



1. INSTALL ANACONDA

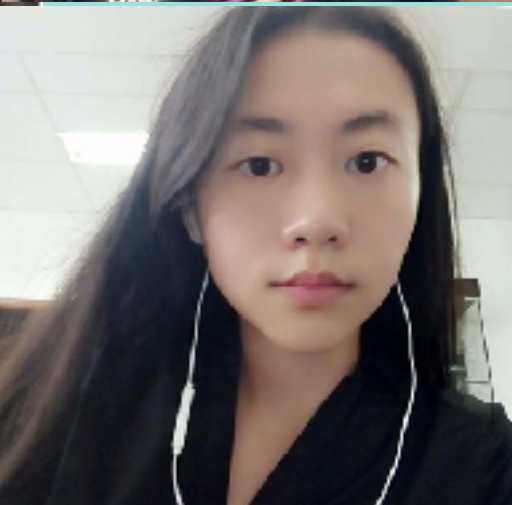
OR

PYTHON3 & JUPYTER NOTEBOOK

---

2. CLONE OR DOWNLOAD:

[HTTPS://GITHUB.COM/WWCODESG/DATASCRAPING](https://github.com/WWCODESG/DATASCRAPING)



# WOMEN WHO CODE<sup>®</sup> SINGAPORE

@wwcodesingapore

singapore@womenwhocode.com

[www.womenwhocode.com/singapore](http://www.womenwhocode.com/singapore)

[facebook.com/groups/wwcodesingapore](https://facebook.com/groups/wwcodesingapore)

[wwcodesg.slack.com](https://wwcodesg.slack.com)

[wwcodesg mailing list](#)



# OUR MISSION

Inspiring women to  
excel in technology  
careers.

WOMEN WHO  
CODE



# OUR GOALS

To Provide women with an **avenue** into tech

To Empower women with the **skills** they need for professional advancement.

To Build environments where **networking** and **mentorship** are valued.

To Create a global community to **support** women in tech wherever they live.

“DEAR WOMEN, WE NEED YOU” -The Tech Industry

## WHAT WE DO . . .

Organise technical events

Code Review (Newsletter)

#ApplaudHer

Scholarships

Conference Tickets

Job board



**THANK YOU TO OUR HOST**



# WOMEN WHO CODE

YUE LIN CHOONG

---

# DATA SCRAPING

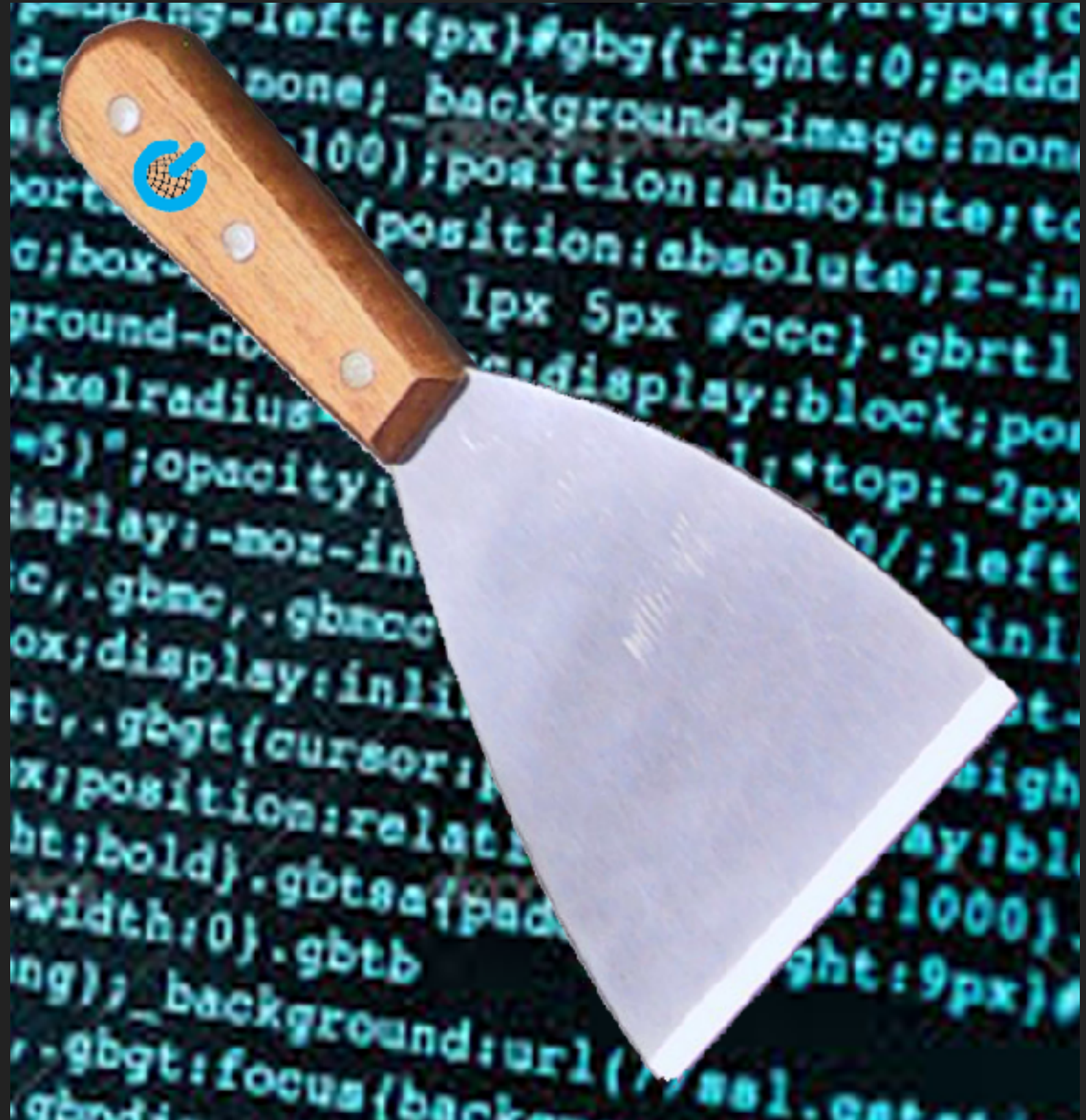


# DATA SCRAPING

---

## ALSO KNOWN AS

- ▶ Web scraping
- ▶ Screen scraping
- ▶ Data mining
- ▶ Web harvesting





# WHY DATA SCRAPING

- ▶ Gathering and processing large amounts of data
- ▶ Text, information and images

## PROCESS

- ▶ Retrieve data
- ▶ Parse data for target information
- ▶ Storing information
- ▶ (optional) moving to another page to repeat the process

```
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
```

```
<div class="container">
  <div class="carousel-caption">
    <h1>One more for good measure.</h1>
    <p>Cras justo odio, dapibus ac facilisis in, egestas eget quam. Donec id elit non mi porta ante dapibus.
    </p>
    <p><a class="btn btn-lg btn-primary" href="#" role="button">Button</a>
  </div>
</div>
</div>
<a class="left carousel-control" href="#myCarousel" role="button" data-slide="prev">
  <span class="glyphicon glyphicon-chevron-left" aria-hidden="true">
  <span class="sr-only">Previous</span>
</a>
<a class="right carousel-control" href="#myCarousel" role="button" data-slide="next">
  <span class="glyphicon glyphicon-chevron-right" aria-hidden="true">
  <span class="sr-only">Next</span>
</a>
</div><!-- /.carousel -->

<!-- Featured Content Section -->

<div class="container">
  <div class="row">
    <div class="col-md-4"></div>
    <div class="col-md-4"><h2>FEATURED CONTENT</h2></div>
    <div class="col-md-4"></div>
  </div>
</div>
```

# AGENDA

- ▶ urllib
- ▶ File download
- ▶ APIs
- ▶ BeautifulSoup (bs4)
- ▶ Selenium



# WOMEN WHO CODE

OLGA KUZMINA

---

# DATA CLEANING



# TRY, EXPLORE, PRACTICE . . .

Thank you 😊

[yuelin@womenwhocode.com](mailto:yuelin@womenwhocode.com)

[oskuzm@gmail.com](mailto:oskuzm@gmail.com)

WOMEN WHO  
CODE

**Coming soon ...**

**Our next event:**

**Coding with Java: Spring Boot**

**Weekly session:**

**Social Coding Monday**

[www.codesg.slack.com](https://www.codesg.slack.com)

[www.codesg mailing list](#)

[workshop assistant volunteer](#)