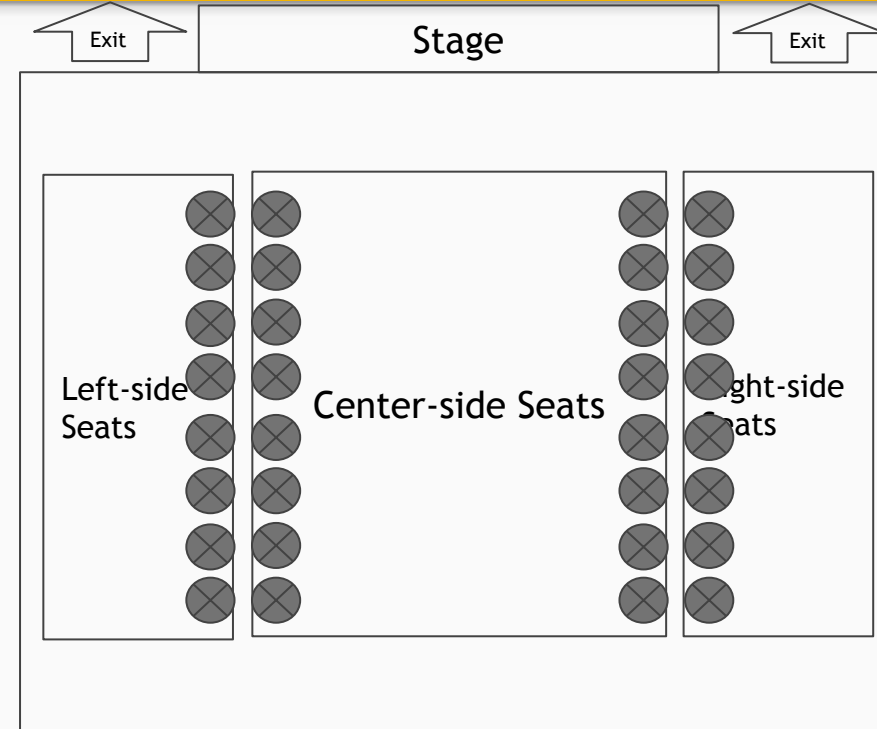


Yasumoto International Academic Park - YIA LT6

Limited Mobile Signal. Please use [on-campus wifi](#).
32 Sockets. But please bring your own charger.



⊗ = sockets

Web Scraping, API

CUHK MSc Data Science & Biz Stat. Program
STAT5106 - Programming Techniques for Data Science
Week 5 @ 10 Oct 2024

Some knowledge on internet

- Reference: [Dr. Chuck](#) free courses
 - [Internet History, Tech, Security](#)
 - [Python for Everybody Ch. 13, 14](#)

- [TCP/IP](#)

- Protocol

A set of rules that all parties follow so we can predict each other's behavior

- [HTTP](#) - HyperText Transfer Protocol
- [HTTPS](#) - (+) Secure
- [Request](#) - GET / PUT / DELETE ...

- Protocol + Host + Files

`https://www.sta.cuhk.edu.hk/default.aspx`

protocol

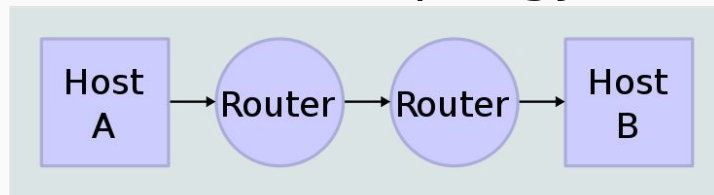
host

file

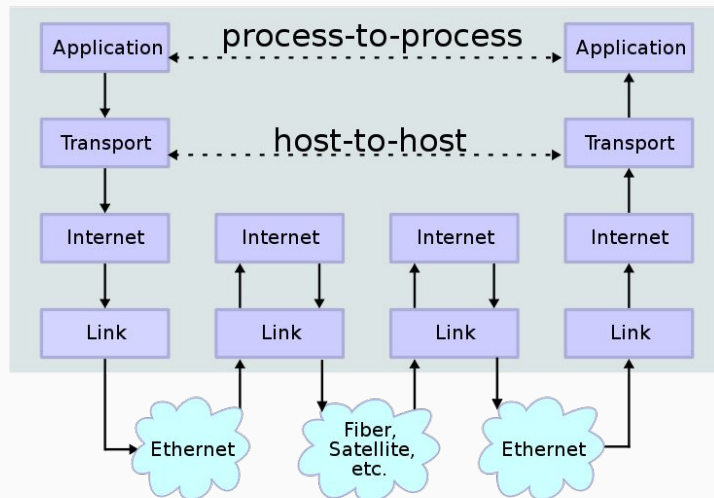
`https://example.com:80/blog?search=test&sort_by=created_at#header`

↓	↓	↓	↓	↓	↓
Protocol	Domain	Port	Path	Query Parameters	Fragment/Anchor

Network Topology



Data Flow



tracert / traceroute

P.S. 你果度我方睇錯係黃大仙黎，要搵你住邊計下角度就知，IP 都唔洗check

<= what does this mean?

```
C:\Users\kyala>tracert www.wikipedia.org
```

在上限 30 個躍點上

追蹤 dyna.wikimedia.org [103.102.166.224] 的路由:

1	4 ms	2 ms	2 ms	192.168.0.1
2	6 ms	6 ms	6 ms	061093127001.ctinets.com [61.93.127.1]
3	7 ms	7 ms	6 ms	10.239.98.1
4	10 ms	9 ms	9 ms	014199253001.ctinets.com [14.199.253.1]
5	7 ms	7 ms	6 ms	061244224065.ctinets.com [61.244.224.65]
6	41 ms	44 ms	42 ms	14907.sgw.equinix.com [27.111.228.186]
7	42 ms	42 ms	41 ms	text-lb.eqs.in.wikimedia.org [103.102.166.224]



My IP Address Is:

IPv4: **61.93.127.1**

IPv6: Not detected

My IP Information:

ISP: Hong Kong Broadband Network

City: Central

Region: Central and Western District

Country: Hong Kong

Surf Privately

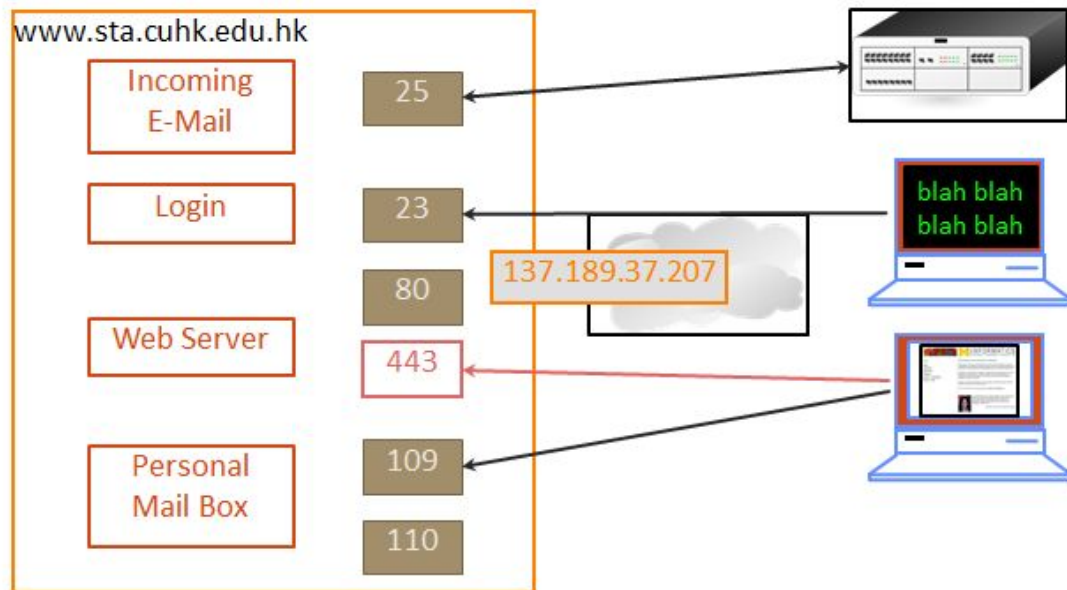
[Click Here](#)



Leaflet | © OpenStreetMap Terms

Location not accurate? [Update my IP location](#)

[Show Complete IP Details](#)



Telnet (23) – Login

SSH (22) – Secure Login

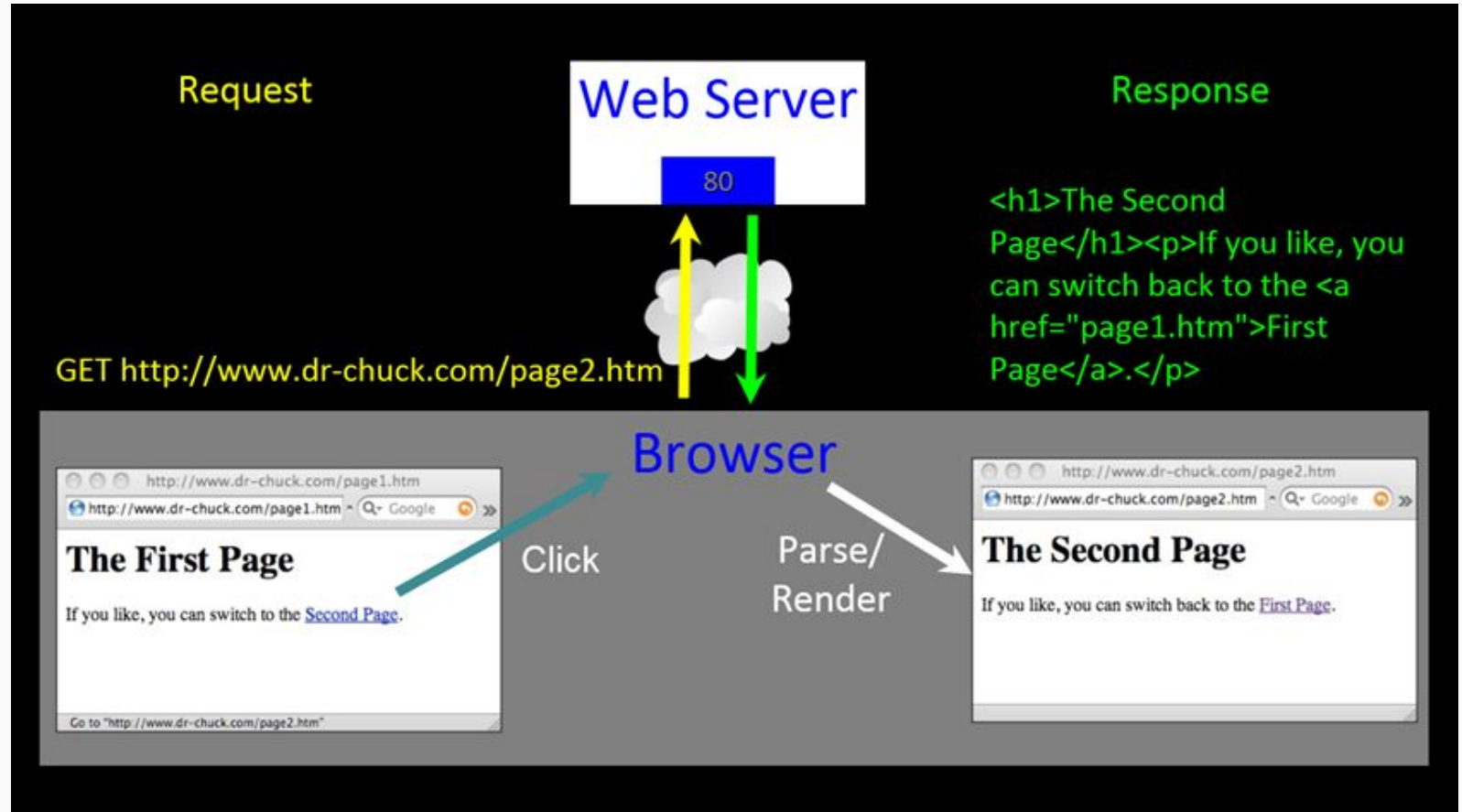
HTTP (80)

HTTPS (443)

SMTP (25) (Mail)

FTP (21) – File Transfer

How to get something on the web



Ctrl+U: View Source Code / F12 - Network Information

```
1 <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
2 <html>
3 <head>
4 <title>Index of /lectures3</title>
5 </head>
6 <body>
7 <h1>Index of /lectures3</h1>
8 <table>
9 <tr><th align="top"></th><th align="left">Name</th><th align="right">Size</th><th align="right">Description</th></tr>
10 <tr><th colspan="5"><hr></th></tr>
11 <tr><td align="top"></td><td align="right"><a href="/">Parent Directory</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
12 <tr><td align="top"></td><td align="right"><a href="/00-Master.ppt">00-Master.ppt</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
13 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-01-Intro.pptx">Pythonlearn-01-Intro.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
14 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-01/>Pythonlearn-01/</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
15 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-02-Expressions.pptx">Pythonlearn-02-Expressions.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
16 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-03-Conditional.pptx">Pythonlearn-03-Conditional.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
17 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-04-Functions.pptx">Pythonlearn-04-Functions.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
18 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-05-Iterations.pptx">Pythonlearn-05-Iterations.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
19 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-06-Strings.pptx">Pythonlearn-06-Strings.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
20 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-07-Files.pptx">Pythonlearn-07-Files.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
21 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-08-Lists.pptx">Pythonlearn-08-Lists.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
22 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-09-Dictionaries.pptx">Pythonlearn-09-Dictionaries.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
23 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-10-Tuples.pptx">Pythonlearn-10-Tuples.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
24 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-11-Regex-Handout.txt">Pythonlearn-11-Regex-Handout.txt</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
25 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-11-Regex.pptx">Pythonlearn-11-Regex.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
26 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-12-HTTP.pptx">Pythonlearn-12-HTTP.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
27 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-12-HTTP/>Pythonlearn-12-HTTP/</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
28 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-13-WebServices.pptx">Pythonlearn-13-WebServices.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
29 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-14-Objects.pptx">Pythonlearn-14-Objects.pptx</a></td><td align="right"></td><td align="right"></td><td align="right"></td></tr>
30 <tr><td align="top"></td><td align="right"><a href="/Pythonlearn-15-Database-
```

Elements Console Sources Network Performance Memory Application >> >

View: [Icons] Group by frame [x] Preserve log [x] Disable cache [x] Offline No thro

Filter [x] Hide data URLs [All] XHR JS CSS Img Media Font Doc WS Manifest Other

10 ms	20 ms	30 ms	40 ms	50 ms	60 ms	70 ms	80 ms	90 ms	100 ms	11
Name x Headers Preview Response Cookies Timing										
lectures3/										
blank.gif										
back.gif										
unknown.gif										
folder.gif										
text.gif										
favicon.ico										
General										
Request URL: https://www.py4e.com/lectures3/										
Request Method: GET										
Status Code: 200										
Remote Address: 104.27.158.166:443										
Referrer Policy: no-referrer-when-downgrade										
Response Headers										
age: 0										
cf-ray: 47f4ec1d097094fd-NRT										
content-type: text/html; charset=UTF-8										
date: Tue, 16 Jul 2019 15:13:50 GMT										
expect-ct: max-age=604800, report-uri="https://report-uri.cloudflare.com										
dn-cgi/beacon/expect-ct"										
server: cloudflare										
status: 200										
vary: Accept-Encoding										
via: 1.1 varnish-v4										
x-varnish: 50616032										

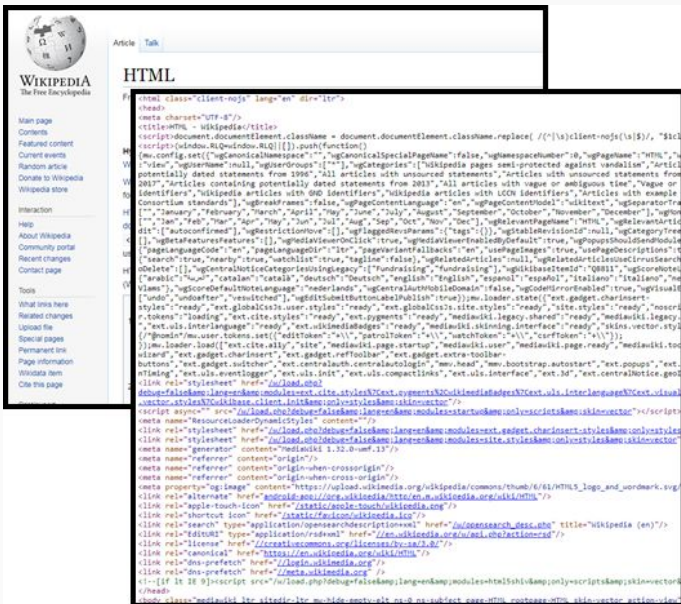
HTTP response

100 Continue	300 Multiple Choices	407 Proxy Authentication Required
101 Switching Protocols	301 Moved Permanently	408 Request Timeout
103 Early Hints	302 Found	
	303 See Other	
200 OK	304 Not Modified	500 Internal Server Error
201 Created		501 Not Implemented
202 Accepted	400 Bad Request	502 Bad Gateway
203 Non-Authoritative Information	401 Unauthorized	503 Service Unavailable
204 No Content	402 Payment Required	504 Gateway Timeout
205 Reset Content	403 Forbidden	505 HTTP Version Not Supported
206 Partial Content	404 Not Found	511 Network Authentication Required
	405 Method Not Allowed	
	406 Not Acceptable	

Start Coding...

Please access [Week 5 - Web Scraping...](#)

Web Scraping: using urllib + BeautifulSoup



DataFrame

BeautifulSoup

```
import urllib.request
from bs4 import BeautifulSoup
```

```
html = urllib.request.urlopen('https://www.sta.cuhk.edu.hk').read()
# "<html lang="en-US">
# <head id="Head">
# <title>
# Statistics > Home
# </title>... "
soup = BeautifulSoup(html, 'html.parser')
```

```
# Retrieve all of the anchor tags
tags = soup('a')
for tag in tags:
    print(tag.get('href', None))
```

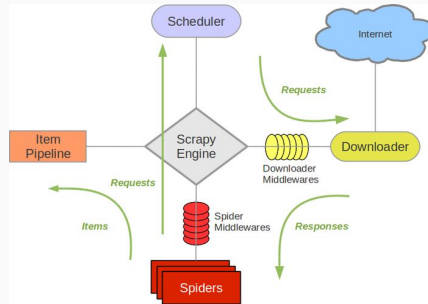
The extreme example (unwire)

A fellow who goes by the handle beaston02 wanted to see how unlimited Amazon's "unlimited" cloud storage plan was, so he uploaded 293 years' worth (2 million gigabytes or 2 petabytes) of 😊😊😊 videos to his account.

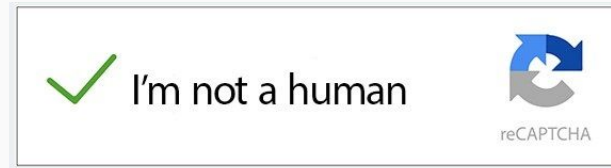


More Reference

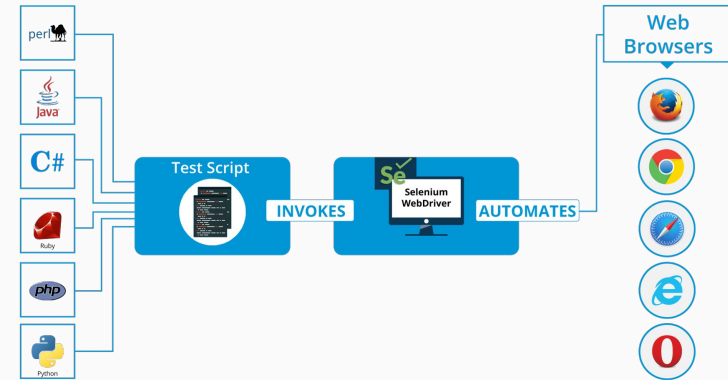
- [scrapy](#)



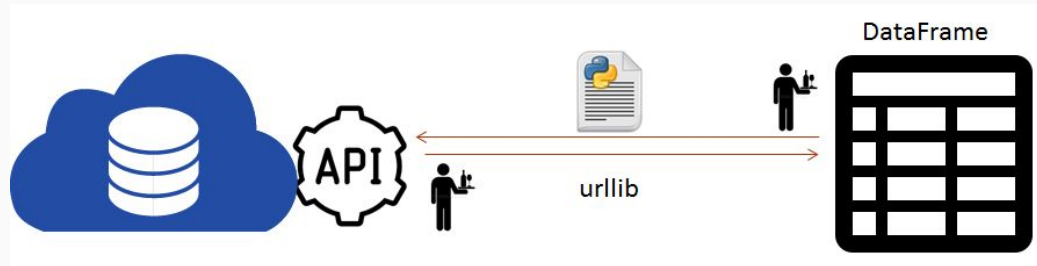
- [reCAPTCHA](#) ([ref](#))



- [Selenium](#) ([Ref](#))



- [Application programming interface](https://youtu.be/zvKadd9Cflc)
<https://youtu.be/zvKadd9Cflc>



- Google Map API
 - Get API key ([Reference](#))
 - At [here](#), Create the credential = your API Key (time limit = 24 hours)



At here , Choose your API you want



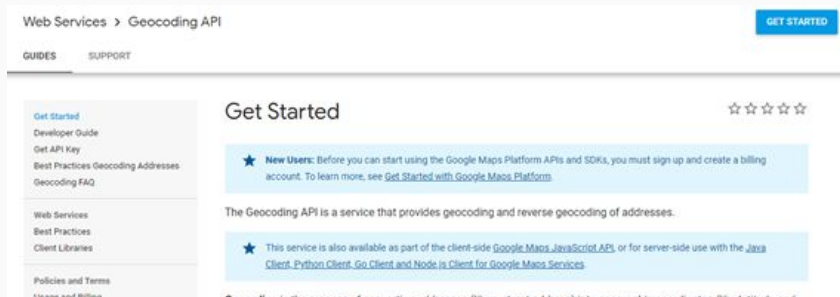
Then choose this



Try this in your browser:

https://maps.googleapis.com/maps/api/geocode/json?address=Lady+Shaw+Building%2C+Hong+Kong&key=YOUR_API_KEY

- [The API documentation](#)



- [Usage and Billing](#)

\$200 USD Google Maps Platform credit is available each month

MONTHLY VOLUME RANGE (Price per REQUEST)		
0-100,000	100,001-500,000	500,000+
0.005 USD per each (5.00 USD per 1000)	0.004 USD per each (4.00 USD per 1000)	Contact Sales for volume pricing

Start Coding...

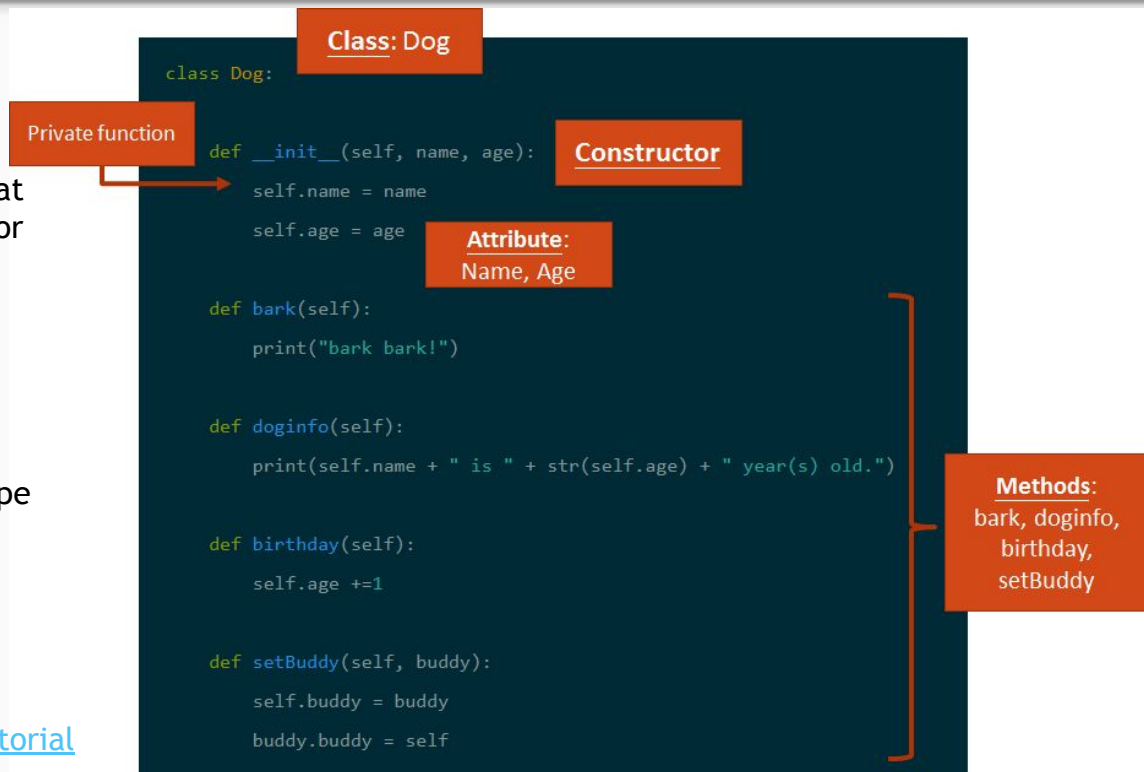
Please access [Week 5 - Google API Example...](#)

Object-Oriented Programming

- **Object-oriented programming (OOP)** is a programming paradigm based on the concept of "objects",
- **Classes** - the definitions for the data format and available procedures for a given type or class of object
- **Objects**, instances of classes, may contain
 - data, in the form of fields, often known as attributes
 - code, in the form of procedures, often known as methods
- Objects are created by calling a special type of method in the class known as a constructor.

Reference:

- <https://www.py4e.com/lessons/Objects>
- <https://www.datacamp.com/community/tutorial/s/python-oop-tutorial>
- <https://ithelp.ithome.com.tw/articles/10161285>
- <http://teddy-chen-tw.blogspot.com/2012/01/2object-class-instance.html>



[Data.gov.hk](https://data.gov.hk)



[CSDI](https://cspi.hk)



To be continue...
if we have no time for API...