DeepTrack: Monitoring and Exploring Spatio-Temporal Data – A Case of Tracking COVID-19 –

Yuyu Luo† Wenbo Li† Tianyu Zhao† Xiang Yu† Lixi Zhang† Guoliang Li† Nan Tang‡
†Department of Computer Science, Tsinghua University ‡Qatar Computing Research Institute, HBKU
{luoyy18@mails., li-wb17@mails., zhaoty17@mails., x-yu17@mails., zhanglx19@mails., liguoliang@}tsinghua.edu.cn, ntang@hbku.edu.qa

ABSTRACT

Spatio-temporal data analysis is very important in many time-critical applications. We take Coronavirus disease (COVID-19) as an example, and the key questions that everyone will ask every day are: *how does Coronavirus spread? where are the high-risk areas? where have confirmed cases around me?* Interactive data analytics, which allows general users to easily monitor and explore such events, plays a key role. However, some emerging cases, such as COVID-19, bring many new challenges: (C1) New information may come with different formats: basic structured data such as confirmed/suspected/serious/death/recovered cases, unstructured data from newspapers for travel history of confirmed cases, and so on. (C2) Discovering new insights: data visualization is widely used for storytelling; however, the challenge here is how to automatically find "interesting stories", which might be different from day to day. We propose DeepTrack, a system that monitors spatio-temporal data, using the case of COVID-19. For (C1), we describe (a) how we integrate and clean data from different sources by existing modules. For (C2), we discuss (b) how to build new modules for ad-hoc data sources and requirements, (c) what are the basic (or static) charts used; and (d) how to generate recommended (or dynamic) charts that are based on new incoming data. The attendees can use DeepTrack to interactively explore various COVID-19 cases.

1. INTRODUCTION

Monitoring and exploring spatio-temporal data is important for many applications. For example, monitoring taxi movements for detecting and predicting traffic jams, and analyzing Twitter data for tracking or predicting the development of certain events such as US election. In 2020, the Coronavirus disease 2019 (COVID-19) is a disaster that significantly impacts everyone on the planet and is leading to great losses of wealth, health and life. Evidently, there is an emerging need to track COVID-19 cases.

Although visualizing spatio-temporal data has been widely studied [1, 4], there are some new challenges for COVID-19. (C1) New information may come with different formats: besides basic data such as confirmed/suspected/serious/death/recovered cases in CSV files, there are also text data from newspapers, JSON files for travel history of confirmed cases, and so on. (C2) Discovering new insights: data visualization is widely used for storytelling, however, the challenge here is how to find "interesting stories" in terms of charts, which might be different from day to day.

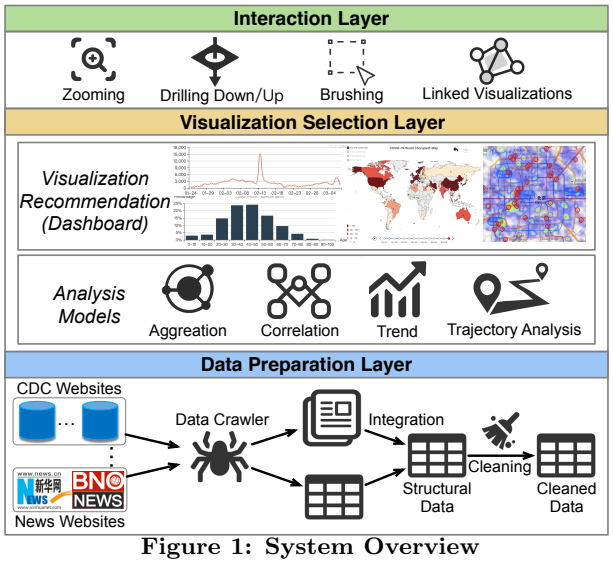

Figure 1: System Overview

An Overview of DeepTrack. We present DeepTrack, an end-to-end framework to prepare data, select visualizations and allow easy-to-use interactions. An overview of DeepTrack is given in Figure 1, which consists of three layers: *data preparation layer*, *visualization selection layer*, and *interaction layer*. Data preparation is responsible for crawling daily updated data from different sources, cleaning them when needed, which is to tackle C1 (see Section 2.1). Visualization selection describes the process of both which charts will be shown (*e.g.,* a heat map on a world map showing new cases of every country), and how to automatically recommend charts that are "interesting" *w.r.t.* new incoming data, for handling C2 (see Section 2.2). Interaction allows a user to explore various and (maybe) new COVID-19 stories in an interactive fashion (see Section 2.3).

DeepTrack-COVID-19: We have implemented a DeepTrack instance for COVID-19, which has attracted a broad range of interest from general users, public health authorities, and researchers who want to explore the COVID-19 data and track the outbreak. Since the system was launched

Proceedings of the VLDB Endowment, Vol. 12, No. xxx
ISSN 2150-8097.
DOI: https://doi.org/10.14778/xxxxxxx.xxxxxxx

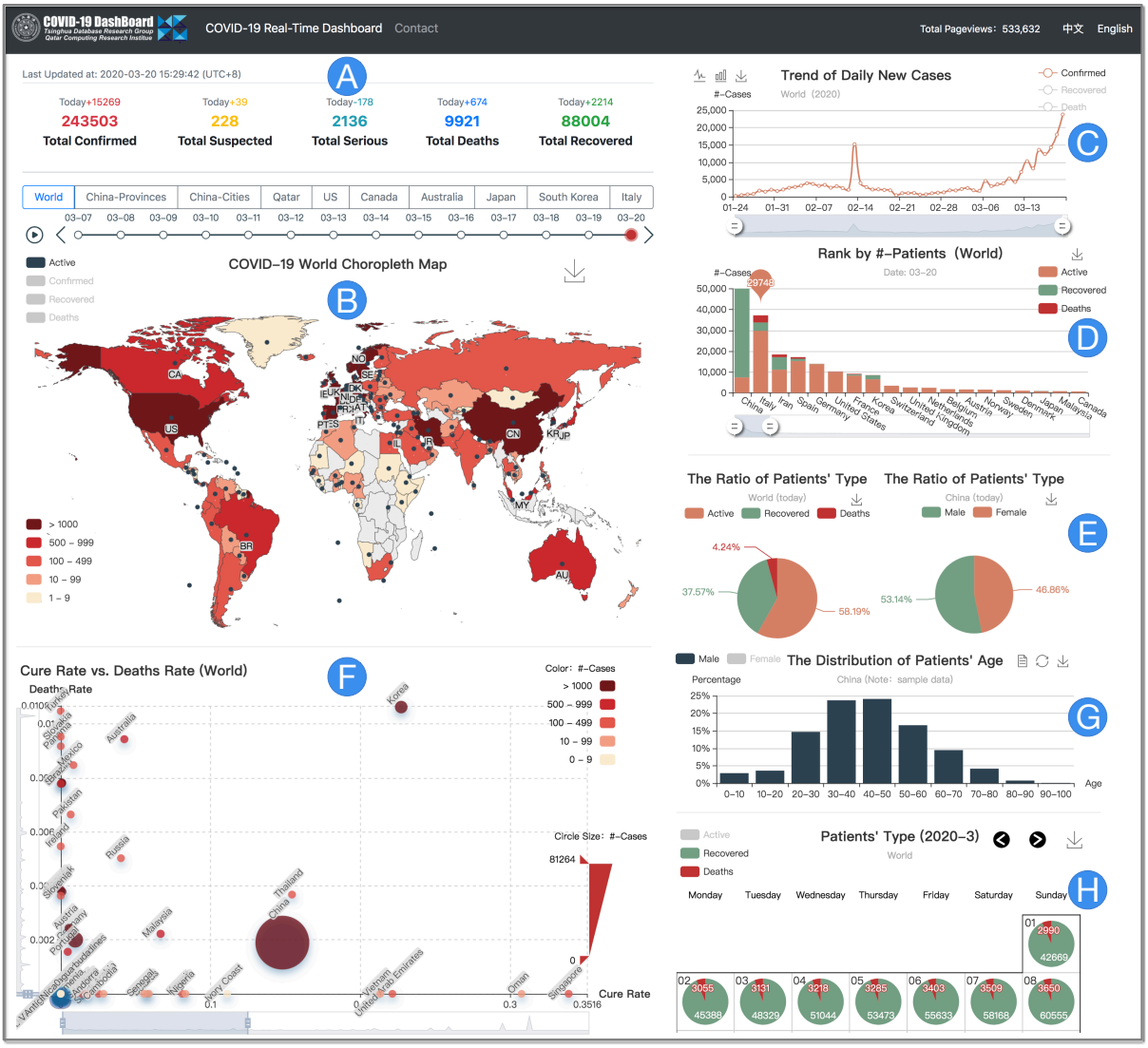

Figure 2: Demonstration of DeepTrack-COVID-19 (https://ncov.deepeye.tech/en)

on 07/02/2020, we have accumulated more than 1 million visits. Some news media in China and Qatar have reported our system, and we have launched online public courses to present our system, which attracted more than 100,000 people to participate. Moreover, some research institutions (*e.g.,* CMU, The University of Hong Kong) also conduct preliminary processing and analysis of COVID-19 epidemiological data based on our platform.

The base view of DeepTrack-COVID-19 is shown in Figure 2, which consists of a choropleth map for the total confirmed/recovered/died cases for all countries, line charts for the trend of daily increased cases, calendar charts for visualizing the types of patients, bar charts for distributions of patients' ages, bubble charts for cure rate - death rate, and so on. Besides the basic view page, it also has ad-hoc features, such as tracking infection path and high-risk areas discovery using the trajectory data of infected persons, which will be discussed in Section 3.

2. THE DEEPTRACK SYSTEM

DeepTrack has three layers (see Figure 1): (1) data preparation, (2) visualization selection, and (3) interaction. Next, we will explain each step using the COVID-19 case.

2.1 Data Preparation Layer

This layer has a predefined pipeline to prepare data, which will run periodically, with the following three steps.

Data Collection. We collect data from the following data sources. (1) We download hourly the official data from the Chinese Center for Disease Control and Prevention (CDC) and other countries' CDCs. (2) We crawl infected cases' age and gender from authoritative news websites. (3) We also connect to other data sources like Chinese population statistics. (4) We are also provided with trajectory data of (potentially) infected persons from China Mobile Limited[1].

Data Integration. Next, we need to integrate different types of data into a predefined relational table (*i.e.,* a global view). For example, we need to extract report date, location, patients' type, #-cases from each country's CDC's reports, and perform schema alignment into $S1$: (*Date, Country, State/Province, City, Total Cases, Active Cases, Total Deaths, Total Recovered, Deaths Rate, Recovered Rate, Gender*), a typical ETL-based data integration process.

[1]We are collaborating with the company and got mobile phone location data under privacy protection.

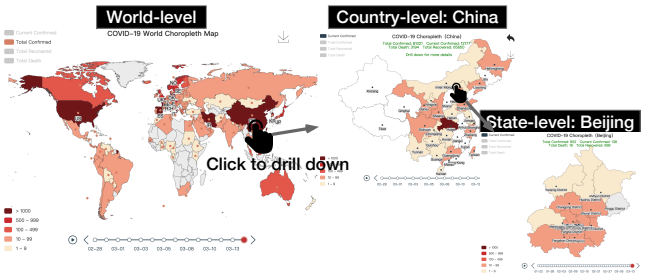

Figure 3: Drill Down Operation

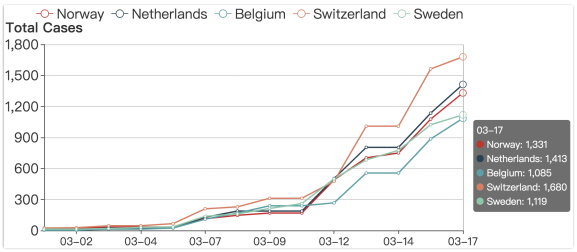

Figure 4: Similar Trend of Confirmed Cases

Data Cleaning. After integrating data from multiple sources, there are typical data errors such as duplicates, missing values, synonyms, and so on. Because data cleaning is known to be tedious and error-prone, we employ our recently proposed technique VisClean [2] for visualization-driven data cleaning, which is way cheaper than cleaning the entire dataset. This is doable only after the charts to display have been selected, as discussed below.

2.2 Visualization Selection Layer

Visualization selection generates three categories of charts: *linked* common visualizations, *ad-hoc* visualizations, and *recommended* visualizations.

Linked Common Visualizations. There are common visualizations for spatio-temporal data exploration, such as a choropleth map (a heat map on a map), line charts to show various trends, bar charts to show the comparison between various groups, scatter charts (or bubble charts) to quantify the relationship between two quantitative variables (*e.g.,* death rate vs. cure rate). We carefully selected charts (see Figure 2) that can attract a wide range of interest, and make them "linked", *e.g.,* when one zooms in from a world level to a country level, all the other charts will be zoomed in, so as to provide a synchronized view from multiple charts.

Ad-hoc Visualizations. We also design ad-hoc visualizations to answer specific questions. In terms of COVID-19, besides publicly available datasets, we also have private trajectory data of potentially infected persons. Based on which we have designed two map-based visualizations, one to show infection paths of these patients (see Figure 5), and the other to show the level of risk for each area (see Figure 6) and thus suggest the authorities to take different anti-epidemic policies for different areas.

Recommended Visualizations. The above common visualizations are fixed, with data being periodically updated. However, as the data keeps changing, some interesting stories cannot be captured by predefined common visualizations. Hence, it requires some mechanism to discover these new interesting visualizations, either manually or automatically. We leverage our previous work, a visualization recommendation system called DeepEye [3], to recommend interesting visualizations, such as finding cities in China that share similar trends as Wuhan in terms of death rate. The basic idea of DeepEye is to take a table as input, enumerates the valid visualizations of the dataset, select those good visualizations by a supervised classification model, and ranks top-k good visualizations by a learning-to-rank model.

2.3 Interaction Layer

This layer interacts with the users. When a user visits DeepTrack, he/she can further explore visualizations by interactive module for finding more interesting insights. DeepTrack supports popular interactions such as drilling down/up, zooming in/out and linked visualizations.

Take drilling down as an example (see Figure 3), when a user clicks a country (*e.g.,* China) on the world-level map, the map will drill down into the country-level map for more details. Note that, DeepTrack provides linked visualizations of the analytical results. That is, when a user performs a drilling down operations, other visualizations will also drill down into certain level automatically. In addition, the user can zoom in/out the map by rolling up/down the mouse.

3. DEMONSTRATION SCENARIOS

This section provides a walk-through demonstrating how our system can be used to monitor and explore the outbreak of COVID-19 interactively. Note that we only release parts of the system for public access[2], and other parts of the system only serve authorities (government) due to the privacy policy. However, we will demonstrate the full version of the system to VLDB 2020 attendees.

Overview of COVID-2019 via Linked Visualizations. The user can get high-level situations of COVID-19 from Figure 2. For example, the user can catch the overall information of the reported cases from Figure 2(A). The choropleth map in Figure 2(B) shows the location and number of confirmed cases, deaths and recoveries for all affected countries. It also provides a timeline toolbar for the user to look back upon previous situations, and a user can click the "▶" button to show an animation. The user can click a country, *e.g.,* China, to drill down into the country-level (province-level or city-level) for more details. Since we apply the linked visualization techniques, the rest of the visualizations will also drill down into the country-level. Figure 2(C), a line chart, illustrates the daily increased cases of the selected location. The stacked bar chart in Figure 2(D) depicts the number of cases for the selected location. The pie charts in Figure 2(E) show the proportion of patients' type. Figure 2(F), a bubble chart, illustrates the relationships across #-cases, deaths rate, and cure rate. The bar chart in Figure 2(G) shows the distribution of patients' age, and the calendar chart in Figure 2(H) illustrates the proportion of types of reported cases for each day.

Similar Trend Search. DeepTrack also supports the similar trend search functionality for finding similar trends. This feature is supported by DeepEye [3] in the back-end. For example, if the user wants to find those trends of confirmed cases that are similar to *Switzerland*, the similarity search functionality will return top-k similar trends about *Switzerland*. The running example is shown in Figure 4. Besides line charts, the similar trend search also supports

[2]Online System: `https://ncov.deepeye.tech/en`

Figure 5: Tracking Infection Path

other charts (*e.g.,* bar chart and pie chart). Thanks to this functionality, users can perform comparative analysis easier.

Tracking Infection Path. Based on the trajectory data of (potentially) infected persons, we can support to visualize and track the infection path at the high-level. Taking Figure 5 as an example, a person started from *Beijing Haidian Hospital* at 13:28 on 2020-02-28, and walked through several streets and finally arrived at his/her neighborhood. Therefore, we cluster those trajectories of infected persons to find a group of high-risk roads. Moreover, we devise a trajectory similarity search technique to find other trajectories similar to those trajectories of infected persons. It will help us to find and report the potentially high-risk groups. Thus, the authorities can take different anti-epidemic policies against people at different risk levels.

High-risk Areas Discovery. To further explore the trajectory data of (potentially) infected persons, we also design a visualization for the trajectory points using a heatmap. Figure 6 gives an at-a-glance understanding of the spatial distribution of the (potentially) infected persons. Those "hot" areas mean there have more (potentially) infected persons visit. We also indicate those neighborhoods that have several confirmed cases in the heatmap by the symbol 88. Based on the history data, we can know that there is a high overlap between areas where high-risk populations are active and neighborhoods containing confirmed cases. Therefore, we can infer that other places (*e.g.,* the red rectangles) visited by high-risk groups may also be dangerous. Therefore, the authorities can take different anti-epidemic policies against people at different risk levels. Thus, the authorities can take precautions against these areas in advance and take different anti-epidemic policies against different areas to achieve refined and effective management. For the general public, DeepTrack provides the module of finding confirmed cases in nearby neighborhoods. Take Figure 7 for example, users can understand the COVID-19 situations near *Tsinghua University, Beijing, China* by a location search box. Note that this module only supports for the mainland China area currently.

4. CONCLUDING REMARKS

There exist many systems for monitoring and analyzing spatio-temporal data, such as a dashboard for visually tracking the outbreak of COVID-19 [1] and a tweet stream sentiment analysis system for US election 2016 [4]. One lesson from the existing systems is that they are usually designed on a case-by-case basis and built from scratch, which cannot fully leverage the recent techniques for data integration and automatic visualization.

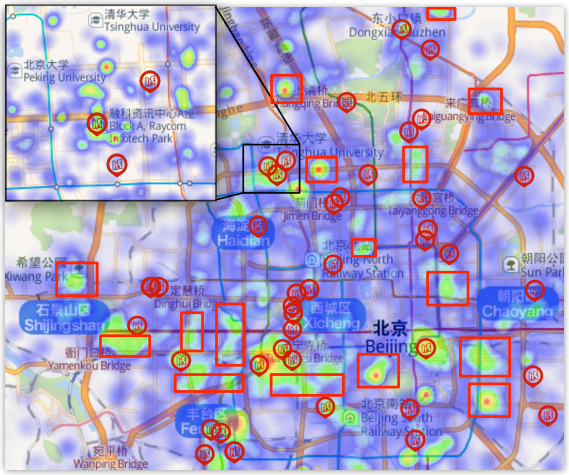

Figure 6: High-risk Area Discovery

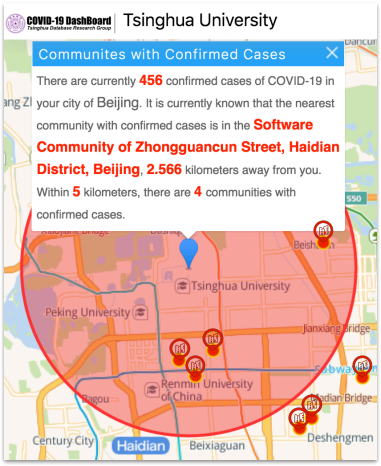

Figure 7: Find Confirmed Cases Around You

On the one side, DeepTrack-COVID-19 shares many common visualizations as the other popular websites for tracking COVID-19 cases. On the other side, it differs from the others in (1) DeepTrack-COVID-19 is based on a general end-to-end framework DeepTrack, and leverages recent techniques for data preparation (*e.g.,* VisClean [2]) and for visualization recommendation (*e.g.,* DeepEye [3]); (2) it supports linked visualization for the users to easily zoom in/out multiple visualizations by a single click; and (3) it also obtains some private data that is not publicly available, so it can demonstrate some unique features. Hopefully we can survive the war of fighting COVID-19 with the minimum cost, and by the time of VLDB 2020, we will have much more to demonstrate.

5. REFERENCES

[1] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 2020.

[2] Y. Luo, C. Chai, X. Qin, N. Tang, and G. Li. Interactive cleaning for progressive visualization through composite questions. In *ICDE*, 2020.

[3] Y. Luo, X. Qin, N. Tang, and G. Li. DeepEye: Towards Automatic Data Visualization. In *ICDE*, 2018.

[4] D. Paul, F. Li, M. K. Teja, X. Yu, and R. Frost. Compass: Spatio temporal sentiment analysis of US election what twitter says! In *SIGKDD*, 2017.