

# DEEPEYE: A Data Science System for Monitoring and Exploring COVID-19 Data

Yuyu Luo<sup>†</sup> Nan Tang<sup>‡</sup> Guoliang Li<sup>†</sup> Tianyu Zhao<sup>†</sup> Wenbo Li<sup>†</sup> Xiang Yu<sup>†</sup>

<sup>†</sup>Department of Computer Science, Tsinghua University    <sup>‡</sup>Qatar Computing Research Institute, HBKU

{luoyy18@mails., liguoliang@, zhaoty17@mails., li-wb17@mails., x-yu17@mails.}tsinghua.edu.cn, ntang@hbku.edu.qa

## Abstract

The COVID-19 pandemic is a global health crisis of our time that has significantly affected almost every single person on earth in just several months. Even worse, we do not know when it will slow down and how long it will last. Analogous to fighting the COVID-19 pandemic in the physical world, data scientists need to deal with the infodemic of COVID-19 data to discover useful insight in order to guide wise and informative decisions, where the COVID-19 infodemic refers to all (messy) data relevant to COVID-19. In this paper, we present DEEPEYE, an end-to-end data science system for monitoring and exploring COVID-19 data, which ranges from (task-driven) data preparation, (descriptive, diagnostic, and prescriptive) data analytics, user interactions through (linked) spatio-temporal data visualizations, and applications in different use cases.

## 1 Introduction

### 1.1 Where do we stand today on COVID-19?

**The history of pandemics.** Nothing has killed more people than infectious disease throughout the human history<sup>1</sup>. Many pandemics changed history, for example, the Antonine Plague (165-180) and the Black Death (1346-1353) have changed the history of Europe. Situations are getting worse in the last two decades, because epidemics happened much more frequently than what we have seen in history: SARS (2002-2003), Swine Flu (2009-2010), MERS (2012-present), Ebola (2014-2016), and now the COVID-19 (2019-present).

**The history of COVID-19.** Shortly after the first confirmed case in early January 2020, and a statement at January 21 from WHO's mission to China saying that there was evidence of human-to-human transmission, COVID-19 quickly spread out to almost every corner of the world. WHO officially named it a pandemic at March 11 2020. Till the date of June 1, 2020, there are more than 5.5 million confirmed cases and it has caused more than 350K deaths worldwide.

**No system to stop a pandemic.** Bill Gates envisioned, during his TED Talk on 2015<sup>2</sup>, that if something will kill more than 10 million people in the next few decades, it will be infectious disease rather than wars. He has also

---

Copyright 2004 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

<sup>1</sup><https://www.visualcapitalist.com/history-of-pandemics-deadliest/>

<sup>2</sup>[https://www.ted.com/talks/bill\\_gates\\_the\\_next\\_outbreak\\_we\\_re\\_not\\_ready?language=en](https://www.ted.com/talks/bill_gates_the_next_outbreak_we_re_not_ready?language=en)

questioned: “Are we ready for the next outbreak?” Sadly, the answer was not that the current health systems do not work for pandemics. Instead, there is *no global health system at all* for such pandemics.

**Where do we stand today?** Undoubtedly, COVID-19 has changed and will keep changing our history from many different dimensions, such as living style, studying style, the way of traveling, among many others. Indeed, we are not even at the peak of the 1st wave of COVID-19, and there might have the 2nd and the 3rd waves, such as the 1918 influenza pandemic and the 2009-2010 H1N1 pandemic, where the 2nd and 3rd waves were much more deadly than the 1st wave – we need to buckle up because the ride is just beginning. In order to stop or get better prepared to stop pandemics, global public health systems need to be ready, like military is ready for wars. Meanwhile, to help stop pandemics, *data science* has played a key role in discovering insights to guide decision makers and general people to make wise and informative decisions.

## 1.2 Do we have good data science systems for monitoring and exploring COVID-19?

**From the real world pandemics to the virtual world infodemics.** As the WHO Director-General Tedros Adhanom Ghebreyesus said<sup>3</sup>: “We’re not just fighting an epidemic; we’re fighting an *infodemic*.” Here, the term “infodemic” generally refers to an excessive amount of information about a problem, which makes it difficult to identify a solution. Similar to the pandemics in the real world, infodemics are in the virtual world, where the goals of scientists in many domains are to deal with the information from the infodemics of the virtual world, so as to find meaningful insights that can be used to fight against the pandemics of the real world.

**No data science system to stop an infodemic.** Unfortunately, similar to the case there is no global health system that is ready to stop a pandemic, there is no data science system that is ready to deal with infodemics, even if we have seen the fruitful successes from many communities for data science, such as database, data mining, machine learning, natural language processing, bioinformatics, and many others. Essentially, all scientific methods are based on empirical or measurable *evidence* that is subject to the principles of logic and reasoning. In terms of infodemics, the evidence is the data that has been collected. However, the central problems are: (i) *Data is inaccurate*: the real confirmed cases and the number of death are conjectured to be much higher than the reported numbers. (ii) *Data is missing*: nobody knows precisely when and where did COVID-19 start, therefore many data is missing from its origin to the date that the data was first collected. (iii) *Data is inconsistent*: there are misinformation, disinformation, and rumors that are widely spread. (iv) *Data is not directly comparable*: different countries calculate mortality rate in different measures, *e.g.*, the total deaths compiled by US CDC include ‘probable’ cases starting from April 16, 2020<sup>4</sup>, while UK mainly considers the cases from the hospitals from the department and health of social care (DHSC) data<sup>5</sup>.

**The goal of an ideal data science system for infodemics.** The *dark side* of infodemics is that *we never have the data well prepared for deriving the truth in any context*. Note, however, that the only certainty in (data) science is uncertainty, just like every single decision we have ever made. Hence, the *bright side of the dark side* is that this infodemics dilemma forces us to re-evaluate and re-design existing data science systems, for fighting against infodemics. Broadly speaking, the main goal of an ideal data science system for infodemics is about giving that data a purpose, which can provide hints to guide wise decisions. For example, although the reproduction number ( $R_0$ , pronounced R-nought or r-zero) of COVID-19 keeps changing, we are certain that *social distancing* is important. More concretely, an ideal data science system should be able to tackle all traditional data analytical tasks but under the situation of infomedics – data is very messy (the above i–iv) and new (and conflicting) data keeps coming – that can: (1) effectively collect, integrate and clean data from different sources; (2) make *descriptive analytics* to tell what happened in the past; (3) conduct *diagnostic analytics* to help

---

<sup>3</sup><https://www.who.int/dg/speeches/detail/munich-security-conference>

<sup>4</sup><https://edition.cnn.com/2020/04/15/health/us-coronavirus-deaths-trends-wednesday/index.html>

<sup>5</sup><https://www.health.org.uk/news-and-comment/blogs/understanding-the-data-about-covid-19-related-deaths>

understand why something happened in the past; (4) do *predictive analytics* to predict what will happen in the future; and (5) perform *prescriptive analytics* to recommend actions that one can take to affect those outcomes.

### 1.3 DEEPEYE: A small step towards a (ideal) data science system for infodemics

In this paper, we present DEEPEYE, an end-to-end data science system for collecting, cleaning, analyzing, and visualizing COVID-19 data. Since the system was launched in 2020/02/05, we have accumulated more than 2 million visits in a month. Some news media in China have reported our system<sup>67</sup>, and we have launched a series of online public courses to give tutorial for the system<sup>8</sup>, which attracted nearly 100,000 people to participate. Besides, as a research-oriented platform, 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.

Generally speaking, DEEPEYE consists of three layers: (task-driven) data preparation layer, (smart) data analytics layer, and user interaction layer (Section 2).

In (*task-driven*) *data preparation layer*, a key observation we made is that it is impossible to prepare the data in the traditional way for some tasks, simply because data preparation is expensive and error-prone, especially when the data keeps changing every day. In particular, we will discuss task-driven data preparation, with the basic intuition that it is much cheaper to prepare the data that is needed for a given task (Section 3).

The (*smart*) *data analytics layer* contains several components. For *descriptive analytics*, we use linked data visualization to provide sufficient context for a user to understand what happened in the past. We also use visualization recommendation techniques that can automatically discover interesting stories (Section 4). For *diagnostic analytics*, we show that by combining with other (domain) knowledge, we can test our hypotheses (*e.g.*, the effect of urban (population) density or temperature to COVID-19) about why something happened (Section 5). For *predictive analytics*, we have tried some predictive model, in particular Susceptible-Exposed-Infectious-Recovered (SEIR) [3], that has been widely used for epidemiology. However, our empirical results show that the prediction for COVID-19 is typically inaccurate that may due to the messy data of infodemics, hence we will not discuss further about it in this paper. For *prescriptive analytics*, we will discuss our collaboration with China mobile that help the government to recommend actions (Section 6).

The discussion of *user interaction layer* will be blended with the discussion of data analytics layer, while describing various use cases.

## 2 An Overview of DeepEye

### 2.1 System Architecture

We present DEEPEYE, an end-to-end framework to prepare data, select visualizations and allow easy-to-use interactions. An overview of DEEPEYE is given in Figure 1, which consists of three layers: (*task-driven*) *data preparation layer*, (*smart*) *data analytics layer*, and *user interaction layer*. Data preparation is responsible for crawling daily updated data from different sources, and cleaning them when needed (Section 2.2). Data analytics describes the process of both which charts will always be shown (*e.g.*, a heat map on a world map showing new cases of every country), and how to automatically recommend visualizations that are “interesting” *w.r.t.* new incoming data, for visual analytics (Section 2.3). Interaction allows a user to explore various and (maybe) new COVID-19 stories in an interactive fashion (Section 2.4).

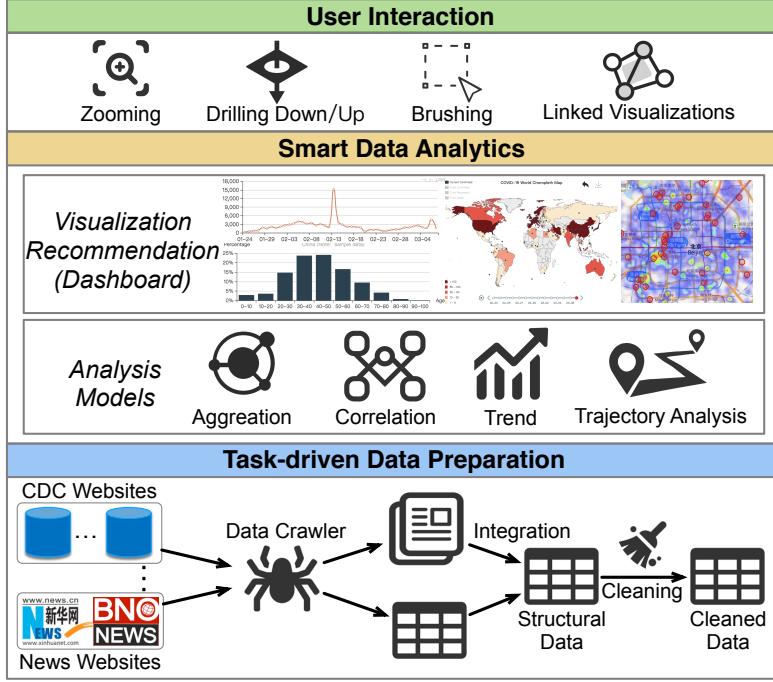


Figure 1: System Overview

## 2.2 Task-driven 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 population statistics, temperature data, and so on. (4) We are also provided with trajectory data of (potentially) infected persons from China Mobile Limited<sup>9</sup>.

**Data Integration.** Next, we need to integrate different types of data into predefined relational tables (*i.e.*, global views). 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, Gender, Total Confirmed, Active Confirmed, Total Deaths, Total Recovered, Death Rate, Recovered Rate*), a typical ETL-based data integration process.

**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 [8] for visualization-aware 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. We will depict more details about visualization-aware data cleaning in Section 3.

<sup>6</sup><https://news.tsinghua.edu.cn/info/1416/77464.htm>

<sup>7</sup>[https://www.aminer.cn/research\\_report/5e774aa1e34bad84366f1f7e?download=false](https://www.aminer.cn/research_report/5e774aa1e34bad84366f1f7e?download=false)

<sup>8</sup>[https://www.bilibili.com/video/BV1FE411W7p4/?spm\\_id\\_from=333.788.videocard.0](https://www.bilibili.com/video/BV1FE411W7p4/?spm_id_from=333.788.videocard.0)

<sup>9</sup>We collaborate with the company and get mobile phone location data under privacy protection.

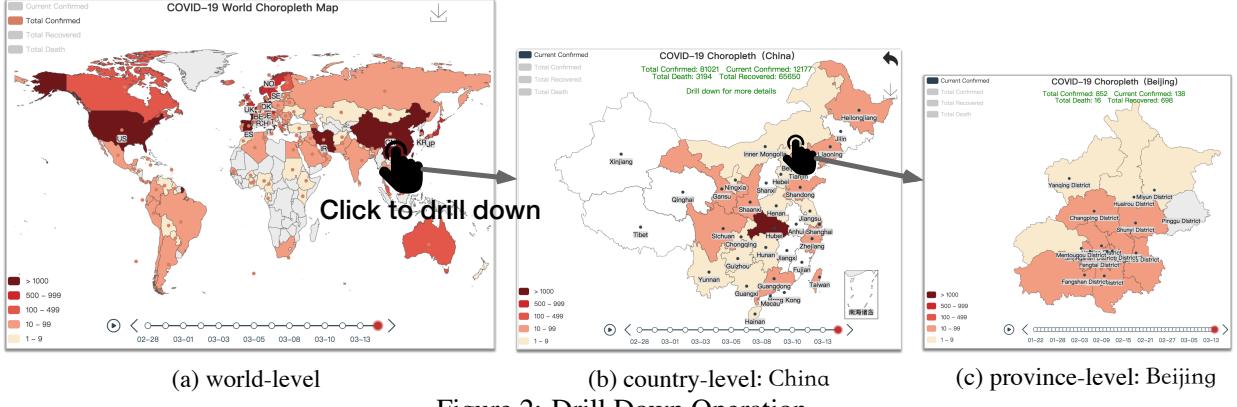


Figure 2: Drill Down Operation

### 2.3 Smart Data Analytics Layer

Based on the availability and reliability of data and meta-data, we have successfully conducted the following types of data analytics.

**Descriptive analytics.** We use linked data visualization and visualization recommendation algorithms to effectively show what happened in the past.

**Diagnostic analytics.** We use maps with different layers to test the spatio-temporal properties of COVID-19 data, especially to show the effect of urban (population) density and temperature to the outbreak of COVID-19.

**Prescriptive analytics.** Based on the collaboration with companies to get private data, we were able to do some meaningful prescriptive analytics that can recommend actions to decision makers.

## 2.4 User Interaction Layer

This is the interface we present to the public. When a user visits DEEPEYE, he/she can further explore visualizations by interactive module for finding more interesting insights. DEEPEYE supports popular interactions such as drilling down/up, zooming in/out and linked visualizations, powered by the visualization library ECharts [6].

Take drill down as an example (see Figure 2), 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, DEEPEYE provides linked visualizations of the analytical results. That is, when a user performs a drill down operation, 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 Task-driven Data Preparation

In the era of big data, from the data perspective, the data can come from governments, business companies, Web, and so on. As shown in Figure 1, in the scenario of COVID-19, the data can come from WHO, each country's Center for Disease Control and Prevention (CDC), news websites, microblog text, and other databases (*e.g.*, population, temperature and international airline data). In addition to various data sources, how to integrate data with different data formats (*e.g.*, table, JSON-like, and text) from different data sources is another problem. One concrete example is how to extract the data from the CDC daily reports and then align such data into predefined relational tables. In this work, we develop an algorithm to crawl, extract data and fill them into relational tables.

Clean data is one of the essential requirements for data analytics and accurate analysis results [1, 2]. However, data collection and integration usually introduce errors (*e.g.*, missing values, mixed formats, and duplicates)

地区 (City)			性别 (Gender)			年龄 (Age)			行程描述 (Trajectory Description)		
蚌埠 Bengbu			男 Male			45			经营一家房地产中介公司，常年居住在武汉市江岸区 Operates a real estate agency company and lives in Jiang'an District of Wuhan City		
蚌埠 Bengbu			女 Female			49			现住址为蚌埠市经开区长淮卫镇卫东村 The current address is Weidong Village, Changhuaiwei Town, Jingkai District, Bengbu		
蚌埠 Bengbu			男 Male			45			患者为上海铁路局集团公司乘务员，住蚌埠市经开区湖滨社区C区4栋。工作线路：蚌埠–北京西(K148, K147) Working for Shanghai Railway and living in Building 4, District C, Hubin Community, Jingkai District, Bengbu City. Journey: Bengbu Station–Beijing West Station(K148, K147)		
蚌埠 Bengbu			女 Female			42			疑似居住于田家庵区 Suspected of living in Tianjiaan District		
蚌埠 Bengbu			男 Male			45			家住蚌埠市经开区湖滨社区C区4栋，上海铁路局集团公司员工。工作线路：蚌埠–北京西 (K148, K147) He lives in Building 4, District C, Hubin Community, Jingkai District, Bengbu City. He works for Shanghai Railway, Journey: Bengbu Station–Beijing West Station(K148, K147)		

(d) duplicates

Figure 3: Examples of Data Errors

in data due to the integration of heterogeneous data sources [7]. As shown in the Figure 3, there are four types of common data errors in the scenario of COVID-19. First, the synonyms usually appear when integrating data from different sources, especially when the data standard of data sources may change. The second data error is missing value, as shown in Figure 3(b). This type of data error is usually caused by the data source not being updated in time. More concretely, on the 05/02, the data source may not have published the total number of deaths on that day, so a missing value was introduced. The missing value imputation is usually estimated using the data of the first few days and later, or imputation by querying other data sources. Similar to the missing value, as shown in Figure 3(c), it may also introduce incorrect values. We use the average values of the data of the first few days and later as a repairing candidate to correct the wrong value. As shown in Figure 3(d), we crawl the travel records of some infected people from the news website. However, because data is crawled from multiple websites and then integrated, it is easy to insert duplicate records. For duplicate records, we utilize the entity matching algorithm to find possible duplicates [5].

Because the analysis scenarios of COVID-19 have high requirements on data quality, we first need to ensure the data quality, and then consider how to reduce the cost of data cleaning. However, it is too expensive and infeasible to resolve all errors. Therefore, we discuss how to conduct problem-oriented (task-driven) data cleaning for COVID-19. The key idea is that we only clean those data relevant to the data analytics tasks, which can reduce the cost of data cleaning. In other words, instead of applying cleaning tools to fully clean each type of data errors before visualization, we aim to clean those data errors that have heavy impact on the accuracy of the visualization results. We devise our recent technique, a framework for visualization-aware data cleaning, called VISCLEAN to achieve this goal. The key idea is that we first generate visualizations based on the possible dirty dataset, and then run cleaning tools in the backend to find possible errors and their repairing candidates. Next, we organize those data errors and their repairing candidates by a graph model. Based on the graph and a predefined benefit model, we aim to select the most “beneficial” questions (*i.e.*, an induced subgraph) to interact with the user. The larger the benefit is, the higher quality of visualization improvement is. Once the user answers the data cleaning questions, the system will fix data errors and refresh the visualization.

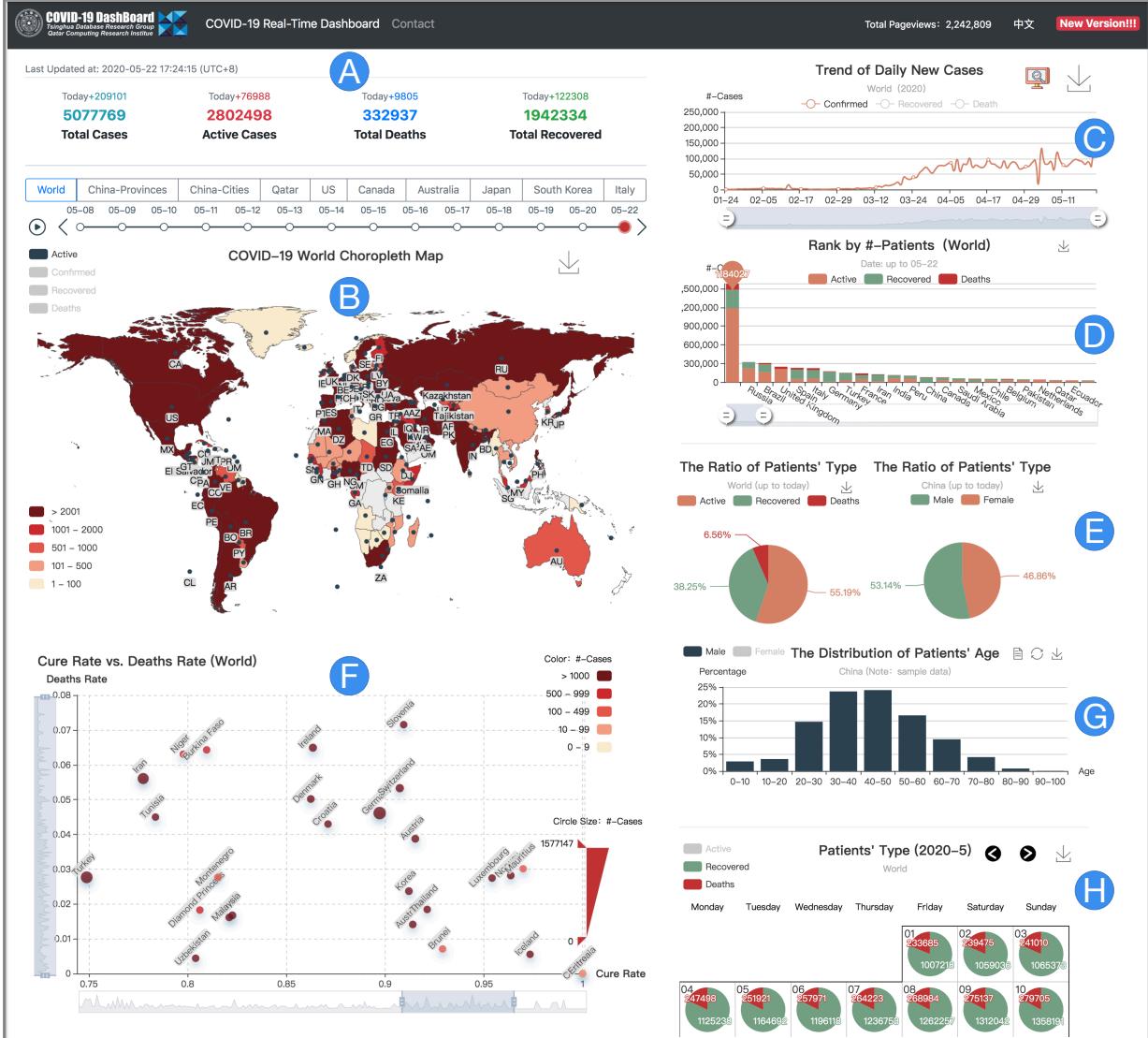


Figure 4: The Frontend of DEEPEYE

## 4 Descriptive Data Analytics of COVID-19

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

**(General) Linked Visualizations.** There are common visualizations for spatio-temporal data exploration, such as a choropleth map (a heat map based 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 4) 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. The user can get high-level situations of COVID-19 from Figure 4. For example, the user can catch the overall information of the reported cases from Figure 4(A). The choropleth map in Figure 4(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.

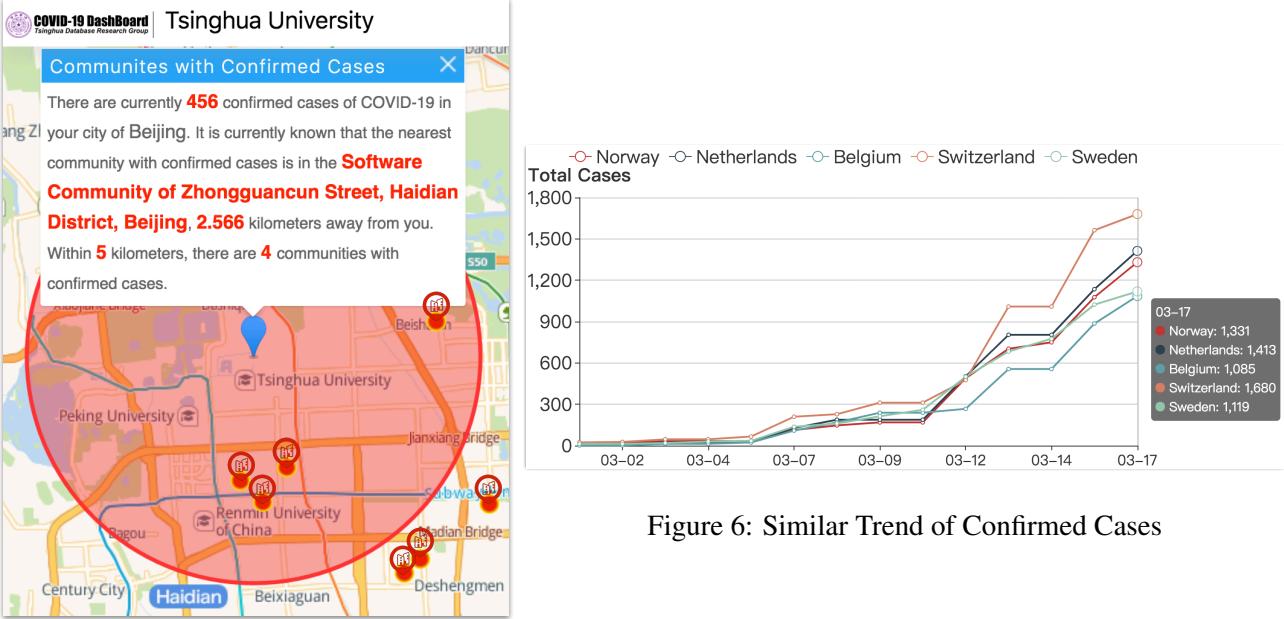


Figure 5: Find Confirmed Cases Around You

Since we apply the linked visualization techniques, the rest of the visualizations will also drill down into the country-level. Figure 4(C), a line chart, illustrates the daily increased cases of the selected location. The stacked bar chart in Figure 4(D) depicts the number of cases for the selected location. The pie charts in Figure 4(E) show the proportion of patients’ type. Figure 4(F), a bubble chart, illustrates the relationships across #‐cases, death rate, and cure rate. The bar chart in Figure 4(G) shows the distribution of patients’ age, and the calendar chart in Figure 4(H) illustrates the proportion of types of reported cases for each day.

**Location-based Search.** For the general public, DEEPEYE provides the module of finding confirmed cases in nearby neighborhoods. Take Figure 5 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.

**Similarity Trends Discovery.** DEEPEYE also supports the similar trend search functionality for finding similar trends. 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 6. Besides line charts, the similar trend search also supports other charts (*e.g.*, bar chart and pie chart). Thanks to this functionality, users can perform comparative analysis easier.

**Automatic News Generation.** Automatic news generation, in other words, automatically extracting insights from data visualization is a promising research and practical direction [11]. Currently, the user usually interacts with the visualization dashboard to get insights and make decisions. For example, the reporter may interact with the dashboard to observe the trend of daily new confirmed cases of each country/state and find a set of similar trends (or find a set of rapidly increasing trends) as news stories. In this scenario, it heavily relies on the user to manually get insights and write a news release. One intuitive idea is whether we can derive insights (news stories) from the visualization dashboard automatically. Roughly speaking, given a set of visualizations  $V$  and a news generation model  $M$ , the automatic news generation problem is to output a set of new stories  $S$ . The key challenge is how to design the news generation model  $M$ . One straightforward approach is to predefined a set of expert knowledge rules to mine insights from the visualization dashboard. Such rules can be similar trends

discovery, outlier trend detection, trends comparison, and so on.

## 5 Diagnostic Data Analytics of COVID-19

Another purpose of data visualization is to perform hypothesis testing, we show how to design visualizations to test two hypotheses – urban (population) density vs. total confirmed cases, and temperature vs. total confirmed cases.

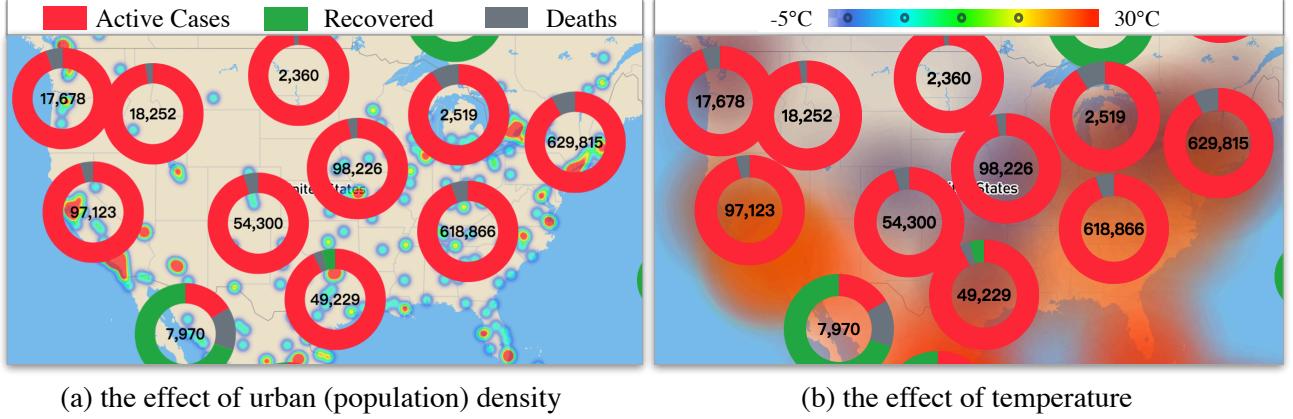


Figure 7: Diagnostic Data Analytics (Case in United States)

**Urban (Population) Density.** One intuitive hypothesis is that whether high population densities catalyze the spread of COVID-19? Since we want to know the relationship (correlation) between the population density and the spread of COVID-19, we first visualize the population density in the map named *population density map*, and then we map the confirmed cases on the top of *population density map*. As shown in Figure 7(a), it takes United States as an example. It shows that in areas with high population density and without lockdown policy, e.g., New York and California, more people are infected with coronavirus. For example, New York with a relatively high population density is likely more vulnerable to the spread of the coronavirus. This conclusion is reasonable, because the intensive contact greatly increases the probability of coronavirus transmission [12].

**Temperature.** We also design visualization to show the relationship between the outbreak of COVID-19 and the temperature factor. Similarly, we first visualize heatmap using temperature data, and then we map the confirmed cases on the top of the heatmap. As shown in Figure 7(b), it is hard for us to make conclusions like the higher temperature, the more infected cases, or the lower temperature, the less infected cases. For example, the average temperature in the central United States is lower than in California, but there are also hundreds of thousands of infected people in the central United States. Comparing Figure 7(a) and Figure 7(b), we can find that under the background of no lockdown, the population density has a stronger correlation with the number of confirmed cases.

## 6 Prescriptive Data Analytics of COVID-19

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 8), and the other to show the level of risk for each area (see Figure 9) and thus suggest the authorities to take different anti-epidemic policies for different areas.



Figure 8: Tracking Infection Path

## 6.1 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 8 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.

**Findings and Insights.** According to the sample trajectory data of (potentially) infected persons provided by China Mobile Limited, we find that most of the trajectories passed through some living places such as restaurants and supermarkets, and finally return home. Some trajectories intersect with public transportation such as trains and subways. There are a few trajectories with several other trajectories that coincide, indicating that they may be traveling together. Since many trajectories have visited supermarkets and other living places, the government can suggest that (1) people go to supermarkets as little as possible; (2) customers go shopping alone if possible; (3) enforcing people's shopping hours; and (4) limiting the number of customers in supermarkets at any one time.

## 6.2 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 9 gives an at-a-glance understanding of the spatial distribution of the potentially infected persons from *Hubei province, China* to *Beijing, China*. Those “hot” areas mean there are more potentially infected persons visit. We also indicate those neighborhoods that have several confirmed cases in the heatmap by the symbol

**Findings and Insights.** According to the location data of potentially infected persons from Hubei province to Beijing provided by China Mobile Limited, we make the following observations: (1) Most of the potential persons arrive in Beijing through transportation hubs such as railway stations ( in Figure 9) and airports. Intuitively, such transportation hubs likely play a significant role in the spread of the coronavirus in Beijing City. Therefore, the government should take corresponding measures for these transportation hubs to minimize the risk of the transmission of coronavirus; (2) Many potentially infected persons visit some commercial places (e.g., supermarkets) and end up staying in their residence; and (3) There is a high overlap between areas where

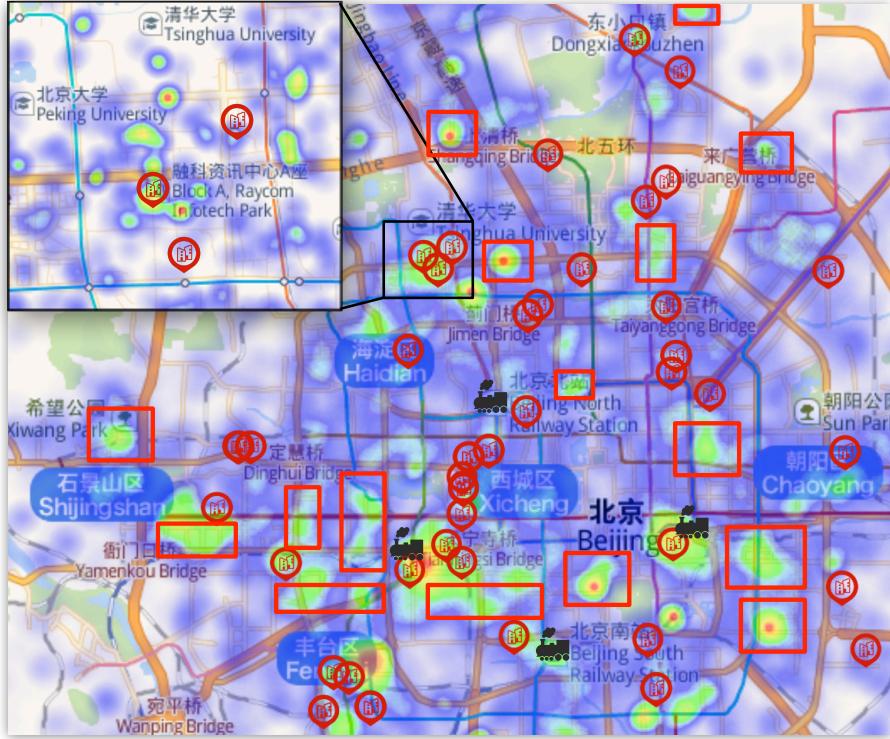


Figure 9: High-risk Area Discovery

high-risk people are active and neighborhoods containing confirmed cases. Intuitively, 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. Note that, for those “hot” areas which do not appear confirmed cases (*e.g.*, the red rectangles), some of them are later reported confirmed cases by the government. 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.

## 7 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 [4] and a tweet stream sentiment analysis system for US election 2016 [10]. 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.

On the one side, DEEPEYE-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) DEEPEYE-COVID-19 is based on a general end-to-end framework DEEPEYE, and leverages recent techniques for data preparation (*e.g.*, VIS-CLEAN [8]) and for visualization recommendation (*e.g.*, DEEPEYE [9]); (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.

## References

- [1] Z. Abedjan, X. Chu, and et.al. Detecting data errors: Where are we and what needs to be done? *PVLDB*, 9(12):993–1004, 2016.
- [2] C. Binnig, L. D. Stefani, T. Kraska, E. Upfal, E. Zgraggen, and Z. Zhao. Toward sustainable insights, or why polygamy is bad for you. In *CIDR*.
- [3] J. M. Carcione, J. E. Santos, C. Bagaini, and J. Ba. A simulation of a covid-19 epidemic based on a deterministic seir model. *arXiv preprint arXiv:2004.03575*, 2020.
- [4] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 2020.
- [5] P. Konda, S. Das, P. S. G. C., A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. F. Naughton, S. Prasad, G. Krishnan, R. Deep, and V. Raghavendra. Magellan: Toward building entity matching management systems. *Proc. VLDB Endow.*, 9(12):1197–1208, 2016.
- [6] D. Li, H. Mei, Y. Shen, S. Su, W. Zhang, J. Wang, M. Zu, and W. Chen. Echarts: A declarative framework for rapid construction of web-based visualization. *Vis. Informatics*, 2(2):136–146, 2018.
- [7] G. Li. Human-in-the-loop data integration. *Proc. VLDB Endow.*, 10(12):2006–2017, Aug. 2017.
- [8] Y. Luo, C. Chai, X. Qin, N. Tang, and G. Li. Interactive cleaning for progressive visualization through composite questions. In *ICDE*, 2020.
- [9] Y. Luo, X. Qin, N. Tang, and G. Li. DeepEye: Towards Automatic Data Visualization. In *ICDE*, 2018.
- [10] 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.
- [11] X. Qin, Y. Luo, N. Tang, and G. Li. Making data visualization more efficient and effective: a survey. *VLDB J.*, 29(1):93–117, 2020.
- [12] J. Rocklöv and H. Sjödin. High population densities catalyse the spread of covid-19. *Journal of travel medicine*, 27(3):taaa038, 2020.