

Psychoinformatics & Neuroinformatics



Week 4 Web Scraper



by Tsung-Ren (Tren) Huang 黃從仁

Tren: How to get these
data from dogpile?

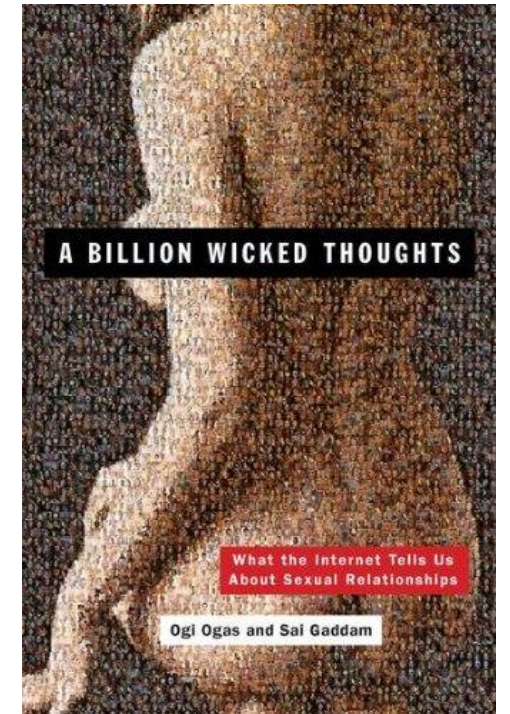


Sai: That's easy!
You can ...



Tren: Thanks for your
generous sharing!

**Sifu opens the door, but
you come & go by yourself!**



Relevant Job Demands

9/20 **Python 爬蟲工程師**

橙航有限公司 | 人力仲介代徵

台北市大安區 | 2年以上 | 專科

●專業技能 1. 2年以上爬蟲開發經驗，熟悉瀏覽器原理、前端 JS、AJAX 2. 熟悉一種主流的爬取技術及爬蟲框架工具，如火車採集器/Scrapy/Selenium等 3. 熟悉多線程、網絡編程、網頁抓取原理、正則表達式、HTTP協議 4. 熟

月薪65,000~130,000元

9/22 **Software engineer/軟體工程師(爬蟲)**

香港商香港球鞋團隊有限公司 | 綜合商品批發代理業

台北市中正區 | 3年以上 | 大學

welcome proposals for new tooling stacks. *** 軟體工程師 (爬蟲) Data Pipeline / ETL: from Crawler to DB 1. Familiar with Python 2.

月薪60,000元以上 員工100人 距捷運台北車站340公尺

9/22 **Python 爬蟲工程師**

德采有限公司 | 電腦系統整合服務業

台北市大安區 | 2年以上 | 大學

●專業技能 1. 2年以上爬蟲開發經驗，熟悉瀏覽器原理、前端 JS、AJAX 2. 熟悉一種主流的爬取技術及爬蟲框架工具，如火車採集器/Scrapy/Selenium等 3. 熟悉多線程、網絡編程、網頁抓取原理、正則表達式、HTTP協議 4. 熟

月薪60,000~140,000元 員工30人



glassdoor



All Job TypesPosted Any Time\$63K-\$138K

More

25 MilesAll CitiesCompany RatingsAll Industries

Create Job Alert

≡ Most Relevant

114 web scraping jobs Jobs



Recorded Future

Senior Engineer, Open Web Collections - Python and Open Web Scraping

Somerville, MA

\$104K - \$159K (Glassdoor est.) ⓘ

Viewed on September 25

Easy Apply 22d



Recorded Future

Data Engineer II, Dark Web Collections, Python & Web Scraping

Somerville, MA

\$73K - \$107K (Glassdoor est.) ⓘ

Viewed on September 25

Easy Apply 22d



Legalist

Web Scraping Engineer

San Francisco, CA

\$79K - \$113K (Glassdoor est.) ⓘ

Easy Apply 30d+

The PRACTIAL uses for web scrapping

Market/marketing research

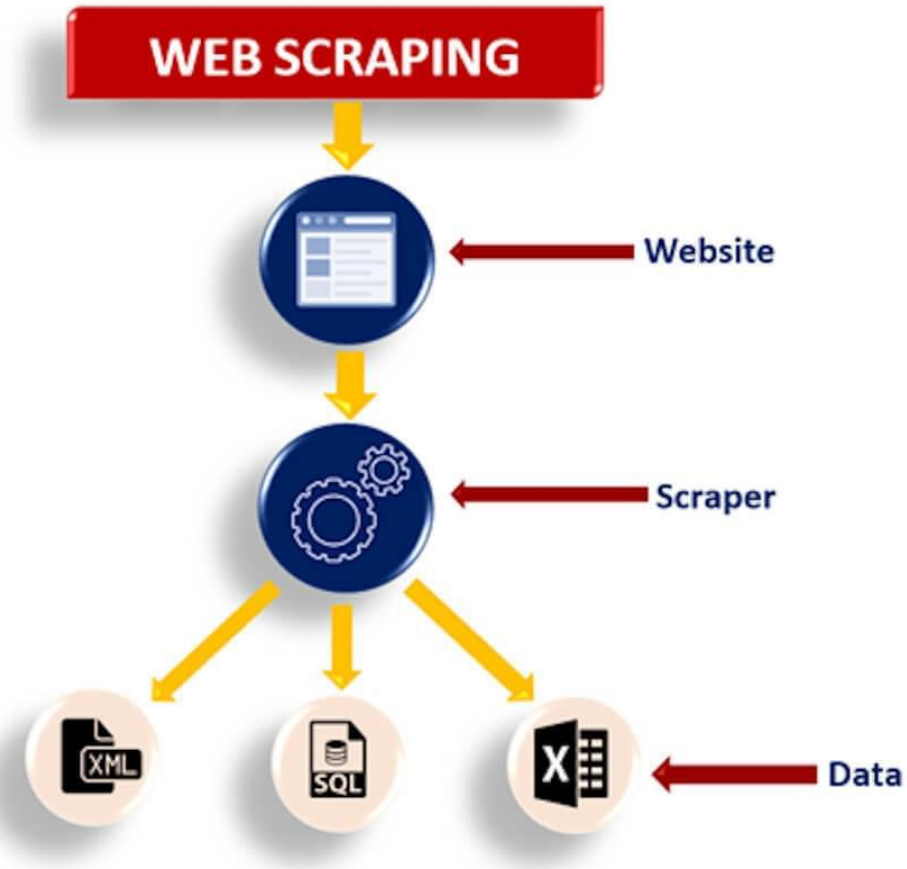
Understand your market/marketing

Building (real-time, big) data products

E.g., Google services & PayPal Honey

Building AI products

Your shiny AI (e.g. LLMs), needs more training data



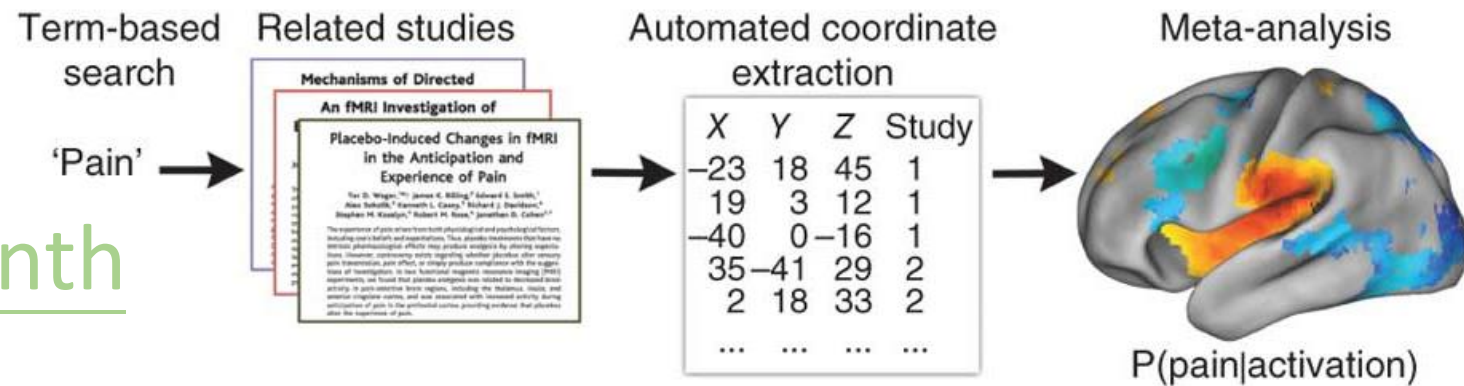
The ACADEMIC uses for web scrapping

(Offline) big data research

For a review, see [this article](#)

Large-scale meta-analysis

E.g., [metaBUS](#) & [NeuroSynth](#)



Building AI products for further analyses

E.g., gender prediction for Twitter data like [this](#) & [this](#)

Goals for today

Learning web basics

HTML & CSS

Learning basic scrapping skills

Beautiful Soup & Scrapy

Learning advanced scrapping skills

Form, cookie, Selenium, CAPTCHA



Goals for today

Learning web basics

HTML & CSS

Learning basic scrapping skills

Beautiful Soup & Scrapy

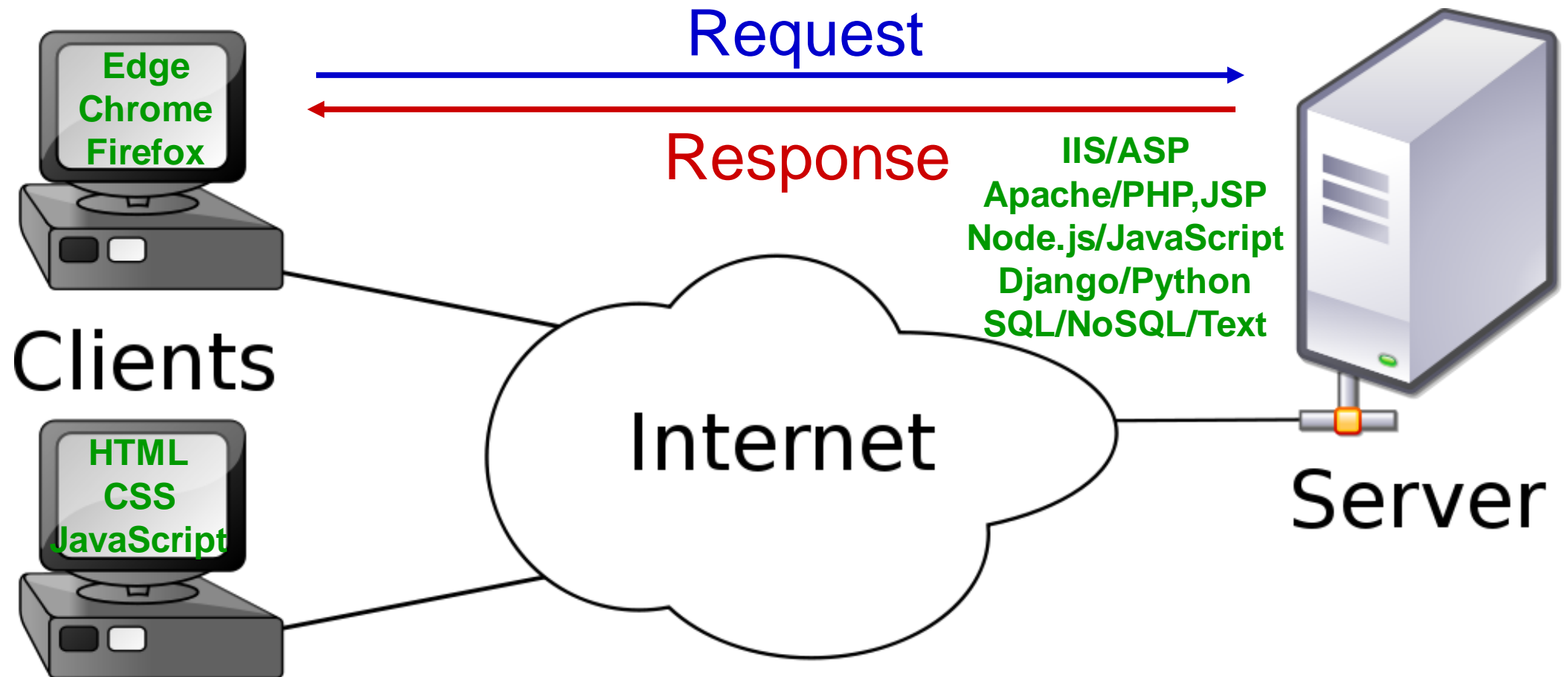
Learning advanced scrapping skills

Form, cookie, Selenium, CAPTCHA



Frontend vs. Backend

Client=Frontend for getting data; Server=Backend for giving data



Source codes of a webpage

The screenshot shows a web browser window with the Google search results for the query "the difference between data engineer and data scientist". The search result from "datascience.virginia.edu" is highlighted, showing the title "Data Science vs Data Engineering" and a paragraph describing the roles of data scientists and data engineers. The browser's developer tools are open, displaying the HTML source code of the page. The code is a snippet from a larger document, showing various HTML elements like

,

, , and with their respective attributes and classes. The text within the tags matches the content of the search result snippet. the difference between data engineer #text 977 x 100 A data scientist cleans and analyzes data, answers questions, and provides metrics to solve business problems. A data engineer, on the other hand, develops, tests, and maintains data pipelines and architectures, which the data scientist uses for analysis. <https://datascience.virginia.edu/news/data-science-vs-data-engineering/> Data Science vs Data Engineering 其他人也問了以下問題 Which is better data scientist or data engineer? Who gets paid more data engineer or data scientist? Which is easier data engineer or data scientist? Can a data engineer become a data scientist? ``` <div class="ULSxyt"> <div class="MjjYud"> <block-component> <div class="g wF4fFd JnWd g-blk" lang="zh-Hant-TW" data-hveid="CAQQAA" data-ved="2ahUKEwjC9_2-vrD6AhXTa4KH5kzBTkQjDY0AHOECAQAA"> <div class="dG2XIIf XzTjhb"> <div class="c2x2Tb"> <div> <div> <div class="xpdopen"> <div class="ifM90"> <h2 class="Uo8X3b 0hScic zSYMMe">網路上的精選摘要</h2> <div> <div> <div class="yp1CPe wDYxhc NFQFxe vi0Shc LKPCQc" data-md="471" lang="zh-Hant-TW"> <div class="V3FYCf"> <div class="wDYxhc" data-md="61" lang="zh-Hant-TW" style="clear:none"> <div class="LG0jhe" data-attrid="wa:/description" aria-level="3" role="heading" data-hveid="CAQQAA"> "A data scientist cleans and analyzes data, answers questions, and provides metrics to solve business problems. A data engineer, on the other hand, develops, tests, and maintains data pipelines and architectures, which the data scientist uses for analysis." == $0 2021年9月23日 </div> </div> <div class="g"></div> </div> </div> </div> </div> </div> </div> </div> </div> </div> </div> </div> ```

HyperText Markup Language (HTML)

`<h1>h1</h1><hr><h2>h2</h2>link`

`You
can break lines`

`<center>Center & change colors</center>`

``

`Bold`

`<i>Italics</i> <u>Underscore</u>`

``

`<table border=1>`

`<tr><td>11</td> <td>12</td></tr>`

`<tr><td>21</td> <td>22</td></tr>`

`</table>`



Cascading Style Sheets (CSS)

```
<style>
body {color:white; background-color:black;}
h1 {color:red; font-size:20pt}
.yy {color:yellow}
span#gg {color:green}
div#bb {color:blue}
</style>
<h1>Hi!</h1>
This is <span class=yy>test1</span><hr>
This is <div class=yy>test2</div><hr>
This is <span id=gg>test3</span><hr>
This is <div id=gg>test4</div><hr>
This is <div id=bb>test5</div><hr>
```



Goals for today

Learning web basics

HTML & CSS

Learning basic scrapping skills

Beautiful Soup & Scrapy

Learning advanced scrapping skills

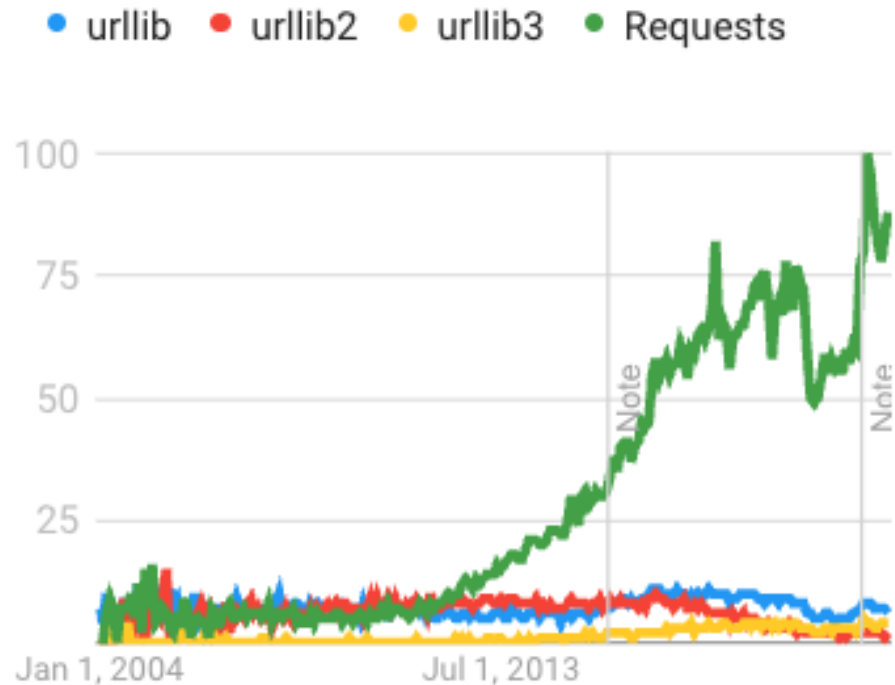
Form, cookie, Selenium, CAPTCHA



Tools of the trade: Popularity

Interest over time

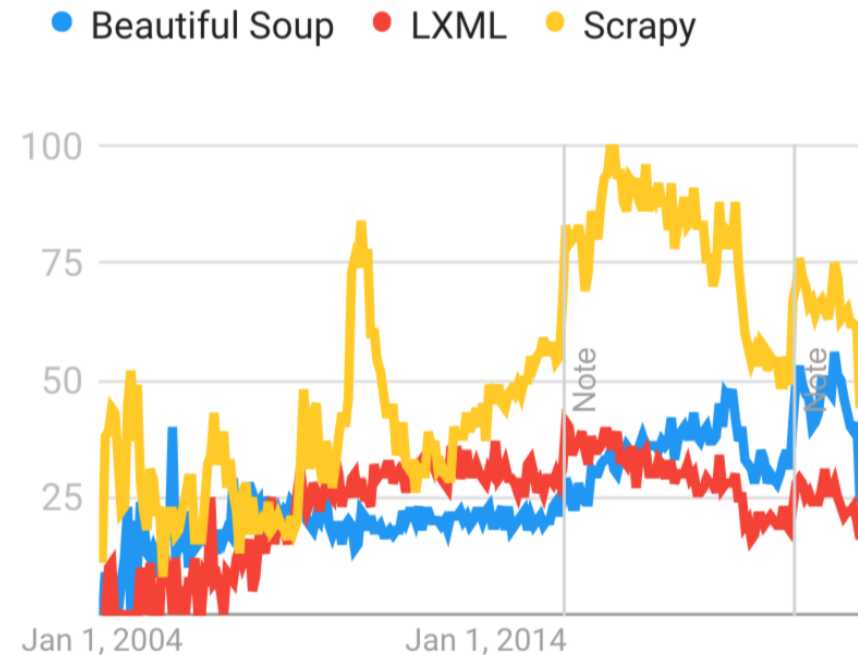
Worldwide. 2004 - present.



Google Trends

Interest over time

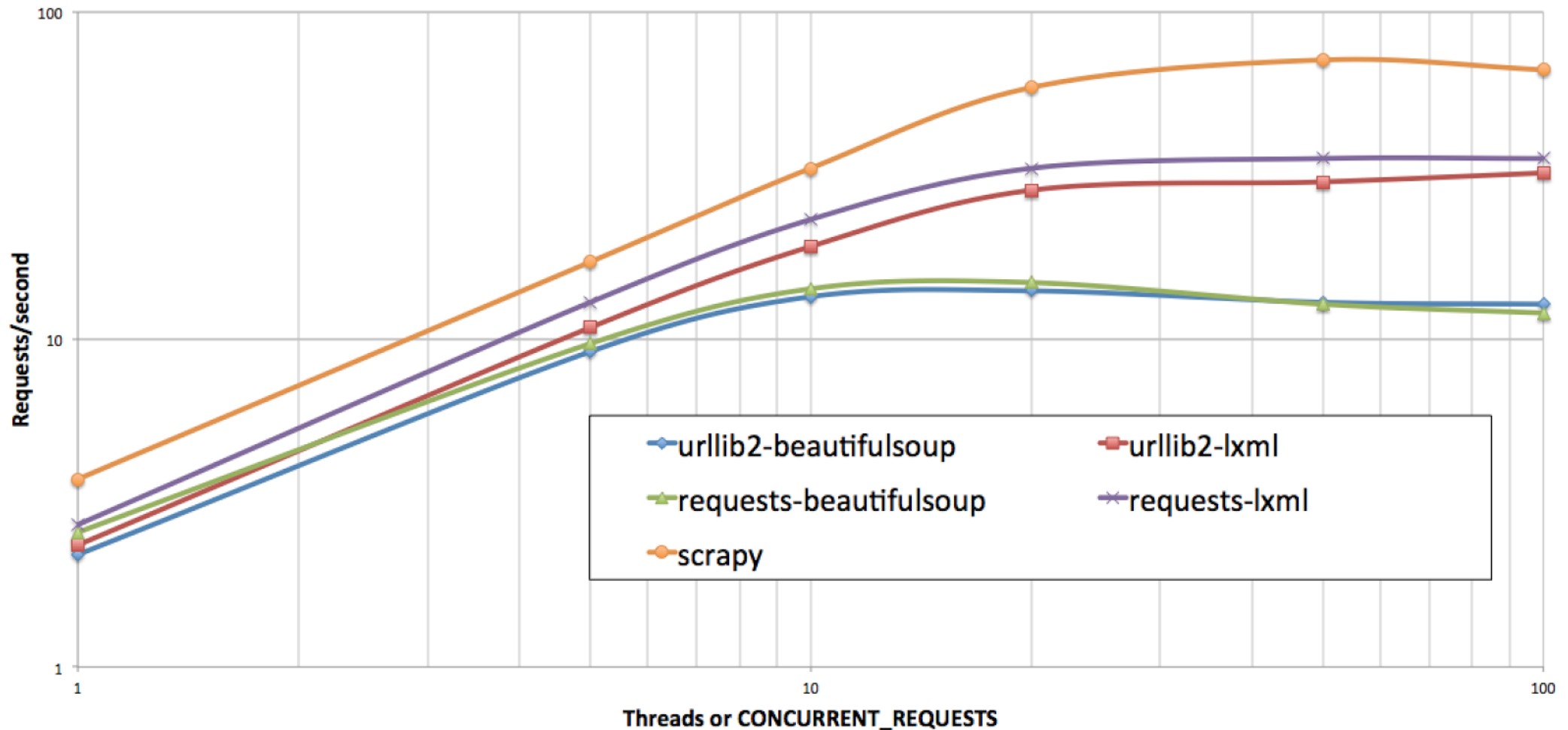
Worldwide. 1/1/04 - 10/5/23.



Google Trends

Tools of the trade: Performance (1/2)

The feature-richer, the slower

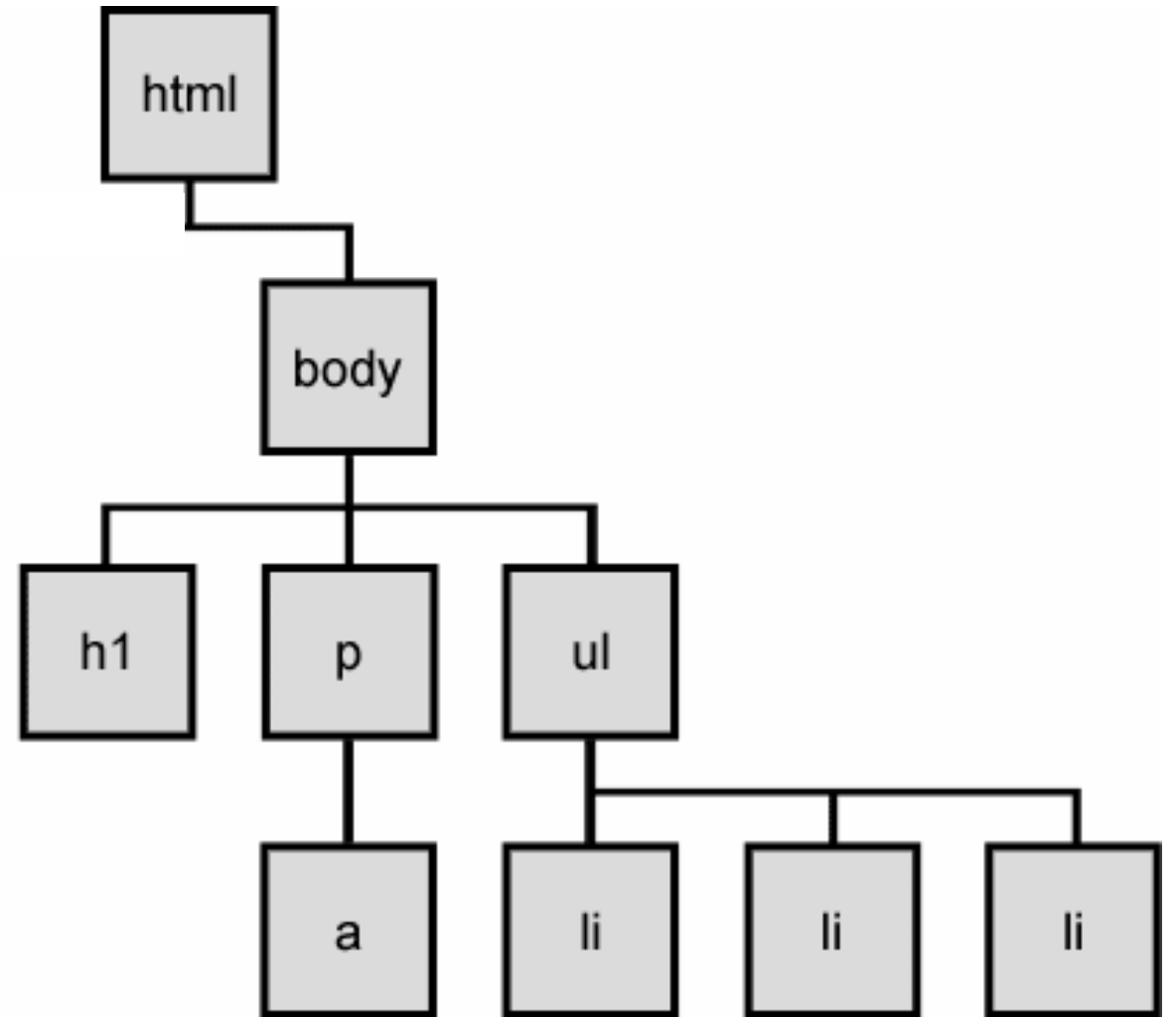


Tools of the trade: Performance (2/2)

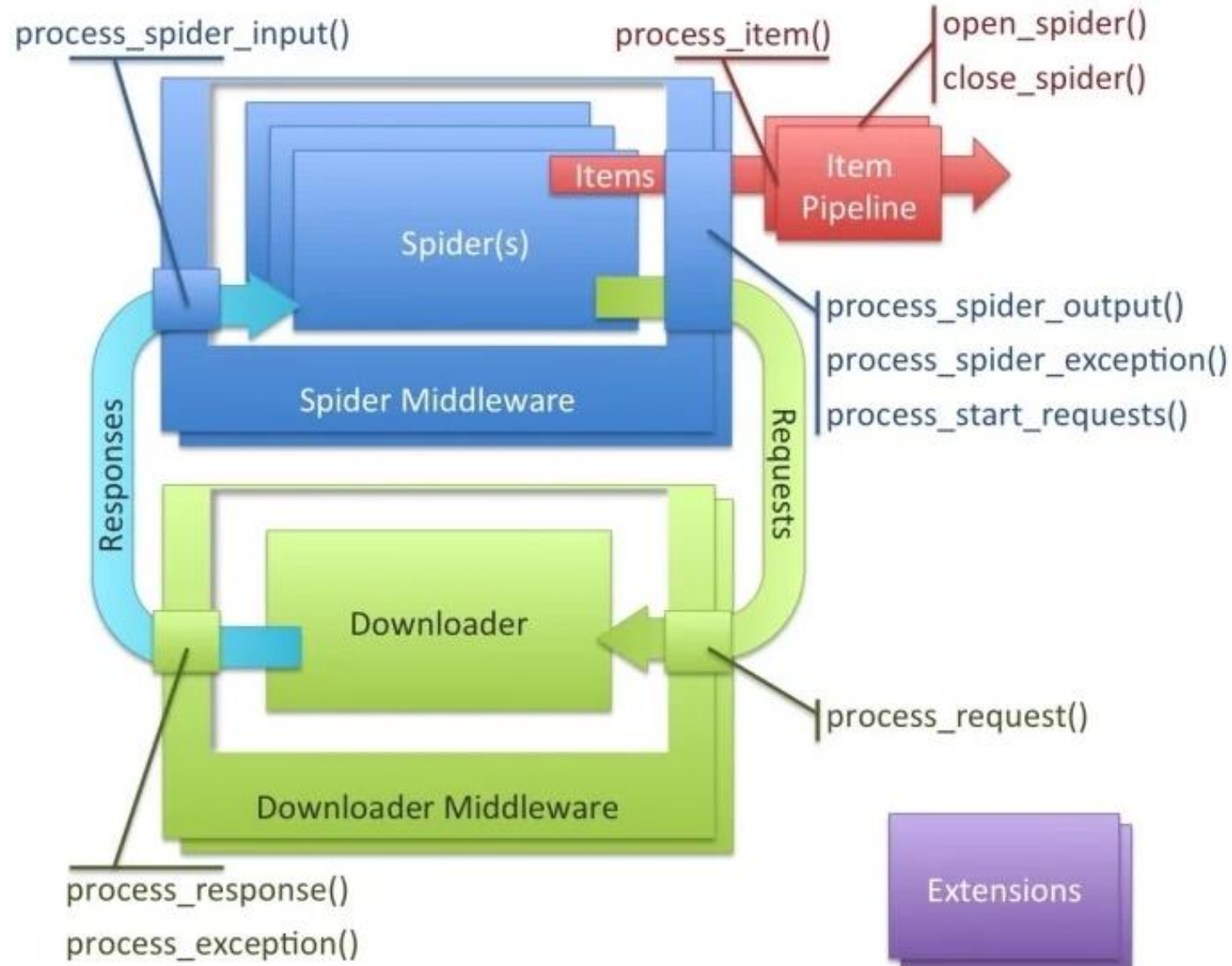
The backend parser can be switched in BeautifulSoup 4

Parser	Typical usage	Advantages	Disadvantages
Python's html.parser	<code>BeautifulSoup(markup, "html.parser")</code>	<ul style="list-style-type: none">• Batteries included• Decent speed• Lenient (As of Python 3.2)	<ul style="list-style-type: none">• Not as fast as lxml, less lenient than html5lib.
lxml's HTML parser	<code>BeautifulSoup(markup, "lxml")</code>	<ul style="list-style-type: none">• Very fast• Lenient	<ul style="list-style-type: none">• External dependency C
lxml's XML parser	<code>BeautifulSoup(markup, "lxml-xml")</code> <code>BeautifulSoup(markup, "xml")</code>	<ul style="list-style-type: none">• Very fast• The only currently supported XML parser	<ul style="list-style-type: none">• External dependency C
html5lib	<code>BeautifulSoup(markup, "html5lib")</code>	<ul style="list-style-type: none">• Extremely lenient• Parses pages the same way a web browser does• Creates valid HTML5	<ul style="list-style-type: none">• Very slow• External dependency Python

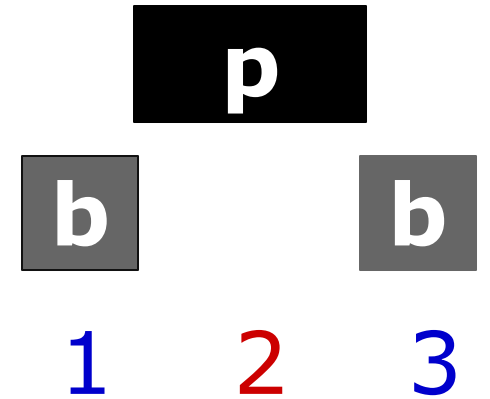
B-Soup: A HTML parser



Scrapy: A heavy/feature-rich scrapper



Beautiful Soup 101 (1/2)



Helping us to navigate in a HTML document

```
from bs4 import BeautifulSoup as BS
data='<p><b>1</b>2<b>3</b></p>'
s=BS(data,'lxml')
print(s) #<html><body><p><b>1</b>2<b>3</b></p></body></html>
print(s.find('2')) #2
print(s.find('p')) #<p><b>1</b>2<b>3</b></p>
print(s.b.string) #1
print(s.find_all('b')) [<b>1</b>, <b>3</b>]
```

Beautiful Soup 101 (2/2)

```
data="""<table>
```

```
<tr><td>11</td><td>12</td></tr>
```

```
<tr><td>21</td><td><a href="http://ptt.cc">22</a></td></tr>
```

```
<tr><td>31</td><td>32</td></tr>
```

```
</table>"""
```

```
from bs4 import BeautifulSoup as BS
```

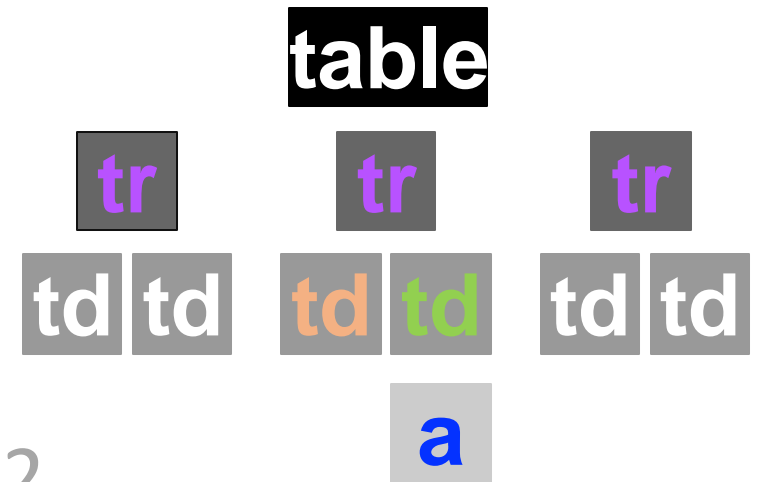
```
s=BS(data)
```

```
print(s.a,'\n',s.a.string) # <a href=...>22</a> & 22
```

```
print(s.a.parent.previous_sibling) #<td>21</td>
```

```
for child in s.tr.children:
```

```
    print(child) #<td>11</td> & <td>12</td>
```



Goals for today

Learning web basics

HTML & CSS

Learning basic scrapping skills

Beautiful Soup & Scrapy

Learning advanced scrapping skills

Form, cookie, Selenium, CAPTCHA



PTT: Age-restricted boards

<http://ptt.cc/bbs/Gossiping>

This website has been processed in accordance with the website content classification regulations

Warning: You must be at least 18 years of age to view the billboard content you are about to enter.

If you are not yet 18 years old, click Leave. If you are at least 18 years of age, you may not distribute, circulate, sell, rent, give or lend the contents of this area to a person under the age of 18 to browse, or present, broadcast or screen the contents of this website to such person.

I agree that I have entered at least eighteen years of age

Leave under the age of 18 or do not agree to these Terms

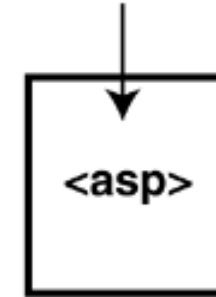
Batch kick kick industrial workshop			> Kanban Gossiping		About us
Kanban	Essence area	Oldest	< Previous	Next >	Latest
搜尋文章...					
	Lei Za did not play volleyball, and ran to Korea to become a dog trainer???	love	9/25	...	
	[News] The new mobile phone has only been bought for half a year, and the third child wants to change the iPhone 14!	b33	9/25	...	
21	[News] Send 4 hearts! Zhou Yuke confessed that "I just love Ah Zhong"	jiern	9/25	...	
5	Buffet guess the price (north chopinmozart		9/25	...	
3	Are cats and cats afraid that I will stink to death?	ghost90331	9/25	...	
2	Strong interest rates in the United States suck up money	psw	9/25	...	
4	Wu Bai did not sing last dance at the concert		9/25	...	
https://www.ptt.cc/bbs/ 566					

Sending requests via POST vs. GET

More convenient:

Using GET

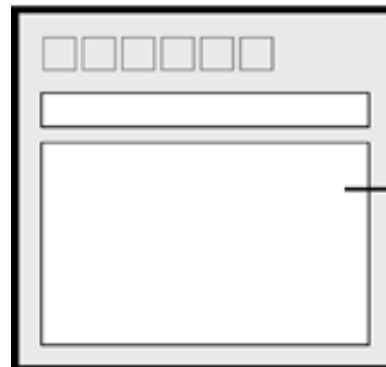
`http://www.somedomain.com/register.asp?name=jobe&email=jobe@electrotank.com`



More secure:

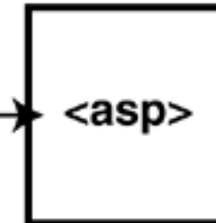
Using POST

`http://www.somedomain.com/register.asp`



HTTP Request

`name=jobe&
email=jobe@
electrotank.com`



Requests & Responses

```
<div class="bbs-screen bbs-content center clear">
  <form action="/ask/over18" method="post">
    <input type="hidden" name="from" value="/bbs/Gossiping/index.html">
    <button class="btn-big" type="submit" name="yes" value="yes">我同意， 我已年滿十八歲<br><small>進入</small></button>
    <button class="btn-big" type="submit" name="no" value="no">未滿十八歲或不同意本條款<br><small>離開</small></button>
  </form>
</div>
```

神 Selenium 神

Selenium can automate all browsing actions

```
from selenium import webdriver
```

```
URI='https://www.ptt.cc/bbs/Gossiping/'
```

```
driver=webdriver.Chrome() # try Firefox() or Edge()
```

```
driver.get(URI)
```

```
btn=driver.find_element_by_name('yes')
```

```
driver.save_screenshot('before_click.png')
```

```
btn.click()
```

```
driver.save_screenshot('after_click.png')
```

```
print(driver.page_source)
```



Solving CAPTCHA by OCR

[PLAN & BOOK](#)[MANAGE BOOKING](#)[ENGLISH](#)

Book THSR Tickets Have a pleasant and safe journey!

Book Tickets

One-Way ▾ Standard Car ▾ No Preference ▾

Search: ☒ by Date/Time ☐ by Train No.

From
Select

To
Select

Depart Date
9月26日 (Mon)

Depart Time
Select

Adult
1

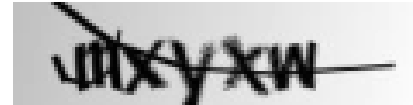
Child(6-11)
0



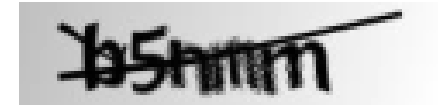
☐ Discount trains only

SEARCH

mxyxw



b5nmm



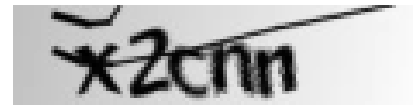
x3deb



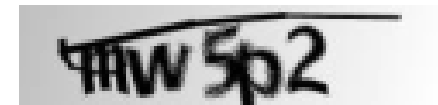
befbd



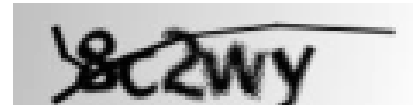
x2cnn



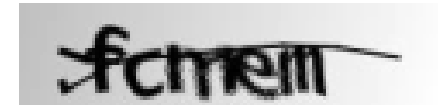
mw5p2



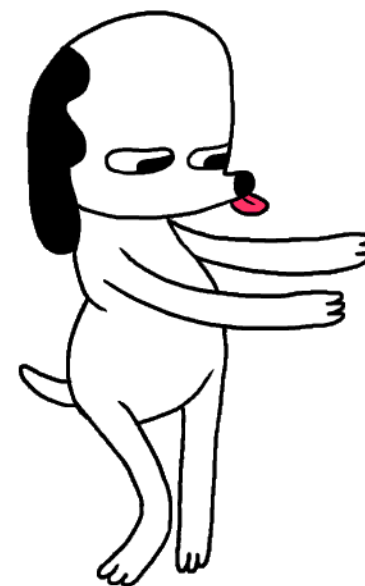
8c2wy



fcmem



輕輕鬆鬆，
打完收工！



GAME Over

