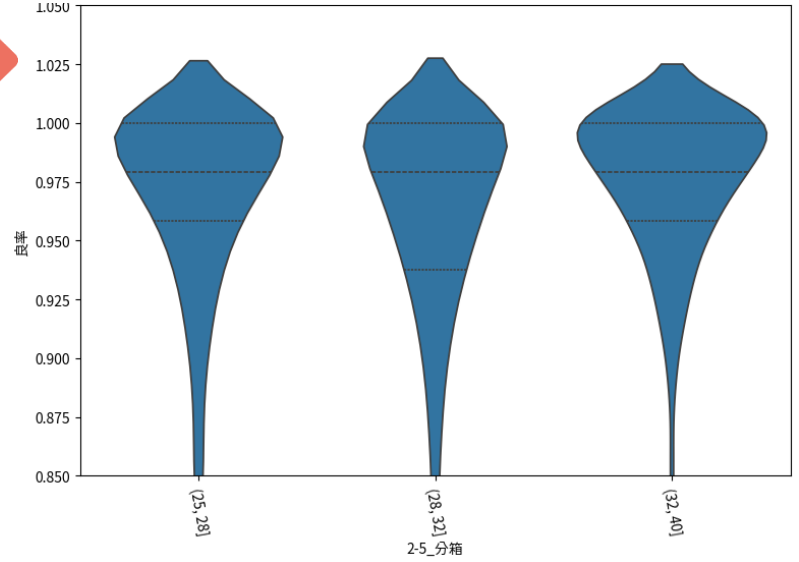
實務應用 – 案例

分析結果

x 因子與良率分布的差異性







學員問題回饋

如果未來想**成為資料科學家**，

- 應該如何針對**技能**和**知識**進加強準備？
  或有哪些需要特別學習或加強的領域？
- 如何找到**最適合**自己發展的產業？
- 在資料科學領中，大型企業和新創公司工作各有哪些優缺點？
  應該如何選擇？

如果未來想**成為資料科學家**，

◆應該如何針對**技能**和**知識**進加強準備？
或有哪些需要特別學習或加強的領域？

技能

- 基本的**程式能力**
  Python、R、SQL
- **數學統計**
  資料科學的基礎是數學，能幫助我們理解演算法的原理。
  統計在實務應用中，很常用來觀察與探尋資料的內涵。
  * 建議：**統計優先於數學**
- **演算法**
  瞭解建模的步驟
  理解模型的用途與適用性
  能解釋模型的結果。

知識

- **溝通**與**可視化**
  學習如何用圖表和簡單的語言，解釋複雜的數據。
- **領域知識**
  了解自己感興趣的產業背景知識，能幫助我們將技術更好地應用於實際問題。

如果未來想**成為資料科學家**，

◆如何找到**最適合**自己發展的產業？

- 產業的**類別**：零售、金融、保險、電子、石化、醫療、.... ← 我們要面對的事物跟什麼有關？
- 產業的**需求** ←我們能做什麼？或可以做什麼?
- 產業的**機會** ← 我們可投遞的職缺有哪些? 有多少職缺? 競爭者的多寡?
- 產業的**前景**與**趨勢** ← 企業的願景與未來

個人	產業
能力 > 興趣 生活 > 工作 (或 工作 > 生活)	需求 > 類別 發展 > 起薪

如果未來想**成為<u>資料科學家</u>**，

◆在資料科學領中，大型企業和新創公司工作各有哪些優缺點？
應該如何選擇？

企業	新創
資源豐富 分工細 技術深	資源有限 多工 技術廣

選擇時的考量面向

- 個人的職涯規劃與目標
- 個人的個性與喜好
- 工作與生活的權衡
- 現實問題

  薪資 (產業間有上限的區別)

  升遷/發展空間

Q & A