



# CDR: COVID-19 DATA SCIENCE PROJECT

Albert Sze

05/29/2020



# PURPOSE

- Research the spread of Covid-19 in the USA and perform initial machine learning algorithms to further understandings
- Determine the spread over time to various states
- Apprehend the change in number of people tested over time



# OVERVIEW

- COVID-19 is a virus caused by a coronavirus called SARS-CoV-2.
- Virus began in November 17 2019 in Wuhan China, which lead to a global pandemic.
- The first US case was identified in January 19, 2020 in Snohomish County, Washington
- Older adults and people who have severe underlying medical conditions like heart or lung disease or diabetes seem to be at higher risk for developing more serious complications from COVID-19 illness.

# DATA SET

- Data was procured from [Kaggle.com](https://www.kaggle.com), with the original source from COVID-19 Tracking project and NYTimes.
  - Contains three tables
    - us\_covid19\_daily: covid-19 data at country level
    - us\_states\_covid19\_daily: covid-19 data at state level
    - us\_counties\_covid19\_daily: covid-19 data at county level



# MOVING FORWARD PLAN

- Design database system for covid-19 data set
- Perform initial data analysis on data set to understand virus spread rate
- Use machine learning algorithms to create similarities and potential correlations of the spread of the virus

# REFERENCES

- [https://en.wikipedia.org/wiki/COVID-19\\_pandemic](https://en.wikipedia.org/wiki/COVID-19_pandemic)
- <https://www.nejm.org/doi/full/10.1056/NEJMoa2001191>
- <https://www.kaggle.com/search>