4-1 資料與資料檔案

4-1-1 資料的概念

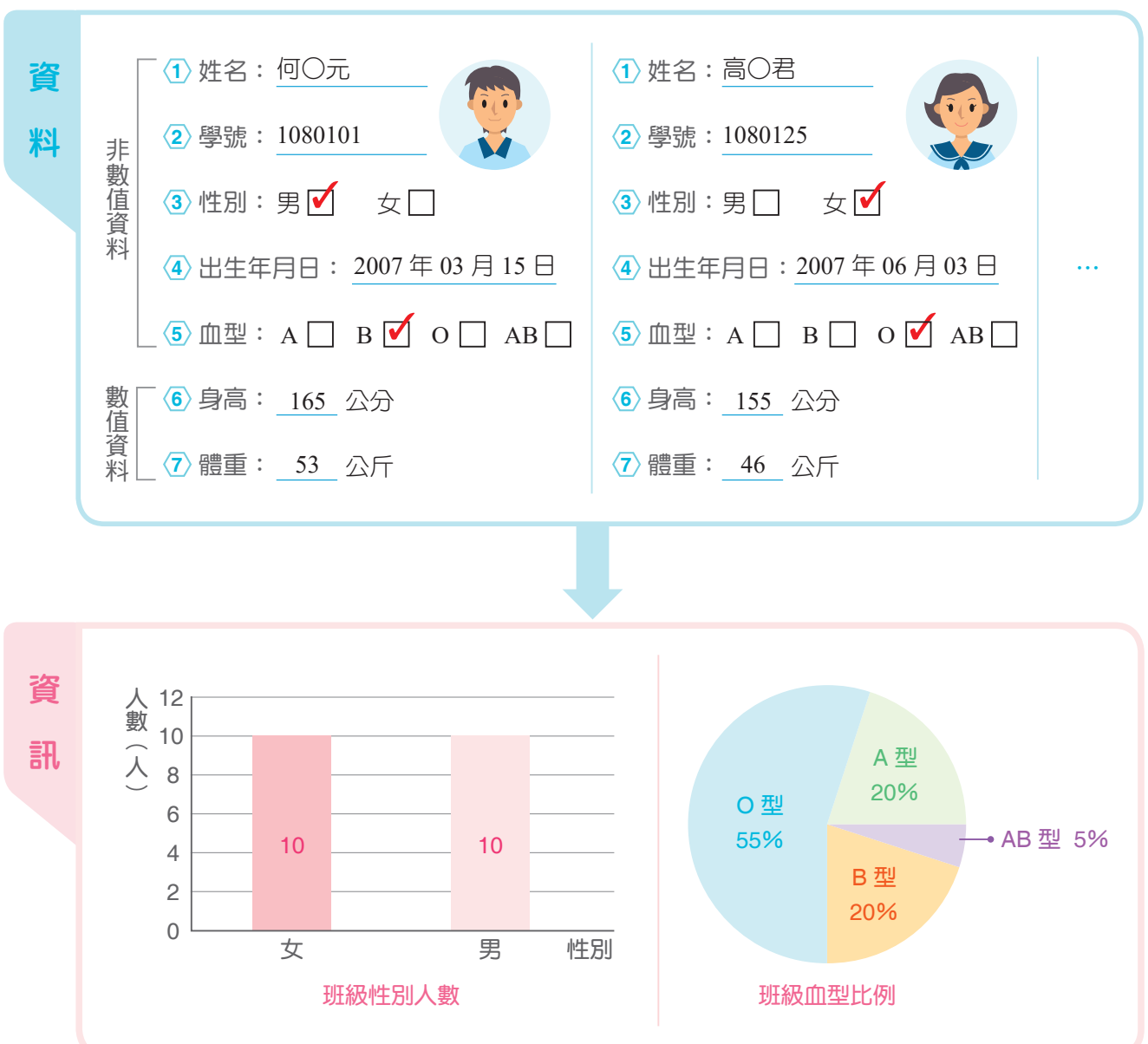
我們的生活周遭存在各式各樣的資料（data），例如：班上每天的點名單、學校登錄的段考成績，或是餐飲外送訂單、手機通訊錄、同學臉書的打卡紀錄、城市每日的空氣品質、超商每月銷售紀錄、全國每年會考成績等，這些都是和你我生活息息相關的「資料」。

生活案例

隨著網路與科技產品的發展，生活越來越便利，資料也無時無刻都在產生。而這些資料透過合適的運用，對使用者或店家有所助益。

資料是指依照某種法則，透過觀察、測量或思想表達，以數字、符號或文字等，給人、事、物、現象而留下的紀錄。資料大致可分為數值資料及非數值資料，要對前者（如體重）進行處理與分析時，可用算術四則運算計算；後者（如符號、文字）通常不用算術四則運算，可以用分類、排序或描述的方式來處理。經由分析整理過後的資料，可以讓使用者了解資料所代表的意義，這樣的資料就稱為資訊（information），如圖 4-1。

▼圖 4-1 多筆身心特徵的資料，可以分析整理成資訊。



4-1-2 資料檔案的形成

如表 4-1 所示，調查並登錄班上每個人的身心特徵，如姓名、學號、性別、出生年月日、血型、身高、體重等，每項特徵結果均由一次單一測量（single measurement）而來，稱為資料值（data value）。當我們登錄每個人的各項資料值後，就形成每個人的紀錄（record）。逐步完成 20 位同學的紀錄，若干筆紀錄即形成一個資料檔案（file）。

從右表的資料內容與格式來看，每一橫列（row）代表一筆紀錄，也就是一個人的資料。而每一直行 / 欄（column）代表一個項目（item）或變數（variable），一種身心特徵就是一個變數。我們常用的試算表軟體，資料大致就是以這種形式呈現。

認識資料的意義與性質，了解其結構或組織，將有助於後續的處理與分析。當提供分析的資料量越大，能分析出的知識也越有價值，這便是「巨量資料（big data）」（或稱大數據）的思維。

小知識

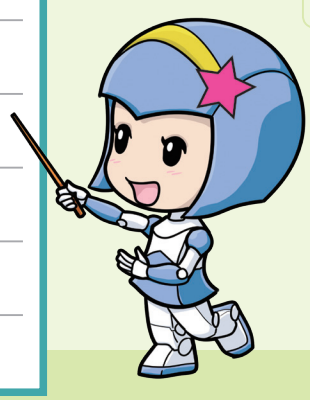
巨量資料 5V 特性

1. 資料量（Volume）：資料量龐大，通常是以 TB 以上的資料量為基本單位。
2. 多樣性（Variety）：因資料來源多樣故資料型態複雜，資料型態包含文字、數值、影音、電子郵件、網址等多種不同形式。
3. 即時性（Velocity）：由於資料的生成速度快，資料的處理分析速度也需相對提升。
4. 真實性（Veracity）：要考慮資料本身的真實性，排除刻意造假或因設備故障導致的錯誤資料。
5. 價值（Value）：巨量資料因數量龐大、其價值密度也相對較低，如何從中找出重要的部分是一個重要議題。

▼表 4-1 班上 20 位同學身心特徵的檔案。

一個資料值　直欄為行　標題列　橫列為列

姓名	學號	性別	出生年月日	血型	身高	體重
王○杰	1080291	男	2007/02/27	B	158	47
林○尹	1080529	男	2007/06/27	O	160	48
張○凡	1080346	女	2006/10/02	A	152	45
李○音	1080623	女	2007/05/18	O	156	43
何○元	1080101	男	2007/03/15	B	165	53
張○柏	1080679	男	2007/04/12	B	150	46
白○怡	1080253	女	2006/12/07	O	156	45
林○偉	1080828	男	2007/01/15	A	148	52
陳○琪	1080202	女	2007/02/25	O	153	46
周○傑	1080655	男	2006/11/28	O	172	65
詹○仕	1080712	男	2007/05/02	O	168	61
楊○玲	1080331	女	2007/08/13	O	146	42
陳○雯	1080419	女	2007/05/20	A	151	46
蘇○俊	1080925	男	2007/02/06	B	149	52
蔡○瑜	1080422	女	2007/07/19	O	147	47
高○君	1080125	女	2007/06/03	O	155	46
謝○毓	1080278	女	2007/08/12	AB	158	50
張○君	1080633	女	2006/09/23	O	154	48
王○昌	1080917	男	2007/04/03	O	157	52
邱○榮	1080384	男	2006/11/16	A	153	53



4-2 資料來源

我們周遭處處都有資料，一般來說，使用者搜集資料都有其目的，例如：本章前面的例子，學校為了解全校學生的性別、年齡、血型、身高、體重等生理特徵，就可以透過測量（如體重）及調查（如血型）得到這些資料。而個人或機構為了解某種現象或問題，測量、調查（圖 4-2）或實驗等都是常用的方法，例如：戶口普查，或近年來新冠病毒肺炎（COVID-19）肆虐全球，生技醫療界研製疫苗，都要透過實驗，搜集人體對疫苗的生理反應資料，以檢證疫苗的藥效等。

▼圖 4-2 透過測量與調查，得到相關的操作資料。

新冠病毒肺炎體溫調查表

① 記錄日期：110/03/31

② 姓名：何○元

③ 體溫：36.5 ℃

④ 聯絡電話：0922XXX123

⑤ 症狀：☑無症狀 ☐發燒 ☐咳嗽 ☐其他：＿＿＿＿

⑥ 共同居住人自境外返回：☑無 ☐有，地點：＿＿＿＿，日期：＿＿＿＿

近年來，政府也建置資料開放平臺（圖 4-3），提供公開的資料。此類資料大部分是公家的組織或機構所提供，開放給民眾使用。開放資料大部分可免費獲得，且是真實的資料，為了方便傳遞和使用，通常會將資料儲存為 CSV、XML、JSON 等檔案格式。使用者可以透過各種分析方法，讓資料活化或加值，變成有用的資訊。

▼圖 4-3 政府資料開放平臺（https://data.gov.tw/）提供很多類別的開放資料。

小知識

常見的資料交換格式

CSV	XML	JSON
最常見的資料交換檔案格式，以逗號作為不同資料欄位的分隔符號，每筆紀錄以換行符號分隔。	一種「描述資料」的語言，檔案內容除了資料本身，還描述資料的結構與組成關係。	以人類及機器都易於理解閱讀的方式來記錄資料，以利資料在不同系統間交換使用。

```
site,county,pm25,datacreationdate,itemunit
大城,彰化縣,18,"2023-02-03 12:00",μg/m3
富貴角,新北市,13,"2023-02-03 12:00",μg/m3
麥寮,雲林縣,21,"2023-02-03 12:00",μg/m3
關山,臺東縣,0,"2023-02-03 12:00",μg/m3
馬公,澎湖縣,17,"2023-02-03 12:00",μg/m3
金門,金門縣,20,"2023-02-03 12:00",μg/m3
馬祖,連江縣,19,"2023-02-03 12:00",μg/m3
```

```
<?xml version="1.0" encoding="UTF-8"?>
<aqx_p_02>
  <data>
    <SITE>大城</SITE>
    <COUNTY>彰化縣</COUNTY>
    <PM25>18</PM25>
    <DATACREATIONDATE>2023-02-03 13:00</DATAC
    <ITEMUNIT>μg/m3</ITEMUNIT>
```

```
{
  "fields": [
    { "id": "site", "type": "text", "info":
    { "id": "county", "type": "text", "info"
    { "id": "pm25", "type": "text", "info":
    { "id": "datacreationdate", "type": "tex
    { "id": "itemunit","type": "text","info"
  "resource_id": "c1f31192-babd-4105-b880-a4c2e
  "records": [
```