Real-time Epidemic Surveillance Management for Supporting COVID-19 Response Workflows

---

A

Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Master of Science

by

Akhil Sai Peddireddy

May 2021

APPROVAL SHEET

This

Thesis

is submitted in partial fulfillment of the requirements
for the degree of

Master of Science

Author: Akhil Sai Peddireddy

This Thesis has been read and approved by the examing committee:

Advisor: Madhav Marathe

Advisor: Srinivasan Venkatramanan

Committee Member: Anil Vullikanti

Committee Member: Henning Mortveit

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

May 2021

Abstract

The COVID-19 outbreak caused by SARS-CoV-2 has disrupted the lives of people globally. It has had a huge impact on health, economies, and society in general, undoubtedly making it the pandemic of the century. As of April 1, 2021, the cumulative number of confirmed COVID-19 cases exceeded 130 million worldwide, with almost 2.85 million deaths. Global Surveillance is one of the important tools of epidemiological response to allow the public to be informed of the pandemic and to provide insights to the policymakers.

With the experience from assessing the US influenza surveillance prior to COVID-19, we identified 6 key metrics, called 6Cs, which we propose as a standard for the design and evaluation of real-time epidemic science data portals. We describe our work building the COVID-19 Surveillance Dashboard, visited by over a million users, its underlying architecture, the multitude of the data present and the data pipelines that conform to the 6Cs standard. As fluid as the pandemic remains, the biggest challenge is to adapt to the real world changes and continuously maintain the dashboard by updating the data sources, be consistent with changes in reporting and adding new data as it becomes available. We also demonstrate a number of utilities of surveillance data that we built such as Q&A based Analytics, PatchViz tool to seed, run and visualize mechanistic simulations and Ensemble Kalman Filter for high-resolution forecasting of COVID-19 cases as part of a Bayesian ensemble that is integrated each week into CDC initiated COVID-19 Forecast Hub.

We hope this work presents a framework for maintaining epidemic surveillance in real-time, discusses the use-cases and lessons learnt, and equips us with the tools & insights needed in the unfortunate event of a future pandemic.

Acknowledgments

I want to thank my advisors Srini Venkatramanan and Madhav Marathe for allowing me to work on some of the exciting projects that have a real-world impact during the tough times of a pandemic. They motivated & inspired me, taught many things and were always available to help even in their busy schedule. Thanks to my thesis committee members, Anil Vullikanti and Henning Mortveit, for dedicating their valuable time to give feedback, helping shape this thesis better, and ensuring the procedures of defense go smoothly.

I would like to extend my sincere thanks to all the members of the NSSAC division and the Biocomplexity Institute. I am proud of being a part of this great team of experts across different fields coming together as 'Team Science' to solve some of the complex problems of the society. Special thanks to Dawen Xie for the valuable mentoring, guidance and all the assistance in the Surveillance Dashboard work and Pramod from Persistent Systems for developing the UI. Sincere thanks to my mentors and collaborators in the forecasting and other projects - Aniruddha, Mandy, Bryan, Przemek, Lijing and Ben.

I would also like to thank the Computer Science department and the professors for sharing their knowledge. I am fortunate to be a part of the University of Virginia, which also taught me many valuable life skills besides the academics, and for instilling morals and values that helped me understand the importance of empathy, equality, diversity, sense of community, hard work and dedication. Thanks to my extended family who were available when in need. Finally, thanks to my parents Kishore Peddireddy and Vijaya Peddireddy, for supporting me and taking care of me throughout the time. I would not be able to do any of this without their support and motivation.

Contents

List of Figures

List of Tables

Operationalizing Epidemic Data Management

1 Introduction

With the uncertainty surrounding the COVID-19 pandemic and its impact, there has been a great need for surveillance of the pandemic in order to make informed and fact-based decisions. The tracking of the pandemic has been of prime importance for policymakers, public health officials, and academic researchers attempting to understand and respond to this public health crisis, along with every lay person concerned about how the pandemic will affect their daily life. The primary goal of this thesis is to describe our work that allows the policymakers to gain insights and the general public be informed of the pandemic by using surveillance and other epidemic workflows. When the past and the present is known from the surveillance, and the future can be forecasted based on this surveillance, it helps to clearly understand and assess the situation of the pandemic in the region of interest. It allows to better manage it both for the public at an individual level and for the policymakers to form the right policies that helps to alleviate the resulting public health and economic consequences. This thesis also aims to offer a detailed look and present the case study of a

Some of the material is adapted from:

A. S. Peddireddy et al., "From 5Vs to 6Cs: Operationalizing Epidemic Data Management with COVID-19 Surveillance," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 1380-1387, doi: 10.1109/BigData50022.2020.9378435.

real-time epidemiological response that happened during one of the worst pandemics in the recent times. We hope that the dashboards, data pipelines, models and other tools we built along with the insights and results we obtained will help be pandemic-prepared and respond effectively in the event of a future pandemic.

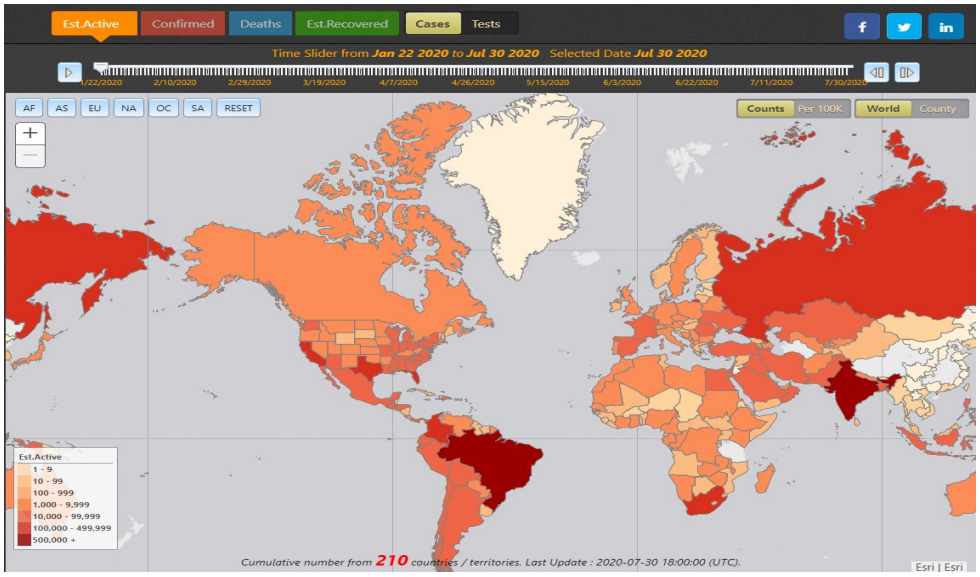
Many dashboards have been developed to help the public better understand the current status, each focused on a different aspect of the pandemic. However, there is no gold standard set for epidemic data management and visualization. There is a very clear need for a user-centric, one-stop solution backed by a reliable source of data and coupled with rich visualizations and an easy-to-use interface that is accessible to both the public and researchers. These issues, along with the prior experience dealing with influenza surveillance, led us to identify 6 metrics that might serve to define a standard for epidemic surveillance data management that we call *the 6Cs standard*. The 6Cs standard proposes that epidemic surveillance data should be *Consistent, Correct, Current, Comprehensive, Curated, and Computer-readable*. With this standard in mind, we created a COVID-19 surveillance dashboard which offers data exploration and visualization features designed to assist researchers, but which even a normal user, unfamiliar with the technical details, can understand. While we sincerely appreciate the contributions and efforts put forth by teams working on similar dashboards, we have observed that these sources are either limited in scope or miss some of the characteristics of the 6Cs standard. Our goal is to provide an insightful view into COVID-19 incidence data through spatio-temporal surveillance visualizations conforming to the 6Cs standard.

The Biocomplexity Institute & Initiative's COVID-19 Surveillance Dashboard, which was released on February 3, 2020, is available at `https://nssac.bii.virginia.edu/covid-19/dashboard/`. It is a single-page, interactive and responsive web application that is dynamically updated; it allows end users to view and explore COVID-19 case counts at both temporal and spatial resolutions with just a few mouse clicks. To the best of our knowledge, our dashboard is one of very few that presents historical data in three visualization formats: choropleth map (Fig. 1a), time series charts (Fig. 5a), and a data

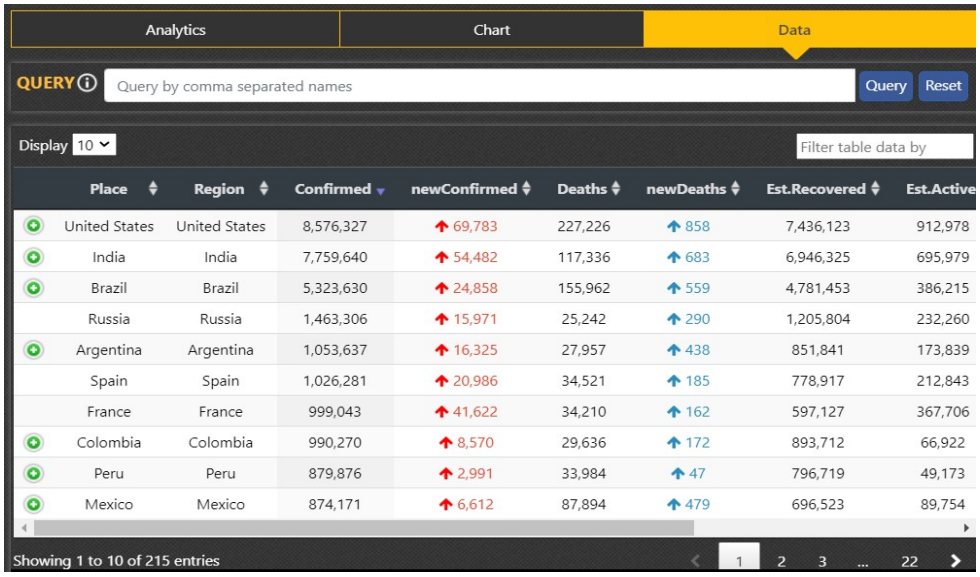
table (Fig. 1b), all of which are interactive. Our charts include Cumulative and Incident Epicurves for all regions, and our unique use of a movie-style time slider makes it easy to simultaneously explore both temporal and spatial evolution of the pandemic. In its current form, the dashboard supports data rendering at the country level for 220 countries in the world; at the state and province level for 20 countries; and county-level statistics for the United States of America (USA). This data can be easily accessed either through the query tool, by filtering the data, or by clicking on the zoomable map. As of April 1, 2021, more than 1.2 million users from over 220 countries have used our dashboard, and more than 60 million requests were processed on the main feature layer hosted on ArcGIS Online.

2 Background: Influenza Surveillance

Prior to COVID-19, we collected and organized influenza surveillance data for about 8 states in the USA from their respective publicly available health department webpages. These states include Alabama, Arizona, Delaware, Idaho, North Dakota, Pennsylvania, Tennessee and Virginia. This was challenging due to multiple reasons such as presence of different strains and subtypes, the seasonality of the disease and hence data availability, importance of age groups in surveillance, huge size of the data spanning multiple years, inconsistent reporting format across different states, retrospective updates of the data, missing data, and irregular update times and intervals. Apart from these data management challenges, there were also other challenges in data collection such as different types of data sources - PDF reports, Tableau dashboards, Choropleth maps, Webpages, APIs etc, lack of computer readability & easy access, and frequently varying formats over the years for the same state. These issues showed us the need for a robust data management framework that requires certain characteristics to be present, for a hassle free access and use of the surveillance data. This guided us to build a database for our collected data by formulating and applying these principles, thereby leading us to our first steps towards the 6Cs. To handle the above mentioned challenges and ensure consistent format, we used the following



(a)



Place	Region	Confirmed	newConfirmed	Deaths	newDeaths	Est.Recovered	Est.Active
United States	United States	8,576,327	69,783	227,226	858	7,436,123	912,978
India	India	7,759,640	54,482	117,336	683	6,946,325	695,979
Brazil	Brazil	5,323,630	24,858	155,962	559	4,781,453	386,215
Russia	Russia	1,463,306	15,971	25,242	290	1,205,804	232,260
Argentina	Argentina	1,053,637	16,325	27,957	438	851,841	173,839
Spain	Spain	1,026,281	20,986	34,521	185	778,917	212,843
France	France	999,043	41,622	34,210	162	597,127	367,706
Colombia	Colombia	990,270	8,570	29,636	172	893,712	66,922
Peru	Peru	879,876	2,991	33,984	47	796,719	49,173
Mexico	Mexico	874,171	6,612	87,894	479	696,523	89,754

(b)

Figure 1: Dashboard screenshots. (a) Map panel. Choropleth map of the world, rendered with the estimated active count, but with the option of switching to other layers: state/province layer, county layer, or different attributes for rendering. (b) Information panel - Data Table. This includes an interactive data table with data of different attributes like Region, Confirmed, Deaths, Est.Recovered, Last Update, etc.

fields in our centralized data repository to cover all the data collected from different states and time periods - Year, Epi Week, State, Location Type, Location, Surveillance, Sub Type, Age Group & Value. This experience came in handy when the COVID-19 pandemic emerged and similar issues were observed in then existing sources. We decided to take up this challenge for COVID-19 surveillance and build a system that conforms to the 6Cs.

3 The Importance of the 6Cs Standard

3.1 Consistent

Consistency can be viewed from two perspectives - consistency in the format of the data, and consistency in the content of historical data. One of the major goals of collecting and managing epidemic data is to support informed health and public policy decisions. This is applicable in various contexts, including policymakers and the academic researchers who support them, individual businesses, and the general public. This decision-making process often depends on projections or forecasts of the pandemic spread produced by epidemic modeling simulations, which, in turn, depend on surveillance data. Modeling applications expect that the data format will be consistent over time; frequent format changes would necessitate frequent revisions to model implementations, causing untimely delays in forecast production. This uncertainty may lead modelers to seek more consistent data sources. Policymakers and public health officials also depend on websites and dashboards to guide their decisions. Many of these online tools, including ours, depend on open data sources. If the underlying format of these sources changes, it can delay dashboard updates, which may, in turn, prevent the policymakers from having the most current information when they need it.

Another perspective is consistency of the historical data. Once published, historical data should be updated as little as possible, except in the case when a previous entry is discovered to be invalid. To reduce the risk of data contamination, data should be collected from proven reliable sources; collation and correction of the data needs to be performed via

a consistent and predictable process; and frequent validation must be a critical part of the curation process. In the event that a correction is made, the downstream impact on forecasts and projections could be high, so it is necessary to provide a record of data updates that is accessible to all consumers of the data. This also leads to our next C, which is Correctness.

3.2 Correct

Epidemic data is sensitive and inaccuracies can have a catastrophic impact on public health. This raises the importance of data correctness. Large scale data curation is prone to errors that occur for a variety of reasons, including incorrect data entry, improper access to the data, or errors in processes performing calculations and data wrangling. Steps must be taken to reduce data error to the maximum extent possible. This can be achieved through proper validation and data checks; for example, the cumulative case counts should not decrease unless there have been upstream revisions of historical data by the original reporting source, or the total count of a region should be equal to the summation of its subregions, etc. Any uncertainty that cannot be resolved should be explicitly accounted for and documented.

3.3 Current

A pandemic such as COVID-19 evolves quickly. We have all seen situations where a region with no prevalence of infections suddenly emerges as a hotspot with a short case doubling time. Hence, the frequency of data updates is vital, as stale data does not capture the current status, and is a poor guide for making decisions under rapidly changing conditions. This emphasizes the need for timely updates, which, in turn, requires automated data collection and maintenance of historical snapshots of the data. It also amplifies the need for storing the data in a temporal representation with a clear indication of when data updates have occurred.

3.4 Comprehensive

A dashboard tracking an epidemic should be comprehensive in providing detailed visual analysis with charts, geospatial mapping, and time series visualizations, as well as with summary statistics. Apart from that, several other metrics can be derived from the core data of cases and deaths to help users understand the present situation in a clear manner. For instance, active cases is an important metric in assessing the current status of the disease, and the number of infections normalized by the population demonstrates the density of the spread. Other data, like laboratory testing, hospitalizations, and vaccines, have a direct influence on interpreting the pandemic and add additional context. Having a comprehensive, complete picture of the pandemic can help researchers understand which factors would curtail the spread of the disease.

3.5 Curated

The epidemic data should be curated from diverse sources in order to cater to the large and disparate needs of the population. First, the data should be available for as many regions and subregions as possible to get a more global picture of the disease spread. This not only makes the dashboard complete and serves the majority of the users, but also helps improve decision-making at the local level. The kind of interventions required for a nation with multiple hotspots and a nation with a single adversely affected hotspot are quite different, which can only be captured if data is available at multiple levels of spatial resolution.

3.6 Computer-readable

To cater to downstream tasks like modeling, analysis or visualization, data should be Computer-readable, meaning it should be easily accessible in the form of CSV files, databases, or through an API endpoint. Data provided in a textual format, such as a report or an article, is good for human consumption, but requires a lot of preprocessing and manual work to prepare it for other computational tasks. Similarly, data should also have standard

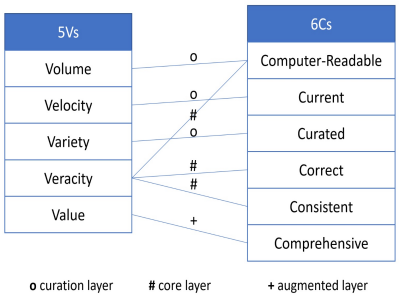


Figure 2: Relationship between the 6Cs and the 5Vs of Big Data, and the data pipeline layers used to achieve them in our surveillance dashboard.

geospatial mapping and naming conventions in order to easily identify or differentiate between the regions and subregions. In addition, hierarchical organization of the spatial and temporal components in the data also facilitates easy and efficient retrieval.

3.7 The 5Vs of Big Data

Volume, Velocity, Variety, Veracity, and Value are popularly known as the 5Vs of Big Data [1]. Although the 6Cs of epidemic data management bear some resemblance to the 5Vs, the two standards actually complement each other quite a bit (See Fig. 2)

Although the size of epidemic data, specifically for an emerging infectious disease, is not as large as that in a typical “Big Data" setting, epidemic data still has a lot of spatial and temporal components. As a pandemic progresses, the size of the data increases rapidly, and when additional data types like mobility, tests, and hospitalizations are added to the set, scalability becomes an important factor. Computer-readability plays an important role in promoting efficient handling of the high Volume of multidimensional data, while still supporting flexibility and easy accessibility. The Velocity of such data is matched with the Current attribute, specifically with the temporal component where the data flows continuously from multiple sources into the application for real-time updates. Several program optimization techniques, along with a minimal amount of human intervention, can help to handle the Velocity of data. Epidemic data has a wide Variety of potential data

sources and formats, ranging from a structured form like CSV, to a semi-structured form like Rest API endpoints and dashboards, to, finally, unstructured forms like webpages or reports. This is ideally what Curated in the 6Cs aims to handle, specifically for the spatial component. Veracity refers to reconciliation of inconsistencies and uncertainty in the data. Collected epidemic data is typically unstructured and messy, so it has to be cleaned and validated to attain Consistency and Correctness, with the right amount of data because the data has no Value by itself unless it is refined down to include only the useful information. The Comprehensive characteristic helps useful information to be conveyed through rich visualizations and analysis.

4 Related Work

In this section, we present an overview of some well-known COVID-19 dashboards and efforts from various groups for epidemic data management and visualization.

When we started our dashboard in late January, the Center for Systems Science and Engineering at Johns Hopkins University (JHU CSSE) [2] was one of the few organizations that gathered and shared COVID-19 data through a dashboard. As critical and influential as their data collection and curation efforts were, especially at the beginning of the pandemic, their data format and sharing platform has changed frequently, which has made it difficult for downstream users to adapt. This drove home the need for data Consistency as one of the most important goals to pursue. Another limitation of the JHU CSSE dashboard is the lack of temporal data and region-level visualizations, along with the inability to query or search for a specific region. The data for each category of cumulative data (cases, deaths, recovered etc.) is organized into separate tabs and panels, which makes it difficult to assess the full picture for a particular region.

1Point3Acres' Global COVID-19 Tracker & Interactive Charts [3] dashboard initially focused on providing near real-time case information for North America. We were one of the early adopters of their data, and have used it for the USA since early March until

December 2020. 1Point3Acres has added rich visualizations and expanded data coverage to more countries over time, but they still don't cover all of the countries which have cases. Although 1Point3Acres provides spatial rendering for USA and Canada, this component is missing for other countries. The temporal data and visualization is also unavailable at sub-national levels, i.e. for states, provinces, and counties. Because of these shortcomings, this dashboard does not meet the 6Cs standards for Comprehensive and Curated.

Worldometer's COVID-19 website [4] collects national-level data from every country and makes it available in a data table for three recent days. For the USA, they provide detailed case counts down to the county level. They also display charts of historical data, which allow users to understand the current status of the pandemic, as well as some aspects of how it evolved over time. However, their data presentation is largely text-based (with the exception of the historical charts), and does not provide spatial visualizations, like maps, that would allow users to visualize differences between contiguous regions. For these reasons, Worldometer falls short of our standard for Comprehensive. Similar shortcomings were observed in Our World in Data [5] and USAFacts [6].

Public health organizations like the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), and the European Centre for Disease Prevention and Control (ECDC) [7–9] provide data that is Consistent, Correct, Current, and Computer-readable, but not completely Curated because they support data at only a single spatial resolution. WHO and ECDC provide the data for all the countries of the world, but not for states, provinces, or counties, whereas CDC provides data for USA states and counties, but not for other countries and subregions.

A number of dashboards have been developed that are specific to a region or a certain area of study. For example, health departments of many countries, states and counties have their own dashboards to track the local pandemic situation. Wissel et al. [10] used an R Shiny app for surveillance of USA cities using data from JHU CSSE, The New York Times [11], and the COVID Tracking Project. Barone et al. [12] developed a statistical surveillance dashboard based on their analysis of the ratio between cases and the days

since the first case in the countries to determine the average speed of its epidemic motion, analogous to concepts in physics. Hohl et al. [13] built an R Shiny app based on their study of space-time scan statistics to detect daily clusters at the county level. On the topic of data curation, there are several efforts that collate data types other than surveillance data, like BeOutbreakPrepared [14] which provides individual-level epidemiological data, also known as line lists.

In the development of our dashboard, the Single Page Application (SPA) and movie-style time slider for temporal visualization of the surveillance data was inspired by similar functionality that was available in the now-defunct EpiViewer application [15]. The division of the front end into several panels is also similar to the presentation provided by EpiViewer, but we extended that concept with the addition of a geographical component and near real-time data from the current outbreak.

5 Data and Sources

This section describes the data elements our dashboard depends upon. Although our full dataset is not currently downloadable from our dashboard, we do maintain a Consistent data structure across all regions. We rely on reputable, openly available data sources to populate our dashboard; those sources are also described in detail in this section.

5.1 Data Fields

Core Surveillance data: Sub-Region (if any), Region, Confirmed Cases, Deaths, Reported Recovered. (Cumulative numbers are provided for the three numerical measures).

COVID Testing Data: Sub-Region (if any), Region, Positive Tests, Negative Tests, Total Tests, Positivity Rate (%), Data Quality Grade.

COVID Vaccine Data: Sub-Region (if any), Region, Total Doses Administered, People Vaccinated, People Fully Vaccinated, Total Distributed, Vaccine Name.

Augmented Surveillance Data:

- Active cases : Confirmed - Deaths - Reported Recovered;
- ID/FIPS : ISO3 for countries, FIPS for USA counties and ID for states / provinces (hereby referred to as admin1 regions) by mapping these regions with an ISO lookup;
- Coordinates : Latitude and Longitude for GIS is obtained from ID/FIPS;
- Last Update : The UTC time when the data was last fetched and updated;
- Estimated Recovered : Estimate of Recovered case count calculated based on the time series of confirmed cases and deaths (more on our algorithm below);
- Estimated Active : Confirmed Cases - Deaths - Estimated Recovered;
- New Cases, New Deaths, New Recovered, New Estimated Recovered, New Estimated Active : Increase in counts from previous day's cumulative numbers (Incidence Data);
- Per 100K counts : Population-normalized numbers for all of the above relevant data fields.

Estimating Recovered Counts: This feature is unique to our dashboard. A significant number of countries or states do not report the number of people who have recovered from COVID-19, and those that do report these numbers are not always up-to-date. The United Kingdom (UK), Netherlands, and Sweden are some of the countries which don't report recovered statistics, and this data is not available for the majority of the states, provinces and counties performing independent reporting. Without knowing the number of recoveries, it is very difficult to calculate the number of Active cases, which is arguably a more important metric to track than Confirmed cases; for example, many local governments use active case counts to plan their reopening strategies. Furthermore, inaccurate recovery counts will lead to inaccurate active case counts. This raises the need for a well-defined method for calculating the number of recovered cases, and, by extension, active cases, which is

consistent across all regions. Such an approach will minimize differences in reporting, hence allowing for fair comparison across regions.

To this end, we developed an algorithm for calculating the number of recovered cases. A joint study conducted by WHO and China [16] concludes that the median time from onset of COVID-19 to clinical recovery for patients with mild cases is approximately 2 weeks, while the median time is 3 to 6 weeks for patients with more severe or critical disease symptoms. A cohort study by Wu Z, McGoogan JM [17] shows that 81% of cases are mild to moderate, 14% are severe, and 5% are critical. This study is referenced by the official CDC interim clinical guidance [18]. Illinois Department of Public Health follows a similar estimate for their calculation of recovered cases [19]. Based on these studies, we calculate Estimated Recovered as follows:

$$(Est.Rec)_T = [0.81 * (Conf.)_{T-14} + 0.14 * (Conf.)_{T-28} + 0.05 * (Conf.)_{T-42}] - (Deaths)_T \quad (1)$$

where T represents the day in the time series for which Estimated Recovered is calculated. While this is a fairly safe estimate for the number of recovered cases, it is possible that actual recovery counts will vary depending on the region or subregion. In cases where the reported recovery number is higher than our safe estimate, we set Estimated Recovered equal to the reported value. If the CDC or WHO guidelines regarding the recovery estimates are updated, we will adjust our formulas accordingly.

5.2 Data Sources

Epidemic surveillance data: We use several data sources for collating our epidemic surveillance dataset. Table 1 summarizes which data sources we are currently using, how often we poll those sites, and the collection methods.

Demographic data: We use different sources for population data, including Worldometer for country-level population estimates [25], World Population Review for USA state-level population estimates [26], WorldAtlas for China province-level populations [27], Wikipedia

Table 1: Collection Details of Epidemic Surveillance Data.

	Source	Frequency	Mode of collection
USA	New York Times	Multiple times a day	Github
National data (except USA)	Wikipedia [20] and WHO [7]	Multiple times a day	Scraping; CSV file
USA - NYC	USAFacts [6]	Daily	CSV file
Sub-national data (16 countries)	Wikipedia	Multiple times a day	Scraping
India	Covid19India [21]	Multiple times a day	Github
Canada	Gov. of Canada [22]	Multiple times a day	CSV file
Greece	Min. of health [23]	Multiple times a day	Scraping
USA Tests	The COVID Tracking Project [24]	Daily	CSV file
Vaccine	Our World in Data	Multiple times a day	Github

for other state/province-level population counts, and Esri Demographics for USA county-level population estimates.

GIS data: Polygons for the USA counties are provided by Esri Demographics. Source data for other polygons, e.g., all countries and state/province-level administrative regions, are provided by ADCi [28]. We host these polygons as feature layers on ArcGIS Online [29].

6 Back-end Architecture

Fig. 3a shows the overall architecture of our dashboard. We present details on the back end in this section, and the User Interface (UI) in Section 7.

As described in Section 5, the data available on our dashboard is multidimensional. At a high level, it includes the surveillance data, demographic data, and GIS data. A

design decision made at the beginning of this project was to separate storage of surveillance data from other demographic and GIS data. In particular, the surveillance data is stored locally on our web server and is organized in a spatio-temporal hierarchy, while ArcGIS Online [29] is used to store and access demographic and GIS data for the administrative regions.

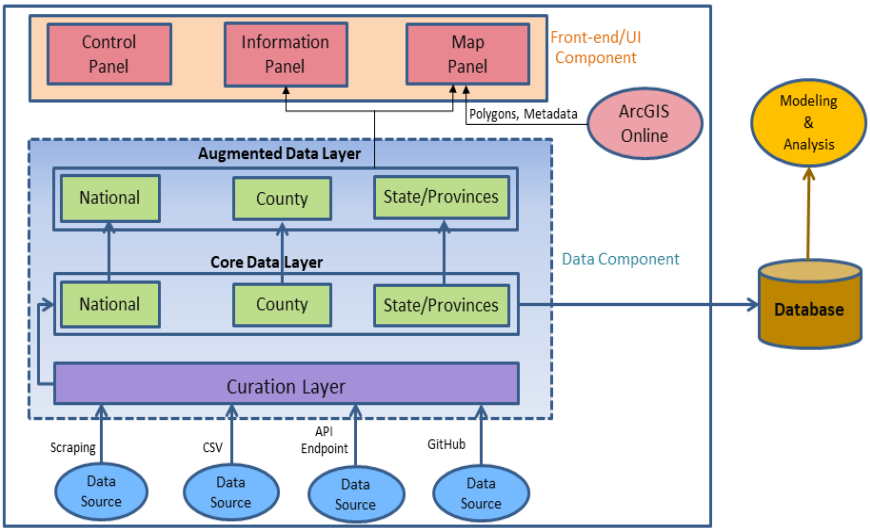
6.1 ArcGIS Online

ArcGIS Online serves as the GIS server that hosts the feature layers needed by our dashboard. The feature layer is a way to organize collections of separate geographic objects, e.g., administrative regions, buildings, roads, etc., available to the web using ArcGIS. In our dashboard, for easy accessibility, we are using three feature layers corresponding to each spatial level, i.e., one for the world map shown by default on the dashboard, one for states and provinces, and one for USA counties; the feature layers include information such as unique identifier, name and population. Our application fetches data from both the web server and ArcGIS Online, performs a join across datasets on the fly, and uses the joined data for the final visualizations.

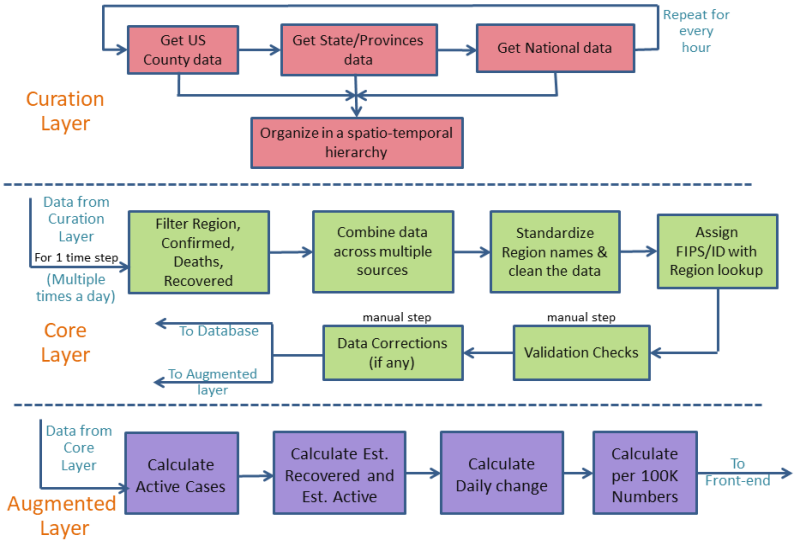
Separating constantly evolving surveillance data and relatively static demographic and GIS data in a GIS application is an efficient approach and allows us to support a large amount of data. There are several other advantages: (*i*) by keeping the map services on ArcGIS Online relatively static, we avoid the need to update feature layers, minimizing service outages. (*ii*) The map data only needs to be fetched once, reducing the data transfer between the application and end user to a small amount of data for each new request. This reduces the load on ArcGIS Online, and makes our application scalable for supporting simultaneous requests.

6.2 Surveillance Data Pipeline

The surveillance data is collated, processed, and augmented with a robust data pipeline which helps to create an all-in-one comprehensive data hub for all temporal and spatial



(a)



(b)

Figure 3: (a) Dashboard architecture: The UI consists of a control panel, information panel and map panel. The data component is a pipeline with three layers: curation layer, core data layer and augmented data layer. (b) Micro services used in the surveillance data pipeline. The series of steps that take place in the data component's layers, from the collection of data to its organization, integration, validation and augmentation in order to facilitate its usage in modeling and the front-end UI.

resolutions. We organized the entire data pipeline in a three-layered approach to effectively achieve the 6Cs as described in the previous sections (See Fig. 2 and Fig. 3b). They are:

Layer 1: Curation Layer: This layer focuses on the Current, Curated, and, partially, the Computer-readable elements of the 6Cs standard. We deployed an automated approach for pulling the data from the different data sources every hour and storing it on our clusters, adding a corresponding timestamp (in UTC) for each entry. This helps us to maintain the historical snapshots, and allows users to have access to data from any time period. This also helps ensure that the displayed data is always the latest. The scraping of data from unstructured sources, like web pages and reports, acts as the first step towards making the data Computer-readable.

Layer 2: Core Layer: The Core layer is an integral part of our workflow where a huge amount of processing, validation and correction takes place. The diverse sets of raw data stored on our clusters is first combined into a standard and consistent format in accordance with the required spatial and temporal resolutions. This includes a hierarchy with three levels of files each for global data, USA county data, and admin1 regions data. In each of these branches, the data is further organized according to temporal variability, where each file corresponds to the data for a specific day.

Furthermore, data generated by the Core Layer is mapped with FIPS/ID in order to standardize and correctly identify the location. This is specifically essential for the county and admin1 levels where different regions might have a subregion with the same name. Data obtained from multiple sources is very noisy because each data source follows their own data format standard. We have manually identified and mapped each region name with its corresponding ISO3 standard, and also encoded the region names into UTF-8, inspired by the suggestions presented in Addressing the EpiData Challenges [30]. This is a challenging step, since most regions have different languages and encoding, especially from the official sources as they are intended for the local population; it involves a significant amount of effort to manually detect when a new admin1 region is added to our data corpus.

With this step, we achieve Consistency of the data in terms of format and historical data. With the standardized Name and ID, Computer readability is possible.

The next step is to make sure the data mined from the trusted sources is indeed correct. We do several sanity checks and validations of our data, including checks for a wide range of edge cases and areas where an error might be possible. We then manually correct the data for the identified alerts and warnings to the maximum extent possible by verifying the potential source of the error. A log is maintained for the entries where a resolution was not reachable, thereby achieving our standard for Correctness. This processed, validated, standardized and corrected data is then moved into a central database which serves as our internal modeling and analysis dataset.

Layer 3: Augmented Layer: We then augment the data by adding several other derived metrics that help present the overall picture of the pandemic. The augmented data is reflected in the dashboard, and includes the calculation of active cases, the daily change for all the metrics, calculating estimated recovered and estimated active cases, and population-based “per 100K" numbers for all of the metrics. All along our data pipeline, we make sure we properly differentiate between unknown and zero values. This ends the data pipeline, and the prepared data is readily available to be loaded into the front-end for visualization and analysis, thereby completing the loop and achieving Comprehensiveness. The data is populated to our production dashboard multiple times per day in order to present visualizations that are as current as possible.

6.3 A Day with the Dashboard : the Data Update Workflow

As shown in Fig. 3b, the data pipeline is a combination of automated data curation procedures that run at a fixed interval, along with some manual steps that are performed each time the data is loaded to production. The data update to the front-end is done multiple times a day, typically starting at 7 AM and extending until 9 PM Eastern time, with a workflow that repeats every 3 to 5 hours (See Fig. 4). During one time step of the update workflow, the logs at the curation layer are checked for potential issues. Since this is an emerging

infectious disease, the format of upstream data sources may not be consistent and there could also be an outage at one of the data sources. All of these issues are automatically logged by the data pipeline, and if any issues are detected, we update the scripts used for automated curation of the data accordingly.

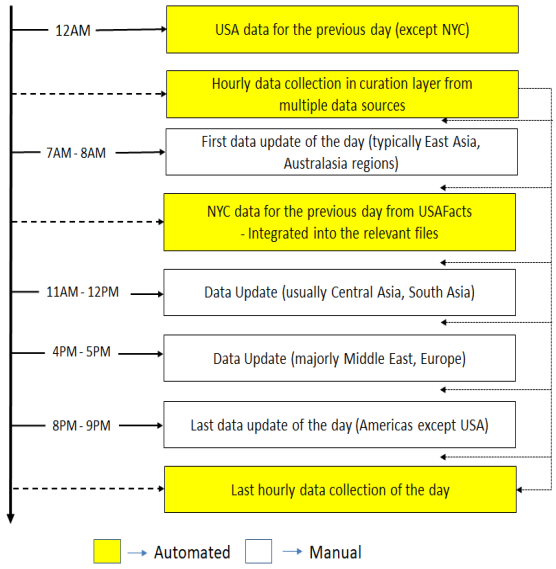


Figure 4: Timeline of daily data updates. This sequence illustrates the 24-hour data update workflow, which is a combination of automated & manual steps. It includes details on which regions' data typically gets updated in each round.

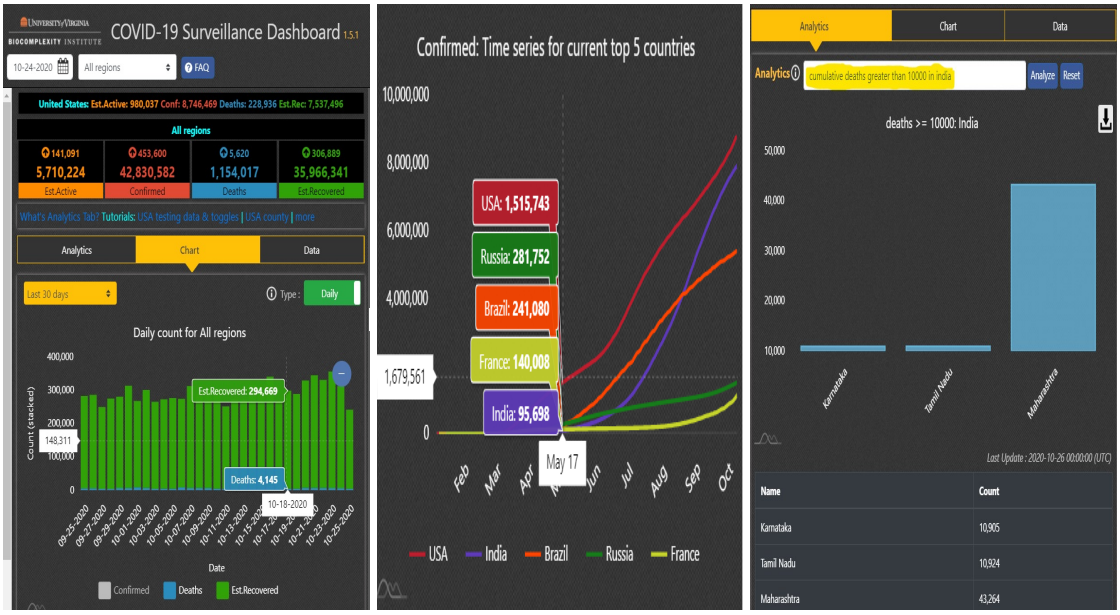
Furthermore, validation and sanity checks are executed which issue warnings for scenarios like: a decrease in confirmed cases, deaths, or recovered counts; a significant increase in cases which are above a certain threshold; the appearance of undocumented regions that may have no FIPS/ID and duplicate entries; and stale data for a certain time period, etc. This is a challenging and time-consuming step, especially when there is an official revision from the health department or administration of a region, as this requires data updates to be performed retroactively. After investigating, fixing, and documenting these potential issues, the data update is pushed to a development site. It is again visually inspected, with some exploration, comparison of total counts with other official sources, and a few other test cases.

This human-in-the-loop visual inspection has proven to be an important step, since the sanity checks occasionally fail to identify some issues, and it is an evolving process to make it robust as we go. It also helps to identify those issues that cannot be validated automatically with the data available to us. As an example, the United Kingdom's COVID-19 cumulative death toll was revised downward on August 13th, 2020, by more than 5000. In this particular case, our data sources did not pick up the downward revision, and this made the total deaths on our dashboard seem much higher than on other COVID-19 dashboards. A human-in-the-loop investigation helped to identify the issue so we could fix it before pushing the data to production. This set of steps are repeated for every update on each day.

7 Front-end UI design

The User Interface is an important component that complements all the efforts put into collecting, cleaning and curating the data by making it accessible to a larger audience. A sophisticated data source which is not easily accessible to its non-technical users would limit its potential. True Comprehensiveness can only be achieved when the vast information present in the data is well-conveyed. Realizing the importance of the UI, we have designed it with the goal of providing a one-stop solution that is easy to use and which has rich visualizations accessible to a variety of users with diverse preferences.

To cater to a variety of user needs, we developed a three-way approach for the spatio-temporal exploration of the data i.e., through a choropleth map, charts, and a data table. The map and charts also provide rich visualizations. The Curated spatial data, which is available at multiple resolutions, is presented using a hierarchical approach that follows the principle of "Overview first, zoom and filter, then details-on-demand" introduced by Ben Shneiderman [31]. This means that, initially, the overall picture of the pandemic across the world is displayed, and from there the user can navigate back and forth from one spatial level to another via multiple paths. Taking all this into consideration, the UI is divided into



(a) (b) (c)

Figure 5: User Interface. (a) Information panel - Charts. This shows interactive charts for cumulative & incidence data for different attributes and summary statistics. (b) Top K charts. The plot to compare the epicurves of the top k regions for different attributes, either by cumulative or incidence data. (c) Information Panel - Analytics. Analytics through Question Answering. Includes results with data and chart.

three panels: Control Panel (Header), Information Panel (Charts, Data and Analytics) and Map Panel.

7.1 Dashboard Panels

Control Panel: The Control Panel (Header) provides temporal exploration functionality through a movie-style time slider or a date selector, spatial exploration through a dropdown list of the countries, and the option to render the choropleth map with different attributes such as Confirmed, Deaths, Estimated Recovered and Estimated Active. It also provides links to more options and toggles to switch between Cases/Testing, Total counts/Per 100K counts, and World map/County map views.

Information Panel: The Information Panel, as seen in Fig. 5a, provides the charts and summary for the selected region. The interactive charts include cumulative and incidence numbers with an option to turn an attribute on or off, i.e., Confirmed, Deaths, Estimated Recovered and Estimated Active. The chart can be zoomed in/out, and detailed information is presented in a tooltip window when a user hovers the mouse pointer over a data point on the plot. The interactive data table, which has all the fields of the Augmented layer in the data pipeline, has a SQL-like query tool which allows users to focus on regions of interest, and a 'Filter' text box to allow them to limit results to records with the selected region name. The region names in the data table are clickable, which will take the user to the next spatial level if supported, i.e., from the national level to the state level, and then to the county level. The advanced analytics, as discussed in Section 9, gives the users the ability to ask questions related to the pandemic and get answers through data and charts.

Map Panel: Our dashboard's landing page shows a world map at the country level, with the exception that USA and China are shown at the state/province level; we also have additional display layers to support county-level rendering for USA, and state/province level rendering for a total of 20 countries. Unlike most other dashboards that use points to represent the regions, we use actual maps/polygons of administrative regions. The Computer-readability in the data has helped to easily do the geospatial mapping. Each spatial level is rendered with a choropleth map using its Estimated Active count by default, and the map can be zoomed in/out. For a selected region, a pop-up window is provided to show its data, along with a navigation option to change the display level of that region.

One unique feature is the multiple ways a user can select a desired region. Spatial exploration can be done through the dropdown list selector, the SQL-like query tool, the filter text box, navigating through the data table, as a question in the analytics, and selection on the map.

7.2 Implementation Details

Our dashboard is a Single Page Application which loads data and default map rendering into an HTML page for the Current data; it is dynamically updated whenever the user interacts with the application. On the application level, our dashboard leverages existing development APIs to provide a basemap layer, and incorporate interactive charts and the data table within a system that is specifically designed to be loosely coupled.

ArcGIS API

Unlike most other dashboards that are built using the configurable ArcGIS Dashboards template, the front end of our dashboard is primarily built on top of the ArcGIS API for JavaScript [32]. Although this was a more challenging approach at the beginning of the project, once we completed that implementation and built the framework, it gave us more flexibility for creating a customizable and functionally rich application. For example, we could add a movie-style time slider to effectively show spatial variation over time, it allowed us to use amCharts 4 [33] for advanced data visualization, and we are able to use DataTables for jQuery [34] to allow users to explore the data in a tabular form. This is the reason why our dashboard looks quite different from most of the other dashboards, and why it has such an extensive set of unique features.

Responsive Web Design

Our dashboard allows users to access the application on any handheld device, such as an iPad, Tablet, or Mobile phone. To tackle this challenge to make the application available on a variety of screen resolutions, we used the Responsive Web Design approach to design and develop our dashboard. This allows the front end to respond to the user's behaviour and environment based on their screen size, platform, and orientation. The practice consists of a mixture of JavaScript, jQuery, Bootstrap, flexible grids and layouts, images, and an intelligent use of CSS media queries.

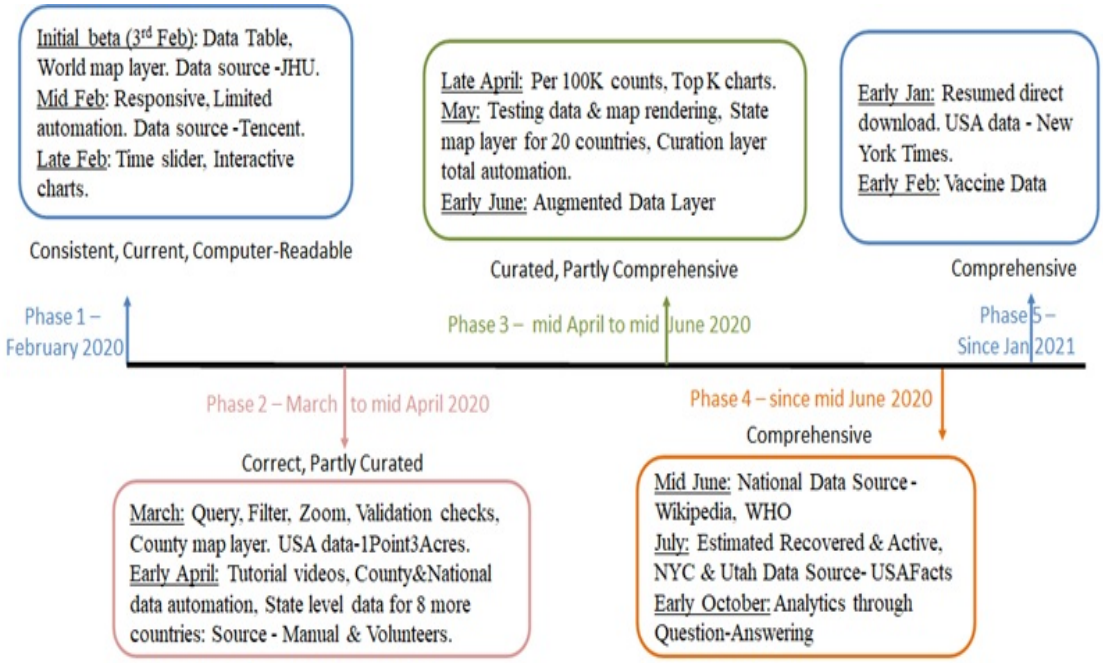


Figure 6: Project Timeline. The course of the development of the project can be broadly divided into five phases; this timeline details which features were added at each phase, and when each of the Cs were achieved.

amCharts

amCharts 4 is a go-to library for advanced data visualization which provides a simple, yet powerful and flexible, drop-in data visualization solution. It includes all basic and advanced chart types, and also supports responsive web design. In particular, we are using line charts and stacked bar charts to visualize incidence and cumulative data, and the comparison plots for the Top K highly affected countries where K can be 5, 10, 15 and 20. (Fig. 5b)

8 Timeline

Maintenance of this system and being flexible and adaptable to the rapid changes in the course of the pandemic is an important challenge in this work. The application in its current form might be easy to rebuild but it is not as straightforward when responding to a fluid and

uncertain pandemic, hence, we discuss the timeline of the project to highlight the efforts that went over time at different phases. Our dashboard was not built in a single release; we went through numerous design iterations and feature improvements to achieve the 6Cs and the other goals. The data sources we depend on have evolved over time as well, requiring us to find new sources as new features were added. Development of this application continues to be an evolving process, and the timeline for our project thus far can be divided into five significant phases as described in Fig. 6.

8.1 First Phase (February 2020)

The initial beta release of our dashboard was deployed on February 3, 2020, with a limited feature set, i.e., only the data table and country-level rendering. USA and China were the only countries which had state/province-level data and geospatial visualizations. The website was not mobile-friendly at this point; however, as we analyzed usage information, we noticed that more than 50% of the users were using mobile devices. This motivated us to make development of a more responsive and mobile-friendly interface our top priority; this release went live in mid-February. Our initial data source was JHU [2], and data curation was a manual process at that time. Starting in mid-February, we switched our primary data source to Tencent [35] as it was more consistent, and implemented a limited automated workflow for the data curation and integration; while we continued to use a manual process for curating the USA, it required a lot of repeated steps to keep the data current. By late February, the time slider and interactive charts had been added, and we ensured that the data was Consistent, Correct, Current and Computer-readable.

8.2 Second Phase (March to mid April)

As the COVID-19 pandemic started to unfold, we focused primarily on three areas: more data coverage, new features, and extensive automation of the data curation process. Starting from early March, we were one of the first dashboards to show USA data at the county level using data from 1point3acres [3] as the source. Querying and filtering the data by

region and enabling zoom in/out functionality on multiple spatial levels helped users to explore the pandemic situation at a granular level, e.g., checking the situation at their loved ones' locales. To allow our users to get the most out of our dashboard, we created a series of tutorial videos explaining the features and how to explore and interpret the data. During this phase, we transitioned to complete automation of the Curation layer, and we added various validation steps that were run before curated data could be moved to the core layer. To ensure data quality, we performed a manual review before each data update. We were also fortunate to have volunteer support for data collection at the state/province level in some countries, including Germany, Chile, Brazil, Colombia, Greece and Argentina – a step forward towards the 'Curated' goal.

8.3 Third Phase (mid April to mid June)

We continued to improve our dashboard with more data and features. We added the data and visualization for COVID-19 Testing, added per 100K counts, added the Top K country charts, and many minor updates. We automated the state/province-level data curation, and extended the list of countries with state/province support to 20, with map rendering through a new third map layer; these countries included Argentina, Austria, Brazil, Belgium, Canada, Chile, China, Colombia, Germany, Greece, India, Italy, Mexico, Peru, Portugal, Saudi Arabia, South Korea, Sweden, Switzerland and USA. We also extended the data pipeline with the augmented data layer to ensure the consistency of core layer data while adding more derived metrics for the application. At this stage, we achieved the Curated and Comprehensive goals, thereby achieving all of the 6Cs.

8.4 Fourth Phase (mid June to December)

In the first half of June, we switched the data source for National data to Wikipedia and WHO. As mentioned in 5.1, we began providing Estimated Recovered and Estimated Active numbers in addition to the much less accurate reported recovered and active counts. We believe this will provide a better picture of the current situation, and will be helpful for

decision-making at the individual level. Finally, we added Advanced Analytics, which allows users to ask questions regarding the pandemic using natural language, and provides the answers through data and plots; we believe this step helped make the dashboard more Comprehensive.

8.5 Current Phase (since January)

Starting from January 1, 2021, the USA data source is changed from 1point3acres to New York Times. This is due to an observation of increasing errors and irregular updates for some counties in 1point3acres data. This switch has also allowed us to open the data for public download. Further, starting from February, we added the data and visualization for COVID-19 vaccinations for all the countries that started administering the doses. This data is at the national level, with the USA also being at the state level and gathered from Our World In Data. These updates allowed us to take a further leap towards the 6Cs and make progress in our efforts to improve the public informedness of the pandemic. It also reflects our commitment to maintain the dashboard while adapting to the real world changes that happen due to the fluid nature of the pandemic.

9 Analytics

The plethora of information present in the data cannot be effectively and completely conveyed on a single web page without cluttering it. Providing users with the ability to quickly and directly get answers to generally asked questions will further minimize the need for them to understand the epidemic terminology, allowing them to spend more time navigating the application. This is how ultimate Comprehensiveness can be achieved, and it also helps to serve the diverse information needs of the public. In order to achieve this, we support interactive queries for analytics, where a user can ask a question in plain conversational English, which the system attempts to interpret, then answers directly with a plot and data as appropriate (See Fig. 5c).

9.1 Methodology for Analytics

We have identified that a question related to an epidemic will typically be composed of 4W1H (which, what, where, when, how) as shown in Table 2. This provides a basis for processing the data and responding effectively to user queries.

Table 2: 4W1H Structure of Epidemic Questions

Which	What	Where	When	How
Confirmed Deaths Est.Recovered Est.Active Recovered Active	count greater than / less than top N / bottom N highest / lowest	India Virginia Queens United States Lombardy World	March 15 Today July 10	Cumulative Incidence

Some examples of the questions supported by the tool include: “Cumulative deaths count in India on July 10", “Cumulative confirmed greater than 100000 in United States today", “Top 5 incidence deaths in Virginia state on April 15". The 5W1H acronym is quite popular in research, journalism, investigations etc, and ‘Who’ replaced with a more relevant ‘Which’ for epidemic data. The other W that is not supported in our 4W1H formulation is ‘Why’. This category of question is not trivial to answer from the surveillance data without the context of the real-world situation, which is often very complex. The default values for each of the 4W1H, if the user doesn’t provide them explicitly, are taken as follows: Which - Confirmed, What - count, Where - World, When - Today, How - Cumulative. The system implementation of 4W1H can search the words in the question for each of the 4W1H keywords in the search space. The search space includes 6 possibilities for ‘Which’, 7 for ‘What’, around 3750 regions (210 countries, 350 states, 3200 counties) for ‘Where’, the no.of days since pandemic start for ‘When’ and 2 for ‘How’.

9.2 Semantic Textual Similarity and The User Query Dataset

To allow users to freely ask questions without restricting them to certain words as shown in Table 2, a state-of-the-art model like SentEval [36] can be used to compute semantic similarity of the tokens in the question against the words within the search space, and to return the result for the match which had the highest similarity score. Whenever a question fails to get a result directly from the search space, this can be invoked and the relevant results can be displayed if the mean similarity score meets an empirically determined minimum threshold. If it fails to meet the minimum threshold, it will be considered a failed query and the user can be alerted that the question could not be answered.

To keep track of the user questions and the status of the results, we save them into a database along with other details like 'Date', 'Time' and 'Result (Pass/Fail)' for each question. This helps us identify which queries are failing, allows us to broaden the scope of our methodology if required, and also guides improvement to the sentence similarity model. Furthermore, there is a feedback option where the users can submit a response if they are satisfied with the results of their question which we can also use to improve the system. This will be a first-of-its-kind dataset containing real-time epidemic questions from users located all over the world. This data could potentially open up new areas of research on understanding the data needs of the public, allowing applications to more closely target their responses during a pandemic.

9.3 Implementation and Future Work

We support spatial questions for the current state of the pandemic in the dashboard, i.e., 3W1H with the exception of 'When'. Whenever a question cannot be resolved within the given search space, the system catches the exception and provides the user with a warning that their question could not be answered. We are expanding the supported input space by manually reviewing the failed queries from the User Query dataset and incorporating changes where feasible. This includes support for sets of synonyms, such as: [greater than, more than, higher than], and [United States, US, USA, America], etc.

A possible future direction is adding support for the temporal 'When' questions, and state-of-the-art sentence similarity models. The expansion of 4W1H input space to add support for Tests, Vaccines to the 'Which' handling, and for Peak, 7-day Moving Average, Test positivity rates, etc. to the 'What' can also be a possible next step.

10 Utility of the Dashboard

10.1 Data storytelling

An important application of the dashboard is in support of *data storytelling*. Data storytelling is the art of developing a narrative based on a data set, incorporating visualizations and analysis tools so viewers can make solid, well-supported interpretations; it is quite popular in the fields of data science [37] and data journalism. The Analytics of the dashboard, along with its historical data and interactive visualizations, make it easy to gain insights that facilitate data storytelling. An excellent illustration of this concept is a blog post by Tomas Pueyo, who has produced an extremely interesting narrative on the COVID-19 pandemic [38].

10.2 Extensive use by various organizations

The application and the back-end data have been used by a large number of analysts, researchers, and laypeople. Our group uses the data to support federal agencies (DoD, CDC), our state (Virginia) and local public authorities (local health districts and our university) as they respond to the pandemic. The data is also used to drive our predictive models, which we use to produce counterfactual analysis, and answer policy questions, including contact tracing, resource allocation and augmentation, and campus reopening and management. See [39] for reports that the Virginia Department of Health (VDH) releases based on our work.

Several other groups use our dashboard and associated data as well. We list just a few to illustrate its broad use: (*i*) it is listed as a part of the NIH MIDAS data catalogue [40], the ESRI COVID-19 GIS Hub [41] and the 2021 Coalition for Academic Scientific Computation (CASC) brochure; (*ii*) it is used by several groups at DoD; (*iii*) it is used by local authorities, including in Bay County [42] and Panama City Beach [43] in Florida, where our active case counts are used as one of their thresholds for allowing vacation rental reservations.

10.3 Web Traffic

During the initial phase, we had around 40,000 users in total, and the top 3 countries for web traffic were the United States, Germany and China. By the end of the second phase, we reached over 750,000 unique users, and the top 3 countries for web traffic were the United States, Canada and Germany. By March 2021, there were a total of over 1.2 million unique users. The top 3 countries that made up the largest portion of users are currently the United States, Canada and India. Overall, two-thirds of the users are from the United States, and the average time spent on our dashboard is around two minutes. The period of maximum engagement was the first week of April, with a peak of 80,000 views and 50,000 unique users on a single day. The overall returning user percentage stands at 19% and it was 11%, 19%, 31%, 23.35% and 25% respectively during our different phases as described in the timeline.

11 Serving Epidemiological Workflows with Surveillance

11.1 PatchViz

The variety of surveillance data available can facilitate counterfactuals such as for projecting the case trajectories under multiple scenarios and also for different constraints such as increase or decrease of transmission at various spatial levels. Such a study requires the counts for Susceptible, Exposed, Infected, Recovered, Current Vaccinations and New

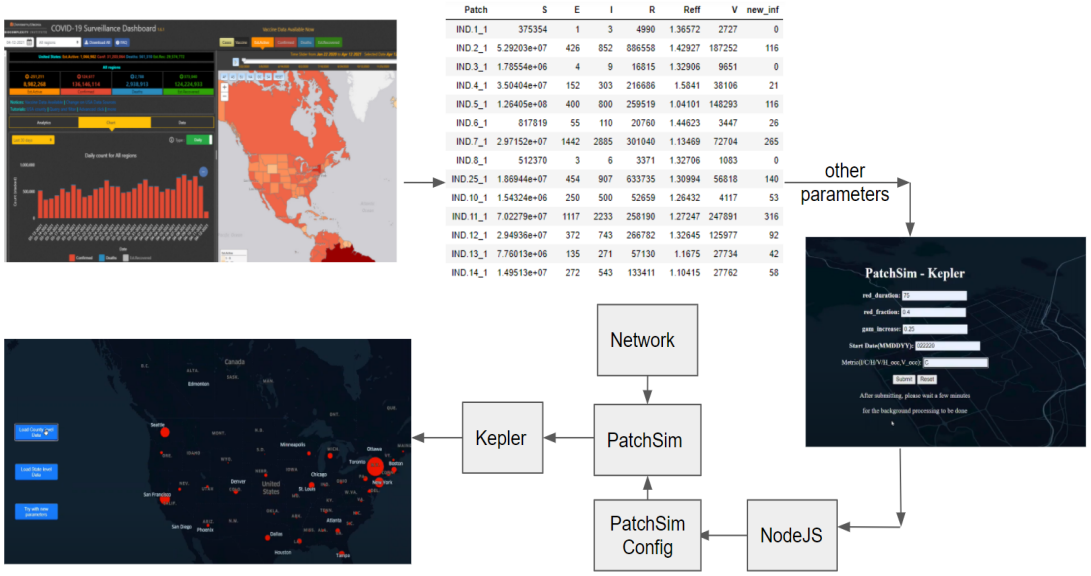


Figure 7: The workflow from Surveillance data to facilitating counterfactuals with S,E,I,R,V formulation, and finally running simulations and visualizing results with PatchViz

infections. The surveillance data can provide all of these where Susceptible is the population of the region, the Exposed and Infected being calculated at a proportion from disease dynamics using the estimated active, and Recovered is taken from estimated recovered. The presence of vaccination data and incidence infections in augmented surveillance data fulfills the initial requirements to run these kinds of studies.

PatchViz is an interactive visualization tool built for the purpose of feeding surveillance to seed and run mechanistic simulations for counterfactual analysis. It is a light-weight web based application that runs with Kepler.gl [44] - an open source geospatial analysis tool, along with ReactJS, Webpack, and the runtime environment as NodeJS. The goal of this tool is to equip the end users with the advanced tools for them to be able to run counterfactuals, in a simplest way possible with minimal technical knowledge required. PatchViz is a 2-page web application with the workflow as follows - the landing page takes all the relevant input parameters (such as Exposure Rate, Recovery Rate, Scaling Factor, Vaccine Efficacy, Start Date etc) along the network file and the surveillance data that are required to run the simulations. These are fed into the backend to the configuration file

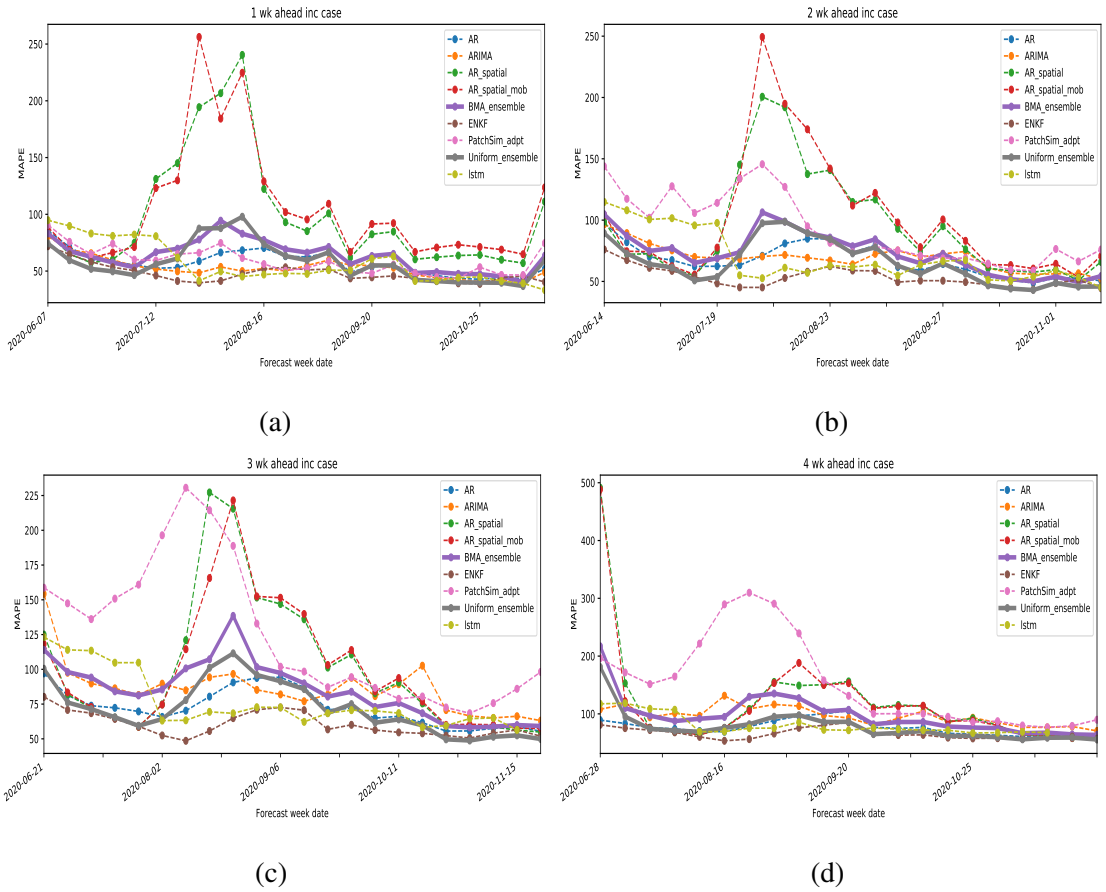
of PatchSim [45],a software that allows modeling SEIR dynamics across sub-populations, and after processing it generates the output. This processed output is communicated with the front-end to a kepler module where the data, often the resulting infections for a given scenario, is visualized in a spatio-temporal map on the second page of the web application.

The input parameters passed on the landing page depends on the study design and the counterfactual questions that we are seeking answers. This hence completes the entire loop, from surveillance data to wrangling it to feed into PatchViz along with the counterfactuals, and finally visualizing the resulting infections on Kepler which resembles the map panel in the dashboard (See Fig 7). These studies can be run by anyone using the data & the PatchViz web application, and could help the policymakers or general public not only be aware of the past and present through surveillance data, but also be able to assess scenarios by running simulations to look forward to the future trajectory.

11.2 High Resolution forecasting with Ensemble Kalman Filter

Time-series forecasting is yet another Epidemiological Workflow that is served by the Surveillance. In this section we demonstrate forecasting with Ensemble Kalman Filter using the surveillance data. The CDC in collaboration with academic partners initiated the COVID-19 Forecast Hub in April 2020, a consortium of modeling teams to coordinate the forecasting efforts. There was a clear need for a comprehensive forecasting framework that combines multiple individual methods from various modeling paradigms within a performance-based ensemble. We used a variety of autoregressive methods (AR, ARIMA) [46] with exogenous variables, an LSTM model [47], ensemble Kalman Filter (EnKF) [48], compartmental SEIR model, and employ Bayesian Model Averaging (BMA) [49] to aggregate the individual forecasts [50]. This has been operational for nearly 10 months, updated weekly at county level resolution. We discuss the contributions to the BMA ensemble with EnKF, which showed in ablation analysis that at individual county level there has been a significant drop in performance of our ensemble in the absence of forecasts that are being produced by this model. It was also observed that EnKF is the best performing

model in comparison to the other models in the ensemble when evaluated using MAPE (See Fig 1.8). We walk through the model training details and the scalability of this model with the weekly forecast submissions for multiple countries to a recent forecasting initiative by European CDC (ECDC).



(a) (b)

(c) (d)

Figure 8: EnKF figures. EnKF is observed to be the best performing model in comparison to other models in the ensemble, evaluated according to MAPE.

Model Selection and Scalability

It is a data-driven model with no exogenous variables, so the choice of the time series will have an influence on the model performance. For model selection, we tried four different models using weekly and daily versions of incident and cumulative case series with sigma

points, $N = 20$. The daily incident version was smoothed with a 7-day moving average to remove within week variations. Evaluation was done using median absolute percent error (APE) for counties with more than 10 cases. The model with weekly cumulative case counts performed consistently well across different forecast duration. The comparison with the mean APE and MSE is also taken into account and is observed to be similar to that of median APE. Using this model, we varied the number of samples from 5 to 2500 over forecasts generated retrospectively for a 25-week time period, with higher N leading to better performance at increasing computational costs. Based on the observed performance-speed trade-off, we set $N = 100$.

For the forecasts at the state and national level we explored two options. In the first approach we aggregated the forecasts from the counties to get them at the state and national level. In the second approach, we trained models for each region directly using its own time series. We did not observe any significant difference in the performance and often the aggregated forecasts seemed to do better than forecasting from the state or national time-series directly. So we stick to the first approach, but it is to be noted that when any region has unknown or undocumented cases in the sub-region level, these results might change. A difference in the ground truth where aggregated cases from its sub-regions is not the same as reported cases of the region, it is preferable to generate the forecasts for that region from its reported time-series directly.

The goal is to be able to serve users and policymakers from multiple countries across the globe. Similar to the global nature of the surveillance dashboard, it is also important for the forecasts to be scalable and adaptable to multiple countries. We started submitting our forecasts to an initiative by ECDC [51] through the BMA ensemble with the similar composition of models as that of the submissions to *COVID-19 Forecast Hub*. EnKF uses similar training mechanisms and supports all the countries in Europe. The simplicity of the model itself with its coarse representation, is very helpful to quickly run the models, one for each region. Due to this, it can also be scaled to produce forecasts for all the regions and sub-regions in the world that have the time-series data available. This is a potential

future work to publish the forecasts for all countries even as part of the dashboard, and let the public be informed of the trajectory of the pandemic under current conditions, in their region of interest. In addition to forecasting the cases, the model can also be used to forecast deaths and hospitalizations by changing the time series and re-training to capture the specific characteristics of the new measure for which the forecasts are being produced.

12 Discussion and Conclusion

12.1 Challenges and Limitations

Getting data from reliable sources and conforming to the 6Cs is the most challenging task in building and maintaining our dashboard, yet worth the effort to maintain the quality. Changes in the Terms of Use at some of our upstream data sources was another big hurdle to overcome, and we had to switch data sources at times as a result. We faced other challenges, especially with the state/province-level data, which lacks a standard format and computer-readability in some of the upstream sources. The difference in the naming of regions also makes it difficult to identify corresponding spatial coordinates, and required a manual effort to map the names. The amount of work involved has limited our ability to expand state-level coverage to additional countries.

With Recovered numbers not being reported consistently across all countries, it has been hard to rely on reported Active case counts, which should be the ideal metric for understanding the current situation. This led us to develop an algorithm for computing estimated recovered values based on available research and studies. Inconsistencies in the definitions of “case" and “death" across the regions also created issues, with some regions reporting presumed case/death counts, while some data sources report only confirmed cases/deaths. This also led to retrospective data corrections in some cases. There have been instances where cases were defined under unknown regions, as well as cases where more than one region was clubbed together. These have imposed some challenges, especially

when the data is intended to be used in modeling; these situations had to be handled on a case-by-case basis.

With the rapidly evolving pandemic and short release cycles, robust testing of our application also turned out to be a challenging task. Building our dashboard on top of the ArcGIS API for JavaScript was a key decision we made at the beginning of the project. This made our dashboard stand out in terms of how it looks and the unique features we are able to deliver. This demanded a significant effort at the beginning to build the framework.

Due to the challenges in the collection and maintenance of the state/province-level data, with the inconsistent formats at upstream sources, our dashboard currently only supports 20 countries at this level of granularity. Also, while a large number of countries do not have this data reported, others provide it in the form of text-based reports or press releases, hence making it difficult to extract on a continual basis; this leaves room for improvement for Curation.

12.2 Lessons

The onset of COVID-19 led to the creation of many special-purpose data collection and visualization efforts that focused on specific aspects of the pandemic, such as current case counts or the status in a particular region. We see the need for a more general-purpose, all-in-one central repository that can provide broader coverage of the pandemic at a granular level of spatio-temporal detail; corresponding data update access could be granted to public health officials, journalists, and researchers to make this a global community-response resource promoting timely decision-making and research.

COVID-19 has shown the importance of easy-to-use and interactive visualizations of the epidemic data to help the public stay informed. We learned that, during an emerging infectious disease, speedy response is essential, and a ready-to-deploy dashboard without much of a development effort could save critical time. Our current design of the application architecture, guided by the 6Cs of epidemic data management, is flexible and adaptable and

could be deployed quickly in case the need arises for another specialized pandemic site in the future.

We also provided a support email for feedback, suggestions, and questions from the users. It turned out to be very useful for understanding public opinion, and over time we performed many rollbacks, upgrades and changes based on the active feedback received. The tutorial videos attracted many viewers and were widely appreciated. We also observed that 6Cs can be scaled for other spatio-temporal datasets that bear a similarity to the surveillance data, through the data collection for a huge, complex airline mobility dataset that facilitated a study to evaluate the impact of international airline suspensions on the early global spread of COVID-19 [52].

12.3 Conclusion

This thesis presents our epidemic surveillance in support of the COVID-19 pandemic planning and response. We proposed 6C metrics as a possible standard for epidemic data management. We discussed the importance of the 6Cs, why it should be treated as a standard, followed by a discussion on its relationship with the 5Vs of big data. We also provided an overview of the related work of other COVID-19 and epidemic dashboards. We then described the procurement and curation of the dashboard data, including the sources used and our algorithm for estimation of COVID-19 recoveries - a very unique feature of our dashboard. We discussed the underlying architecture of our dashboard, along with user interface (UI) details. We then described the timeline of our project and presented Q&A based analytics that aims to help understand & process the multitude of data with plain conversational questions.

We then presented some of the applications of the surveillance data - 1. Data Storytelling, an important utility of surveillance for writers and journalists, who are often on the frontline to communicate with the public and act as a primary source of information for the society. 2. PatchViz tool that allows end users to run and visualize simulations in a web browser, for counterfactual analysis. This completes the loop and equips the users and policymakers

with tools not only to track the pandemic but also to ask complex what-if questions and find out the answers on their own. 3. COVID-19 time-series forecasting, yet another application of the surveillance, highlighting contributions to the Bayesian Model Averaging (BMA) ensemble with the Ensemble Kalman Filter (EnKF), as part of the CDC-initiated *COVID-19 Forecast Hub*. Finally, we discussed the usage of our dashboard, the challenges faced, and lessons learned.

Bibliography

[1] J Anuradha et al. A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia computer science*, 48:319–324, 2015.

[2] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[3] Tong Yang, Kai Shen, Sixuan He, Enyu Li, Peter Sun, Lin Zuo, Jiayue Hu, Yiwen Mo, Weiwei Zhang, Pingying Chen, et al. Covidnet: To bring the data transparency in era of covid-19. *arXiv preprint arXiv:2005.10948*, 2020.

[4] Worldometer: Coronavirus update. https://www.worldometers.info/world-population/population-by-country/.

[5] Esteban Ortiz-Ospina Max Roser, Hannah Ritchie and Joe Hasell. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[6] US Coronavirus cases and deaths. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

[7] WHO Coronavirus disease (COVID-19) dashboard. https://covid19.who.int/.

[8] CDC COVID data tracker. https://www.cdc.gov/covid-data-tracker/.

[9] ECDC COVID-19 data. https://www.ecdc.europa.eu/en/covid-19/data.

[10] Benjamin Wissel and P el al. Camp. An interactive online dashboard for tracking covid-19 in u.s. counties, cities, and states in real time. *Journal of the American Medical Informatics Association : JAMIA*, 27, 04 2020.

[11] The new york times coronavirus (covid-19) data in the united states. https://github.com/nytimes/covid-19-data.

[12] Stefano Barone, Alexander Chakhunashvili, and Albert Comelli. Building a statistical surveillance dashboard for covid-19 infection worldwide building a statistical surveillance dashboard for covid-19 infection worldwide. *Quality Engineering*, 06 2020.

[13] Alexander Hohl, Eric Delmelle, Michael Desjardins, and Yu Lan. Daily surveillance of covid-19 using the prospective space-time scan statistic in the united states. *Spatial and Spatio-temporal Epidemiology*, 34:100354, 06 2020.

[14] Bo Xu, Bernardo Gutierrez, and et al. Mekaru. Epidemiological data from the covid-19 outbreak, real-time case information. *Scientific data*, 7(1):1–6, 2020.

[15] Swapna Thorve, Mandy Wilson, Bryan Lewis, Samarth Swarup, Anil Kumar Vullikanti, and Madhav Marathe. Epiviewer: An epidemiological application for exploring time series data. *BMC Bioinformatics*, 19, 12 2018.

[16] Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19). https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

[17] Zunyou Wu and Jennifer M. McGoogan. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72[U+202F]314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, 323(13):1239–1242, 04 2020.

[18] CDC: Interim clinical guidance for management of patients with confirmed coronavirus disease (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html.

[19] Illinois DPH : COVID-19 Statistics. https://www.dph.illinois.gov/covid19/covid19-statistics.

[20] Wikipedia. https://en.wikipedia.org/wiki/COVID-19_pandemic_by_country_and_territory.

[21] COVID-19 India. https://www.covid19india.org/.

[22] Government of Canada. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html.

[23] Government of Greece. https://covid19.gov.gr/.

[24] The COVID tracking project. https://covidtracking.com/.

[25] Worldometer: Population by country (2020). https://www.worldometers.info/coronavirus/.

[26] World population review: Us states. https://worldpopulationreview.com/states.

[27] Worldatlas: Chinese provinces by population. https://www.worldatlas.com/articles/chinese-provinces-by-population.html.

[28] Adc worldmap. https://www.adci.com/adc-worldmap/.

[29] ArcGIS Online. https://www.esri.com/en-us/arcgis/products/arcgis-online/overview.

[30] Geoffrey Fairchild and Byron el at. Tasseff. Epidemiological data challenges: planning for a more robust future through data standards. *Frontiers in Public Health*, 6:336, 2018.

[31] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IN IEEE SYMPOSIUM ON VISUAL LANGUAGES*, pages 336–343, 1996.

[32] ArcGIS API for JavaScript. https://developers.arcgis.com/javascript/3/.

[33] amCharts, JavaScripts & Map. https://www.amcharts.com/.

[34] DataTables: Table plug-in for jQuery. https://datatables.net/.

[35] Tencent's data hub on COVID-19. https://news.qq.com/zt2020/page/feiyan.htm.

[36] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. 2018.

[37] Brent Dykes. Data storytelling: The essential data science skill everyone needs. *Forbes*, 03 2016.

[38] Coronavirus: Why you must act now. https://medium.com/@tomaspueyo/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca.

[39] VDH COVID-19 weekly report. https://www.vdh.virginia.gov/content/uploads /sites/182/2020/07/UVA-COVID-19-Model-Weekly-Report-2020-07-17.pdf.

[40] MIDAS online portal for COVID-19 modeling research. https://midasnetwork.us/covid-19/.

[41] ESRI COVID-19 GIS hub. https://midasnetwork.us/covid-19/.

[42] Bay county plan for opening short-term rentals, phase-ii. https://www.baycountyfl.gov/CivicAlerts.aspx?AID=156.

[43] Panama city beach chamber COVID-19 updates. https://www.pcbeach.org/news-article/panama-city-beach-chamber-coronavirus-covid-19-updates/.

[44] Kepler.gl. https://kepler.gl/.

[45] Srini Venkatramanan, Parantapa Bhattacharya, Przemek Porebski, and Brian Klahn. Nssac/patchsim: First official release, December 2020.

[46] Prashant Rangarajan, Sandeep K. Mody, and Madhav Marathe. Forecasting dengue and influenza incidences using a sparse representation of google trends, electronic health records, and time series data. *PLOS Computational Biology*, 15(11):1–24, 11 2019.

[47] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[48] Wan Yang, Alicia Karspeck, and Jeffrey Shaman. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol*, 10(4):e1003583, 2014.

[49] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[50] Aniruddha Adiga, Lijing Wang, Benjamin Hurt, Akhil Sai Peddireddy, Przemyslaw Porebski, Srinivasan Venkatramanan, Bryan Lewis, and Madhav Marathe. All models are useful: Bayesian ensembling for robust high resolution covid-19 forecasting. *medRxiv*, 2021.

[51] European covid-19 forecast hub. https://covid19forecasthub.eu/.

[52] Aniruddha Adiga, Srinivasan Venkatramanan, James Schlitt, Akhil Peddireddy, Allan Dickerman, Andrei Bura, Andrew Warren, Brian D Klahn, Chunhong Mao, Dawen Xie, et al. Evaluating the impact of international airline suspensions on the early global spread of covid-19.