

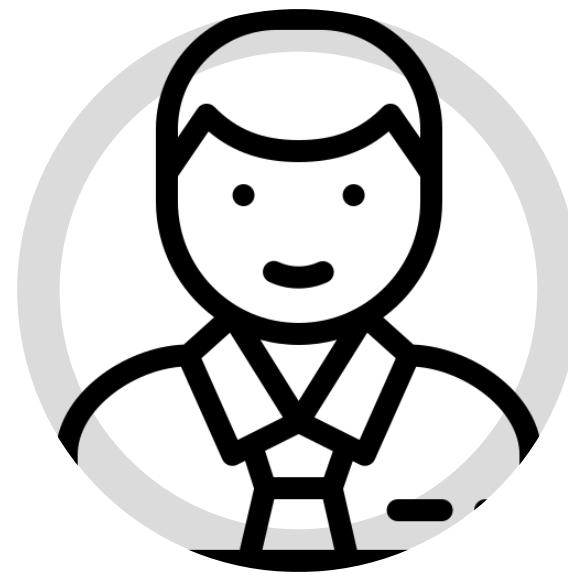


Day 01

cupay

# 資料來源與檔案存取

資料來源與檔案存取



出題教練：張維元



python

# 本日知識點目標

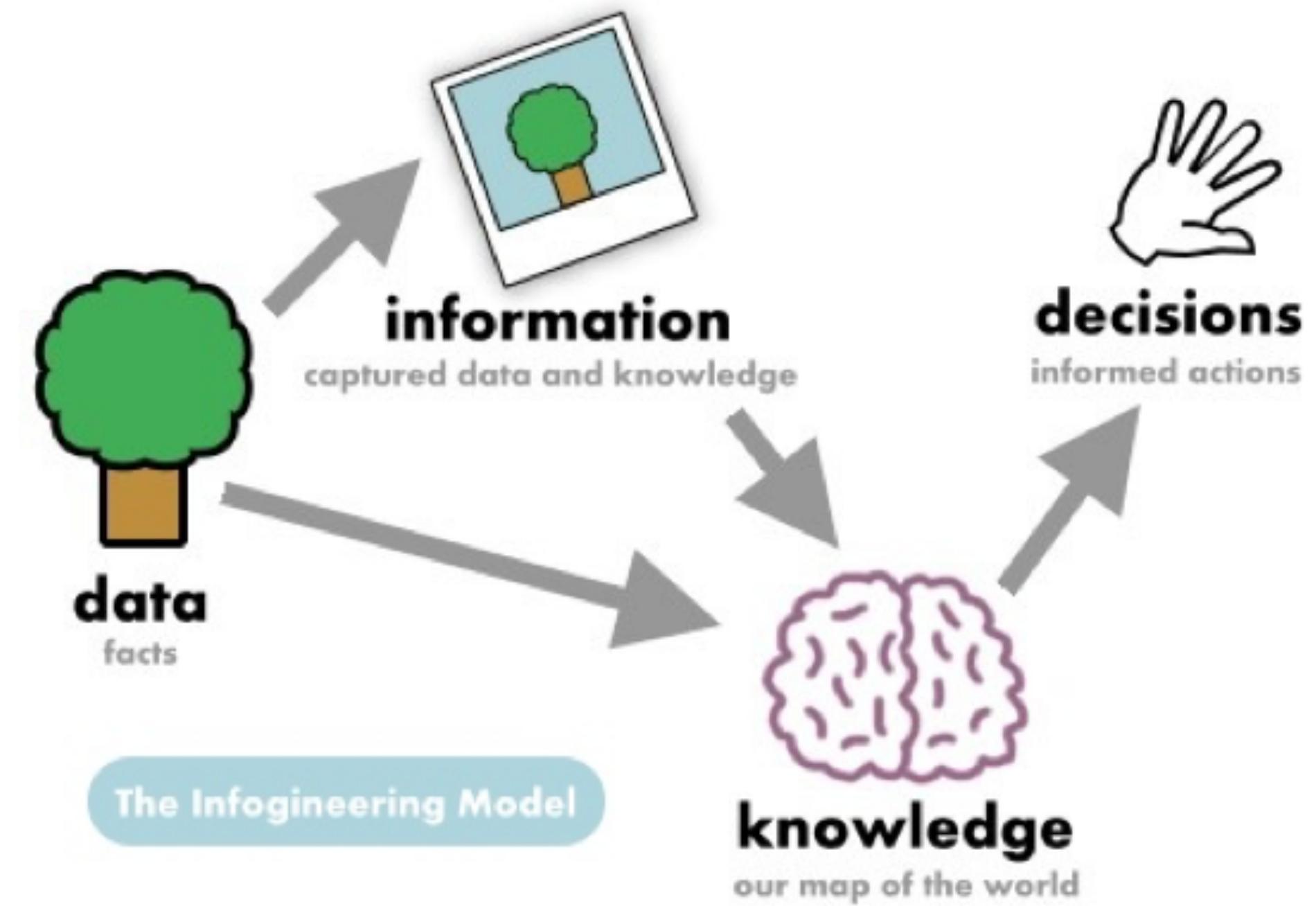
- 資料來源與取得
- 開放資料
- 資料儲存格式
- Python 存取檔案

# 知識管理

根據維基百科中，對於資料的定義：「資料是指未經過處理的原始記錄。一般而言，資料缺乏組織及分類，無法明確的表達事物代表的意義，它可能是一堆的雜誌、一大疊的報紙、數種的開會記錄或是整本病人的病歷紀錄。」

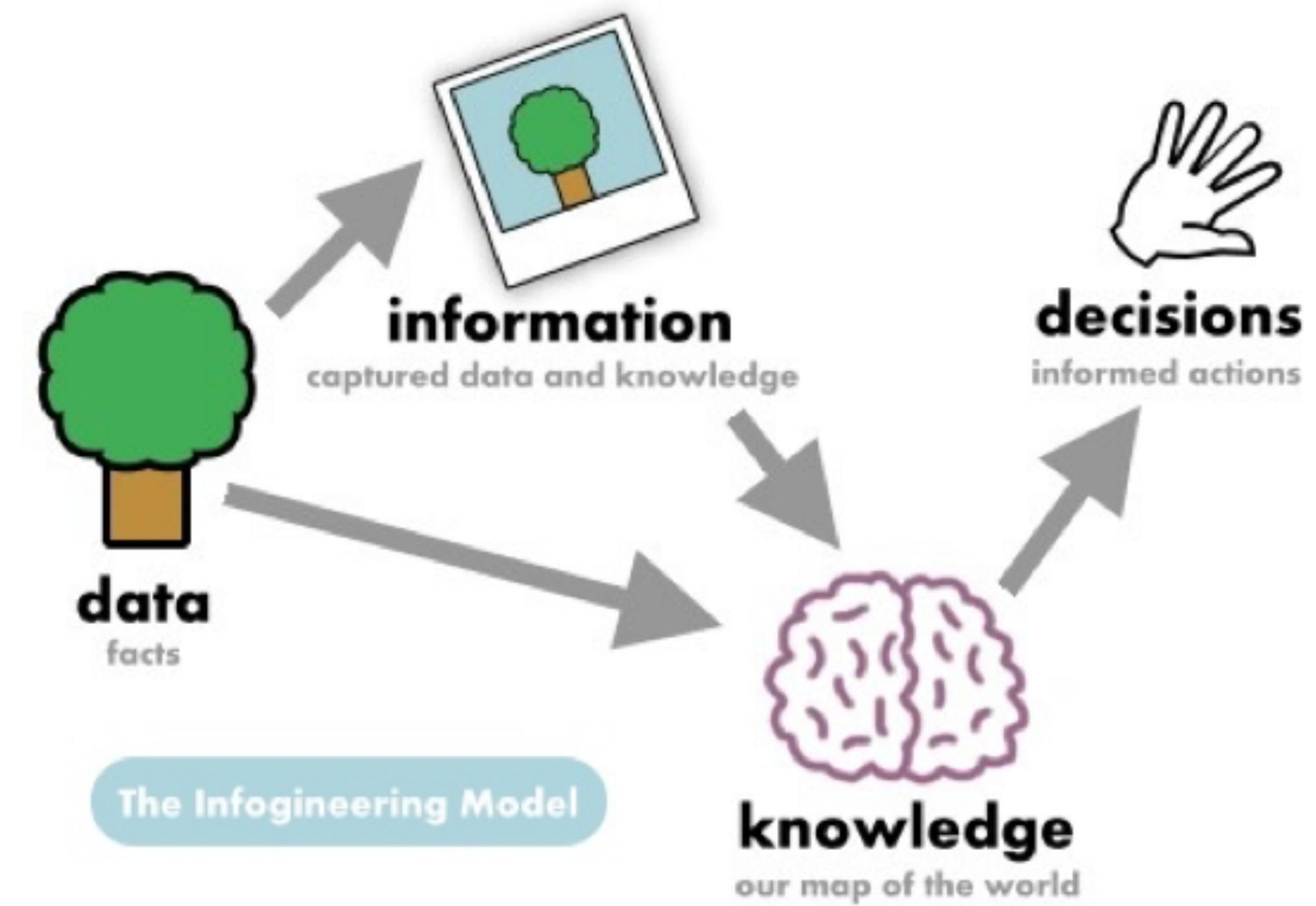
# 知識管理的四個階段

知識管理是管理學中一種討論知識產生的過程，資料累積成資訊，資訊與資料能夠統整出經驗，最終可以輔助決策。



# 數位化與人工智慧

- 數位化：從 Data 到 Information 的過程
- 人工智能：從 Data 到 Knowledge 的過程



# 資料分析的思考流程

知識管理恰巧就與資料科學中，從資料到決策的過程一樣：

「思考目標」→「找資料」→「整理資料」→「應用資料」

知識管理是人類從資料到決策的過程

資料科學是利用計算機輔助人類從資料到決策的過程

# 資料釋出的三個來源

## 1. 檔案

資料會包成檔案提供下載，格式可能包含常用的標準格式，例如「CSV」、「JSON」等等通用的格式。

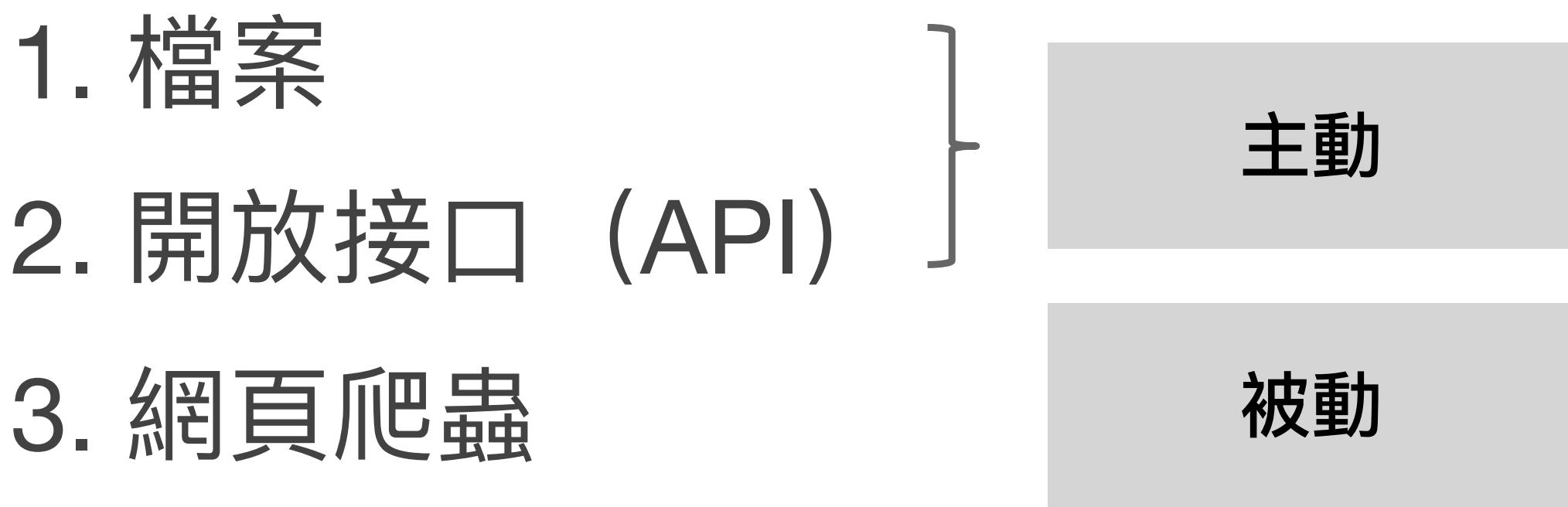
## 2. 開放接口 (API)

提供程式化的連接的接口，讓工程師/分析師可以選擇資料中要讀取的特定部分，而不需要把整批資料事先完整下載回來

## 3. 網頁爬蟲

資料沒有以檔案或 API 提供，但出現在網頁上。可以利用爬蟲爬蟲程式，將網頁的資料解析所需的部分。

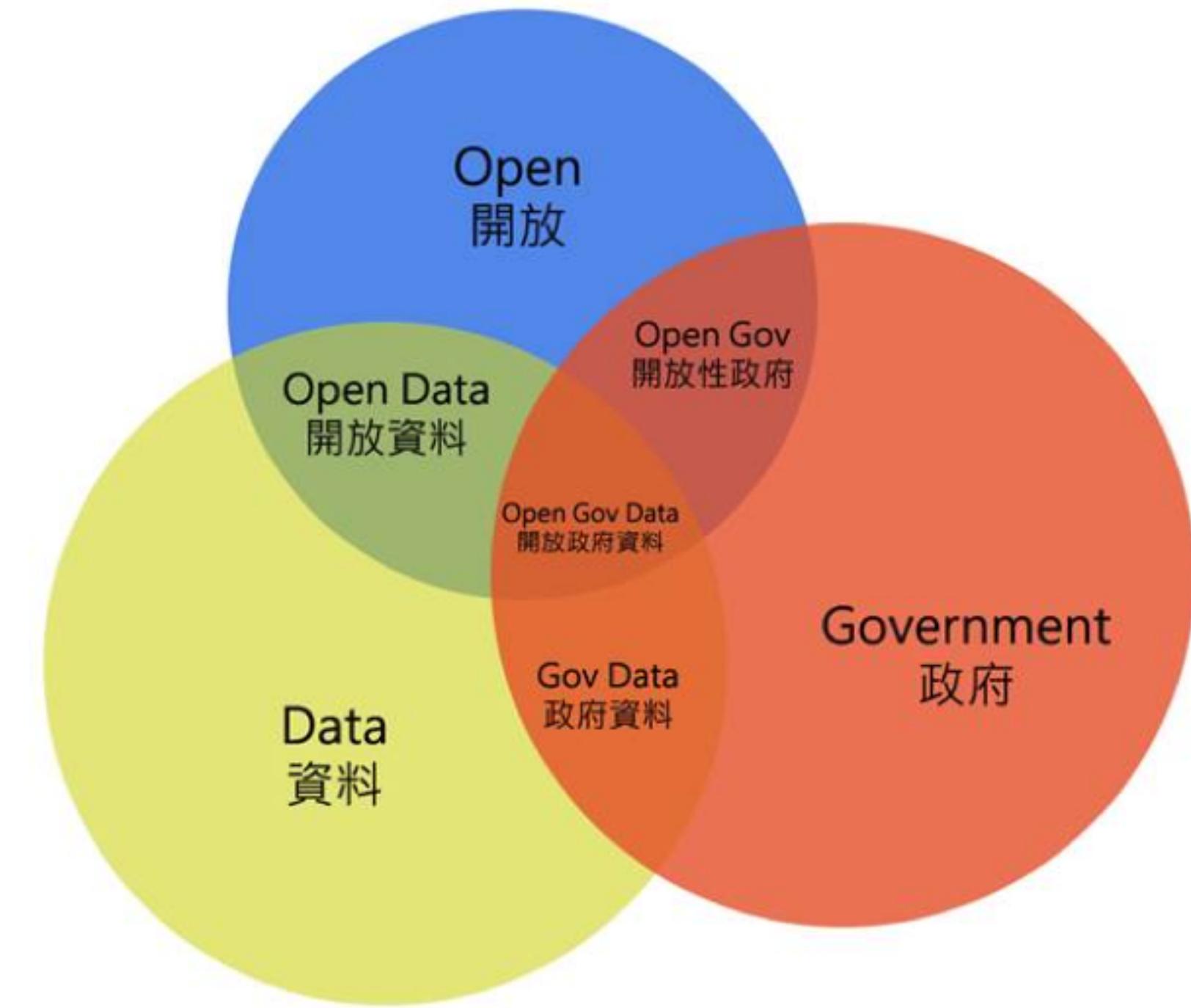
# 資料釋出的三個來源



資料的來源方式很多，檔案 & API 是由資料擁有者主動釋出，爬蟲則是資料擁有者被動公開的。所以需要取得資料的時，通常會先考慮前兩者方法，真的無法才使用網頁爬蟲。

# 什麼是開放資料？

開放資料 (英語：Open data) 指的是一種經過挑選與許可的資料。這種資料不受著作權、專利權，以及其他管理機制所限制，可以開放給社會公眾，任何人都可以自由出版使用，不論是要拿來出版或是做其他的運用都不加以限制。



# 開放資料的五顆星

可得性與可讀性

重新使用與散播



# 開放資料的五顆星

可得性與可讀性

重新使用與散播

機器可讀性是開放資料的一個特色之一。除了讓一般使用者可閱讀之外，如何讓機器可以有效的存取才可以提升資料的應用層面。

# 「公開的資料」不等於「開放資料」

公開的資料泛指的網路上你看到的所有資料，但不一定授權做加工使用。因此，爬蟲的資料在應用上，還是要小心可能會有權限使用的疑慮。

# 政府資料開放應用和政府資訊公開有何不同？

「配合資通訊科技的發展進步，除了資訊公開外，各國逐漸推動資料開放，將政府資料以資料集為基本單位，提供開放格式、便於再利用的原始資料(raw data)，讓民眾在不受到限制的情形下，進行編輯、分析、公開傳輸或為其他利用方式，開發各種產品或應用服務，滿足民眾「用」的「權益」。」

**換句話說，Open Data 將資料的使用權限回給公民，  
提供更多加值應用的可能！**

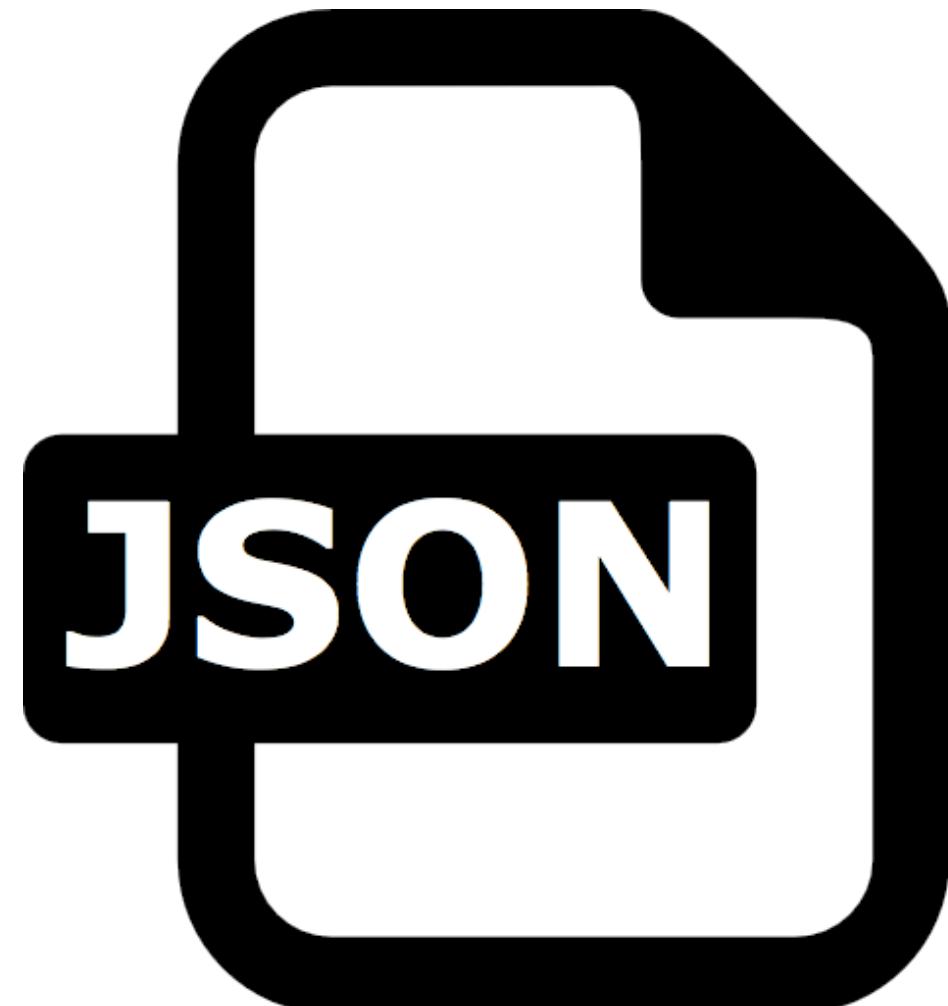
# 開放資料的趨勢

## Public Data → Open Data → Open API

資料公開的被揭露於網站，或是 Open Data 概念興起，提供彈性的原始資料做利用。其目的都是將資料視為一種公共財，還資料於民，在公私協力之下，創造更多的加值應用。Open API 則是更近一步將 Open Data 利用更方便的方式做釋出。

# 資料該怎麼存？

原始的資料在電腦中會利用固定的格式來儲存，常見的格式有以下：



# 資料該怎麼存？

CSV

```
no, name, birth
1, "王小明", "1992/07/30"
4, "林小華", "1994/08/20"
```

JSON

```
[  
  {  
    "no": 1,  
    "name": "王小明",  
    "birth": "1992/07/30"  
  },  
  {  
    "no": 4,  
    "name": "林小華",  
    "birth": "1994/08/20"  
  }]  
]
```

XML

```
<班級>  
  <學生>  
    <no>1</no>  
    <name>王小明</name>  
    <birth>1992/07/30</birth>  
  </學生>  
  <學生>  
    <no>4</no>  
    <name>林小華</name>  
    <birth>1994/08/20</birth>  
  </學生>  
</班級>
```

CSV (Comma Separated Values) 逗號分隔值，是一種常見的資料格式，使用逗號將不同欄位做為分隔。可以使用一般的文字編輯器以原始格式開啟，也可以使用 excel 或 number 等試算表軟體以表格方式開啟。

### 優點

- 結構單純
- 人機皆可讀
- 檔案小

### 缺點

- 未限定編碼(big5, utf-8 ... )
- 值內有逗點或換行可能造成欄位判斷錯誤
- 第一行不一定是欄位名稱

JSON (JSON stands for JavaScript Object Notation) JavaScript 物件格式，是一種延伸自 JavaScript 物件來儲存和交換簡單結構的輕量級純文字資料交換格式。一般格式如下，每一筆資料都會用屬性 + 數值的格式紀錄，也可以是巢狀資料。

## 優點

- 可以存放結構較複雜的資料
- 大部份瀏覽器都支援

## 缺點

- 檔案較大（不過比XML小）
- 不一定適合轉換成表格型式

XML (eXtensible Markup Language) 可延伸標記式語言，是一種標記式語言，利用標籤紀錄資料的屬性與數值，常用來處理包含各種資訊的資料等。

## 優點

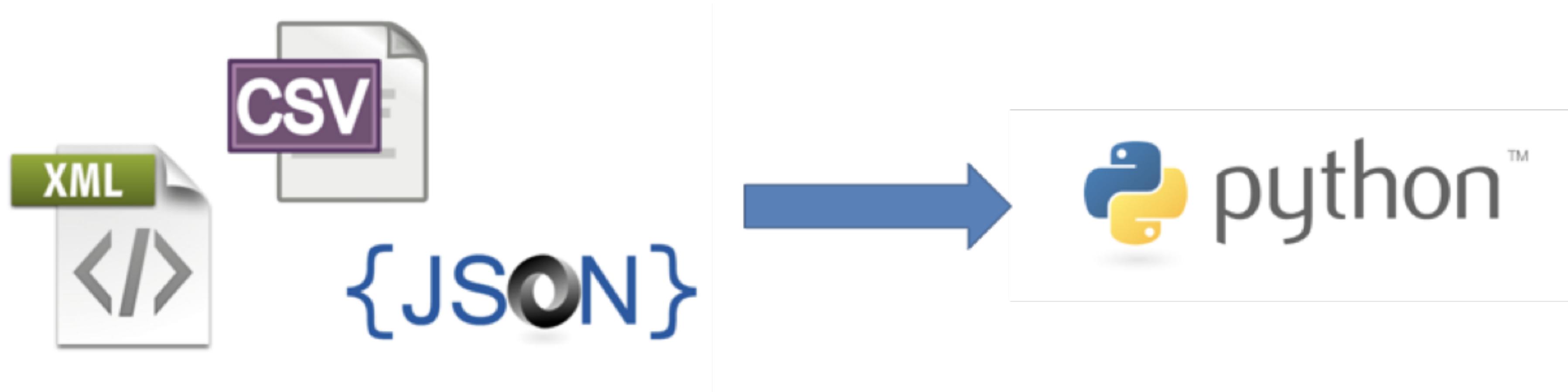
- 可以存放結構較複雜的資料
- 大多瀏覽器可幫忙排版成較易讀格式

## 缺點

- 檔案較大
- 不一定適合轉換成表格型式

# 從檔案中整理資料

面對這種檔案型的資料，我們的目標即是將檔案轉換成程式中的資料變數，且利用一個適合的資料結構做處理。



# 從檔案中整理資料

在 Python 的世界中，每一種常見的格式，都有特定的套件協助我們做存取：

- CSV => csv
- JSON => json
- XML => xmltodict, beatifulsoup

# 下載檔案

- 在 Python 可以使用第三方套件「urllib」中的「urlretrieve」方法來下載檔案。
- urllib 是一個用於網路資源（URL）操作的函式庫（library），例如：檔案下載就是這一類。

```
1 from urllib.request import urlretrieve  
2  
3 urlretrieve ("http://www.example.com/songs/mp3.mp3", "mp3.mp3")
```

檔案網址

存檔檔名

# 路徑的用法

在 Python 程式當中，有兩種表示路徑的方法

## ① 相對路徑：

- 「./data/sample.csv」 => 與程式**相同目錄**下 data 資料夾中的 sample.csv 檔案
- 「../data/sample.csv」 => 程式的**前一層目錄**下 data 資料夾中的 sample.csv 檔案

## ② 絶對路徑：

- 「C:\Users\cupoy\Desktop\sample.csv」 => windows 環境中，桌面的 sample.csv 檔案
- 「/Users/cupoy/Desktop/sample.csv」 => mac 環境中，桌面的 sample.csv 檔案

# 補充：OS Library

- 路徑的寫法有可能因為不同的作業系統而產生歧義，因此我們通常可以搭配 OS 這個內建套件來幫我們處理路徑字串。
- 例如：windows 環境用「\」，mac 環境用「/」

```
1 import os  
2  
3 os.path.join('/hello/','good/boy/','doiido')  
4 # '/hello/good/boy/doiido'
```

# Python File I/O

- File I/O 全名叫 File Input and Output，意思是是如何在程式當中存取一個外部的檔案。在大部分的程式語言當中，這是一個基本功能，不用使用額外的套件。
- 在檔案存取時候，依照權限可以分為以下幾種：

模式	意義
'r'	read，讀取檔案
'w'	write，寫入檔案（會覆蓋原本的內容）
'a'	append，寫入檔案（會對已存在的部分增加內容）

# Python File I/O

## ● 讀取檔案

	檔名	存取權限
1	fh = open("example.txt", "r")	
2	fh.read()	
3	fh.close()	

## ● 寫入檔案

	檔名	存取權限
1	fh = open("example.txt", "w")	
2	fh.write("To write or not to write\nthat is the question!\n")	
3	fh.close()	

# Python File I/O

## ● 讀取檔案

- `read()`：一次將整個文件讀成一個字串
- `readline()`：一次讀取文件中的一行資料成字串
- `readlines()`：將整個文件逐行存成一個列表

```
1  f = open("test.txt", "w")
2
3  print(f.read())
4  print(f.readline())
5
6  for line in f.readlines():
7      print(line)
8
9  f.close()
```

# Python File I/O

## ● 寫入檔案

- `write()`：將一個字串寫進檔案
- `writelines()`：將一個字串列表寫進檔案

```
1  f = open("test.txt", "w")
2
3  s = 'Hello World'
4  f.write( seq )
5
6  seq = ["Hello", " ", "World"]
7  f.writelines( seq )
8
9  f.close()
```

# 從檔案中整理資料

- 一個 File 在 Open 之後，如果沒有 Close，則會將檔案的狀態顯示為被佔用，因此可能造成資源的浪費或是其他人無法使用該檔案。
- Python 提出了一個 With 的語法，稱為資源管理器，用於 File 存取時能夠自動在使用完畢後直接關閉佔用。

```
1  with open("example.txt", "w") as fh:  
2      fh.write("To write or not to write\nthat is the question!\n")  
3  
4  with open("example.txt", "r") as fh:  
5      fh.read()  
6  
7
```

# 重要知識點複習

- 資料來源與取得
- 開放資料
- 資料儲存格式
- Python 存取檔案





- 柯文哲野生官網，用API釣出民間開發高手
  - 「柯文哲野生官網」開放API的初衷，則更偏向於促進政府開放資料(OpenData)。今年高雄氣爆發生以來，呼籲政府開放管線資料的聲音更讓大眾正視OpenData的議題，由資訊人組建的「零時政府G0v」已透過Hackfoldr等共筆工具，利用群眾智慧將公開的醫療資訊，整合成全台醫院待床數圖表，哪間醫院還缺多少病床，一目瞭然。
- 五顆星 ★ 開放資料
  - 全球資訊網 (World Wide Web) 發明者和鏈結資料的創始者，提姆·柏納-李 (Tim Berners-Lee)建議了一個開放資料五顆星的分類架構 在此，網站提供在每一顆星中每一步驟的範例且解釋這些步驟的成本和效益。



- 開放政府資料的法制趨勢與發展探討
  - 作者依據國內的開放政府與開放資料的授權模式進行探討與分析，參考「歐盟公部門資訊再利用指令(the directive on the re-use of public sector information, Directive 2013/37/EU)」與美國參眾兩院的「開放政府資料法(Open, Public, Electronic, and Necessary Government Data Act)」之要點摘錄，進行國內開放資料法未來發展要點的探討與析論，期能為國內相關議題的後續討論提供基礎，並作為未來就此一議題進一步深入研究之起點。
- Python 的 with 語法使用教學：Context Manager 資源管理器
  - 介紹 with 的實作原理，為什麼 File 的存取可以這樣用。With 語法的全名是 Context Manager 資源管理器，該文章有更完整的教學與說明。

# 參考資料



以下整理一些國內政府單位與第三方機構所釋出的公開資料集：

- 台灣政府資料開放平台：<https://data.gov.tw/>
- 台北市開放資料平台：<https://data.taipei/index>
- 新北市開放資料平台：<http://data.ntpc.gov.tw/>
- 台中市開放資料平台：<http://opendata.taichung.gov.tw/>
- 桃園市開放資料平台：<http://data.tycg.gov.tw/>
- 台南市開放資料平台：<http://data.tainan.gov.tw/>
- 高雄市開放資料平台：<https://data.kcg.gov.tw/>

以下整理一些國內外政府單位與第三方機構所釋出的公開資料集：



- 香港政府數據中心：<https://data.gov.hk/en/>
- 英國國家數據中心：<https://data.gov.uk/>
- 日本統計局：<http://www.stat.go.jp/>
- 中國國家數據中心：<http://data.stats.gov.cn/>
- 美國政府開放資料：<https://www.data.gov/>
- 歐盟資料平台：<https://www.europeandataportal.eu/>

以下整理一些國內外政府單位與第三方機構所釋出的公開資料集：



- 世界經濟貿易合作組織資料庫：<https://data.oecd.org/>
- 世界銀行開放資料：<https://data.worldbank.org.cn/>
- 世界衛生組織：<http://apps.who.int/gho/data/node.home>
- Google BigQuery 公開資料集：<https://cloud.google.com/bigquery/public-data/>
- Google 開放資料搜索：<http://www.google.com/publicdata>
- 亞馬遜開放資料：<https://aws.amazon.com/cn/datasets/>

# 解題時間

# LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

