# Visualizing Covid-19 Misinformation on Twitter

**Matthew Seesselberg and Daehan Kwak, Ph.D.**
**School of Computer Science and Technology, Kean University, Union, NJ 07083**

## Abstract

As the Covid-19 pandemic continues, there have been efforts to slow the spread of the virus as well as how to stop the infection altogether. For this effort, attempts are being made by both the public and private sectors to try and contain the spread of the virus as well as spread key information pertaining to how individuals can help. The problem with the public sector helping is the partisan split from municipal to federal level, as well as the supporters of the elected officials, not agreeing to methods of prevention. This has caused some conflicting information into what is the proper protocol and what the public should do. The overall goal of this project is to gain a better understanding of how misinformation spreads by visualizing the misinformation spread about Covid-19. Since twitter is being used as a source of communicating information about the pandemic, it can be used as a source of information to collect data from. The methods of data collection allow for a faster, broader search that can be narrowed down to limit how much information needs to be looked at. The end result will allow us to see where misinformation tweets originate from as well as information such as user account creation date, user statistics such as followers, and geographic location of the user.

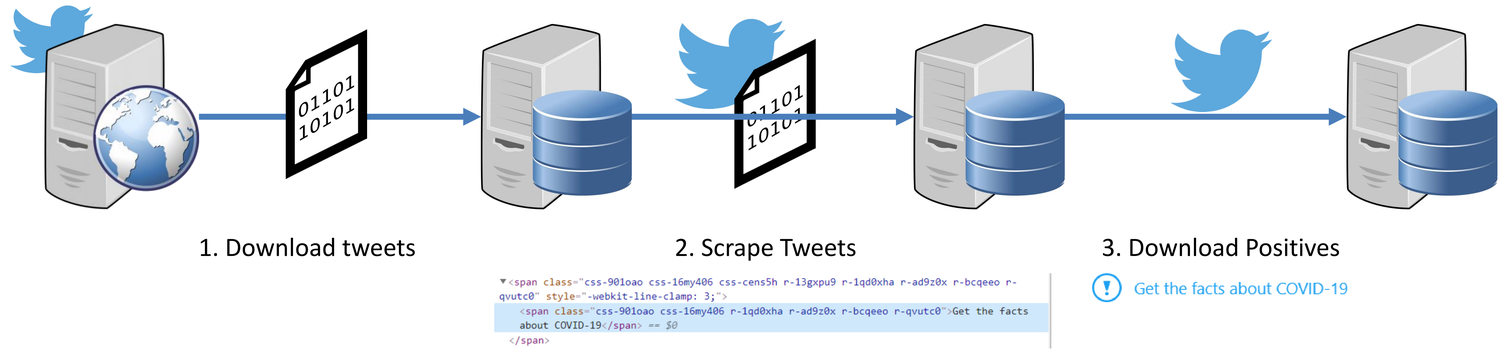
## Introduction

- Is there correlation between misinformation source and location?
- Twitter API could not be implemented directly due to lack of flag
  - Datamining would have to be broken into steps
  - Get Tweets, Scrape HTML, Source Tweets

## Implementation

- An automated script was written in Python using GetOldTweets3 to download tweets that matched certain key words.
- Data collected is between March 1st 2020 and August 31st 2020. The resulting tweets were stored in MySQL
- Tweets are then scraped from twitter.com to locate the misinformation tag.
- A Python script ran on multiple machines to speed up the process. Tweets that contained fewer than 15 characters were ignored for the initial pass.
- Tweets that do contain the misinformation tag were downloaded fully using Tweepy.
- This setup allows for concurrency as well as handling duplicate text data to lower priority of duplicate text being checked through another tweet ID.
- The obtained final data would be visualized on a map to showcase the source of misinformation as well as the content of the misinformation.

## Scrape System




## Results

- Due to the scale of the data collected, the data is still being collected and processed
  - So far, around 13,500 tweets have been scraped out of nearly 330,000 tweets collected.
- Of the tweets scraped, none have the misinformation tag.

| id | permalink | python_state | python_term |
|---|---|---|---|
| 1233496318… | https://twitter.com/ScottyW… | Alaska | Coronavirus |
| 1235086044… | https://twitter.com/ScottyW… | Alaska | Coronavirus |
| 1237116010… | https://twitter.com/ScottyW… | Alaska | Coronavirus |
| 1237385862… | https://twitter.com/bittyann… | Alaska | Coronavirus |
| 1237821064… | https://twitter.com/halimurra… | Alaska | Coronavirus |

- Due to a backend API change on twitter, GetOldTweets3 no longer works. Less tweets are pulled per query but allows for direct scraping.

| url | dump | py_state | py_term | covid_flag |
|---|---|---|---|---|
| https://twitter.com/search… | <head><meta charset="utf-8"> <meta… | USA | Coronavirus | No |
| https://twitter.com/search… | <head><meta charset="utf-8"> <meta… | USA | Coronavirus | No |
| https://twitter.com/search… | <head><meta charset="utf-8"> <meta… | USA | Coronavirus | No |
| https://twitter.com/search… | <head><meta charset="utf-8"> <meta… | USA | Coronavirus | No |
| https://twitter.com/search… | <head><meta charset="utf-8"> <meta… | USA | Coronavirus | No |

## Conclusions

- Twitter has become a source of information spread.
- Some of this information is harmful (misinformation).
- Python was used to collect tweet data to locate misinformation flagged by twitter.
- From what has been collected, misinformation is not as widespread when compared to the total amount of tweets to a subject.

## Future Work

- Optimize the script to speed up data collection.
  - Approach query/min limit
- Look at Twitter's other tag regarding "Manipulated Media"

## Acknowledgements

- This material is based upon work supported by the National Science Foundation under Grant No. HRD-1034620.