



Web Scraping with Python

GDSC NTNU 2023/10/30



Hugo Wang
@whyhugo

<https://github.com/whyhugo/GDSC-NTNU-data-course>

```
def filterStudies(studies, filterByOrg):
    filteredStudies = []
    for study in studies:
        if study.lead_organization == filterByOrg:
            filteredStudies.append(study)
    return filteredStudies

def filterStudiesByStatus(studies, filterByStatus):
    filteredStudies = []
    for study in studies:
        if study.status == filterByStatus:
            filteredStudies.append(study)
    return filteredStudies

def filterStudiesByOrgAndStatus(studies, filterByOrg, filterByStatus):
    filteredStudies = filterStudies(studies, filterByOrg)
    filteredStudies = filterStudiesByStatus(filteredStudies, filterByStatus)
    return filteredStudies
```

whoami

Hugo Wang

- 師大資工 大一
- GDSC NTNU CTM – Speaker
- SITCON 2024 議程組副組長
- 教育大數據微學程 通識 TA



Outline

You will learn...

- ． 資料處理開發環境
- ． 認識 data
- ． 網路爬蟲簡介
- ． requests 套件實作
- ． Selenium 與 BeautifulSoup 實作



資料處理開發環境

#intro

- 區塊化 (cell) 編程
 - 容易 debug
 - 分段處理、觀察
- Preview
- Markdown 筆記
- 保存結果共享

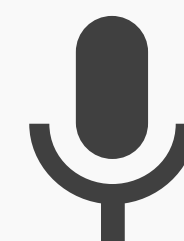
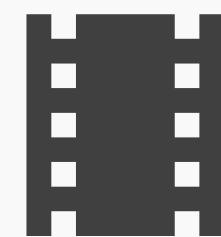


什麼是 Data ?

#intro

結構化資料 vs. 非結構化資料

”



圖像資料前處理

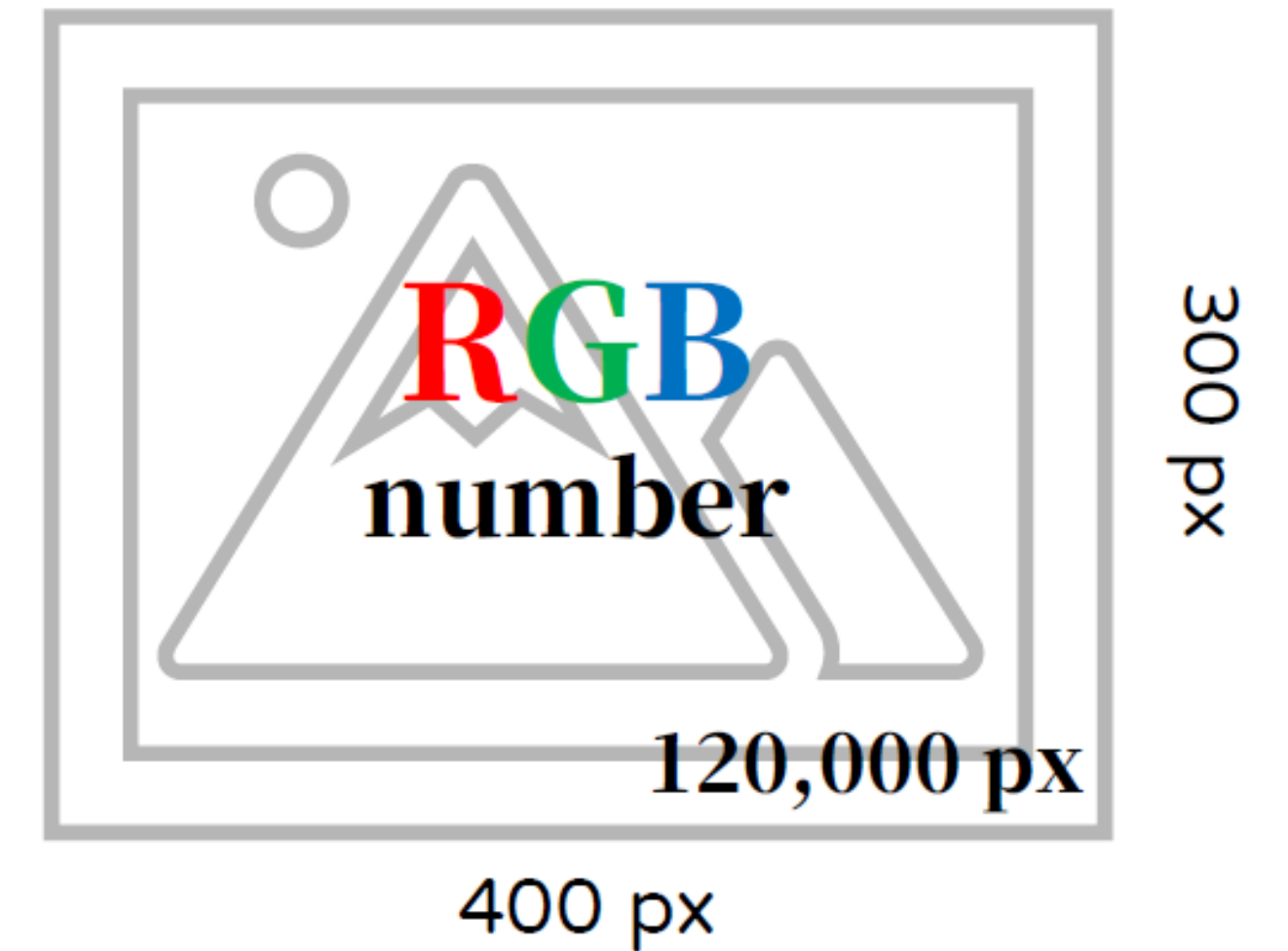
#data

- 圖像資料儲存樣貌
- 扁平化 (Flatten) 影像資料
- 正規化影像資料
- Demo: 以 MNIST 為例

👁️ 人看到的



💻 電腦看到的



影像特徵

#data



<彩色>

- 1 pixel >> (R,G,B)



<黑白>

- 亮度

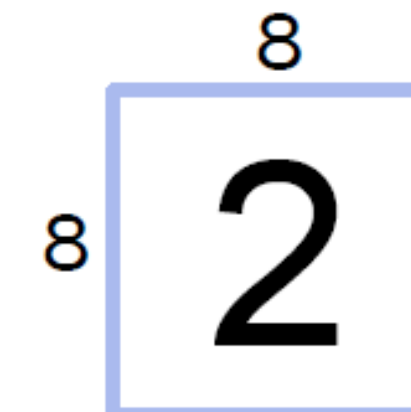
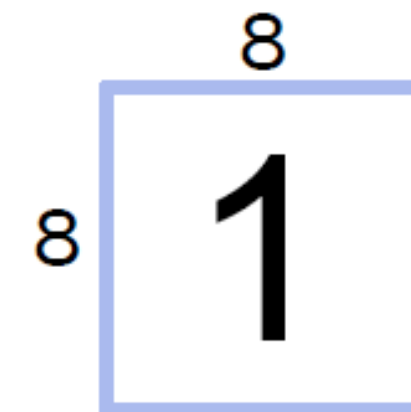
👉 數字辨識：0、1、2……

8x8 image



數位化/平面化

1x64 image



210 | 87 | | 220 | 225

數據正規化

#data

(明顯非數字) 性別：male/female

(物理意義不同) 風力：0~14級

(物理意義不同) 氣溫：-50~50°C

數據化



固定的範圍 (有限)

0, 1

$0, \frac{1}{14}, \frac{2}{14}, \frac{3}{14}, \dots, \frac{13}{14}, 1$

$0, \frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \dots, \frac{99}{100}, 1$

- 數據具有多種可能性或不具數字意義：one-hot encoding

例如：

circle = [1, 0, 0, 0], square = [0, 1, 0, 0]

triangle = [0, 0, 1, 0], rectangle = [0, 0, 0, 1]

網頁架構

#web

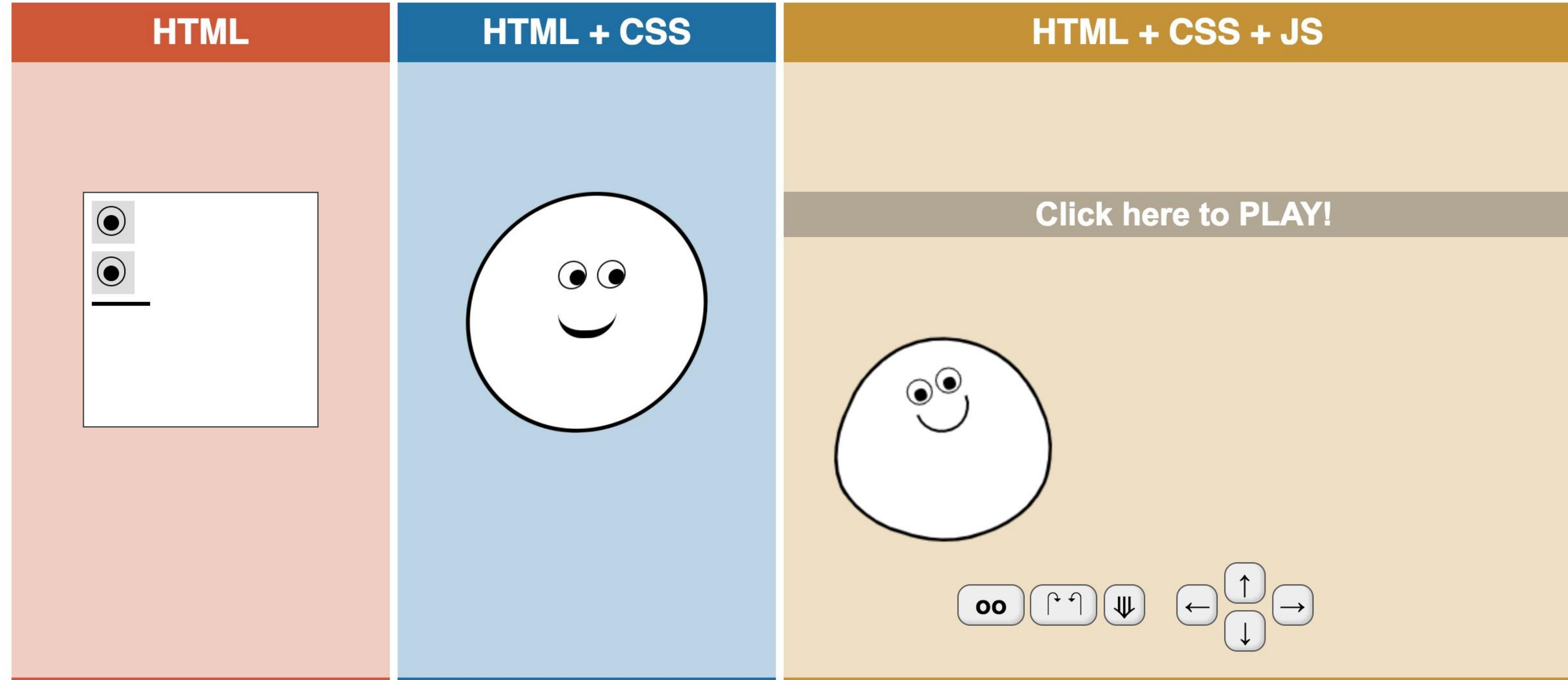
- 「HTML」：一種文件編排的標示語言（ Markup Language ）>> 網頁呈現



CSS



HTML



HTML



HTML+CSS



```
<!DOCTYPE html>
<html>
  <head>
    <title> website title </title>
  </head>
  <body>
    <h1> Welcome!!! </h1>
    <div>
      <p> busy day~~~ </p>
      

      <h2> title h2 </h2>
      <p> hello, GDSC NTNU. </p>
      <a href="https://www.example.com"> click me </a>
    </div>
  </body>
</html>
```

何謂網路爬蟲



#intro

- 自動化蒐集資料
- 常見工具：
 - Requests
 - BeautifulSoup
 - Selenium



如何爬蟲

#How

- 確立目標資料
 - 觀察網址、網頁結構
 - 選擇工具
 - 打扣打扣
 - 建立欲保存的格式 (csv, json...)
- 
- 
- 對伺服器發出 HTTP 請求
 - 取得回應
 - 透過<tag>等方式取出資料

Requests

#crawler

- GET 請求
 - 回傳屬性
 - text：文字檔案
 - 二進位資料
 - HTTP 狀態碼
- POST 請求：填寫資訊

Beautiful Soup : 網頁剖析

#crawler

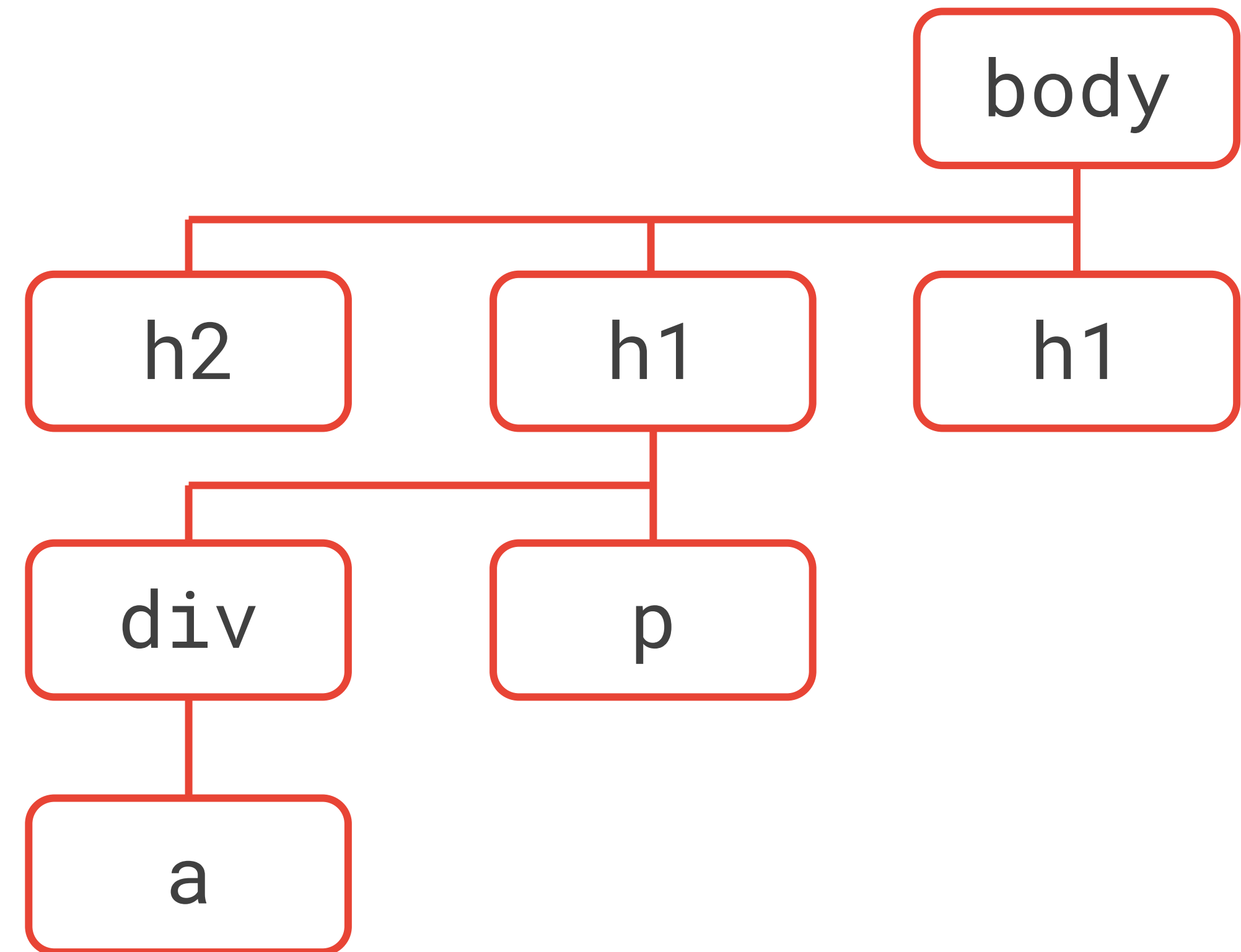
- 從 HTML 中提取內容
- 解析器 : html.parser vs. lxml
- 常用屬性 : tag name 、 text
- 常用方法 :
 - find() : 尋找**第一個**符合條件的 tag , string 回傳
 - find_all() : 尋找**所有**符合條件的 tag , list 回傳
 - select() : CSS 選擇器 (id or class) , list 回傳

Selenium

#crawler

- 模擬自動化操作
 - 按下按鈕
 - 輸入文字
 - 滾動上下移動頁面
 -任何人類操作網站的動作
- 需要 Chrome WebDriver

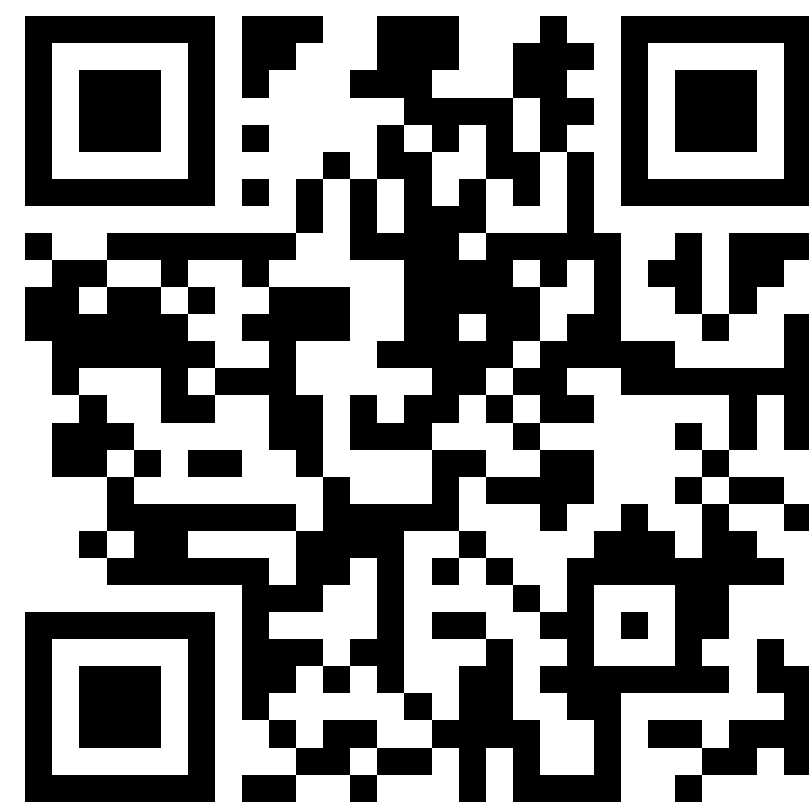
- 什麼是 xpath : DOM 節點階層關係路徑



THANK YOU

:D

匿名回饋表單



今日課程資料

