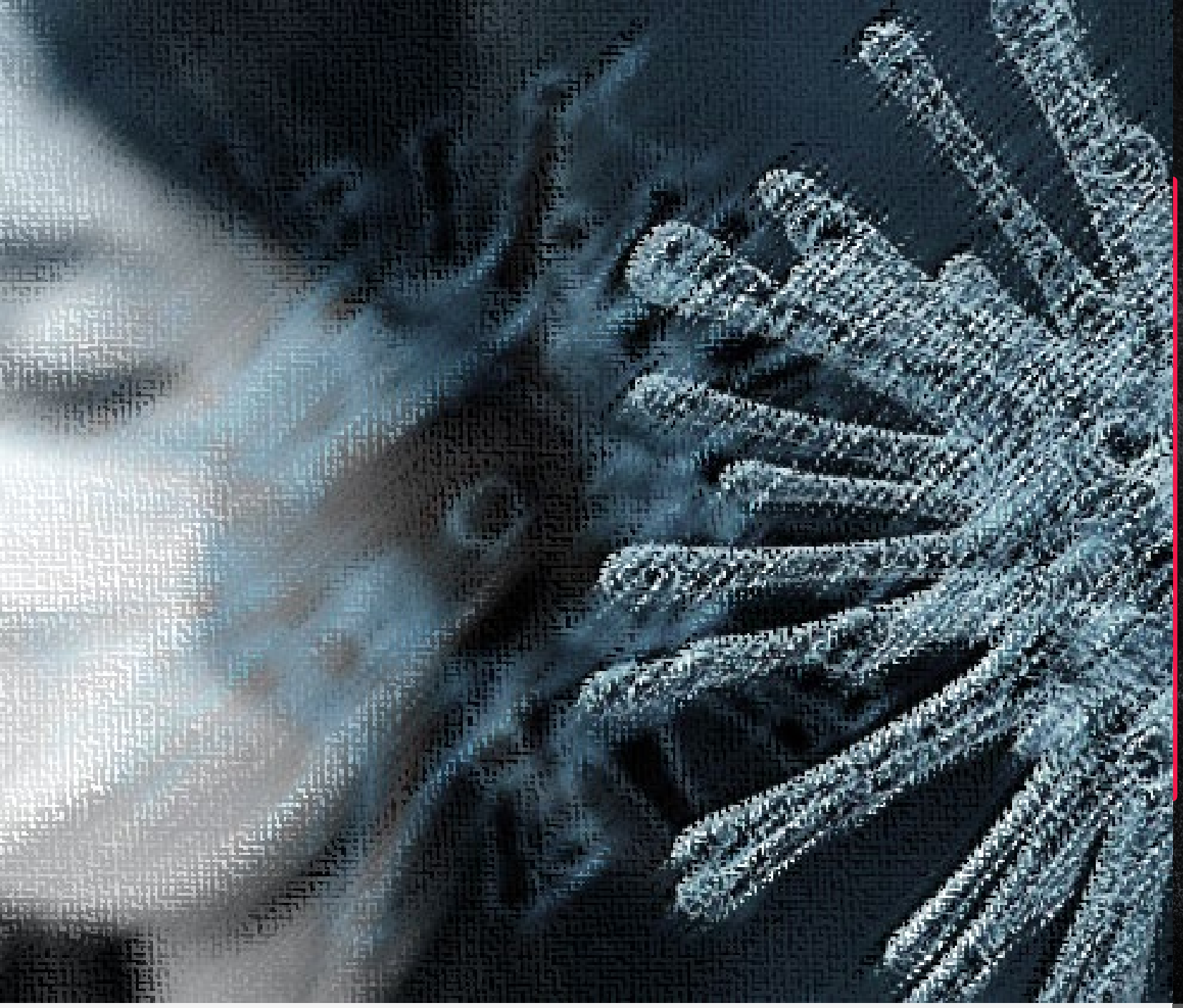


Combating COVID-19 Outbreaks

WQD7005 Data Mining
Master of Data Science | University of Malaya

Part A: Web Crawling of Real-time Data Assignment

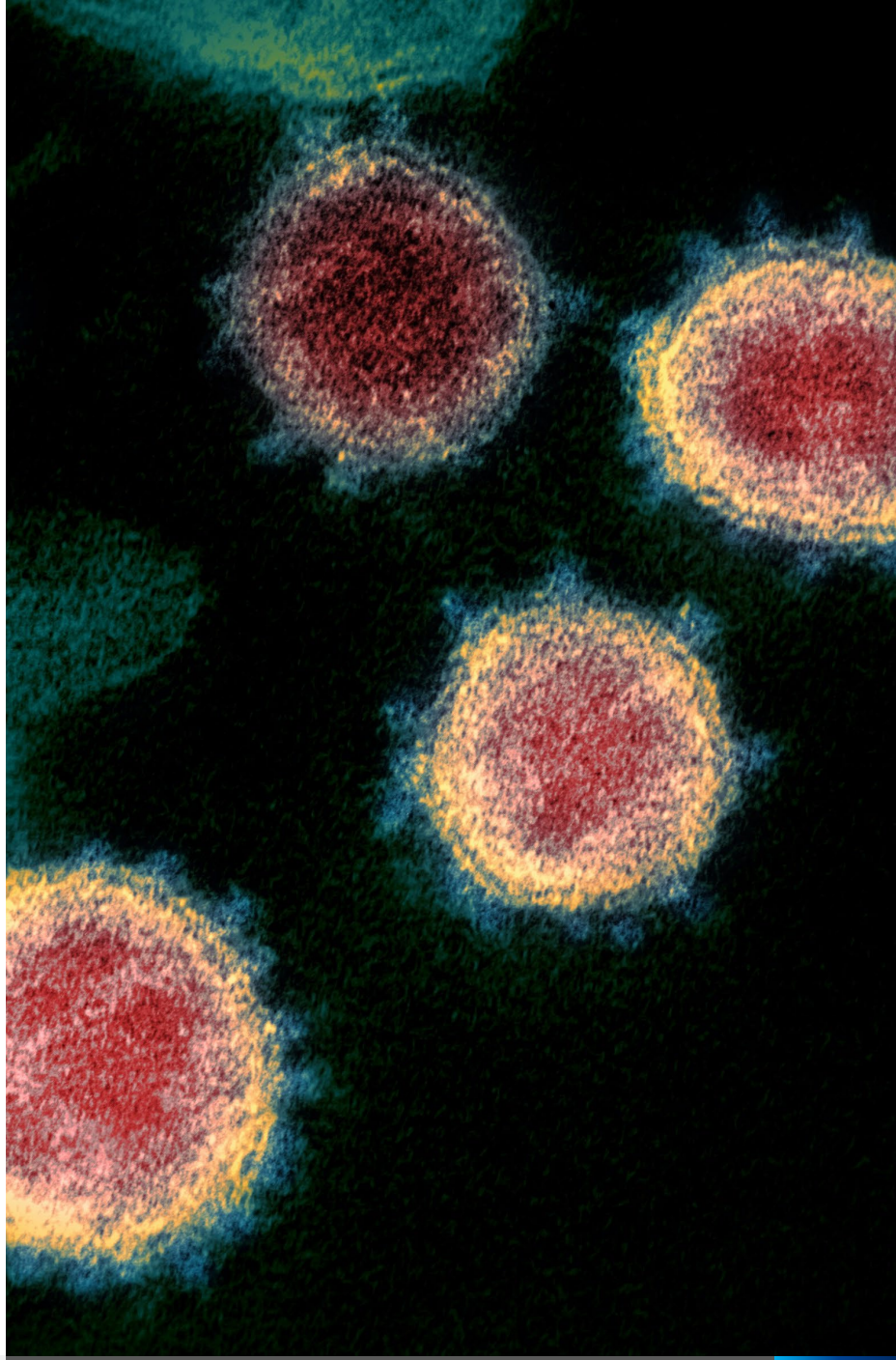
Group Members:
Azwa bin Kamaruddin (WQD170089)
Kok Hon Loong (WQD170086)

UNIVERSITY OF MALAYA

Assignment Milestones

WQD7005 Data Mining | Semester 2 Session 2019/2020

- **Part A: (Group)**
  - **Web Crawling of Real-time Data**
- Part B: (Group)
  - Data Management using Data Lake or Hive Data Warehouse
- Part C: (Group)
  - Accessing and Processing of Data from Data Lake or Hive Data Warehouse
- Part D: (Individual)
  - Interpretation and Communication of Data Insights
- Part E: (Group)
  - Deployment of the Data Mining Results on Web (Flask) and Mobile Application (Kivy)



Assignment Background

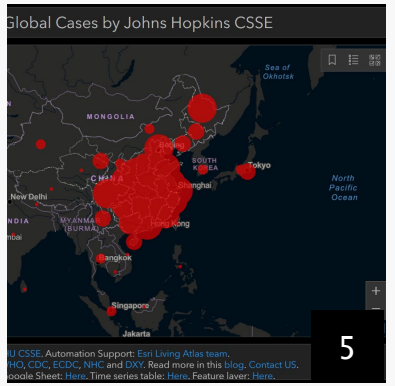
With the COVID-19 virus spreading to dozens of countries globally and still growing, the fear and worries from the people for their own safety and their loved ones are rightly so.

- In this assignment, we plan to predict the spread of the virus in each of the infected regions by crawling to the John Hopkins University GitHub website that contains daily data of COVID-19 outbreaks cases issued by the World Health Oganisation (W.H.O.)
- Our team have chosen Python as the programming language and uses the BeautifulSoup package to pull data out of HTML and XML files
  - The method our group uses to crawl the website to extract the data is also uploaded in this GitHub link: https://github.com/scholarazwa/wqd7005-assignment
  - Based on the data acquired, we have done our interpretation to better understand the daily total cases of confirmed, death and recovered per location.

How did we crawl the website to acquire the data for mining?

Our Approach…

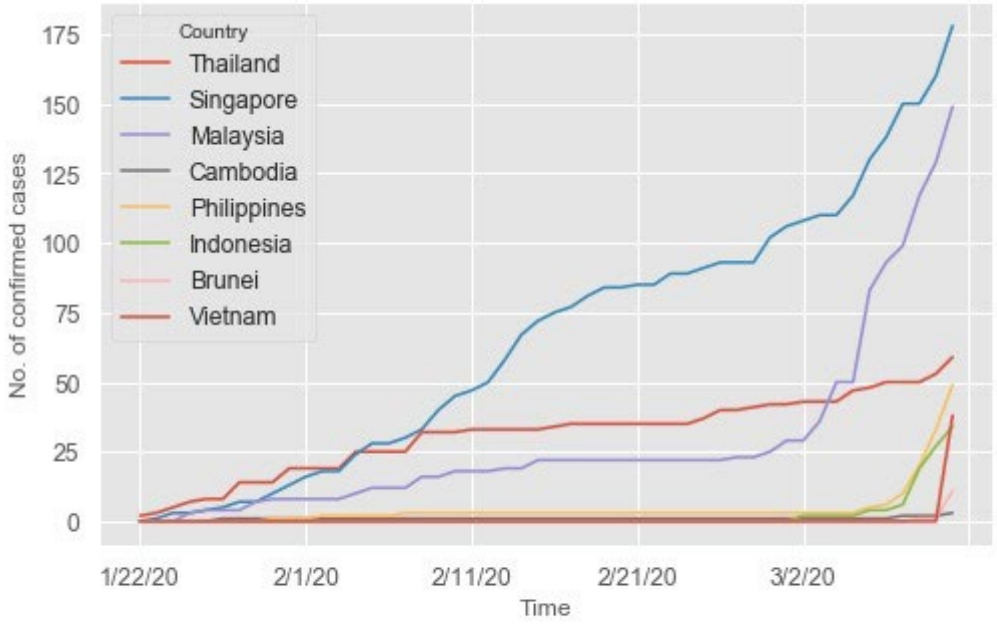
```
### ================================== ###
### NO. OF RECOVERED CASES
### ================================== ###

### ================================== ###
### NO. OF DEATH CASES
### ================================== ###

### ================================== ###
### NO. OF CONFIRMED CASES                                                    red.csv"
### ================================== ###


# In[ ]:


url = "https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_19-covid-Confirmed.csv"

try:
    page = requests.get(url, timeout=5)
    if page.status_code == 200:
        soup = BeautifulSoup(page.content,'html.parser')
        table = soup.find("table", {"class": "js-csv-data csv-data js-file-line-container"})
        df = pd.read_html(str(table))
        print(df)
    else:
        print(str(page.status_code) + " - Error, page not found.")
except requests.ConnectionError as e:
    print('Connection error')
    print(str(e))


# In[ ]:


data = df[0]
print(type(data))
print(data.describe())
data.describe().transpose()
```

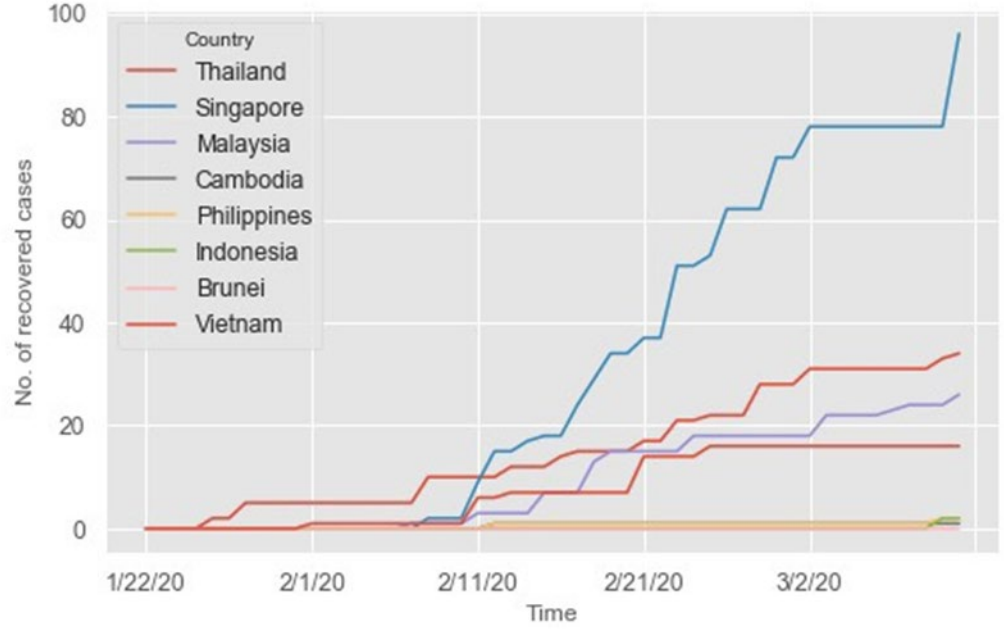


UNIVERSITY OF MALAYA

Initial Visual Plots from the Data Acquired

Reported Cases and Patients Recoveries from COVID-19 Outbreaks in ASEAN





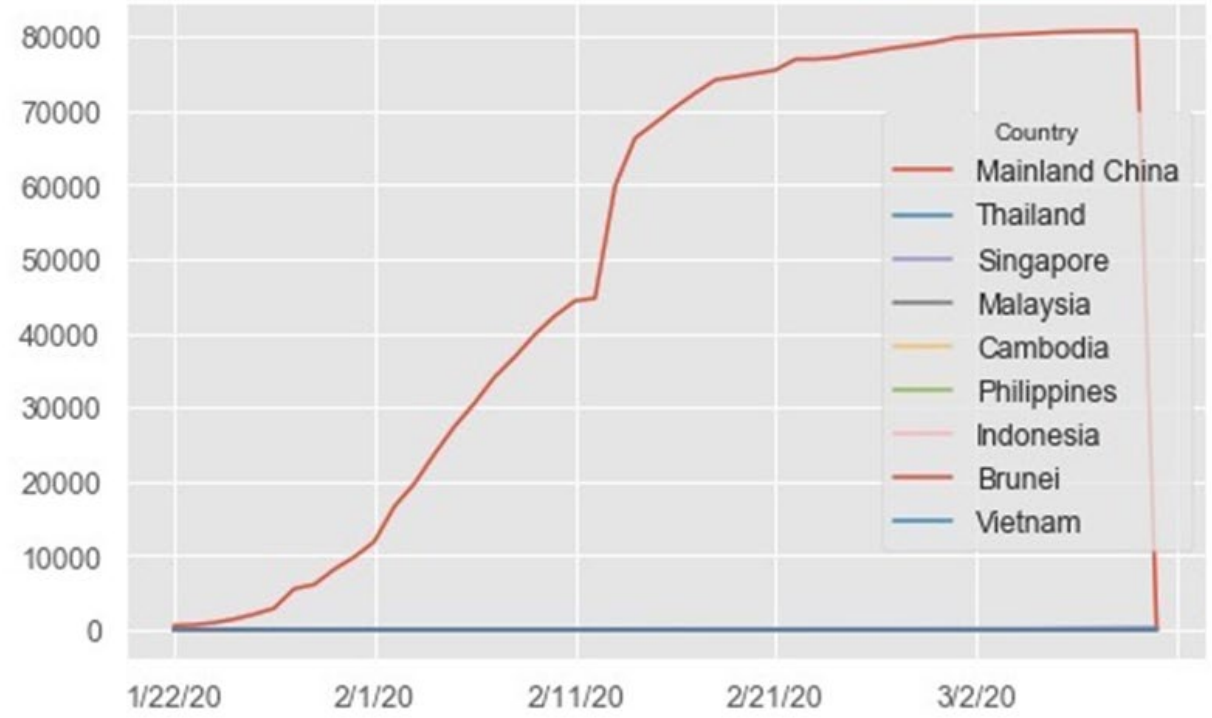

```
# COMMENT: Singapore is leading in the number of covid-19 recovery.
# Followed by Thailand and Malaysia.
```

```
# COMMENT: Singapore has the most number of recorded confirmed cases and is increasing exponentially.
# Malaysia is second and Thailand is 3rd most recorded confirmed cases in ASEAN.
```

UNIVERSITY OF MALAYA

Initial Visual Plots from the Data Acquired

Reported Cases from COVID-19 Outbreaks for ASEAN vs Greater China

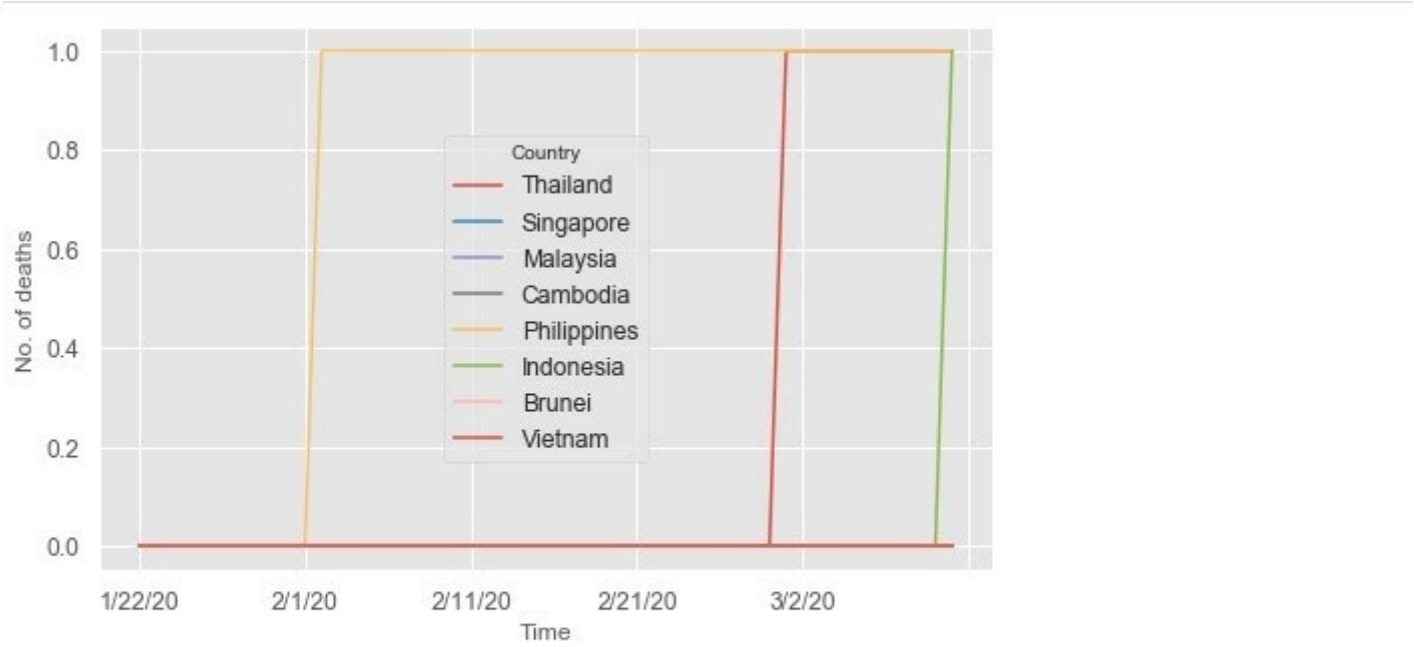



```
# COMMENT: China numbers are too huge to compare with the number of cases in ASEAN.
```

UNIVERSITY OF MALAYA

Initial Visual Plots from the Data Acquired

Death Cases from COVID-19 Outbreaks in ASEAN




```
# COMMENT: The number of deaths from the covid-19 is rare/low in the ASEAN region.
# Only 3 patients have died: 1 from Phillipines, 1 from Thailand, and 1 from Indonesia.
```

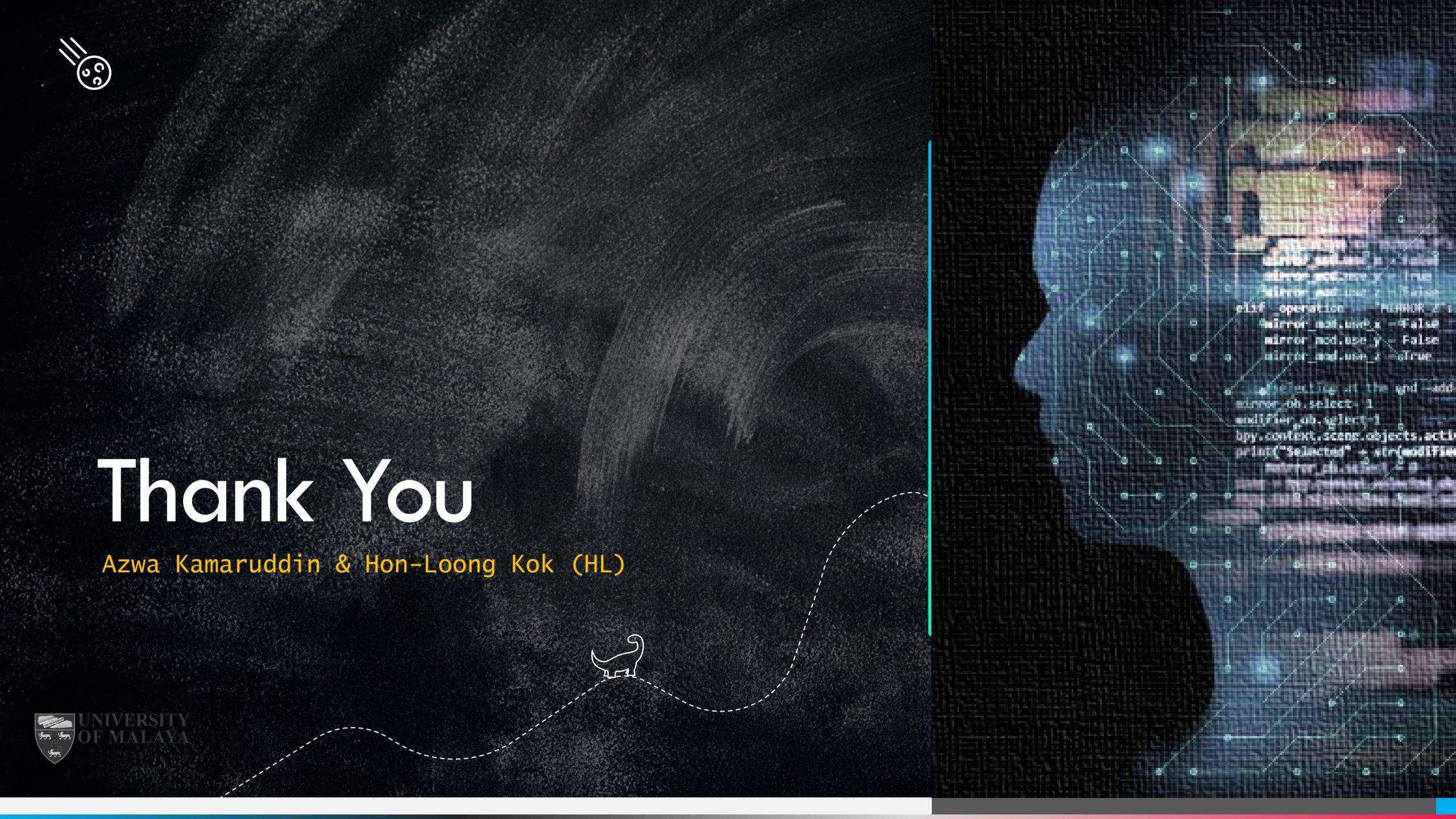
UNIVERSITY OF MALAYA

Summary

- We have use web crawling process to extract data from websites to validate and check for contents.
- The data interpreted by the crawler is in an unstructured CSV or XML format.
- Web crawling skills is getting more important in Data Science, as we are currently in the era of information explosion.
- By acquiring the data from web crawling process, we can then proceed with our Data Mining analytics to provide powerful insights.

Thank You

Azwa Kamaruddin & Hon-Loong Kok (HL)



UNIVERSITY OF MALAYA