Python Web Scraping (Crawling)

Dr. Steve Lai

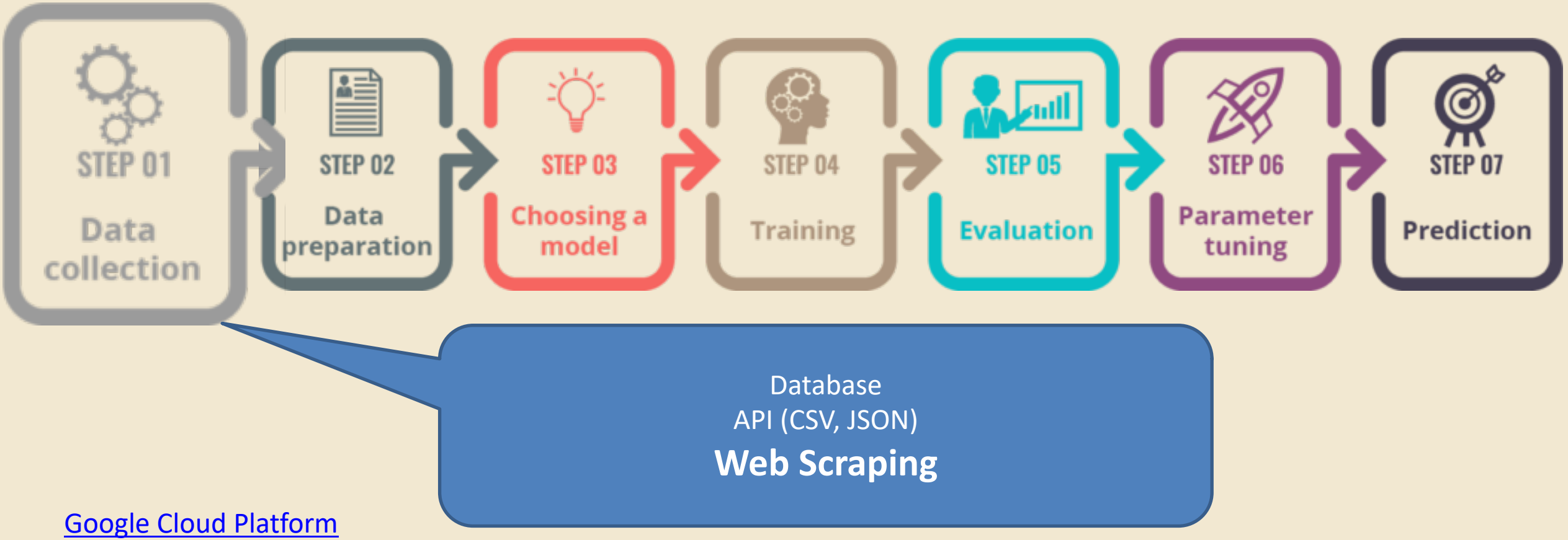
2021/09

Web Scraping (Crawling)

- In theory, web scraping is the practice of gathering data through any means other than a program interacting with an API
  - or through a human using a web browser
  - writing an automated program that queries a web server, requests data and then parses that data to extract needed information
  - programming techniques and technologies,
    - Data analysis
    - Natural language parsing (NLP
    - Information security

- If you can view data in your browser, you can access it via a Python script. If you can access it in a script, you can store it in a database. And if you can store it in a database, you can do virtually anything with that data.

From Data to AI



Google Cloud Platform
https://www.youtube.com/watch?v=nKW8Ndu7Mjw
The 7 steps of machine learning (AI Adventures)

https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d

Outline for The Four Time Slots

1. Basic python
   - Python fundamental
   - Data sturctures
   - Data Analysis
2. Getting data from web API
   - JSON
   - CSV
3. Python crawling modules
   - beautifulshop
   - selenium
4. Case study
   - Ptt crawling
   - TWSE crawling

Steve Lai（賴昭榮）

- Founder & CEO of Mathison Intelligence
- 台灣大學資訊工程學系博士

- 中央研究院資訊科技與創新中心助理
- iOS, Android, Python 講師

- Are you familiar with python?
  - (not familiar) 0 ~ 5 (very familiar)

Quizs

- What is the maximum value of Int in python?
  1. 2^64 -1
  2. 2^32 -1
  3. Infinity

Quizs

- The BIF id() return the identity of an object which the identity is the address of the object in memory. Read the following program

```
1 a = 1
2 id1 = id(a)
3 a = 2
4 id2 = id(a)
5 print(id1 == id2)
```

- What is the output?
  1. An address number
  2. True
  3. true
  4. False
  5. false
  6. Cause an error

Quizs

- What is the output of the following program?

```
1 print = 3
2 print += 3
3 print(2)
```

1. 2
2. 3
3. 6
4. Cause an error

Basic Data Type in Python

- int
- float
- bool
- str

Control Flow and Programming Structure

- if-elif-else
- Loop
  - while loop
  - foreach loop
- Self define function
- Self define class

1. How is the difficulty so far?
   - (easy) 0 ~ 5 (difficult)

2. How do you like the online way of the class?
   - (bad) 0 ~ 5 (nice)

Basic Data Structure

- tuple
- list
- dict
- set

Quizs

- What is benefit of using function?
  - Easy to read
  - Easy to debug
  - Easy to use

Quizs

- Which of the following are synonyms?
  1. instance
  2. attributes
  3. method
  4. object
  5. function
  6. variable
  7. class

Data Analysis Modules

- Numpy
- Pandas
- Matplotlib

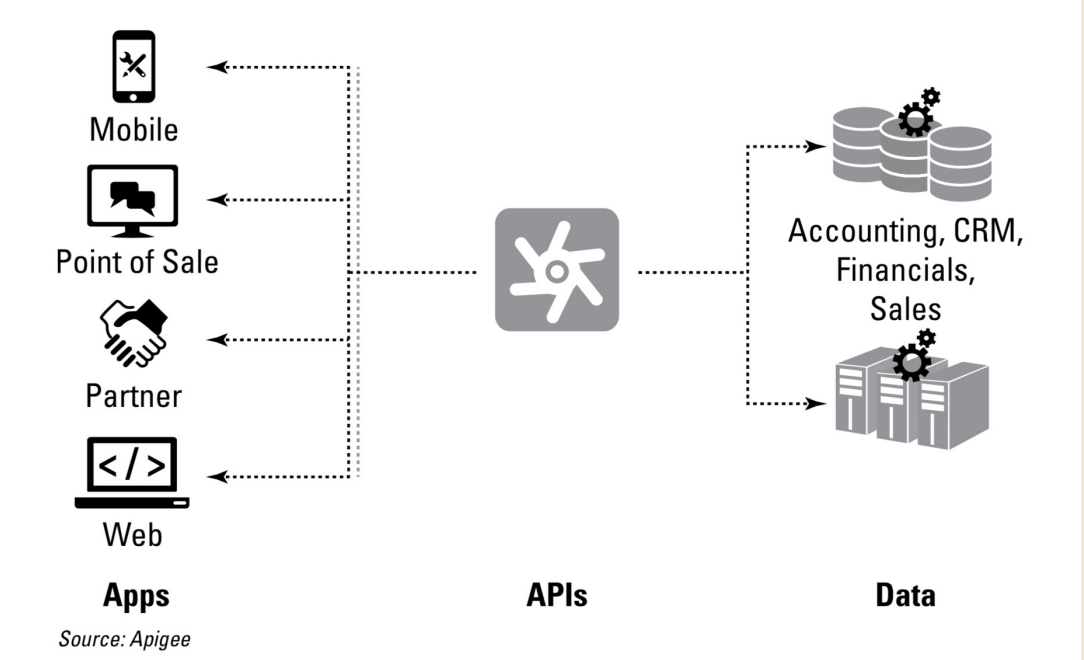
Python
Getting data from web API

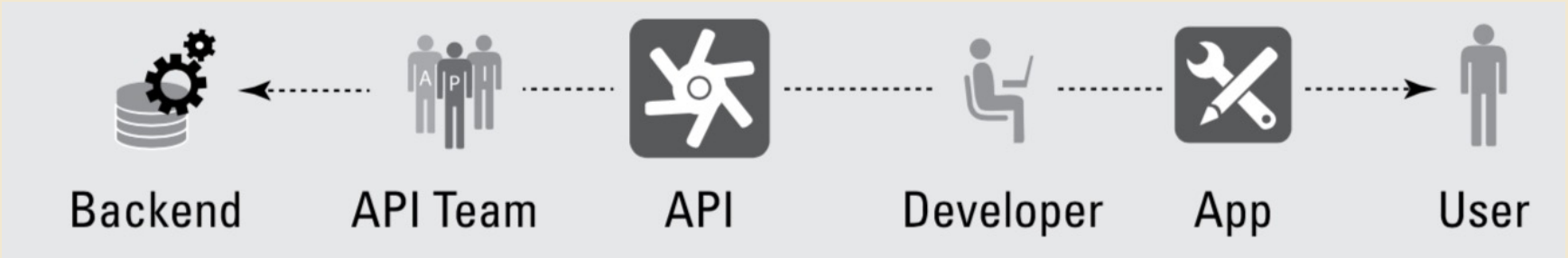
Dr. Steve Lai

2021/09

Connecting to The Appliccation

- API
  - Application Programming Interface
  - Libraries, CSV, JSON, XML, …





JSON

- What is the following similar with JSON Object in Java?
  1. Dictionary
  2. Set
  3. Tuple
  4. List

- What is the following similar with JSON Array in Java?
  1. Array
  2. Set
  3. Map
  4. StringBuilder

What is JSON?

- JSON
  - JavaScript Object Notation
- JSON Object
  - Key-Value pairs
- JSON Array
  - Sequential data with order
- import json
  - load, loads
  - dump, dumps

CSV

- import csv

APIs

- https://api.github.com/search/repositories?q=language:python&sort=stars
- https://www.travel.taipei/open-api/zh-tw/Attractions/All?categoryIds=12&page=1

Python Crawling Modules

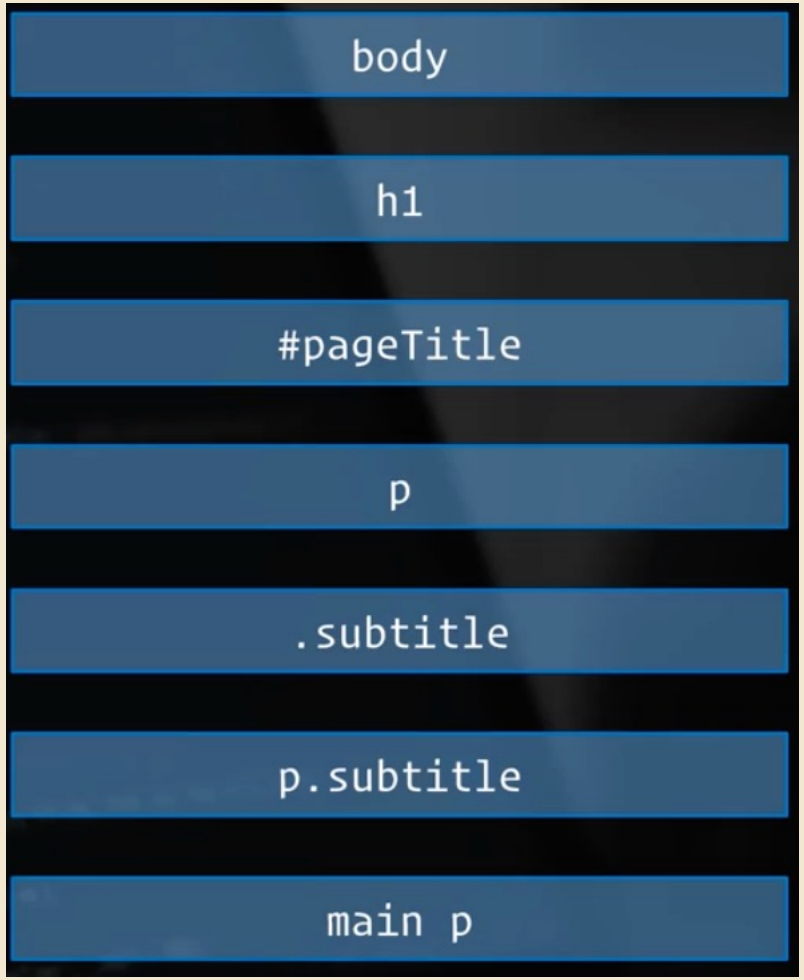
Dr. Steve Lai

2021/09

The Structure of HTML Code

•

```
<body>
    <h1 class="header" id="pageTitle">Hello World!</h1>
    <p class="subtitle">A subtitle.</p>
    <main>
        <p>This is the main text.</p>
    </main>
</body>
```

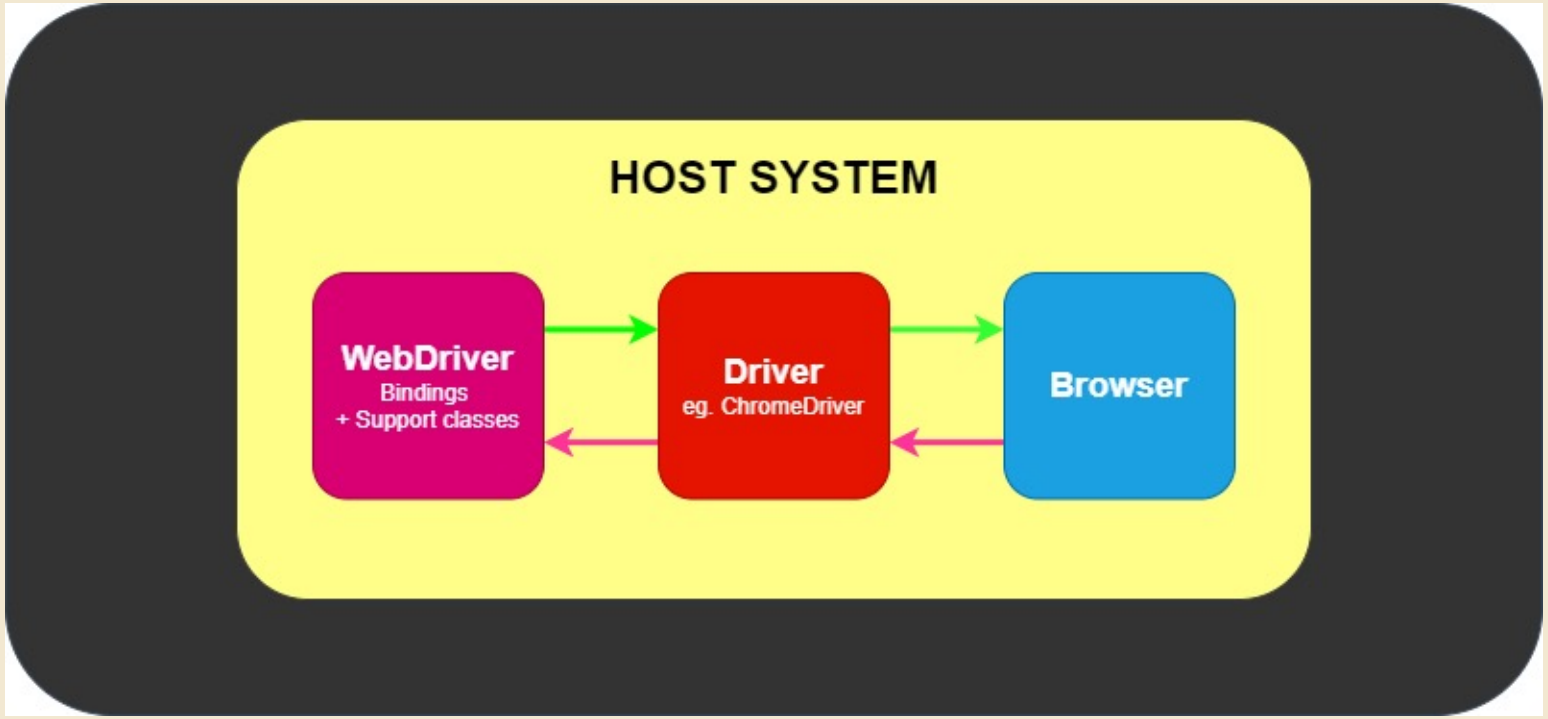


Beautifulsoup4

- Find tag
  - One or all
  - With attributes
- CSS selector
- Parser class

Selenium

- Selenium Python bindings provide a convenient API to access Selenium WebDrivers like Firefox, Ie, Chrome, Remote etc
- Through the auto action of browser to parse the web page



Locating Elements

- Find element in page
  - find_element_by_id
  - find_element_by_name
  - find_element_by_xpath
  - find_element_by_link_text
  - find_element_by_partial_link_text
  - find_element_by_tag_name
  - find_element_by_class_name
  - find_element_by_css_selector
- To find multiple elements (these methods will return a list):
  - find_elements_by_name
  - find_elements_by_xpath
  - find_elements_by_link_text
  - find_elements_by_partial_link_text
  - find_elements_by_tag_name
  - find_elements_by_class_name
  - find_elements_by_css_selector

Case Study

- PTT crawling
- TWSE

Case Study

- NLP
- Word Cloud

Code Reference

- https://github.com/lzrong0203/fin_ios_python