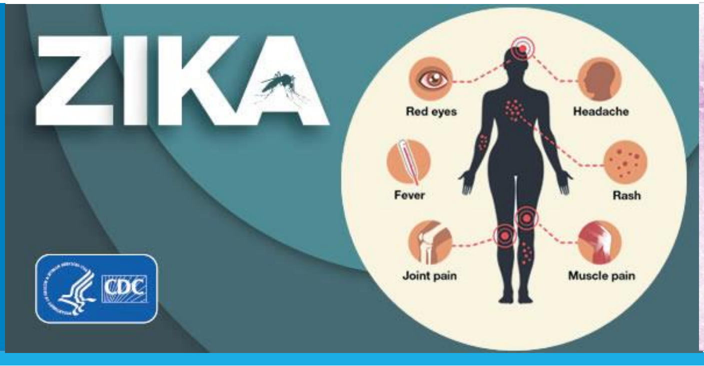
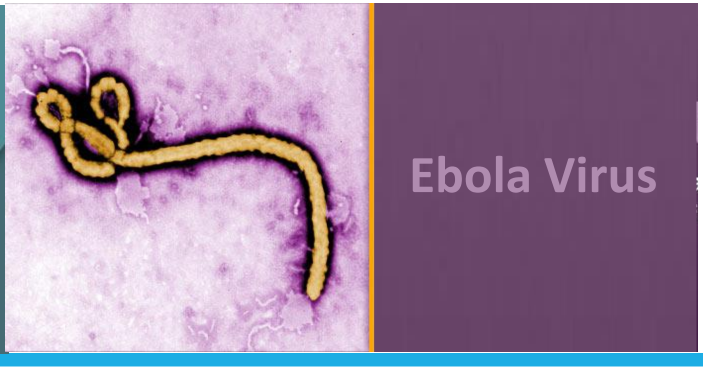
Global Health Challenge

DATA SCIENCE IN INFECTIOUS DISEASES

Background

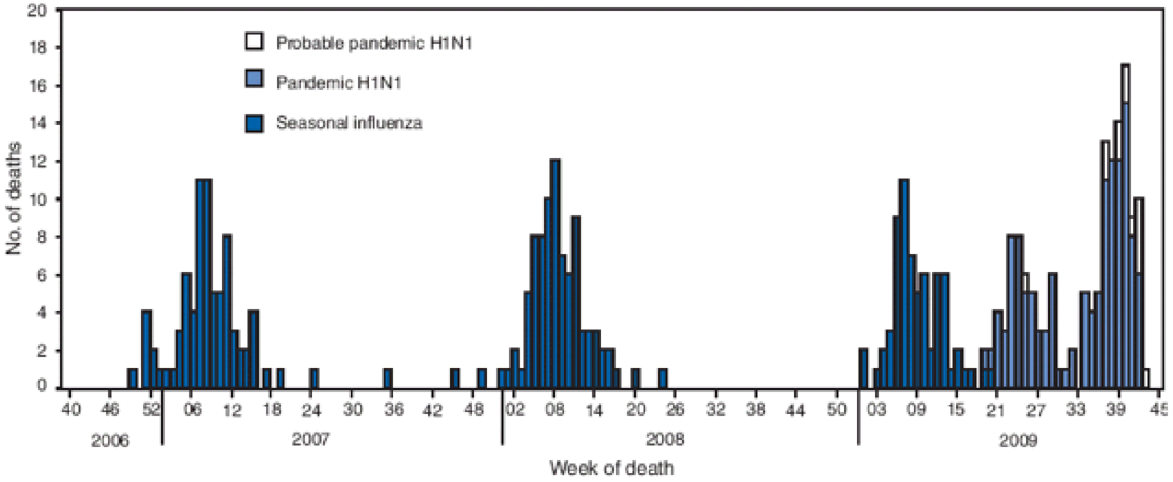
- Infectious diseases are closely related to our human world.
- Influence everybody’s life in a certain way.
- Travel across continents and spread widely.





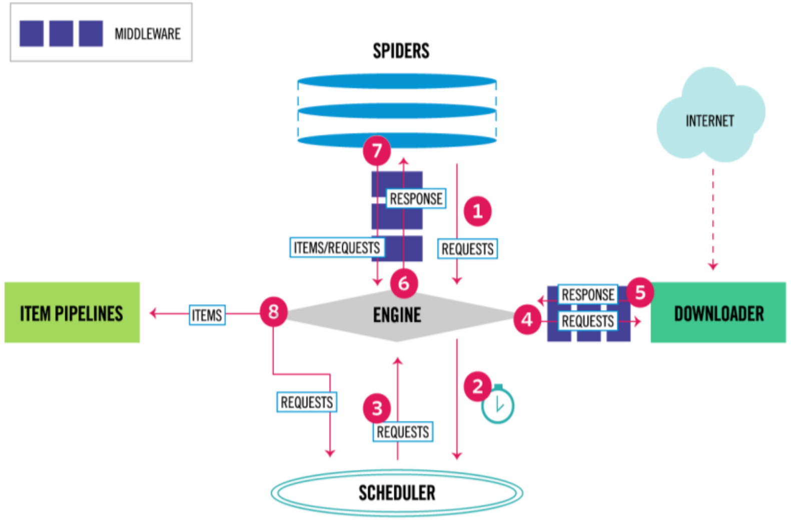
Technical Ideas

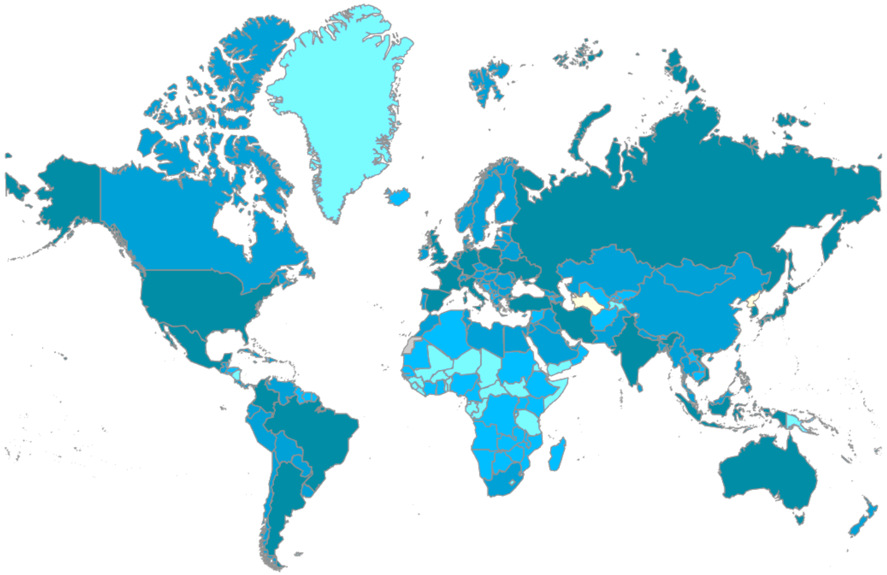
- Infectious diseases have extensive existences.
- Carefully monitored and studied by scientists around the globe.
- Many of their pandemics take place periodically.
- Many of their spread and mutation are strongly featured.

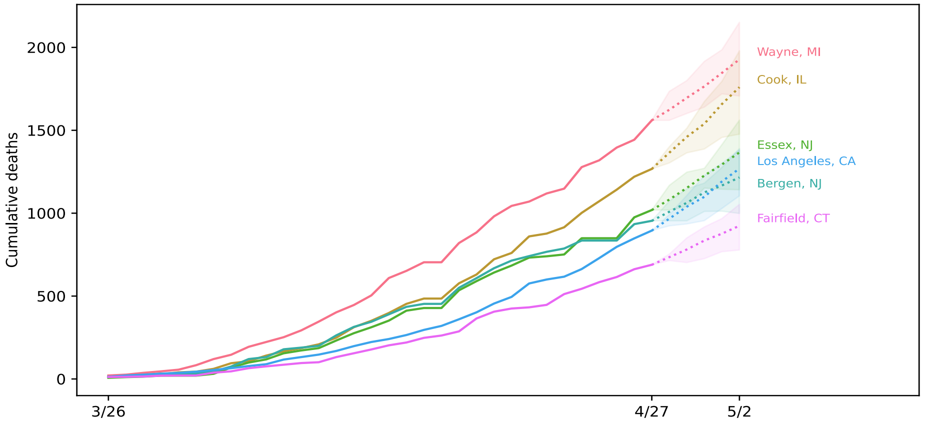


Technical Route

- Sampling and collecting data from the Internet.
- Visualization using the collected data.
- Data mining and further analysis.

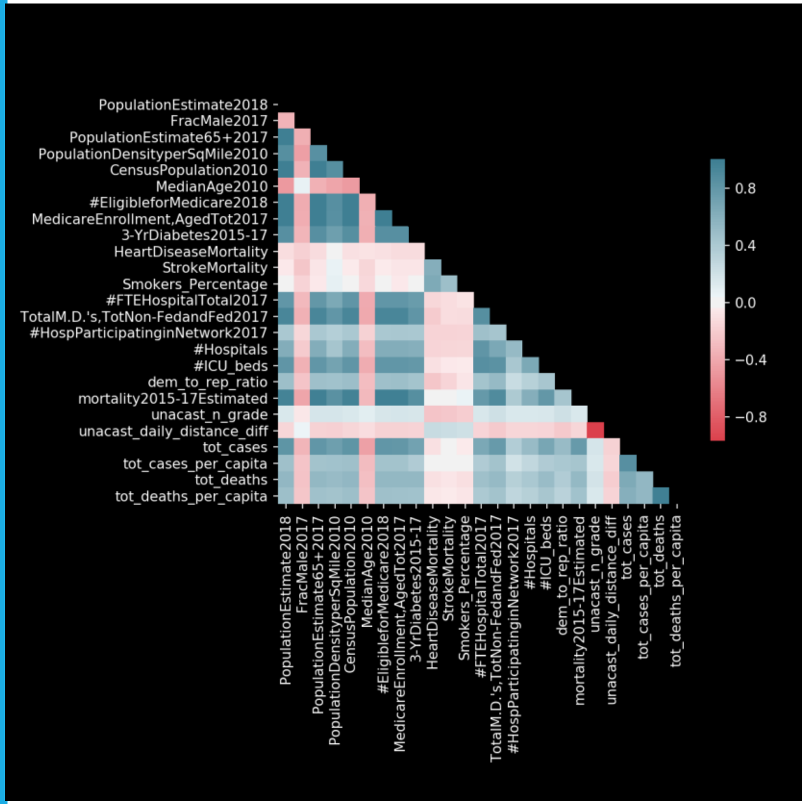


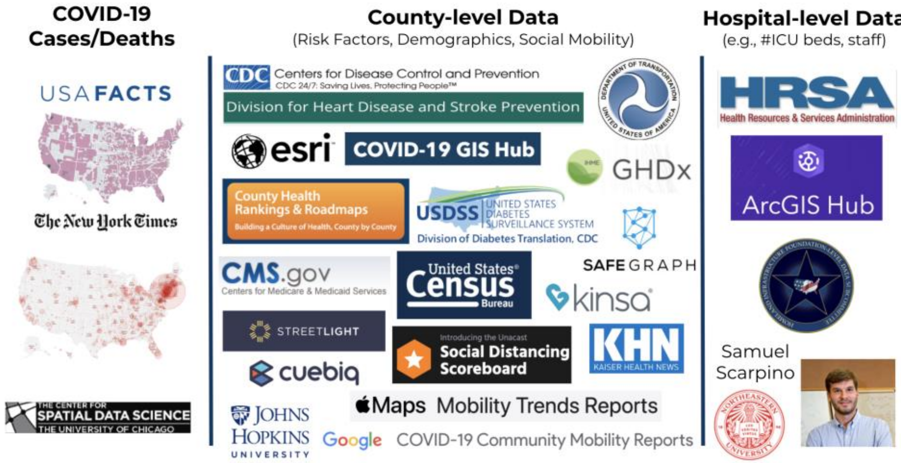


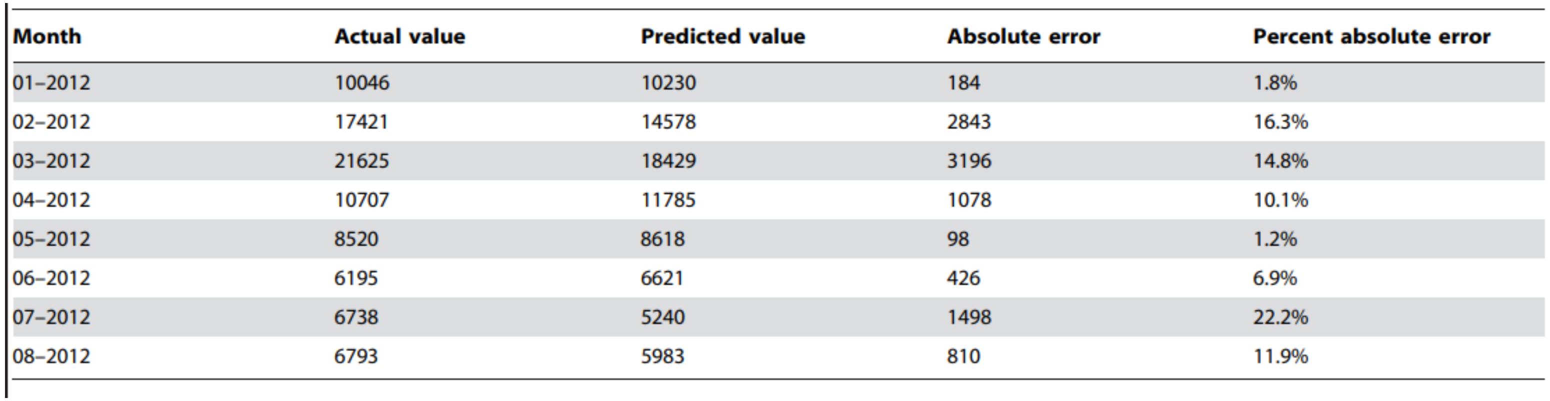


Sampling and Collecting Data

- High dimensional data from multiple sources on the Internet.
- Automatically update and parse data using web spiders.



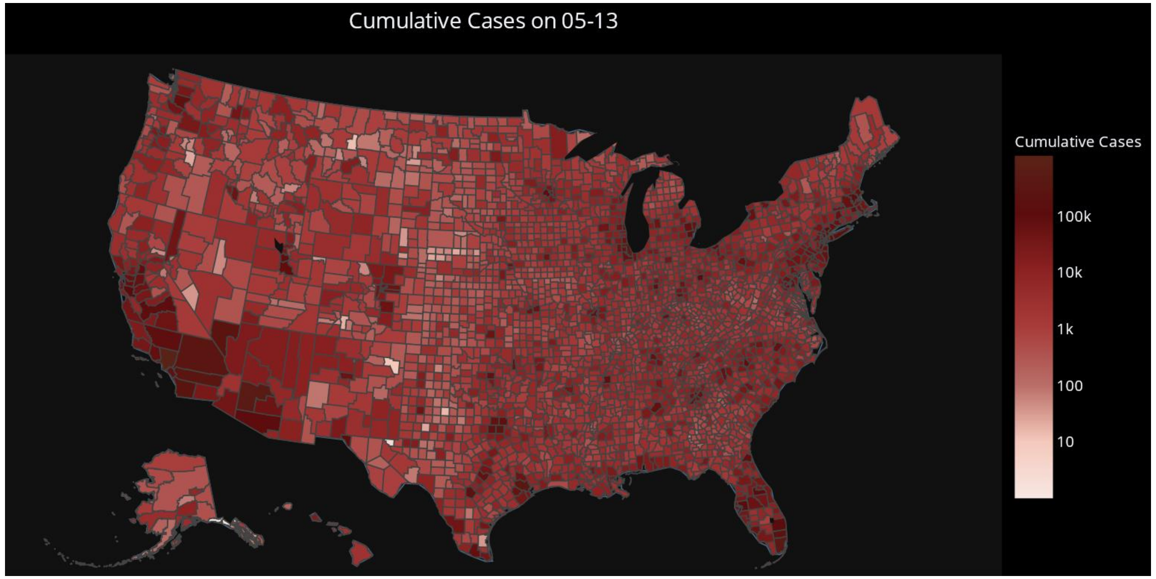

COVID-19 Cases/Deaths
USAFACTS
The New York Times
THE CENTER FOR SPATIAL DATA SCIENCE THE UNIVERSITY OF CHICAGO
County-level Data
(Risk Factors, Demographics, Social Mobility)
CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
Division for Heart Disease and Stroke Prevention
esri
COVID-19 GIS Hub
GHDx
County Health Rankings & Roadmaps
Building a Culture of Health, County by County
USDSS UNITED STATES DIABETES SURVEILLANCE SYSTEM
Division of Diabetes Translation, CDC
SAFEGRAPH
CMS.gov
Centers for Medicare & Medicaid Services
United States Census Bureau
kinsa
STREETLIGHT
cuebiq
Introducing the Unacast Social Distancing Scoreboard
KHN
KAISER HEALTH NEWS
JOHNS HOPKINS UNIVERSITY
Maps Mobility Trends Reports
Google COVID-19 Community Mobility Reports
Hospital-level Data
(e.g., #ICU beds, staff)
HRSA
Health Resources & Services Administration
ArcGIS Hub
Samuel Scarpino

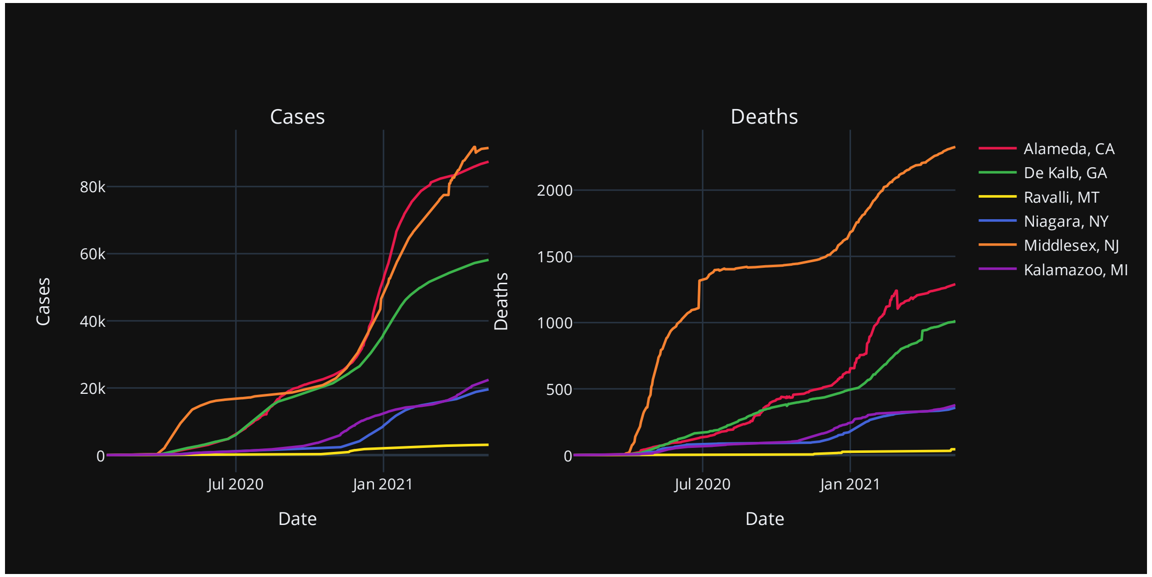


Month	Actual value	Predicted value	Absolute error	Percent absolute error
01–2012	10046	10230	184	1.8%
02–2012	17421	14578	2843	16.3%
03–2012	21625	18429	3196	14.8%
04–2012	10707	11785	1078	10.1%
05–2012	8520	8618	98	1.2%
06–2012	6195	6621	426	6.9%
07–2012	6738	5240	1498	22.2%
08–2012	6793	5983	810	11.9%

Data Visualization

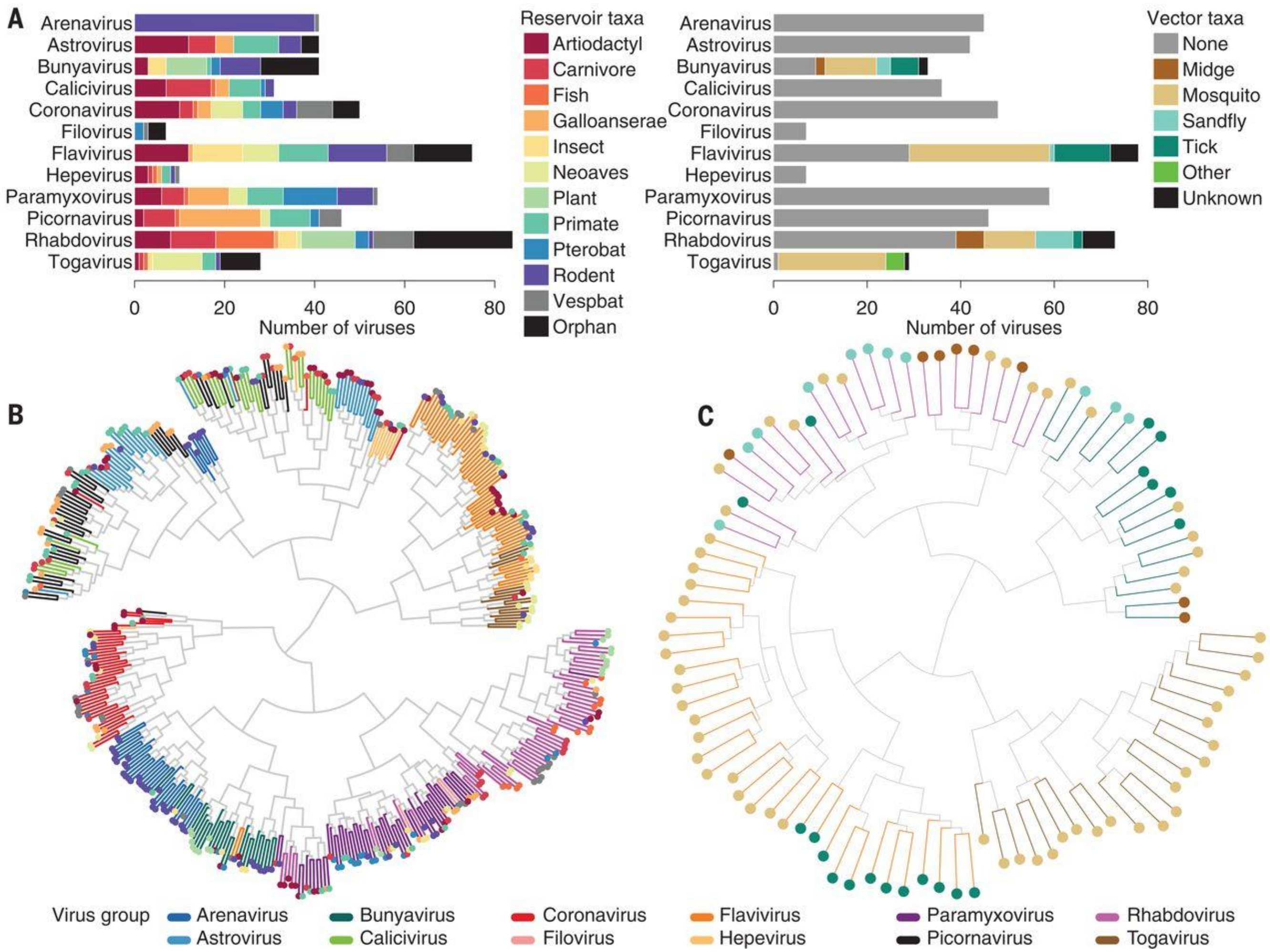
- □ Generate charts from the data automatically.
- □ Provide clear and overall images of pandemics.
- □ Help scientists' research and public education.



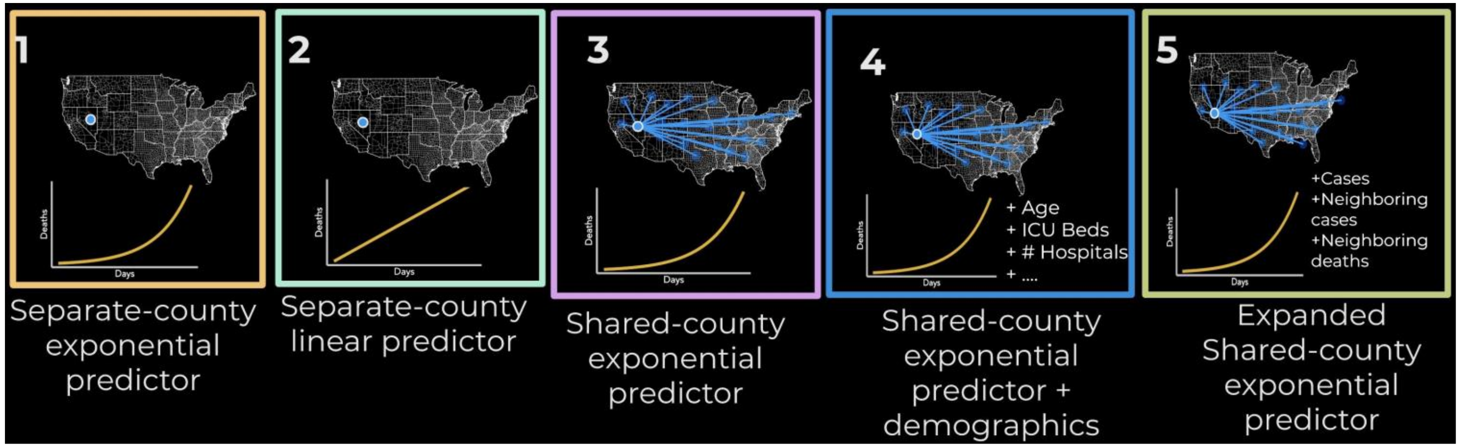


Data Mining and Analysis

- Learn the trend of infectious disease with epidemic models.
- Mine the data using Deep Learning Networks.
- Provide statistic data for other scientific researches.



Difficulties


1
Deaths
Days
Separate-county exponential predictor
2
Deaths
Days
Separate-county linear predictor
3
Deaths
Days
Shared-county exponential predictor
4
Deaths
Days
+ Age
+ ICU Beds
+ # Hospitals
+ ....
Shared-county exponential predictor + demographics
5
Deaths
Days
+Cases
+Neighboring cases
+Neighboring deaths
Expanded Shared-county exponential predictor

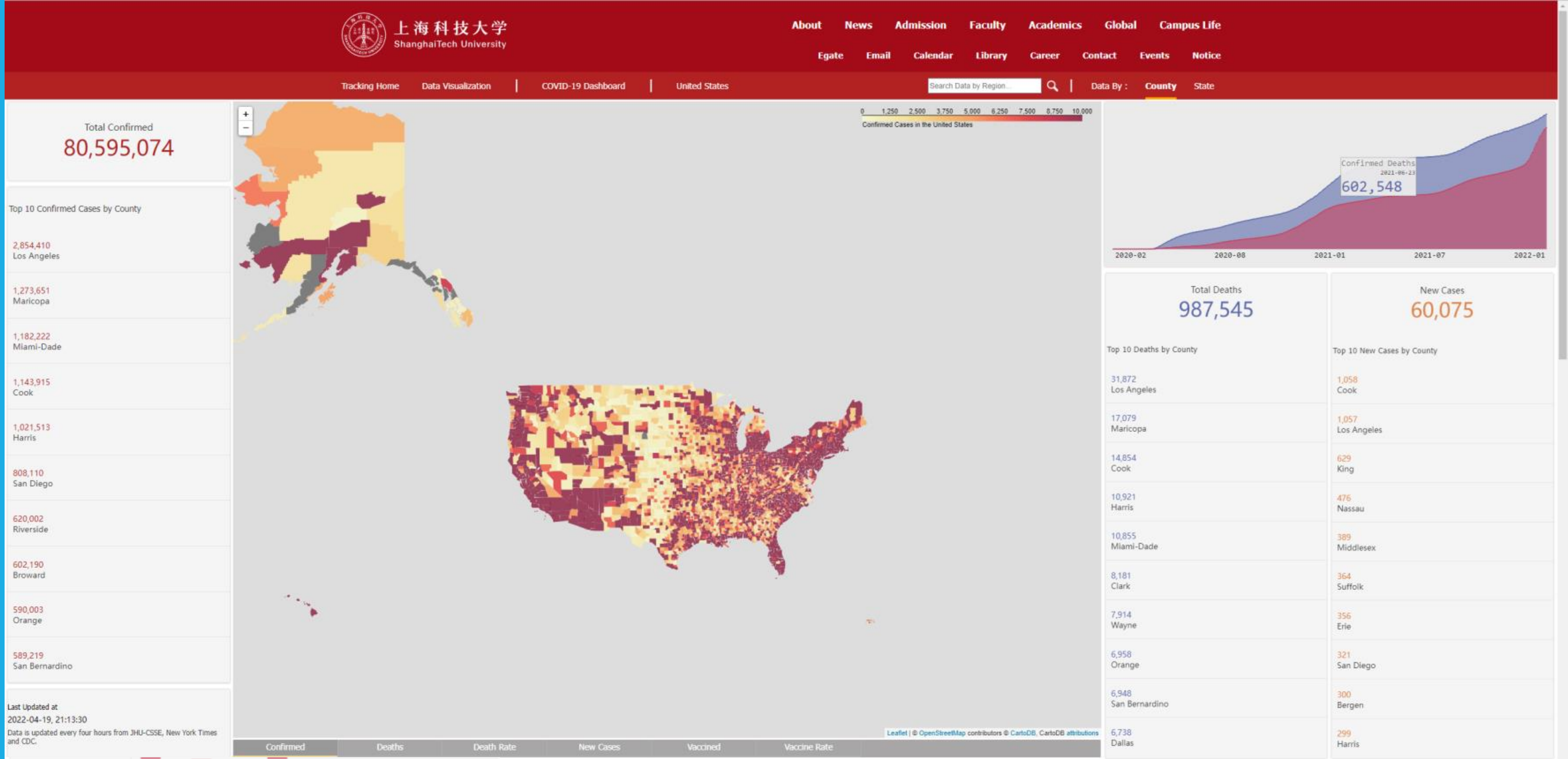
- Inconsistent data form.
- Incomplete or broken data pieces.
- Useless or disturbing data.
- Customize Deep Learning Models.
- Select data with different features.

Example

A prototype build by our team.

Website (in ShanghaiTech campu)
http://10.19.75.90:12345/index.html

Source code:
https://github.com/yanglinshu/covid



上海科技大学 ShanghaiTech University

About News Admission Faculty Academics Global Campus Life

Egate Email Calendar Library Career Contact Events Notice

Tracking Home | Data Visualization | COVID-19 Dashboard | United States | Search Data by Region... | Data By : County State

Total Confirmed
80,595,074

Top 10 Confirmed Cases by County

2,854,410	Los Angeles
1,273,651	Maricopa
1,182,222	Miami-Dade
1,143,915	Cook
1,021,513	Harris
808,110	San Diego
620,002	Riverside
602,190	Broward
590,003	Orange
589,219	San Bernardino

Last Updated at
2022-04-19, 21:13:30
Data is updated every four hours from JHU-CSSE, New York Times and CDC.

Total Deaths
987,545

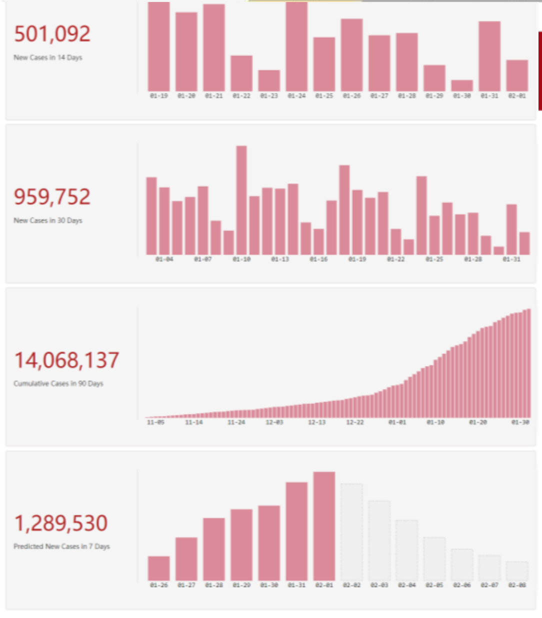
Top 10 Deaths by County

31,872	Los Angeles
17,079	Maricopa
14,854	Cook
10,921	Harris
10,855	Miami-Dade
8,181	Clark
7,914	Wayne
6,958	Orange
6,948	San Bernardino
6,738	Dallas

New Cases
60,075

Top 10 New Cases by County

1,058	Cook
1,057	Los Angeles
629	King
476	Nassau
389	Middlesex
364	Suffolk
356	Erie
321	San Diego
300	Bergen
299	Harris

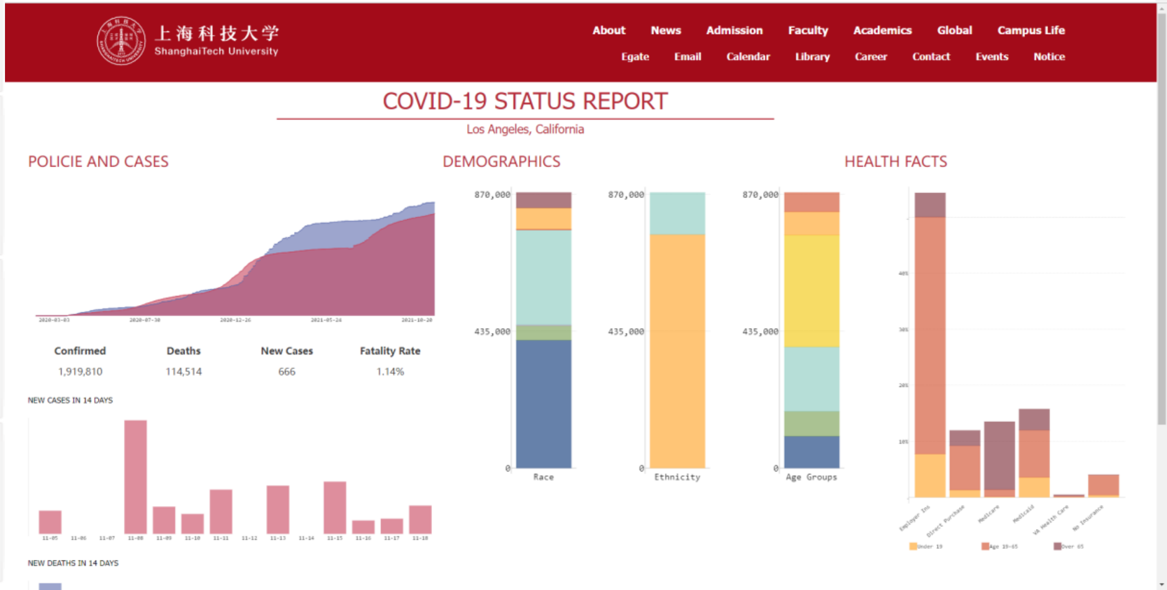


501,092
New Cases in 14 Days

959,752
New Cases in 30 Days

14,068,137
Cumulative Cases in 90 Days

1,289,530
Predicted New Cases in 7 Days



上海科技大学 ShanghaiTech University

About News Admission Faculty Academics Global Campus Life

Egate Email Calendar Library Career Contact Events Notice

COVID-19 STATUS REPORT

Los Angeles, California

POLICIE AND CASES

Confirmed	Deaths	New Cases	Fatality Rate
1,919,810	114,514	666	1.14%

NEW CASES IN 14 DAYS

NEW DEATHS IN 14 DAYS

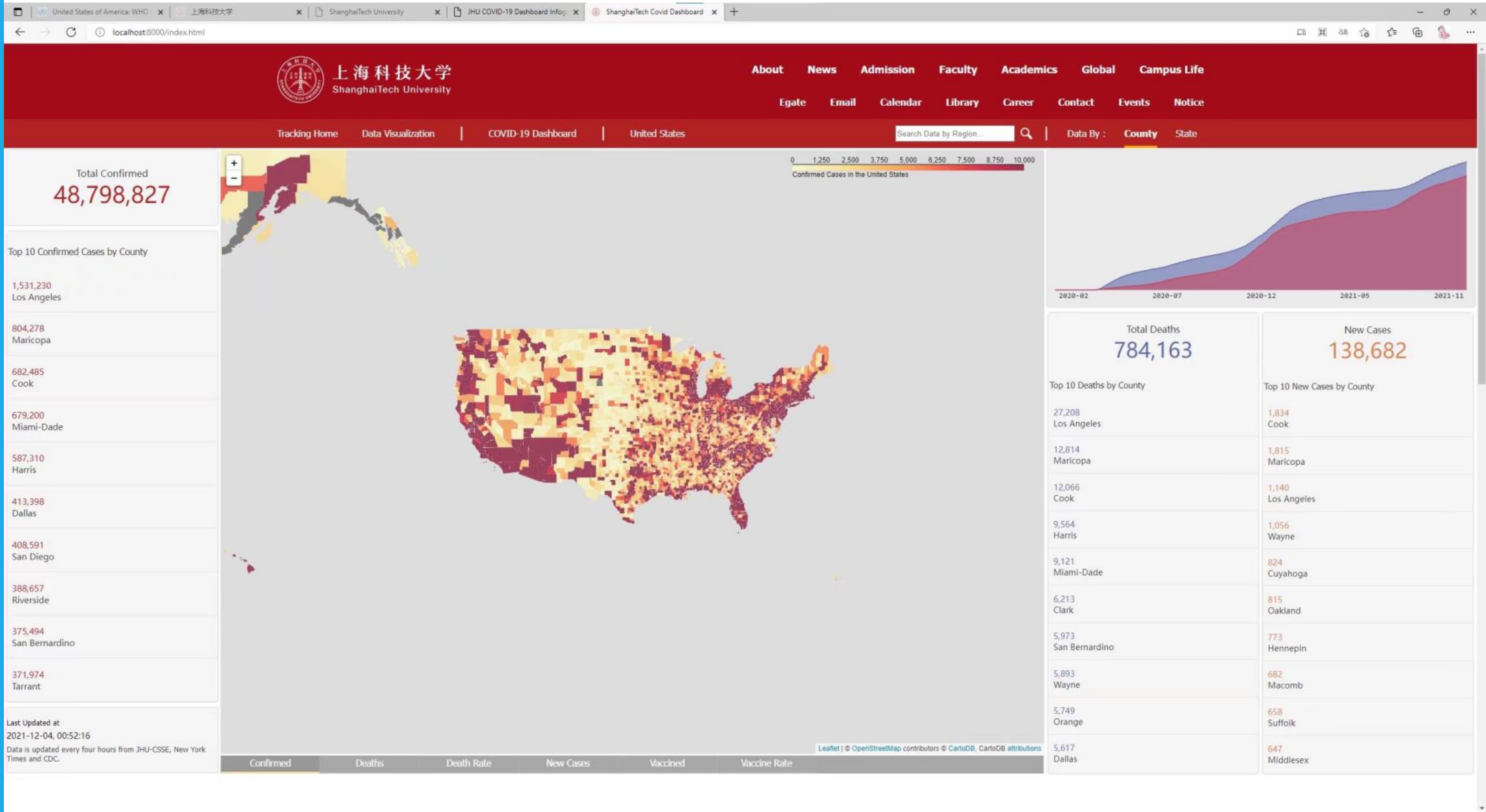
DEMOGRAPHICS

HEALTH FACTS

Example

- Our Data

collected from Johns Hopkins University, CDC, Census Bureau, New York Times, using a web spider based on python. Parsed and cleaned using Dataframe and Pandas.



Total Confirmed
48,798,827

Top 10 Confirmed Cases by County

1,531,230	Los Angeles
804,278	Maricopa
682,485	Cook
679,200	Miami-Dade
587,310	Harris
413,398	Dallas
408,591	San Diego
388,657	Riverside
375,494	San Bernardino
371,974	Tarrant

Last Updated at
2021-12-04, 00:52:16
Data is updated every four hours from JHU-CSSE, New York Times and CDC.

Total Deaths
784,163

Top 10 Deaths by County

27,208	Los Angeles
12,814	Maricopa
12,066	Cook
9,564	Harris
9,121	Miami-Dade
6,213	Clark
5,973	San Bernardino
5,893	Wayne
5,749	Orange
5,617	Dallas

New Cases
138,682

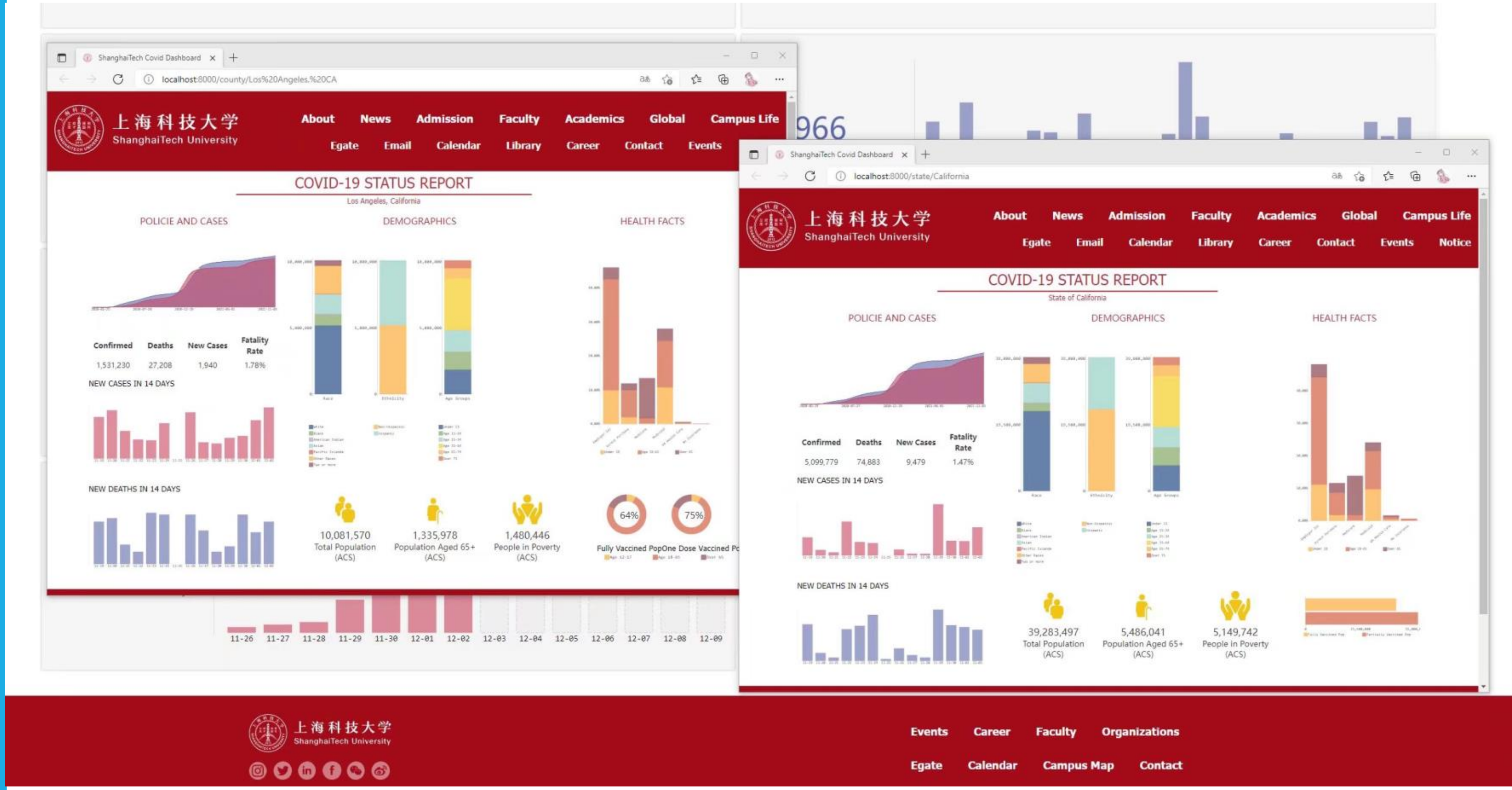
Top 10 New Cases by County

1,834	Cook
1,815	Maricopa
1,140	Los Angeles
1,056	Wayne
824	Cuyahoga
815	Oakland
773	Hennepin
682	Macomb
658	Suffolk
647	Middlesex

Example

- Visualization

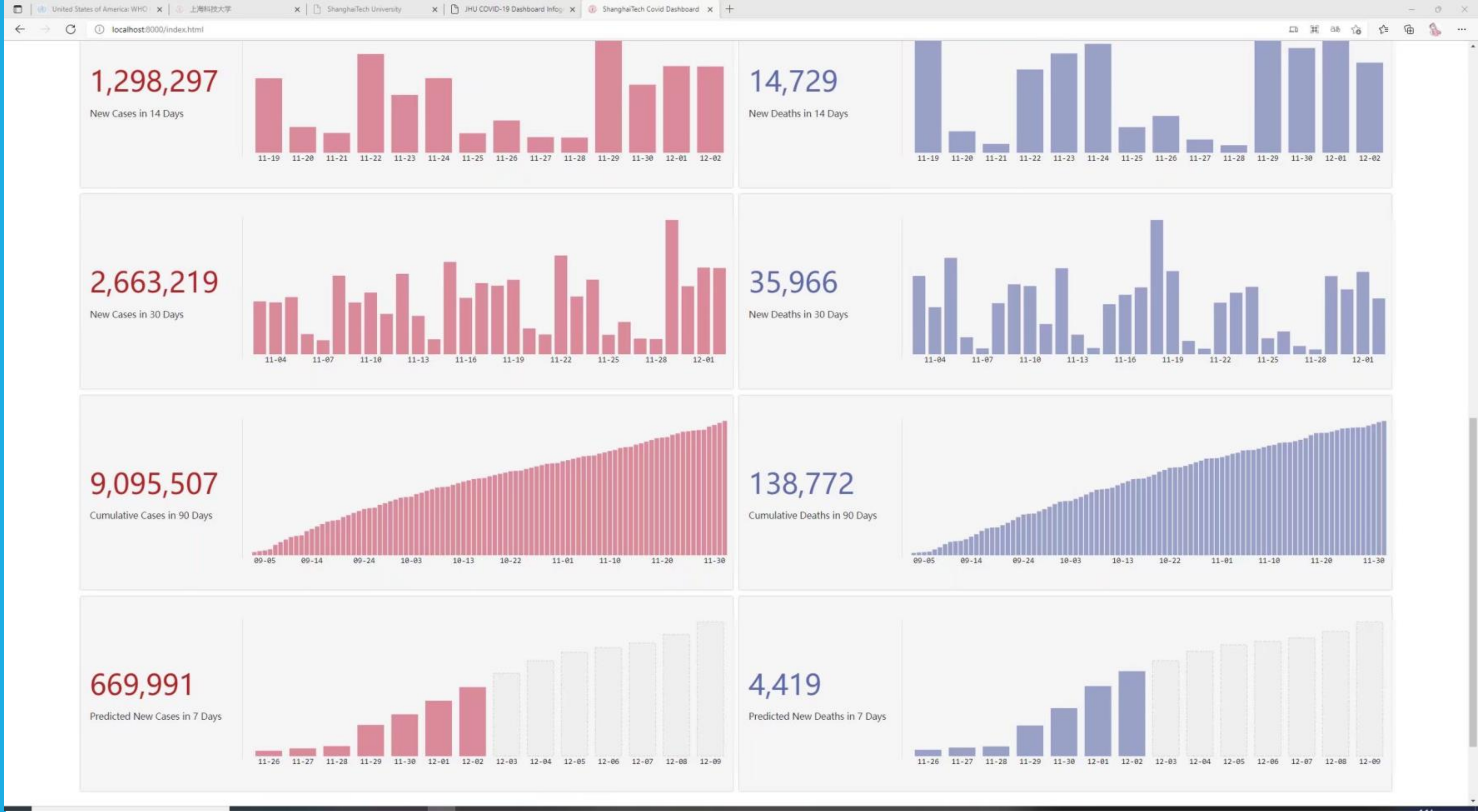
Generated using Pygal and Folium with geographical data from opendatasoft.com.



Example

- Data Mining

A simple moving average model using the data of the last 14 days.



1,298,297
New Cases in 14 Days

14,729
New Deaths in 14 Days

2,663,219
New Cases in 30 Days

35,966
New Deaths in 30 Days

9,095,507
Cumulative Cases in 90 Days

138,772
Cumulative Deaths in 90 Days

669,991
Predicted New Cases in 7 Days

4,419
Predicted New Deaths in 7 Days

Example

- JieYingWu/COVID-19_US_County-level_Summaries: Attempt to find correlation between a region's demographic/economic factors with its ability to manage disease spread (github.com)
- Yu-Group/covid19-severity-prediction: Extensive and accessible COVID-19 data + forecasting for counties and hospitals. 📈 (github.com)
- facebookresearch/CovidPrognosis: COVID deterioration prediction based on chest X-ray radiographs via MoCo-trained image representations (github.com)