

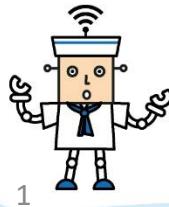


第1章

破冰！資料科學觀念養成

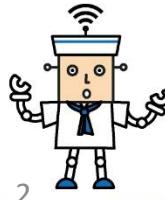
資料科學是一種以理性的數據分析，對充滿情感性的數據進行探索與決策的藝術，資料科學家也被評為21世紀最性感的職業。

懷抱著對大數據分析的憧憬情懷，一起跟著本章掌握其中奧妙，看如何以神祕的機器學習模型展開對未知的浪漫想像吧！



1-1 資料科學的概念

- 美國國家標準技術研究所 (NIST) 將**資料科學** (Data Science) 列為第四個科學典範
 - 理論科學、實驗科學、計算科學與資料科學
- 跟**資料** (Data，又稱**數據**) 有關的科學就是資料科學，包含**資料取得**、**資料處理**到**資料分析**的過程
- 資料經過處理後稱為**資訊** (Information)，最後從這些資訊中分析出來的訊息，就稱為**知識** (Knowledge)
- 再經過不斷的行動及驗證，逐漸形成**智慧** (Wisdom)。
- **大數據** (Big Data) 風起雲湧後，資料科學這門學問就顯得更加重要了。





▲ 資料科學的步驟 (後續各節一一簡述)



資料科學 × 機器學習

實戰探索

Practical Exploration



1-2 資料取得

- 資料通常以**表格**的方式呈現，若我們想知道「誰才是年度滾球大王」，就備妥滾球大賽的成績資料

▼ 滾球大賽分數

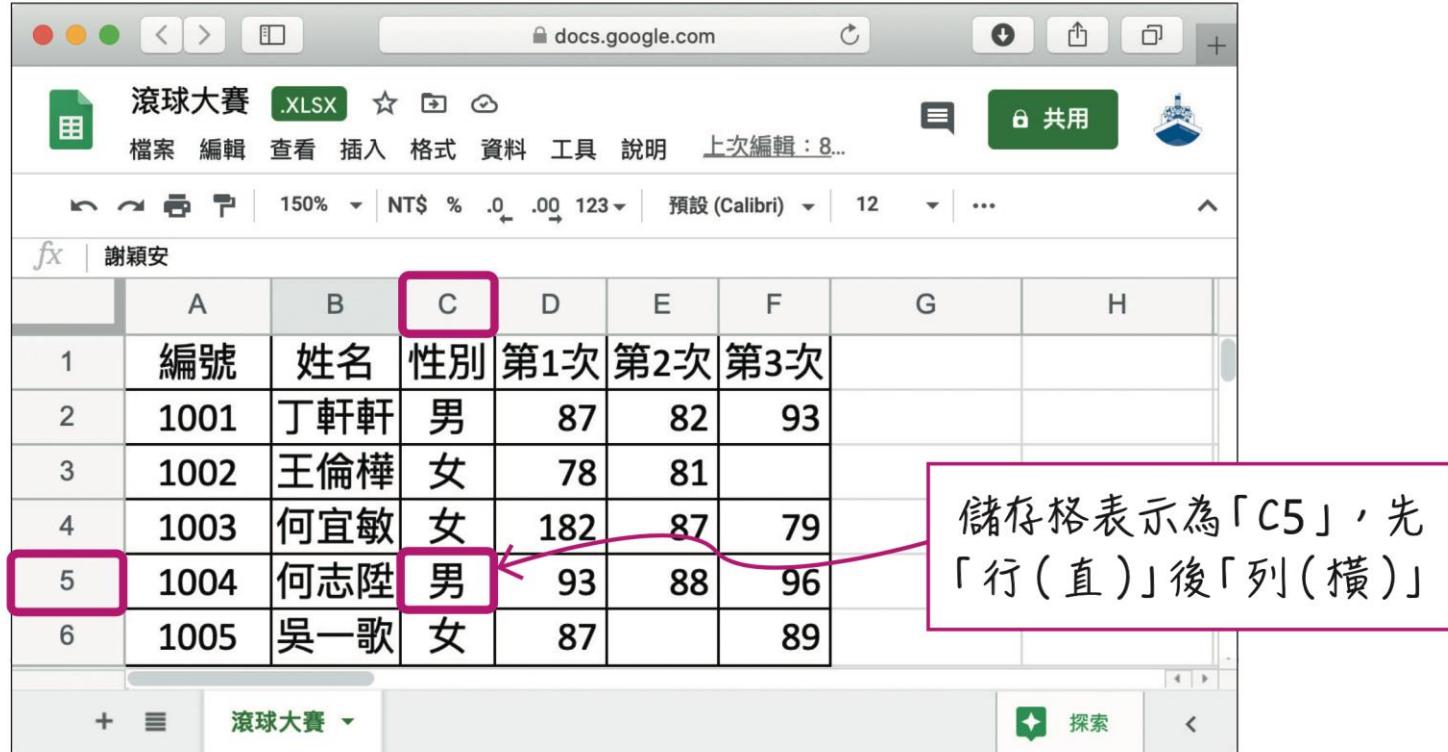
| 編號 | 姓名 | 性別 | 第1次 | 第2次 | 第3次 |
|------|-----|----|-----|-----|-----|
| 1001 | 丁軒軒 | 男 | 87 | 82 | 93 |
| 1002 | 王倫樺 | 女 | 78 | 81 | |
| 1003 | 何宜敏 | 女 | 182 | 87 | 79 |
| 1004 | 何志陞 | 男 | 93 | 88 | 96 |
| 1005 | 吳一歌 | 女 | 87 | | 89 |

直行為欄位

橫列為記錄

1-2 資料取得

- 我們會使用如 Microsoft Excel 或 Google 試算表等軟體來處理



docs.google.com

滾球大賽 .XLSX 共用

| | A | B | C | D | E | F | G | H |
|---|------|-----|----|-----|-----|-----|---|---|
| 1 | 編號 | 姓名 | 性別 | 第1次 | 第2次 | 第3次 | | |
| 2 | 1001 | 丁軒軒 | 男 | 87 | 82 | 93 | | |
| 3 | 1002 | 王倫樺 | 女 | 78 | 81 | | | |
| 4 | 1003 | 何宜敏 | 女 | 182 | 87 | 79 | | |
| 5 | 1004 | 何志陞 | 男 | 93 | 88 | 96 | | |
| 6 | 1005 | 吳一歌 | 女 | 87 | | 89 | | |

儲存格表示為「C5」，先「行(直)」後「列(橫)」

1-2 資料取得

- 資料科學界常用的則是 CSV 或 JSON 格式
- CSV (Comma-Separated Value，逗號分隔值) 格式中，每個欄位之間以逗號隔開，每筆資料之間則以換行來分隔。

| 編號 | 姓名 | 性別 | 第1次 | 第2次 | 第3次 |
|------|-----|----|-----|-----|-----|
| 1001 | 丁軒軒 | 男 | 87 | 82 | 93 |
| 1002 | 王倫樺 | 女 | 78 | 81 | |
| 1003 | 何宜敏 | 女 | 182 | 87 | 79 |
| 1004 | 何志陞 | 男 | 93 | 88 | 96 |
| 1005 | 吳一歡 | 女 | 87 | 89 | |
| 1006 | 宋緯挺 | 男 | 91 | 92 | |

▲ CSV 檔案的資料格式

1-2 資料取得

- JSON 的儲存方式為「{ 屬性 : 值 }」
- 每組資料以大括號包起來，字串需加上雙引號 (") 標註，數字則不用



```
[{"編號": "1001", "姓名": "丁軒軒", "性別": "男", "第1次": "87", "第2次": "82", "第3次": "93"}, {"編號": "1002", "姓名": "王倫樺", "性別": "女", "第1次": "78", "第2次": "81", "第3次": ""}, {"編號": "1003", "姓名": "何宜敏", "性別": "女", "第1次": "182", "第2次": "87", "第3次": "79"}, {"編號": "1004", "姓名": "何志陞", "性別": "男", "第1次": "93", "第2次": "88", "第3次": "96"}, {"編號": "1005", "姓名": "吳一欽", "性別": "女", "第1次": "87", "第2次": "", "第3次": "89"}, {"編號": "1006", "姓名": "宋緯挺", "性別": "男", "第1次": "91", "第2次": "92", "第3次": ""}, {"編號": "1007", "姓名": "李宇綸", "性別": "女", "第1次": "91", "第2次": "87", "第3次": "89"}, {"編號": "1008", "姓名": "李憲勝", "性別": "男", "第1次": "94", "第2次": "100", "第3次": ""}, {"編號": "1009", "姓名": "杜以潔", "性別": "女", "第1次": "82", "第2次": "88", "第3次": "81"}, {"編號": "1010", "姓名": "沈程隆", "性別": "男", "第1次": "78", "第2次": "82", "第3次": "81"}, {"編號": "1011", "姓名": "沈慈惠", "性別": "女", "第1次": "86", "第2次": "88", "第3次": "82"}, {"編號": "1012", "姓名": "林絜峰", "性別": "男", "第1次": "192", "第2次": "80", "第3次": "95"}, {"編號": "1013", "姓名": "林保苓", "性別": "女", "第1次": "82", "第2次": "87", "第3次": "86"}, {"編號": "1014", "姓名": "林宏銘", "性別": "男", "第1次": "91", "第2次": "92", "第3次": ""}, {"編號": "1015", "姓名": "邱堂儀", "性別": "女", "第1次": "93", "第2次": "", "第3次": "92"}, {"編號": "1016", "姓名": "施帆蓉", "性別": ""}]
```

▲ JSON 檔案的資料格式

1-2-1 開放資料和資料集網站

- **開放資料 (Open Data)** 是一種可以開放和允許任何人自由存取、使用、修改以及分享的資料。

The screenshot shows the 'Datasets' search results page for 'PM2.5'. A search bar at the top contains 'PM2.5' with a magnifying glass icon. Below the search bar are sorting options ('排序') and a dropdown for search results ('搜尋結果匯出'). The main content area displays two datasets:

- 臺南市空氣品質微型感測器監測資料**: This dataset is from the Central Government Agency (行政院). It includes a 'JSON' download button. The description states it's from the Environmental Protection Agency (行政院環保署) and provides API details. It was last updated on April 8, 2021.
- 臺南市空氣品質監測小時值**: This dataset is from the Local Government (臺南市政府). It includes a 'CSV' download button. The description states it's from the Environmental Protection Agency (行政院環保署) and provides API details. It was last updated on April 8, 2021.

On the left side, there are filters for '中央機關' (Central Government), '地方機關' (Local Government), and '法人機關' (Legal Person). There are also sections for '主題分類' (Topic Category) and '服務分類' (Service Category).

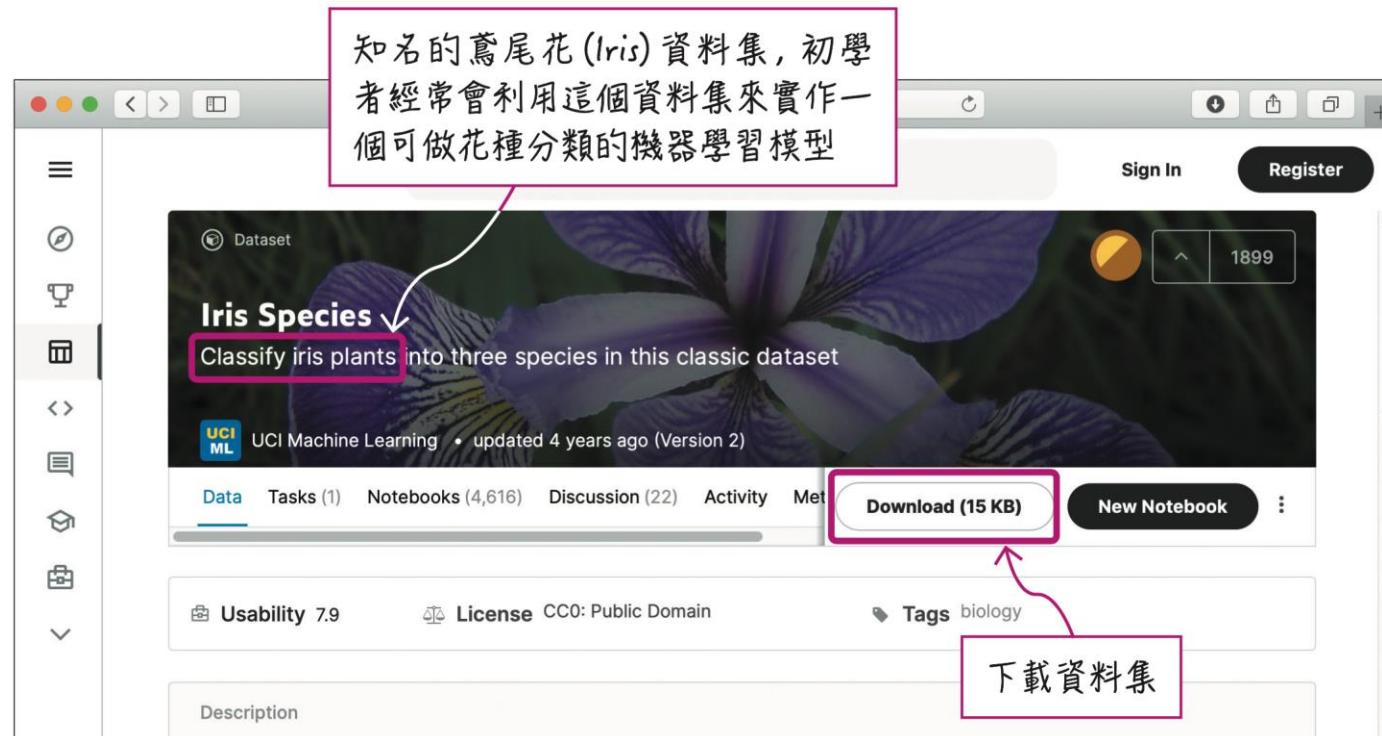
A callout box highlights the 'PM2.5日均值(每日提供)' dataset, noting its daily provision and providing a detailed description of its fields (SiteId, SiteName, County, ItemId, ItemName, ItemEngName) and source (Environmental Protection Agency).

Another callout box points to the download buttons for the PM2.5 dataset, stating: '可供下載的資料集格式，若想做PM2.5相關研究就可加以利用' (The data set format can be downloaded, which is useful for PM2.5 related research).

At the bottom, a note indicates: ▲ 在政府資料開放平臺中搜尋「PM2.5」 (Search for 'PM2.5' on the Government Open Data Platform).

1-2-1 開放資料和資料集網站

- 資料集網站「**kaggle**」是坊間頗受歡迎的資料科學競賽平台



▲ 資料集網站「kaggle」中的 Iris 資料集 (Iris.csv)

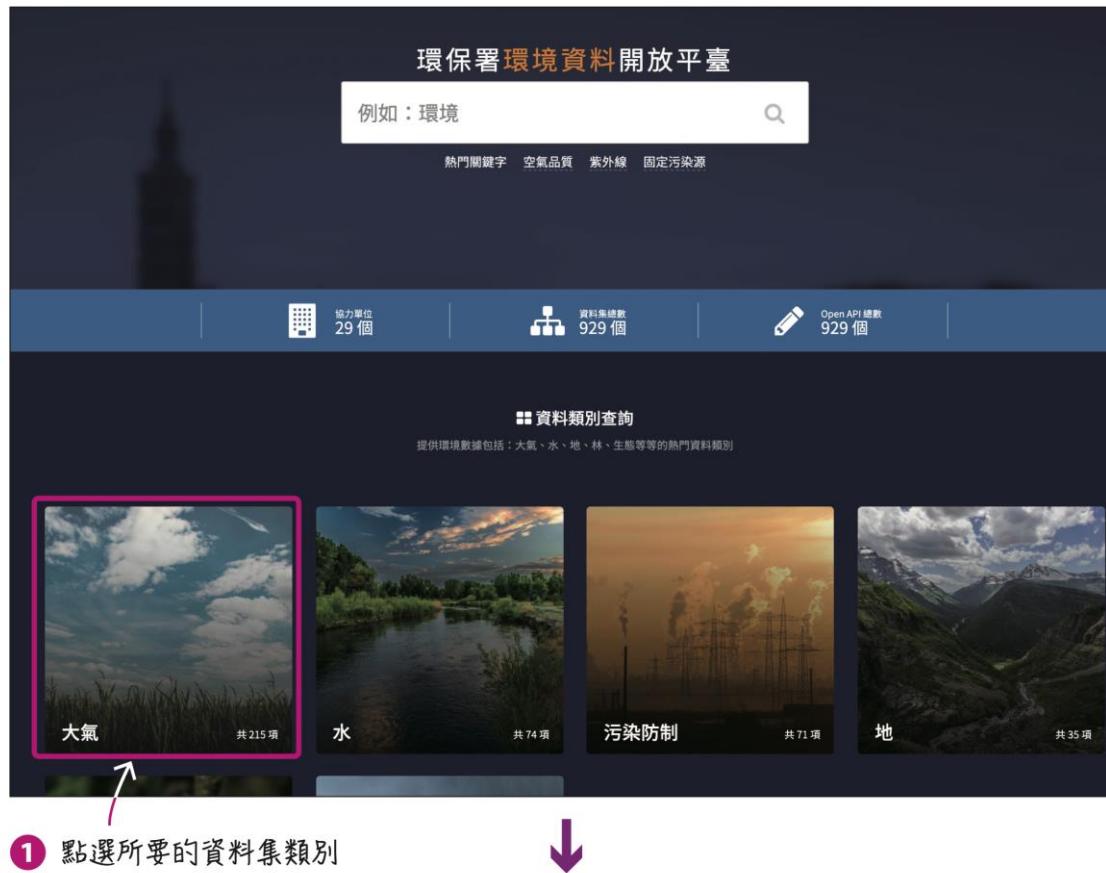
1-2-2 手動下載資料檔

- 開放資料平台所提供的資料，通常可以透過「手動下載」自行找到並下載相關的檔案
- 或者是利用「**網路爬蟲**」的方式自動擷取我們想要的資料。



1-2-2 手動下載資料檔

- 從行政院環保署的「環境資料開放平臺」手動下載每日的「PM2.5日均值」資料



1-2-2 手動下載資料檔

- 從行政院環保署的「環境資料開放平台」手動下載每日的「PM2.5日均值」資料

The screenshot shows a search interface for environmental data. In the search bar at the top left, the text "pm2.5" is entered. To the right of the search bar is a "Q 搜尋" button. Below the search bar, a message says "根據\"pm2.5\"找到13個資料集". On the right, there is a dropdown menu labeled "排序依照: 關聯". Underneath, a filter "資料集類別: 大氣" is applied. The main area displays a table of 13 datasets. The columns are "資料集名稱", "資料集類別", "資料格式", "最後更新", and "詳細". The first dataset listed is "空氣品質即時污染指標(含PM2.5)", which is categorized under "環境監測及資訊處" and has "大氣" as its type. It supports "CSV, JSON, XML" formats and was last updated on "2021/04/06". The second dataset is "細懸浮微粒資料 (PM2.5)", also from "環境監測及資訊處" and "大氣" type, updated on "2021/04/10". The third dataset, highlighted with a pink border and a pink bracket, is "PM2.5日均值(每日提供)", also from "環境監測及資訊處" and "大氣" type, updated on "2021/04/09". The fourth dataset is "PM2.5化學成分監測數據", also from "環境監測及資訊處" and "大氣" type, updated on "2021/04/09". A pink arrow points from the number 2 in the bottom left to the third dataset, and a pink arrow points from the bottom right to the text "搜尋所需要的資料集".

| 資料集名稱 | 資料集類別 | 資料格式 | 最後更新 | 詳細 |
|--------------------------------|-------|----------------|------------|-----|
| 空氣品質即時污染指標(含PM2.5) 環境監測及資訊處 | 大氣 | CSV, JSON, XML | 2021/04/06 | (1) |
| 細懸浮微粒資料 (PM2.5) 環境監測及資訊處 | 大氣 | CSV, JSON, XML | 2021/04/10 | (1) |
| PM2.5日均值(每日提供) 環境監測及資訊處 | 大氣 | CSV, JSON, XML | 2021/04/09 | (1) |
| PM2.5化學成分監測數據 環境監測及資訊處 | 大氣 | CSV, JSON, XML | 2021/04/09 | (1) |

② 搜尋所需要的資料集

1-2-2 手動下載資料檔

- 從行政院環保署的「環境資料開放平台」手動下載每日的「PM2.5日均值」資料



1-2-2 手動下載資料檔

- 從行政院環保署的「環境資料開放平台」手動下載每日的「PM2.5日均值」資料

PM2.5日均值(每日提供)

根據資料集摘要

提供PM2.5日均值(每日提供)

來源：PM2.5日均值(每日提供)

Q 資料瀏覽器 D 資料字典

三 打開篩選面板

下載篩選後資料 JSON CSV XML

3 可選擇所要下載的檔案格式

78 records « 1 - 78 »

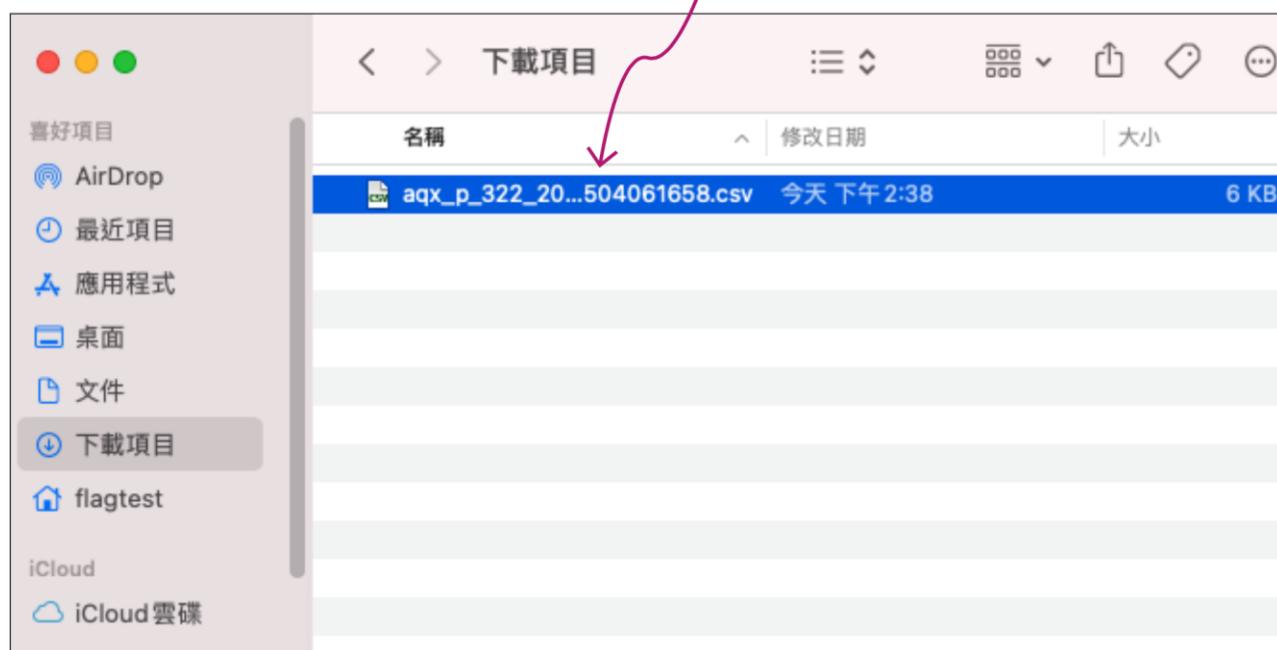
| 測站代碼 | 測站名稱 | 縣市 | 測項代碼 | 測項名稱 | 測項英... | 測項單位 | 監測日期 | 數值 |
|------|------|-----|------|--------|--------|--------------------------|-----------|----|
| 1 | 基隆 | 基隆市 | 33 | 細懸浮... | PM2.5 | $\mu\text{g}/\text{m}^3$ | 2021-0... | 14 |
| 2 | 汐止 | 新北市 | 33 | 細懸浮... | PM2.5 | $\mu\text{g}/\text{m}^3$ | 2021-0... | 15 |



1-2-2 手動下載資料檔

- 從行政院環保署的「環境資料開放平台」手動下載每日的「PM2.5日均值」資料

④ 此例是下載 CSV 檔並儲存在電腦內



▲ 環境資源資料開放平台中的「PM2.5 日均值」資料

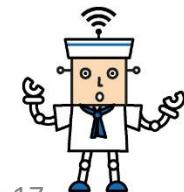
1-3 資料處理

- 常見**資料處理** (Data Processing) 項目：
- **資料清理**:刪除不必要或重複的紀錄、刪除異常資料及補缺失值等
- **資料轉換**:將分類的資料轉為數值
 - 例如:「女 / 男」轉換為「0/1」，「甲、乙、丙」等級轉換為「1、2、3」等
- **資料統計**:進行運算產生新的數據
 - 例如:計算總和、平均、最大或最小值、排序等。



1-3-1 資料清理

- 資料清理 (Data Cleaning) 就是去除不需要的資料，或是補上殘缺的資料
- 常見的清理動作有刪除異常資料、刪除不必要的欄位及補值等。



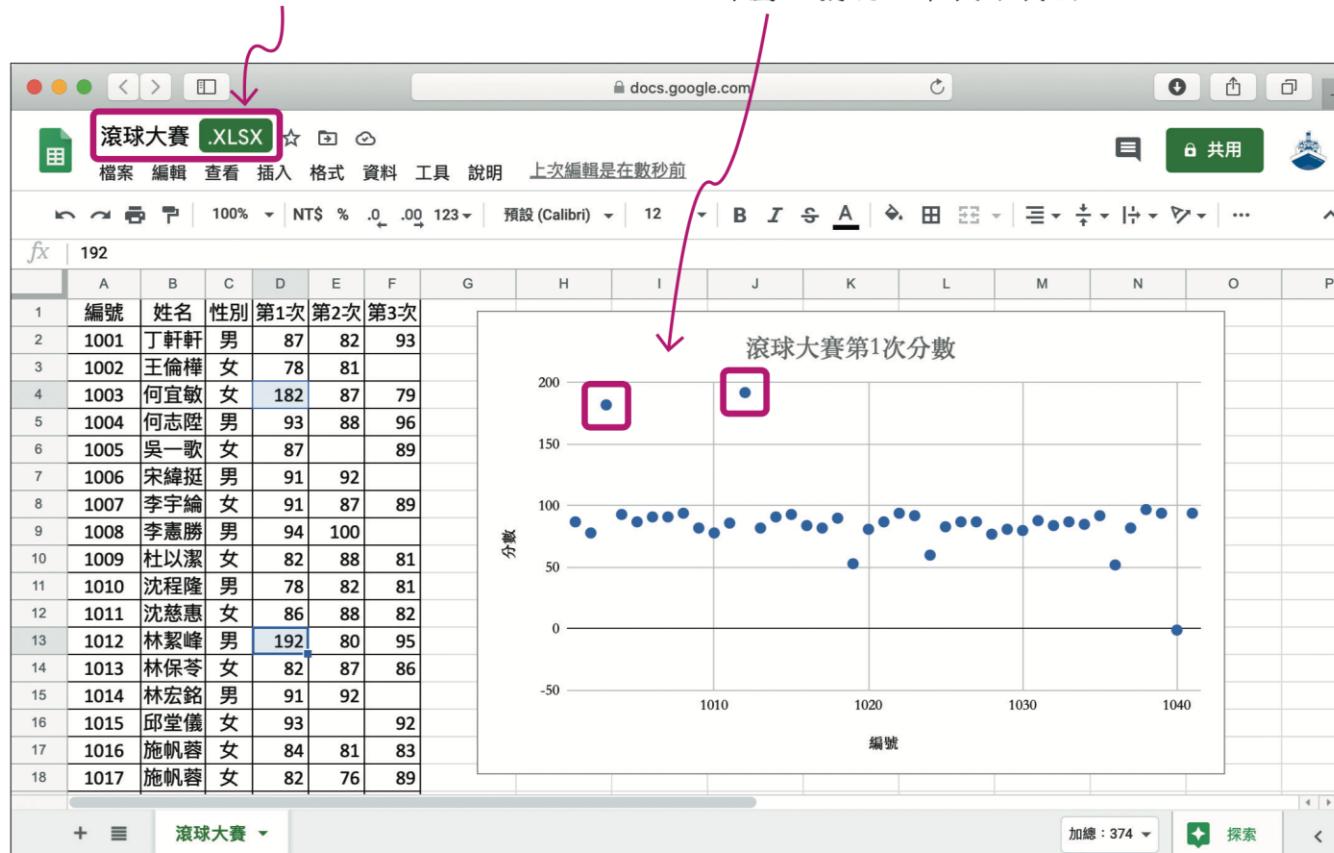


刪除異常資料

舉個例子，這裡以 Google 試算表^{註 7}開啟滾球大賽的統計數據。滾球大賽每次滿分為「100」分，從下圖的散佈圖中發現有 2 筆異常的分數，應該要將之更正或刪除，以免日後統計分數時產生錯誤的結果。

以 Google 試算表處理滾球大賽的數據

在散佈圖上發現 2 筆異常資料





刪除不必要的欄位

在下圖的滾球大賽統計數據中，如果想製作第 1 次比賽分數的英雄排行榜，分析時用不到的「性別」、「第 2 次」、「第 3 次」等欄位便可將之刪除。

不必要的欄位，
可刪除

| | A | B | C | D | E | F | G |
|----|------|-----|----|-----|-----|-----|---|
| 1 | 編號 | 姓名 | 性別 | 第1次 | 第2次 | 第3次 | |
| 2 | 1001 | 丁軒軒 | 男 | 87 | 82 | 93 | |
| 3 | 1002 | 王倫樺 | 女 | 78 | 81 | | |
| 4 | 1003 | 何宜敏 | 女 | 182 | 87 | 79 | |
| 5 | 1004 | 何志陞 | 男 | 93 | 88 | 96 | |
| 6 | 1005 | 吳一歌 | 女 | 87 | | 89 | |
| 7 | 1006 | 宋緯挺 | 男 | 91 | 92 | | |
| 8 | 1007 | 李宇綸 | 女 | 91 | 87 | 89 | |
| 9 | 1008 | 李憲勝 | 男 | 94 | 100 | | |
| 10 | 1009 | 杜以潔 | 女 | 82 | 88 | 81 | |
| 11 | 1010 | 沈程隆 | 男 | 78 | 82 | 81 | |
| 12 | 1011 | 沈慈惠 | 女 | 86 | 88 | 82 | |
| 13 | 1012 | 林絜峰 | 男 | 192 | 80 | 95 | |

▲ 資料清理：依需求刪除不必要的欄位



補值

若資料表中有缺少資料的部份，可進行補值的動作以減少統計結果出現太大的誤差。例如在下圖滾球大賽分數表中有部分參賽者沒有當次的比賽分數，應該進行補值；如果無法補值，則可以考慮刪除這筆記錄或用平均數（或眾數）來取代。

| 1 | 編號 | 姓名 | 性別 | 第1次 | 第2次 | 第3次 | |
|----|------|-----|----|-----|-----|-----|--|
| 2 | 1001 | 丁軒軒 | 男 | 87 | 82 | 93 | |
| 3 | 1002 | 王倫樺 | 女 | 78 | 81 | | |
| 4 | 1003 | 何宜敏 | 女 | 182 | 87 | 79 | |
| 5 | 1004 | 何志陞 | 男 | 93 | 88 | 96 | |
| 6 | 1005 | 吳一歌 | 女 | 87 | | 89 | |
| 7 | 1006 | 宋緯挺 | 男 | 91 | 92 | | |
| 8 | 1007 | 李宇綸 | 女 | 91 | 87 | 89 | |
| 9 | 1008 | 李憲勝 | 男 | 94 | 100 | | |
| 10 | 1009 | 杜以潔 | 女 | 82 | 88 | 81 | |
| 11 | 1010 | 沈程隆 | 男 | 78 | 82 | 81 | |
| 12 | 1011 | 沈慈惠 | 女 | 86 | 88 | 82 | |
| 13 | 1012 | 林絜峰 | 男 | 192 | 80 | 95 | |

空值時，應
進行補值

◀ 資料清理：缺少資料的部份進行補值

1-3-2 資料統計

- 如果要探索「誰才是年度滾球大王」，絕對少不了需要知道**總分**、**排名**等資訊

| | A | B | C | D | E | F | G | H |
|----|------|-----|----|-----|-----|-----|-----|----|
| 1 | 編號 | 姓名 | 性別 | 第1次 | 第2次 | 第3次 | 總分 | 排名 |
| 2 | 1001 | 丁軒軒 | 男 | 87 | 82 | 93 | 262 | 18 |
| 3 | 1002 | 王倫樺 | 女 | 78 | 81 | 85 | 244 | 32 |
| 4 | 1003 | 何宜敏 | 女 | 82 | 87 | 79 | 248 | 29 |
| 5 | 1004 | 何志陞 | 男 | 93 | 88 | 96 | 277 | 9 |
| 6 | 1005 | 吳一歌 | 女 | 87 | 85 | 89 | 261 | 19 |
| 7 | 1006 | 宋緯挺 | 男 | 91 | 92 | 76 | 259 | 21 |
| 8 | 1007 | 李宇綸 | 女 | 91 | 87 | 89 | 267 | 15 |
| 9 | 1008 | 李憲勝 | 男 | 94 | 100 | 85 | 279 | 7 |
| 10 | 1009 | 杜以潔 | 女 | 82 | 88 | 81 | 251 | 26 |
| 11 | 1010 | 沈程隆 | 男 | 78 | 82 | 81 | 241 | 34 |
| 12 | 1011 | 沈慈惠 | 女 | 86 | 88 | 82 | 256 | 23 |
| 13 | 1012 | 林絜峰 | 男 | 92 | 80 | 95 | 267 | 15 |

▶ 資料統計後的分數資訊：總分及排名

經過資料處理後產生有用的資訊

1-4 資料分析

- 資料分析 (Data Analysis) 是資料科學領域中最核心的工作，它不僅能幫助我們擬定適當的決策（例如：當咖啡銷售量成長趨緩時，再加碼推出與特定甜點合購時有折扣）
 - 另外，還可以協助我們發現問題（例如：某商品銷售量因為標價錯誤形成搶購而爆增，它的營業額卻反而減少）。
- 細分成探索性資料分析，以及近年流行的機器學習 (Machine Learning)

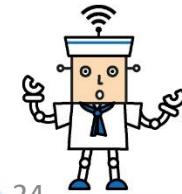
1-4-1 探索式資料分析

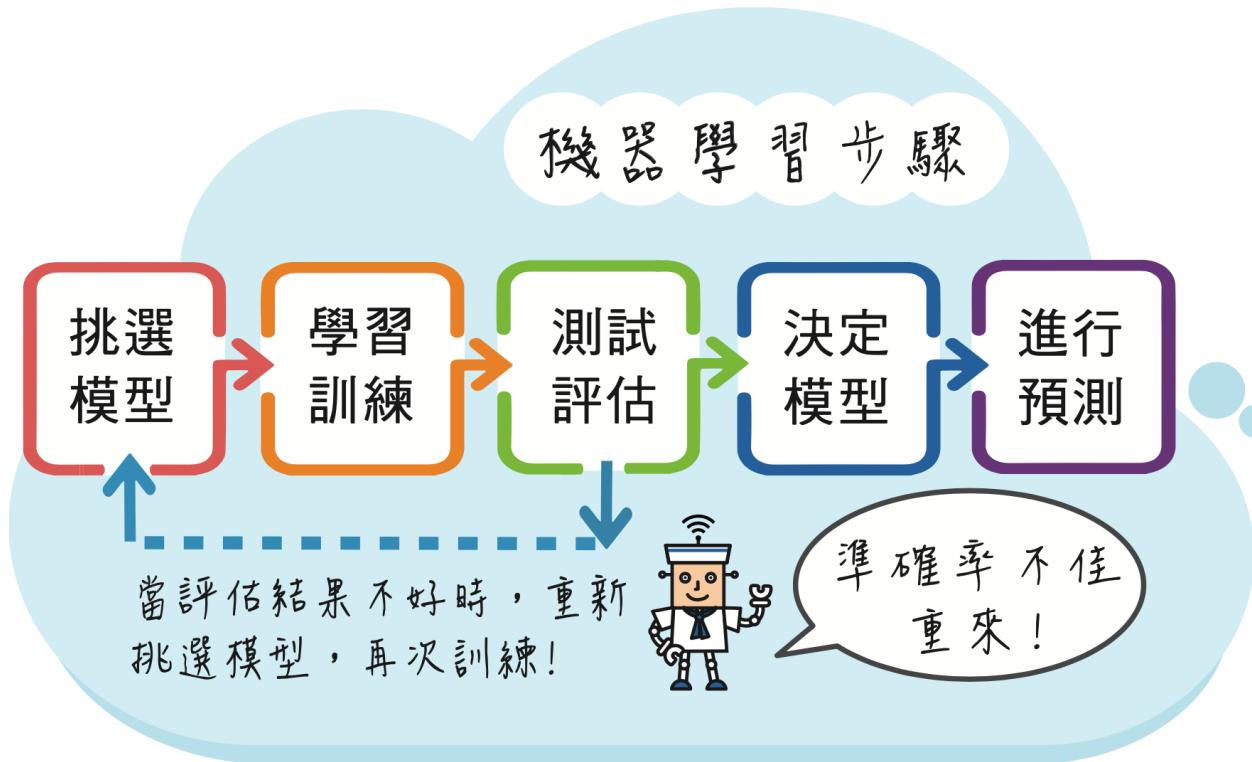
- 探索性資料分析主要的精神是運用如:**統計**、**視覺化**等工具，**反覆探索**資料的特性，獲取資料所包含的資訊、結構，從其中取得重要的「**特徵**(Feature)」
- 找出重要特徵的動作對於下一步的機器學習具有非常關鍵性的影響



1-4-2 機器學習

- 機器學習是實現人工智慧 (AI) 的方法之一
- 運用演算法自我學習，自動改進電腦演算法的效能（如準確度），讓本身能更加進步
- 藉由機器學習，希望從既有資料中找出隱藏的規則性和關聯，也就是建立出模型 (Model)。





▲ 機器學習的主要工作



資料科學 × 機器學習

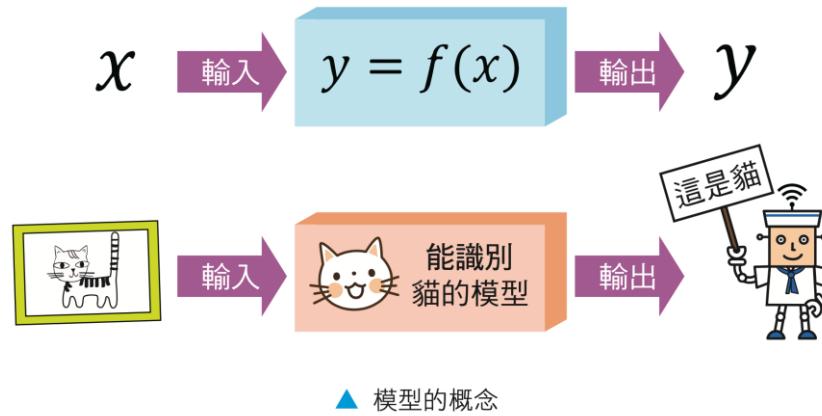
實戰探索

Practical Exploration



1-4-2 機器學習

- 模型想成是 $y = f(x)$
- 只要代入 x ，就能得到 y
- 例如：輸入一張貓 (x) 的照片到模型 $f(x)$ 中，模型會輸出（識別）這是「貓」的答案 (y)



▲ 模型的概念





機器學習的應用

趨勢預測

經由觀察現有的資料，預測未來可能的狀況。



例

隨著天氣溫度的變化，
冷熱飲的銷售量會不會
有所增減呢？



資料科學 × 機器學習

實戰探索

Practical Exploration

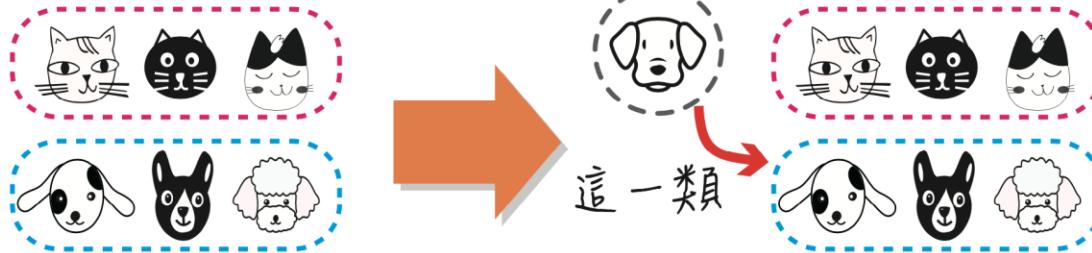




機器學習的應用

分類

將搜集到的資料先人工分類好，接著將定義好的分類以及觀察資料的「特徵」給予電腦，選定模型並訓練好後，模型就可進行物體辨識。



例

將資料分成貓跟狗兩類，當有未知的資料加入時，可以自動將它分配到貓或狗其中一個類別。





機器學習的應用

分群

針對搜集到的資料，我們不事先定義資料各屬於哪一群，讓模型根據特徵進行分群。分群的目標就是在找出不同群資料之間的關係。



例

模型將蒐集到的一些樹葉分為三群，雖然不知道是哪些樹的葉子，但是模型會自動把特徵相似的樹葉放在同一群。





機器學習的應用

關聯

找出資料之間隱藏的關聯性。



例

模型分析出買咖啡豆或
咖啡粉的同時很有可能
會同時購買牛奶，可以
應用在銷售時的推薦系
統上。





各種資料分析工具

