Spatial-Temporal Data Science of COVID-19 Data

Deyu Deng, Carson K. Leung(✉), Chenru Zhao, Yan Wen, Hao Zheng
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
✉ kleung@cs.umanitoba.ca

2021 IEEE 15th International Conference on Big Data Science and Engineering (BigDataSE) | 978-1-6654-0038-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIGDATASE53435.2021.00011

***Abstract*—Big data are emerging paradigm that can be applied to huge volume of valuable data, which are often generated or collected at a fast velocity from a wide variety of rich data sources. These data can be of a wide variety of formats and/or type; they can be at different levels of veracity. Embedded in these data is implicit, previously unknown and useful information and knowledge that can be discovered by data science. Healthcare and medical data such as epidemiological data for disease like coronavirus disease 2019 (COVID-19) are examples of big data. Analyzing and mining these data led to discovery of knowledge and information about the disease, which in turn help people to get better understanding of the disease so that they could take parts in preventing or slowing down the spread of the disease, and/or protecting themselves from the disease. Hence, in this paper, we present a data science engine to analyze and mine COVID-19 data. As COVID-19 cases may not evenly distributed among spatial locations and/or evenly distributed throughout the entire period of pandemic, our engine conducts spatial-temporal data science to reveal important information and knowledge about epidemiological characteristics of the disease across different spatial locations and its temporal trends. Evaluation on real-life COVID-19 data demonstrates the effectiveness of our engine in conducting spatial-temporal data science of COVID-19 data.**

Keywords*—*data science, coronavirus disease, COVID-19, big data, data mining, data analytics, data visualization, big data applications, epidemiological data

I. Introduction and Related Works

Big data [1-3] are emerging paradigm that can be applied to data whose *volume* is so huge that is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. These big data, which can be of various formats and/or types, are often generated or collected at a fast *velocity* from a wide *variety* of rich data sources. Examples of big data include:

- audio and video such as music data [4, 5];
- biodiversity data [6];
- biomedical and healthcare data, together with disease reports [7-10];
- census data [11];
- meteorological data [12];
- patent records [13, 14];
- social media and social network data [15-18];
- time series [19-22];
- transportation and urban data for smart city [23-26]; and
- web content and web logs [27].

Moreover, these big data can be of a high *value* and at different levels of *veracity* in establishing trust in it for business decision making. Hence, massively parallel processing databases, scalable storage systems, cloud computing platforms, and/or high-performance computing techniques (e.g., MapReduce, edge computing, fog computing, dew computing) have supported the handling of big data.

Besides big data handling, *big data science* [28, 29] is also in demand for discovering implicit, previously unknown and useful information and knowledge embedded in the big data. In general, data science makes good use of data analytics [30-32], high-performance computing [33-35], visual analytics techniques [36, 37], and/or data mining and machine learning. Analyzing and mining these big data can be for social good. For instance, analyzing and mining the healthcare data and disease reports helps people to get a better understanding of diseases.

Over the past century, there have been some notable diseases including 1918 "Spanish flu" pandemic (1918-1920), 1957-1958 "Asian flu" pandemic, 1968 "Hong Kong flu" pandemic (1968-1970), 2009 "Swine flu" pandemic (2009-2010), and coronavirus disease 2019 (COVID-19). The latter broke out in 2019, became pandemic in 2020, and is still prevailing in 2021.

Since the COVID-19 pandemic, researchers from different disciplines have explored various aspects of COVID-19. For instance, there have been studies on:

- managing risks and crises faced by individuals and businesses (including employers and employees) due to the COVID-19 outbreak [38];
- analyzing the social and economic impacts of COVID-19 [39];
- building mathematical and statistical models—such as the susceptible-infectious-recovered (SIR) compartmental infectious disease model in epidemiology—to predict the spread of COVID-19 [40];
- developing data science solutions to analyze and mine COVID-19 data (e.g., epidemiological data) [41, 42];

978-1-6654-0038-1/21/$31.00 ©2021 IEEE
DOI 10.1109/BigDataSE53435.2021.00011

- conducting systematic reviews on literature about medical and health science research on COVID-19 [43, 44];
- focusing on clinical and treatment information [45];
- developing vaccines [46] such as (a) messenger ribonucleic acid (mRNA) vaccines (e.g., Moderna, Pfizer-BioNTech), (b) adenovirus vector vaccines (e.g., AstraZeneca, Janssen), (c) inactivated virus vaccines (e.g., CoronaVac, Covaxin, CoviVac) and (d) subunit vaccines.

Although COVID-19 vaccines have been developed, the vaccination process may take time. In the meantime, new COVID-19 cases are continuing to be reported. As a concrete example, there have been 426,453 new daily cases and 6,601 new daily deaths reported worldwide on July 10, 2021. Of which, 31,625 new daily cases and 527 new daily deaths were reported from North America (where 4.73% of total world population inhabited). These account for 7.42% and 7.98% of worldwide new cases and deaths, respectively. On the same day, 317 new cases and 6 deaths in Canada (which respectively account for 1.00% and 1.14% of new daily cases and deaths in North America, or equivalently 0.07% and 0.09% of new daily cases and deaths worldwide) where 6% of North American population or 0.5% of the world's total population inhabited.

Moreover, like many other viruses, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)—which causes COVID-19—also changes over time. Its mutations led to the reporting of SARS-CoV-2 variants[1], which can be classified as:

- variants of concern (VOC), which include alpha (lineage B.1.1.7), beta (lineages B.1.351, B.1.351.2 and B.1.351.3), gamma (lineages P.1, P.1.1, P.1.2, P.1.4, P.1.6, and P.1.7) and delta (lineages B.1.617.2, AY.1, AY.2, AY.3 and AY.3.1), which have first been reported in the UK. South Africa, Brazil and India, respectively. Among them, delta variants started as a VOI in April 2021 and rapidly evolved into a VOC in May 2021;
- variants of interest (VOI), which include eta (lineage B.1.525), iota (lineage B.1.526), kappa (lineage B.1.617.1) and lambda (lineage C.37), which have first been reported in multiple countries, USA, India and Peru, respectively; as well as
- other designated alerts for further monitoring, which include
    - lineages B.1.417 and B.1.429 (latter was parts of the former VOI epsilon with B.1.429), which have first been reported in the USA;
    - B.1.466.2, which has first been reported in Indonesia;
    - B.1.621 and B.1.1621.1, which have first been reported in Colombia; and
    - B1.1.318, B.1.1.519, C.36.3, C.36.3.1, and R.1, where have been reported in multiple countries.

Hence, there are demands for data science engines, which help discovery of knowledge and information about COVID-19. The knowledge discovery then allows people to get better understanding of the disease so that they could take parts in preventing or slowing down the spread of the disease, and/or protecting themselves from the disease.

By July 10, 2021, there have been cumulatively 186,492,674 COVID-19 cases worldwide, of which 4,029,920 deaths (i.e., approximately 2.16% of all cases).

For North America, there have been 40,047,453 COVID-19 cases (i.e., 21.47% of worldwide cases), of which 907,963 deceased (i.e., approximately 2.27% of all North American COVID-19 cases, or 22.54% of cumulative worldwide deaths). In Canada, there have been 1,427,899 COVID-19 cases (i.e., 3.57% of North American cases, or 0.77% of worldwide cases), of which 26,389 deceased (i.e., approximately 1.85% of all Canadian COVID-19 cases, 2.91% of cumulative North American deaths, or 0.65% of worldwide deaths).

Observed that many existing works focused on reporting the number of confirmed cases and mortality on a daily basis. Moreover, these cases are not evenly distributed among different spatial locations and/or time interval period during the pandemic. Hence, there are demands for spatial-temporal data science, which helps reveal important information and knowledge about epidemiological characteristics of the disease across different spatial locations and its temporal trends.

In response to the demands, we design and develop a data science engine that conducts spatial-temporal data science of textual-based COVID-19 epidemiological data. Our engine aims to discover common characteristics (beyond just the numbers of confirmed cases and mortality) among COVID-19 cases in a certain geographic location at a specific time interval, and compares them with those in other geographic locations and/or other time intervals. Such an engine also helps in answering questions like:

- What is the most common transmission method at a specific geographical location and/or during a certain time interval?
- How do transmission methods change among different geographical locations and/or over time?
- Similarly, what are the most common VOC, sets of symptoms, hospitalization status, and clinical outcomes at a specific geographical location and/or during a certain time interval?
- How do VOC, sets of symptoms, hospitalization status, and clinical outcomes change among different geographical locations and/or over time?

[1] https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

Our *key contributions* of this paper include the design and development of our data science engine that conducts spatial-temporal data science of COVID-19 epidemiological data.

We organize the remainder of this paper as follows. We first present our data science engine for spatial-temporal data in the next section, and then show our evaluation results on real-life COVID-19 data in Section III. Finally, we draw the conclusions in Section IV.

II. Our Spatial-Temporal Data Science Engine for COVID-19 Data

In this section, we describe our spatial-temporal data science engine for analyzing, mining and visualizing COVID-19 epidemiological data to reveal interesting spatial-temporal characteristics the data.

A. Data Integration

In general, statistics and details of COVID-19 cases in different geographical locations are generated and collected from a wide variety of data sources. As such, they can be captured in different formats and/or types. Different characteristics may be captured.

Let us consider a concrete example. As health care is a responsibility of provincial governments in Canada, Canadian COVID-19 data are gathered from different provinces. The provincial data are, in turn, gathered from their health administrative units called health authorities (which are also known as health regions) within the province.

- As an example, there are five regional health authorities (RHAs)—such as Winnipeg RHA—in the Canadian province of Manitoba. For instance, COVID-19 data are collected from more than 200 health service facilities (e.g., clinics, community health offices, health centers, hospitals, long-term care centers, personal care homes) within Winnipeg RHA.
- As another example, there are 14 local health integration networks (LHINs) in the province of Ontario. For instance, COVID-19 data are collected from more than 170 health service providers (including 18 hospitals) across Toronto Central LHIN.
- As a third example, there are five RHAs in the province of British Columba. Within Vancouver Coastal Health (one of the five RHAs), there are three health service delivery areas, which in turn are divided into 14 local health areas.

These data are usually reported and updated on a daily basis. However, sometime, due to weekends and/or holidays, this information may be delayed.

Moreover, due to different factors, availability of detailed characteristics of COVID cases may vary from one data source to another, and may vary from one level of spatial granularity to another. Factors include privacy concerns. For instance, to preserve privacy of individuals for data publishing (i.e., privacy-preserving data publishing), some data may be merged into a cluster (e.g., a spatial cluster by grouping data from close-by geographical locations, a temporal cluster by grouping data from consecutive time intervals).

In addition to COVID-19 epidemiological data, some other relevant data are also integrated. These include population statistics and timelines of other information (e.g., travel restrictions, social gathering restrictions, lockdown measures, reopening schedule). This information may help in the computation (e.g., for relative percentage).

B. Data Preprocessing

Once relevant data are integrated, our data science engine then preprocesses the integrated data. For instance, as mentioned earlier, given that data are integrated from different sources, their formats and/or types can be different. Some characteristics (e.g., symptoms, occupation) of COVID-19 may be available from some sources, but some may not. Classification of occupation may also vary. To address these issues, our engine provides users with options of:

- focusing mostly on common characteristics, or
- representing the absent characteristics (i.e., attribute values) by NULL.

In terms of common characteristics, most COVID-19 epidemiological data contain:

1. geographical location
2. date
3. age
4. gender
5. hospitalization status, ranges from Boolean status (i.e., hospitalized or not) to different types of hospitalization (e.g., intensive care unit (ICU), semi-ICU, regular ward, not hospitalized)
6. clinical outcome (e.g., recovered or death)

Here, age can be grouped into age groups (e.g., 20s, 30s, 40s, ...) or categories (e.g., child, youth, adult, senior). Some of the data may include the following as well:

7. occupation
8. transmission methods
9. symptoms (for symptomatic cases)

Moreover, partially due to weekends or holidays, reporting of numbers may be delayed. For instance, due to closure of laboratories and/or government offices, the detailed data may be reported after the weekend or holidays. To address these issues, our engine provides users with options of:

- using the reported date, or
- taking an average of several days (e.g., a *7-day average*) of the data.

For privacy-preserving data publishing, spatial data can be grouped into clusters by merging close-by data. Similarly, temporal data can be grouped into clusters by merging consecutive time intervals.

C. Spatial-Temporal Hierarachy

As describing in Section II-A, COVID-19 data are integrated from various data sources, which can be put in a hierarchical fashion. For instance, data can be gathered from health service facilities to report to RHAs, which then report to the province to become provincial data. These provincial data can be collected to form some regional data (e.g., data for the west coast, the prairie, etc.). Regional data can be aggregated to from national data. Along this direction, national data can be aggregated to form data for the continent and then the worldwide data. See Fig. 1 for spatial hierarchy.

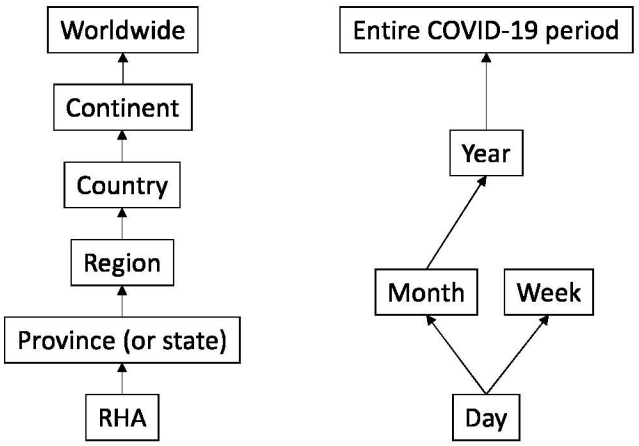

Fig. 1. Our spatial-temporal hierarchy

Similarly, as describing in Section II-A, COVID-19 data are usually reported and updated on a daily basis. So, it is logical to use daily figures. They can be aggregated to form a weekly or monthly figures. Along this direction, weekly or monthly figures can be aggregated to form yearly data, and then the grand total for the entire COVID-19 period. See Fig. 1 for temporal hierarchy.

With the spatial-temporal hierarchy, we are not confined to discovering interesting patterns from the worldwide statistics over the entire COVID-19 period. We have the flexibility to discover interesting patterns from any combinations of 6 granularity levels in the spatial components and 4 (or 5) granularity levels in the temporal components of the hierarchy for a total of 6×4=24 (or 6×5=30) possible combinations—including a (worldwide, entire COVID-19 period)-combination, ..., a (region, week)-combination, and a (province, week)-combination.

There can be multiple instances of some combinations. Take a (province, week)-combination as an example. For 10 provinces in Canada, there can be 10×(52+27) = 790 weekly instances covering the entire period from the beginning of 2020 to Week 27 (July 04-10) of 2021.

D. Spatial-Temporal Data Analytics

With the 6 common characteristics (including geographical location and date), it is tempting to set up a 6-dimensional data cube. For some of these dimensions, the numbers of values can be large. For example, for geographical location, there can be more than 200 countries.

An alternative way to find interesting patterns is to apply *frequent pattern mining* to some spatial-temporal combinations of interest. With long lists of interesting patterns, our data science engine lists them in descending order of frequency (e.g., absolute frequency or relative frequency).

To compute relative frequency, our engine sometimes uses the integrated non-COVID-19 data. For example, to compute relative frequency of COVID-19 (e.g., the percentage of COVID-19 cases per 1M inhabitants in that geographical location), it makes good use of the population statistics.

Moreover, our engine compares and contrasts the interesting patterns computed for each values for the spatial-temporal combinations of interest. For example, we can mine frequent patterns from the (Manitoba, Week 27 of 2021)-combination and compare them with those computed from other combinations like the (P, Week 27 of 2021)-combinations where $P \in$ {BC, Alberta, Saskatchewan, ..., Newfound} for *spatial data analytics*—i.e., spatial comparisons among cases from the other nine Canadian provinces.

Similarly, we can mine frequent patterns from the (Manitoba, Week 27 of 2021)-combination and compare them with those computed from other combinations like the (Manitoba, W)-combinations where $W \in$ [Week 1 of 2020, Week 25 of 2021] for *temporal data analytics*—i.e., temporal comparisons among weekly cases in the Canadian province of Manitoba.

In addition, our engine also applies *time-series analysis* to examine time series representing for each geographical location. It helps identify similarities and differences among different time series.

III. Evaluation

To evaluate our data science engine, we applied it to different datasets. Common rich data sources for COVID-19 cases include: World Health Organization (WHO) [47], European Centre for Disease Prevention and Control (ECDC)[2], Johns Hopkins University (JHU) Coronavirus Resource Center[3], Statistics Canada[4], Wikipedia[5], and media (e.g., Canadian national TV networks[6, 7]).

[2] https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html

[3] https://coronavirus.jhu.edu/map.html

[4] https://www150.statcan.gc.ca/n1/pub/13-26-0003/132600032020001-eng.htm

[5] https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data/Canada_medical_cases

[6] https://www.ctvnews.ca/health/coronavirus/tracking-every-case-of-covid-19-in-canada-1.4852102, https://beta.ctvnews.ca/content/dam/common/exceltojson/COVID-19-Canada-New.txt

[7] https://www.ctvnews.ca/health/coronavirus/tracking-variants-of-the-novel-coronavirus-in-canada-1.5296141 https://beta.ctvnews.ca/content/dam/common/exceltojson/COVID-Variants.txt

When mining the (worldwide, entire COVID-19 period)-combination, our data science engine discovered that there have been cumulatively 186,492,674 COVID-19 cases worldwide, of which 4,029,920 deaths (i.e., approximately 2.16% of all cases) as of July 10, 2021.

Then, we mined the (continent, entire COVID-19 period)-combination. The mined frequent patterns in Table I reveal that:

- Asia and Europe have the highest number of cumulative cases (with 57,430,886 cumulative cases in Asia and 49,110,565 in Europe) as of July 10, 2021. In contrast, Oceania have had the lowest number—with 61,221 cumulative cases.
- Although Asia has the highest number of COVID-19 cases, it is ranked the fourth in terms of death tolls. Europe, South America and North America are the top-3 continents—with 1,114,964 deaths, 1,033,722 deaths and 908,001 deaths.
- However, in terms of percentages of deaths among the COVID-19 cases, South America has the highest percentage—with ~3.05% of South American cases passed away.

TABLE I. RESULTS FROM THE (CONTINENT, ENTIRE COVID-19 PERIOD)-COMBINATION AS ON JULY 10, 2021: ABSOLUTE NUMBERS OF CUMULATIVE CASES & DEATHS

	cases	deaths	%deaths wrt cases
Asia	57,430,886	820,914	1.43%
Europe	49,110,565	1,114,964	2.27%
N. America	40,047,453	908,001	2.27%
S. America	33,927,048	1,033,722	3.05%
Africa	5,914,774	151,175	2.56%
Oceania	61,221	1,167	1.91%

TABLE II. RESULTS FROM THE (CONTINENT, ENTIRE COVID-19 PERIOD)-COMBINATION AS ON JULY 10, 2021: NUMBERS OF CUMULATIVE CASES & DEATHS PER 1 MILLION POPULATION

	cases per 1M pop'n	deaths per 1M pop'n
S. America	78,760.95	2,399.76
N. America	67,639.48	1,533.59
Europe	65,596.20	1,489.24
Asia	12,377.75	176.92
Africa	4,412.04	112.76
Oceania	1,434.49	27.34

As population in these continents are different, our engine integrated COVID-19 data with population statistics to examine the percentage of cases and deaths per 1M inhabitants. Mining results in Table II reveal that:

- Despite South America was ranked the fourth in terms of the absolute number of COVID-19 cases, it has the highest relative number: 78,760.95 cases per 1 million people inhabited in South America.
- The next three continents on the list (for the number of cases per 1M population) are North America, Europe and Asia. Hence, the top-4 continents in terms of relative number are in reverse order of their ranking in terms of absolute number of cases.
- The ranking for deaths per 1M population is identical to the ranking for cases per 1M population.

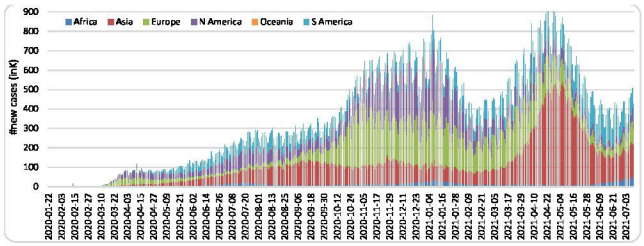

Fig. 2. A stacked column chart showing the results from the (continent, day)-combination: absolute numbers of new daily cases for six continents

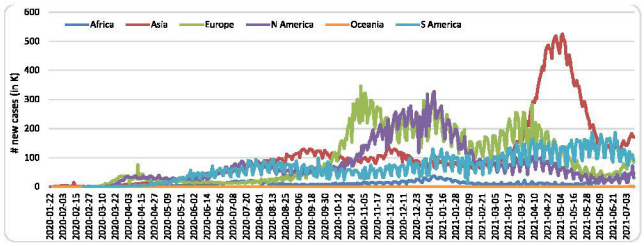

Fig. 3. Line curves showing the results from the (continent, day)-combination: absolute numbers of new daily cases for six continents

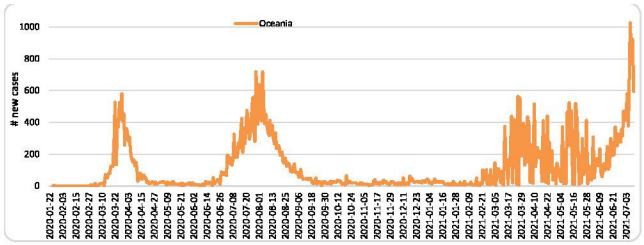

Fig. 4. A line curve showing the results from the (Oceania, day)-combination: absolute numbers of new daily cases

Next, we mined the (continent, day)-combination. The time series for each continent show that the numbers of new cases have not been evenly distributed throughout the COVID-19 period. Moreover, as observed from the aforementioned mining results on the number of cumulative cases, we infer that the new cases have not been evenly distributed among the continent. Figs. 2-4 confirm this inference. Furthermore, we observed the following from the mining results shown in the two figures:

- First reported cases started on different days in different continents. For instance, first 98 Asian cases were reported on January 23, 2020. The first North American cases and the first two European cases were reported a day later (on January 24). First four Oceanian cases were reported on January 26. The first African case was reported on February 14. South America was infected the latest, with its first two cases reported on February 23 (i.e., a month after the first 98 cases reported in Asia).
- Absolute numbers of new daily cases in Oceania have been relatively low (e.g., < 2,000 daily cases) when compared with other continents. More new daily cases were observed in the fall & winter (in the southern

hemisphere, i.e., March-September) than summer. It also experienced with a few waves (e.g., peaked in March 2020, August 2020, March-July 2021). See Fig. 4.

- There have been a few waves in Europe (e.g., started November 2020, January 2021, April 2021).
- There have been a peak (with 525,129 new daily cases) in Asia during April-May 2021.

Afterwards, we analyzed a different combination. Specifically, we mined the (country, entire COVID-19 period)-combination. The mined frequent patterns about cumulative cases in Table III reveal that:

- USA has had the highest cumulative COVID-19 cases. It has been the key contributor to the North American cumulative cases. The 33,847,784 US cases have accounted for 84.52% of all cumulative North American cases.
- Similarly, India has had the second-highest cumulative COVID-19 cases. It has been the key contributor to the Asian cumulative cases. The 30,837,222 Indian cases have accounted for 53.69% of all cumulative Asian cases.
- Moreover, Brazil was ranked the third in terms of absolute numbers of cumulative COVID-19 cases. It has been the key contributor to the South American cumulative cases. The 19,069,003 Brazilian cases have accounted for 56.21% of all cumulative South American cases.
- In Europe, no single country has dominated the numbers of cumulative cases. For instance, France, Russia, UK and Italy all contributed. With about 5.8 million, 5.6 million, 5.1 million and 4.2 million cumulative cases (in France, Russia, UK and Italy, respectively), they accounted for 11.95%, 11.58%, 10.40% and 8.69%—for a total of 42.63%—of all cumulative European cases.

TABLE III. RESULTS FROM THE (COUNTRY, ENTIRE COVID-19 PERIOD)-COMBINATION AS ON JULY 10, 2021: ABSOLUTE NUMBERS OF CUMULATIVE CASES

	cases	% in continental cases
USA	33,847,784	84.52%
India	30,837,222	53.69%
Brazil	19,069,003	56.21%
France	5,870,463	11.95%
Russia	5,688,807	11.58%
UK	5,107,780	10.40%
Italy	4,269,885	8.69%

The mined frequent patterns about cumulative death tolls in Table IV reveal that:

- USA once again has been on the top of the list, with the highest cumulative COVID-19 death tolls. It has been the key contributor to the North American cumulative COVID-19 deaths. Despite that the 607,063 US deaths have accounted for ~1.79% of all US cumulative cases, they have accounted for 66.86% of all cumulative North American deaths.
- Similarly, Brazil has had the second-highest cumulative COVID-19 death tolls. It has been the key contributor to the South American cumulative death tolls. Despite that the 532,893 Brazilian deaths have accounted for ~1.73% of all Brazilian cumulative cases, it have accounted for 51.55% of all cumulative South American deaths.
- Moreover, India was ranked the third in terms of absolute numbers of cumulative COVID-19 death tolls. It has been the key contributor to the Asian cumulative deaths. The 408,040 Indian deaths have accounted for 49.71% of all cumulative Asian deaths.
- Mexico and Peru were not in the top-7 in terms of cumulative cases, but they had high death rates. For example, 234,907 Mexican deaths and 193,230 Peruvian deaths accounted for close to 10% (more precisely, 9.08% and 9.30%) of their respective cumulative numbers of cases. Their combined death tolls between the two countries accounted for 44.56% of all cumulative South American deaths.
- Again, no single country has dominated the numbers of cumulative deaths in Europe. For instance, Russia, UK and Italy all contributed. With about 139K, 128K and 127K cumulative deaths (in Russia, UK and Italy, respectively), they accounted for about 2-3% of the cumulative cases in their countries and about 11-12% each (for a total of 35.55%) of cumulative European death tolls.

TABLE IV. RESULTS FROM THE (COUNTRY, ENTIRE COVID-19 PERIOD)-COMBINATION AS ON JULY 10, 2021: ABSOLUTE NUMBERS OF CUMULATIVE DEATHS

	deaths	%deaths wrt cases	% in continental death tolls
USA	607,063	1.79%	66.86%
Brazil	532,893	1.73%	51.55%
India	408,040	1.32%	49.71%
Mexico	234,907	9.08%	25.87%
Peru	193,230	9.30%	18.69%
Russia	139,896	2.46%	12.55%
UK	128,665	2.52%	11.54%
Italy	127,768	2.99%	11.46%

In terms of numbers of cumulative cases per 1 million population, the list is quite different. The top-10 countries with highest cumulative cases per 1 million inhabitants have been Andorra, Seychelles, Montenegro, Bahrain, Czech Republic, San Marino, Maldives, Slovenia, Luxembourg, and Uruguay. The list includes a few small-sized countries.

Next, we examined the (Canada, week)-combination. The time series in Figs. 6-9 show uneven distribution and temporal changes throughout the COVID-19 period:

- Fig. 6 shows the absolute number of Canadian COVID-19 cases. It also shows the third notable waves (in April 2020, December 2020, and April 2021).
- Fig. 7 shows the results of frequent pattern mining (i.e., 1-itemsets {domestic acquisition} for most weeks) that a majority of cases were transmitted through community exposures (i.e., domestic acquisition).

- Fig. 8 shows the results of frequent pattern mining (i.e., 2-itemsets {domestic acquisition, no hospitalization} for most weeks) that a majority of cases were transmitted through community exposures but did not required hospitalization.
- Along this direction, Fig. 9 shows the results of frequent pattern mining (i.e., 3-itemsets {domestic acquisition, no hospitalization, recovered} for most weeks) that a majority of cases—who were transmitted through community exposures but did not required hospitalization—were recovered.

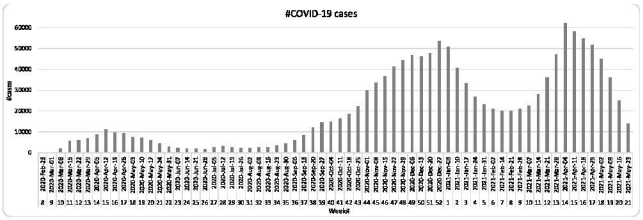

Fig. 5. A column chart showing the results from the (Canada, week)-combination: absolute numbers of new daily cases

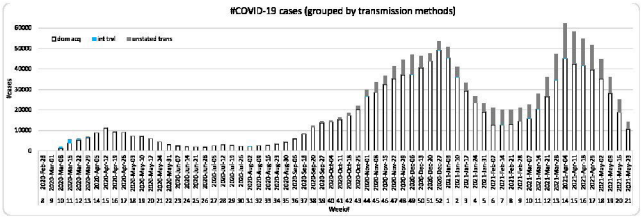

Fig. 6. A stacked column chart showing the results from the (Canada, week)-combination: numbers of new daily cases & tranmission methods

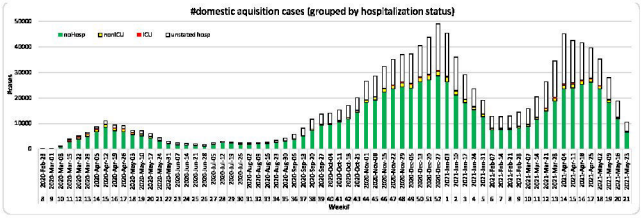

Fig. 7. A stacked column chart showing the results from the (Canada, week)-combination: #new daily cases, transmission methods & hospitalization status

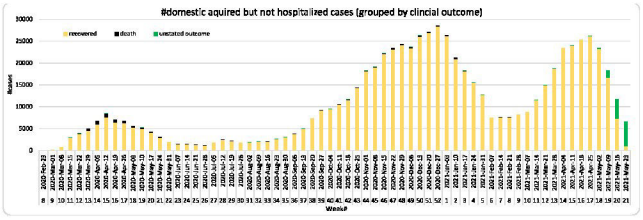

Fig. 8. A stacked column chart showing the results from the (Canada, week)-combination: #new daily cases, transmission methods, hospitalization status & clinical outcome

IV. Conclusions

In this paper, we present a data science engine to analyze and mine COVID-19 data. As COVID-19 cases may not evenly distributed among spatial locations and/or evenly distributed throughout the entire period of pandemic, our engine conducts spatial-temporal data science to reveal important information and knowledge about epidemiological characteristics of the disease across different spatial locations and its temporal trends. Evaluation on real-life COVID-19 data (e.g., on continental, country-wide, Canada data) demonstrates the effectiveness of our engine in conducting spatial-temporal data science of COVID-19 data. As *ongoing and future work*, we transfer learned knowledge to conduct spatial-temporal data science of data from other domains.

Acknowledgment

This work is partially supported by NSERC (Canada) and University of Manitoba.

References

[1] A. Ahmet, T. Abdullah, "Real-time social media analytics with deep transformer language models: a big data approach," IEEE BigDataSE 2020, pp. 41-48.

[2] A. Alsaig, et al., "A critical analysis of the V-model of big data," IEEE TrustCom-BigDataSE 2018, pp. 1809-1813.

[3] R. Mo, et al., "A differential privacy-based protecting data preprocessing method for big data mining," IEEE TrustCom-BigDataSE 2019, pp. 693-699.

[4] K.E. Barkwell, et al., "Big data visualisation and visual analytics for music data mining," IV 2018, pp. 235-240.

[5] S.P., Singh, et al., "Analytics of similar-sounding names from the web with phonetic based clustering," IEEE/WIC/ACM WI-IAT 2020, pp. 580-585.

[6] I.M. Anderson-Grégoire, et al., "A big data science solution for analytics on moving objects," AINA 2021, vol. 2, pp. 133-145.

[7] C.K. Leung, et al., "Big data analysis and services: visualization of smart data to support healthcare analytics," IEEE iThings-GreenCom-CPSCom-SmartData 2019, pp. 1261-1268.

[8] A.S. Shahraki, et al., "A dynamic access control policy model for sharing of healthcare data in multiple domains," IEEE TrustCom-BigDataSE 2019, pp. 618-625.

[9] S. Shang, et al., "Spatial data science of COVID-19 data," IEEE HPCC-SmartCity-DSS 2020, pp. 1370-1375.

[10] J. Souza, et al., "An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics," AINA 2020, pp. 669-680.

[11] C.M. Choy, et al., "Natural sciences meet social sciences: census data analytics for detecting home language shifts," IMCOM 2021, pp. 520-527.

[12] T.S. Cox, et al., "An accurate model for hurricane trajectory prediction," IEEE COMPSAC 2018, vol. 2, pp. 534-539.

[13] W. Lee, et al., "Reducing noises for recall-oriented patent retrieval," IEEE BDCloud 2014, pp. 579-586.

[14] C.K. Leung, et al., "Information technology-based patent retrieval model," Springer Handbook of Science and Technology Indicators, pp. 859-874.

[15] J.D. Hamilton, et al., "Identifying the right person in social networks with double metaphone codes," IEEE ISPA-BDCloud-SocialCom-SustainCom 2020, pp. 794-801.

[16] F. Jiang, et al., "Finding popular friends in social networks," CGC 2012, pp. 501-508.

[17] S.P. Singh, C.K. Leung, "A theoretical approach for discovery of friends from directed social graphs," IEEE/ACM ASONAM 2020, pp. 697-701.

[18] Y. Zheng, et al., "Improved weighted label propagation algorithm in social network computing," IEEE TrustCom-BigDataSE 2018, pp. 1799-1803.

[19] A.K. Chanda, et al., "A new framework for mining weighted periodic patterns in time series databases," ESWA 79, 2017, pp. 207-224.

[20] K.J. Morris, et al., "Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data," IEEE ICMLA 2018, pp. 1486-1491.

[21] K.K. Roy, et al., "Mining sequential patterns in uncertain databases using hierarchical index structure," PAKDD 2021, Part II, pp. 29-41.

[22] H. Yang, et al., "A practical machine learning approach for dynamic stock recommendation," IEEE TrustCom-BigDataSE 2018, pp. 1693-1697.

[23] A.A. Audu, et al., "An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city," CISIS 2019, pp. 224–236.

[24] P.P.F. Balbin, et al., "Predictive analytics on open big data for supporting smart transportation services," Procedia Computer Science 176, 2020, pp. 3009-3018.

[25] C.S.H. Hoi, et al., "Prediction of food preparation time for smart city," IEEE HPCC-SmartCity-DSS 2020, pp. 1203-1210.

[26] C.K. Leung, "Big data mining and computing in a smart world," UIC-ATC-ScalCom-CBDCom-IoP 2015, p. xcviii.

[27] C.K. Leung, et al., "Explainable machine learning and mining of influential patterns from sparse web," IEEE/WIC/ACM WI-IAT 2020, pp. 829-836.

[28] K.E. Dierckens, et al., "A data science and engineering solution for fast k-means clustering of big data," IEEE TrustCom-BigDataSE-ICESS 2017, pp. 925-932.

[29] C.K. Leung, et al., "Big data science on COVID-19 data," IEEE BigDataSE 2020, pp. 14-21.

[30] Y. Chen, et al., "Temporal data analytics on COVID-19 data with ubiquitous computing," IEEE ISPA-BDCloud-SocialCom-SustainCom 2020, pp. 958-965.

[31] C.K. Leung, et al., "A data science model for big data analytics of frequent patterns," IEEE DASC-PICom-DataCom-CyberSciTech 2016, pp. 866-873.

[32] C.K. Leung, et al., "Big data analytics for personalized recommendation systems," IEEE DASC-PiCom-CBDCom-CyberSciTech 2019, pp. 1060-1065.

[33] A. Albakri, et al., "Hierarchical polynomial-based key management scheme in fog computing," IEEE TrustCom-BigDataSE 2018, pp. 1593-1597.

[34] L. Bellatreche, et al., "Advances in cloud and big data computing," CCPE 31(2), 2019, pp. e5053:1-e5053:3.

[35] B. Yin, et al., "A cooperative edge computing scheme for reducing the cost of transferring big data in 5G networks," IEEE TrustCom-BigDataSE 2019, pp. 700-706.

[36] C.K. Leung, C.L. Carmichael, "Exploring social networks: a frequent pattern visualization approach," IEEE SocialCom 2010, pp. 419-424.

[37] C.K. Leung, et al., "PyramidViz: Visual analytics and big data visualization of frequent patterns," IEEE DASC-PICom-DataCom-CyberSciTech 2016, pp. 913-916.

[38] W. Kuo, J. He, "Guest editorial: crisis management – from nuclear accidents to outbreaks of COVID-19 and infectious diseases," IEEE Trans. Reliab. 69(3), 2020, 846-850.

[39] Y. Chen, et al., "A data science solution for supporting social and economic analysis," in IEEE COMPSAC 2021, 1690-1695.

[40] A. Viguerie, et al., "Simulating the spread of COVID-19 via a spatially-resolved susceptible-exposed-infected-recovered deceased (SEIRD) model with heterogeneous diffusion," Appl. Math. Lett. 111, 2021, 106617:1-106617:9.

[41] D.L.X. Fung, et al., "Predictive analytics of COVID-19 with neural networks," in IJCNN 2021.

[42] P. Gupta, et al., "Vertical data mining from relational data and its application to COVID-19 data," in Big Data Analyses, Services, and Smart Data, 2021, 106-116.

[43] W.T. Li, et al., "Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis," BMC Medical Informatics Decis. Mak. 20(1), 2020, 247:1-247:13.

[44] A.S. Albahri, et al., "Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review," J. Medical Syst. 44(7), 2020, 122:1-122:11.

[45] M.B. Jamshidi, et al., "Artificial intelligence and COVID-19:deep learning approaches for diagnosis and treatment," IEEE Access 8, 2020, 109581-109595.

[46] M.J. Mulligan, et al., "Phase I/II study of COVID-19 RNA vaccine BNT162b1 in adults," Nature 586, 2020, 589-593.

[47] J. Hasell, et al., "A cross-country database of COVID-19 testing," Sci Data 7, 2020, 345:1-345:7.