From 5Vs to 6Cs: Operationalizing Epidemic Data Management with COVID-19 Surveillance

Akhil Sai Peddireddy*†, Dawen Xie†, Pramod Patil‡, Mandy L. Wilson†
Dustin Machi†, Srinivasan Venkatramanan†, Brian Klahn†, Przemyslaw Porebski†
Parantapa Bhattacharya†, Shirish Dumbre‡, Erin Raymond†, Madhav Marathe*†
*Dept of Computer Science, University of Virginia
†Biocomplexity Institute & Initiative, University of Virginia
‡Persistent Systems, Pune, India

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378435

***Abstract*—The COVID-19 pandemic brought to the forefront an unprecedented need for experts, as well as citizens, to visualize spatio-temporal disease surveillance data. Web application *dashboards* were quickly developed to fill this gap, but all of these dashboards supported a particular niche view of the pandemic (ie, current status or specific regions). In this paper, we describe our work developing our COVID-19 Surveillance Dashboard, which offers a unique view of the pandemic while also allowing users to focus on the details that interest them. From the beginning, our goal was to provide a simple visual tool for comparing, organizing, and tracking near-real-time surveillance data as the pandemic progresses. In developing this dashboard, we also identified 6 key metrics which we propose as a standard for the design and evaluation of real-time epidemic science dashboards. Our dashboard was one of the first released to the public, and continues to be actively visited. Our own group uses it to support federal, state and local public health authorities, and it is used by individuals worldwide to track the evolution of the COVID-19 pandemic, build their own dashboards, and support their organizations as they plan their responses to the pandemic.**

***Index Terms*—COVID-19, surveillance, dashboard, big data, epidemic data**

I. Introduction

The COVID-19 outbreak caused by the novel coronavirus 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 October 25, 2020, the cumulative number of confirmed COVID-19 cases exceeded 43 million worldwide, with almost 1.15 million deaths. With the uncertainty surrounding the pandemic and its impact, there was an urgent need to collect and visualize pandemic case data to guide informed and fact-based decisions. The pandemic data has been of prime importance for policymakers, public health officials, and academic researchers attempting to interpret and respond to this crisis, along with every layperson concerned about how the pandemic will affect their daily life. Many dashboards have been developed to help the public better understand the current status, each focused on a different aspect of the pandemic. One of the most used dashboard was was developed and is being maintained by the Johns Hopkins University. 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 data source and coupled with rich visualizations and an easy-to-use interface that is accessible to both the public and researchers.

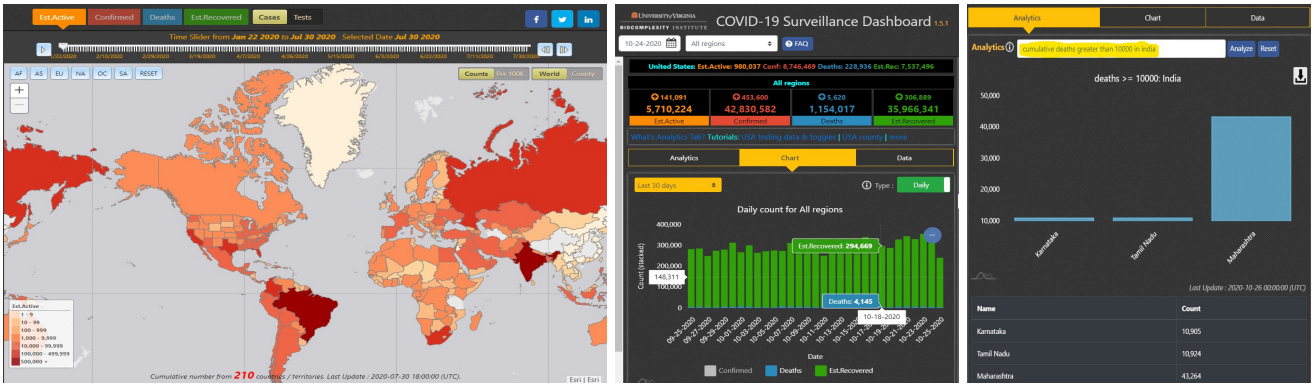
These issues led us to identify 6 metrics that might serve to define a standard for epidemic surveillance data management that we call *the 6Cs standard*. The 6Cs standard proposes that epidemic surveillance data should be *Consistent, Correct, Current, Comprehensive, Curated, and Computer-readable*. 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.

The Biocomplexity Institute & Initiative's COVID-19 Surveillance Dashboard, originally released on February 3, 2020, can be accessed 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. 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), charts (Fig. 1b), and a data table, 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 all countries in the world; at the state and province level for 20 countries; and county-level statistics for the United States (USA). As of October 25, 2020, more than 1.13 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.

II. The Importance of the 6Cs Standard

A. 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



(a) (b) (c)

Fig. 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 or different attributes (b) Information panel. Includes interactive charts, summary statistics, and a data table. (c) Information Panel - Analytics through Question Answering. Includes results with data and chart.

public policy decisions. 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. Models 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. Many of the online tools depend on open data sources. If the underlying format of these sources changes, it can delay their 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 this 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.

B. Correct

Epidemic data is sensitive, and inaccuracies can have a catastrophic impact on public health. This raises the importance of the data correctness. Large scale data curation is prone to error that occurs for a variety of reasons, including incorrect data entry, improper access to the data, or errors in 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 sum of its subregions, etc. Any uncertainty that cannot be resolved should be explicitly documented.

C. 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 quick 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 decision-making in rapidly changing conditions. This emphasizes the need for timely updates, which, in turn, requires automated data collection and maintenance of historical snapshots. It also amplifies the need for storing data in a temporal representation with clear indication of when data updates have occurred.

D. 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, mobility, and interventions, have a direct influence on the pandemic and add meaning to the core data. Having a comprehensive, complete picture of the pandemic can help researchers understand which factors could curtail the spread of the disease.

E. 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

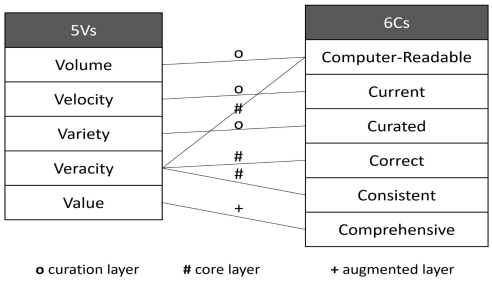


Fig. 2: Relationship between the 6Cs and the 5Vs of Big Data, and the data pipeline layers used to achieve them.

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.

F. 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 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 data retrieval.

G. The 5Vs of Big Data

Volume, Velocity, Variety, Veracity, and Value are popularly known as the 5Vs of Big Data. Although the 6Cs of epidemic data management bear some resemblance to the 5Vs, the two standards actually complement each other quite a bit (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 characteristic Velocity resembles Current, specifically with the temporal component where the data flows continuously from multiple sources into the application for real-time updates. Several 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 – not more and not less – 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.

III. 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, Johns Hopkins University (JHU) [1] was one of the few organizations that gathered and shared COVID-19 data through a dashboard. The system continues to be one of the most influential dashboards to date and is undoubtedly the most widely used dashboards. One limitation of their dashboard was that the data format and sharing platform 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 JHU's 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 (cases, deaths, recovered etc.) is organized into separate tabs and panels, which makes it challenging to assess the full picture for a particular region.

1Point3Acres' COVID-19 Tracker [2] initially focused on providing near real-time case information only for North America. We were one of the early adopters of their data, and have been using it for USA since early March. 1Point3Acres has expanded data coverage to more countries over time. 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 [3] collects national-level data from every country and makes it available in a data table for 3 days. For USA, they provide detailed case counts down to the county level. They also display charts of historical data; 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.

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) [4]–[6] 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. [7] used an R Shiny app for surveillance of USA cities using data from JHU, The New York Times, and the COVID Tracking Project. Barone et al. [8] 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. [9] 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 [10] which provides individual-level epidemiological data, also known as *line lists*.

IV. Data and Sources

This section describes the data elements our dashboard depends upon. We use several data sources for collating our epidemic surveillance dataset. Table I summarizes which data sources we are currently using, how often we poll those sites, and the collection methods.

TABLE I: Collection Details of Epidemic Surveillance Data.

	Source	Frequency	Mode of collection
USA	1Point3Acres [2], USAFacts [11]	Daily	API; CSV file
National	Wikipedia [12] and WHO [4]	Multiple times a day	Scraping; CSV file
Sub-national	Wikipedia	Multiple times a day	Scraping
India	Covid19India [13]	Multiple times a day	Github
Canada	Gov. of Canada [14]	Multiple times a day	CSV file
Greece	Min. of health [15]	Multiple times a day	Scraping
USA Tests	COVID Tracking Project [16]	Daily	CSV file

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

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

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 Geographic Information System (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)
- Est.Active : Confirmed Cases - Deaths - Est.Recovered
- New Cases, New Deaths, New Recovered, New Est.Recovered, New Est.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.

Demographic data: We use different sources for population data, including Worldometer for country-level population estimates [17], World Population Review for USA state-level population estimates [18], WorldAtlas for China province-level populations [19], Wikipedia 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 [20]. We host these polygons as feature layers on ArcGIS Online [21].

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. 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 [22] 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 et al. [23] 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 [24]. Illinois Department of Public Health follows a similar estimate for their calculation of recovered cases [25]. Based on these studies, we calculate Estimated Recovered as follows:

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

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.

V. Back-end Architecture

Fig. 3a shows the overall architecture of our dashboard. As described in Section IV, the data available on our dashboard is multidimensional. A design decision made at the beginning 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 [21] is used to store and access demographic and GIS data for the regions.

A. ArcGIS Online

ArcGIS Online serves as the GIS server that hosts the feature layers needed by our dashboard. For easy accessibility, we are using three feature layers corresponding to each spatial level, i.e., the world map layer shown by default on the dashboard, the states and provinces layer, and the USA counties layer. 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 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 support of simultaneous requests.

B. Surveillance Data Pipeline

The surveillance data is collated, processed, and augmented with a robust data pipeline (See Fig. 3b) which helps to create an all-in-one comprehensive data hub for all temporal and spatial resolutions. We organized the entire data pipeline in a three-layered approach to effectively achieve the 6Cs as described in the previous sections. They are :

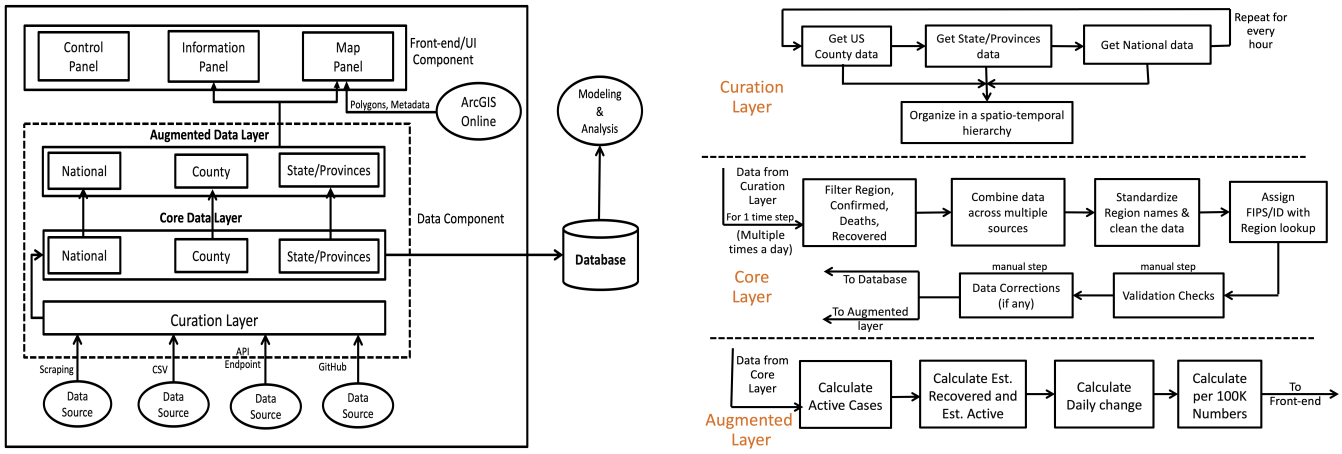
Layer 1: Curation Layer This layer focuses on the Current, Curated, and, partially, the Computer-readable elements of the 6Cs standard. We deployed an automated approach for pulling the data from the different data sources every hour and storing it on our clusters, adding a corresponding timestamp (in 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 webpages 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, state/province data, and USA county 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 [26]. 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 have completely achieved Consistency of the data in terms of format and historical data. With the standardized Name and ID, the loop of Computer-readability is closed.

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 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 helps 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



(a) (b)

Fig. 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.

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.

VI. Front-end UI design

The User Interface (UI) is an important component that complements all the efforts put into 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. It is a Single Page Application that is specifically designed to be loosely coupled. It uses ArcGIS API for JavaScript [27], unlike most other dashboards which are built using the configurable ArcGIS Dashboards template. We used the Responsive Web Design approach to design and develop our dashboard, and amCharts [28] for data visualization.

To cater to a variety of user needs, we developed a three-way approach for the spatiotemporal exploration of the data i.e., through choropleth map, charts, and 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 [29]. 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 three 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. 1b, 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 VII, 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 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 the 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.

VII. 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 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 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. 1c).

TABLE II: 4W1H Structure of Epidemic Questions

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

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 II. This provides a basis for processing the data and responding effectively to user queries. 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". Our implementation searches 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'. (* denotes the default value in Table II)

Semantic Textual Similarity: To allow users to ask questions without restricting them to certain words as shown in Table II, a state-of-the-art language model is used to compute semantic similarity of the tokens in the question against the words within the search space, returning 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 displayed if the mean similarity score meets an empirically determined minimum threshold. If it fails to meet the minimum threshold, it is considered a failed query and the user is alerted that the question could not be answered.

The User Query Dataset: We keep track of the user questions and the status of the results in a database, along with other details like 'Date', 'Time' and 'Result (Pass/Fail)'. This helps us identify which queries are failing, allows us to broaden the scope of our methodology, and also guides improvement to the sentence similarity model. Furthermore, there is a feedback option where users can submit their satisfaction 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.

Current Implementation: We support spatial questions for the current state of the pandemic, i.e., 3W1H with the exception of 'When'. 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.

Future Work: We are adding support for the temporal 'When' questions, and exploring state-of-the-art sentence similarity models to incorporate them in future releases. We are also exploring the expansion of our 4W1H input space to add support for Tests to the 'Which' handling, and for Peak, 7-day Moving Average, Test positivity rates, etc. to the 'What'.

VIII. Utility of the Dashboard

Data storytelling: An important application of our 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 [30] and data journalism. The analytics of the dashboard, along with its historical data and interactive visualizations, promote 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 [31]. Owing to space limitations, we omitted illustrative examples of data storytelling, which can be found in our longer version [32].

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 (Department of Defense (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 counter-factual analysis and answer policy questions, including resource allocation and augmentation, and campus reopening and management. See [33] for reports that the Virginia Department of Health releases based on our work.

Several other groups use our dashboard and associated data as well. We list a few to illustrate its broad use: (i) it is listed

as a part of the NIH MIDAS data catalogue [34], the ESRI COVID-19 GIS Hub [35] 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 [36] and Panama City Beach [37] in Florida, where our active case counts are used as one of their thresholds for allowing vacation rental reservations.

Web Traffic: During the initial phase, we had around 40,000 users in total and by mid-April, we reached over 750,000 users. By late October, there were over 1.13 million unique users. The top 3 countries that made up the largest portion of users are the United States, India and Canada. 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.

IX. Discussion and Conclusion

We presented our COVID-19 Surveillance Dashboard in support of the pandemic planning and response. Our experience suggests developing a standard based on 6Cs metrics to improve epidemic data management and optimization.

As with any software engineering project that builds a data intensive pipeline, we faced a number of challenges; some stemmed from the real-time nature of the problem, as well as our goal of releasing our system as quickly as possible. Important challenges included: (i) getting reliable data, (ii) data sources updating their terms of use, (iii) lack of standards (e.g. file formats, naming conventions, update cycles) across various data sources, (iv) inconsistency in reporting recovered cases that, in turn, affects the number of active cases, and (v) challenging software test cycles due to short sprints in a rapidly evolving environment.

Important lessons learned during development of the dashboard include: (i) there is a clear need to build such systems so that we are better prepared for the next pandemic; (ii) the challenges are not so much in the software technologies, but in data availability and sharing; (iii) further adoption of the 6Cs standard and formal evaluation of emerging data sources would facilitate development of similar systems.

Acknowledgements: This work was partially supported by the National Institutes of Health (NIH) Grant 1R01GM109718, NSF BIG DATA Grant IIS-1633028, NSF DIBBS Grant ACI-1443054, DTRA subcontract/ARA S-D00189-15-TO-01-UVA, NSF Grant No. OAC-1916805, NSF Expeditions in Computing Grant CCF-1918656, CCF-1917819, NSF RAPID CNS-2028004, NSF RAPID OAC-2027541, US Centers for Disease Control and Prevention 75D30119C05935, and a collaborative seed grant from the UVA Global Infectious Disease Institute.

References

[1] 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.
[2] Tong Yang et al. Covidnet: To bring the data transparency in era of covid-19. *arXiv preprint arXiv:2005.10948*, 2020.
[3] Worldometer: Coronavirus update. https://www.worldometers.info/world-population/population-by-country/.
[4] WHO Coronavirus disease (COVID-19) dashboard. https://covid19.who.int/.
[5] CDC COVID data tracker. https://www.cdc.gov/covid-data-tracker/.
[6] ECDC COVID-19 data. https://www.ecdc.europa.eu/en/covid-19/data.
[7] Benjamin Wissel et al. 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.
[8] Stefano Barone et al. Building a statistical surveillance dashboard for covid-19 infection worldwide building a statistical surveillance dashboard for covid-19 infection worldwide. *Quality Engineering*, 06 2020.
[9] Alexander Hohl et al. 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.
[10] Bo Xu et al. Epidemiological data from the covid-19 outbreak, real-time case information. *Scientific data*, 7(1):1–6, 2020.
[11] US Conronavirus cases and deaths. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.
[12] Wikipedia. https://en.wikipedia.org/wiki/COVID-19_pandemic_by_country_and_territory.
[13] COVID-19 India. https://www.covid19india.org/.
[14] Government of Canada. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html.
[15] Government of Greece. https://covid19.gov.gr/.
[16] The COVID tracking project. https://covidtracking.com/.
[17] Worldometer: Population by country (2020). https://www.worldometers.info/coronavirus/.
[18] World population review: Us states. https://worldpopulationreview.com/states.
[19] Worldatlas: Chinese provinces by population. https://www.worldatlas.com/articles/chinese-provinces-by-population.html.
[20] Adc worldmap. https://www.adci.com/adc-worldmap/.
[21] ArcGIS Online. https://www.esri.com/en-us/arcgis/products/arcgis-online/overview.
[22] 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.
[23] 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 72314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, 323(13):1239–1242.
[24] 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.
[25] Illinois DPH : COVID-19 Statistics. https://www.dph.illinois.gov/covid19/covid19-statistics.
[26] Geoffrey Fairchild et al. Epidemiological data challenges: planning for a more robust future through data standards. *Frontiers in Public Health*, 6:336, 2018.
[27] ArcGIS API for JavaScript. https://developers.arcgis.com/javascript/3/.
[28] amCharts, JavaScripts & Map. https://www.amcharts.com/.
[29] 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.
[30] Brent Dykes. Data storytelling: The essential data science skill everyone needs. *Forbes*, 03 2016.
[31] Coronavirus: Why you must act now. https://medium.com/@tomaspueyo/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca.
[32] Akhil Sai Peddireddy et al. From 5vs to 6cs: Operationalizing epidemic data management with covid-19 surveillance. *medRxiv*, 2020.
[33] 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.
[34] MIDAS online portal for COVID-19 modeling research. https://midasnetwork.us/covid-19/.
[35] ESRI COVID-19 GIS hub. https://midasnetwork.us/covid-19/.
[36] Bay county plan for opening short-term rentals, phase-ii. https://www.baycountyfl.gov/CivicAlerts.aspx?AID=156.
[37] Panama city beach chamber COVID-19 updates. https://www.pcbeach.org/news-article/panama-city-beach-chamber-coronavirus-covid-19-updates/.