2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)

2020 IEEE 22nd International Conference on High Performance Computing and Communications | 978-1-7281-7649-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/HPCC-SmartCity-DSS50907.2020.00177

# Spatial Data Science of COVID-19 Data

Siyuan Shang, Carson K. Leung(✉), Yubo Chen, Adam G.M. Pazdor
*Department of Computer Science*
*University of Manitoba*
Winnipeg, MB, Canada
✉ Email: kleung@cs.umanitoba.ca

***Abstract*—Huge amounts of big data can be generated and collected from a wide variety of rich data sources. Embedded in these big data are useful information and valuable knowledge. An example is healthcare and epidemiological data such as data related to patients who suffered from viral diseases like the coronavirus disease 2019 (COVID-19). Knowledge discovered from these epidemiological data via data science helps researchers, epidemiologists and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, control and combat the disease. In this paper, we present a spatial data science system for analyzing big COVID-19 epidemiological data, with focus on the spatial data analytics among different geographic locations. The system helps users to get a better understanding of information about the confirmed cases of COVID-19. Evaluation results show the benefits of our system in spatial data analytics of big COVID-19 data.**

***Keywords—data science, coronavirus disease, COVID-19, data system, data application, big data, spatial data, data mining***

## I. Introduction and Related Works

Nowadays, big data [1-3] are everywhere. To elaborate, huge amounts of big data—with different levels of veracity (e.g., precise data, imprecise or uncertain data [4-8])—have been generated and collected at a rapid rate from a wide variety of rich data sources in numerous real-life applications. These include networks (e.g., social networks [9-11], transportation networks [12-15]), financial time series [16], biomedical data (e.g. omic data [17-19], disease reports [20], epidemiological data [21, 22]). Valuable knowledge and useful information embedded in these big data can be discovered by *data science* [23-25]—which apply data mining algorithms [26-30], machine learning tools [31-34], mathematical and statistical models [35, 36], informatics [37, 38], data analytics [39-44], and visual analytics [45, 46].

The discovered knowledge is useful as it can significantly improve the quality of human life. For instance, knowledge discovered from the epidemiological data helps researchers, epidemiologists and policy makers to get a better understanding of diseases, which may inspire them to come up ways to detect, prevent, and/or control diseases—including viral diseases like coronavirus disease 2019 (COVID-19), which broke out in 2019 and became a pandemic in 2020.

Because of the COVID-19 pandemic, many researchers have focused on different aspects of the COVID-19 disease. For instance, from the social science aspects, there has been works studying on crisis management for the COVID-19 outbreak [47]. From medical and health science aspects, there has been works focusing on clinical and treatment information [48], as well as drug discovery and vaccine development [49]. From the natural science and engineering aspect, researchers have examined artificial intelligence (AI)-driven informatics, sensing, imaging for tracking, testing, diagnosis, treatment and prognosis [50] such as those imaging-based diagnosis of COVID-19 using chest computed tomography (CT) images [51, 52]. Researchers have also come up with mathematical modelling of the spread of COVID-19 [53].

In contrast, we examine COVID-19 epidemiological data because they can be considered as an excellent example of big data, especially characterized by their several V's (namely, high volumes, velocity, variety and veracity). For instance, as of November 15, 2020, there have been high *volumes* of 53M+ cumulative COVID-19 cases globally appear at high *velocity* of mean 400+ new cases per minute (derived from ~594,000 new daily cases) [54]. These cases are associated with a wide *variety* of information (e.g., symptoms, clinical course and outcomes, transmission methods) collected from a wide variety of data sources (e.g., regional health authorities). These cases contain data of different levels of *veracity*. While some data are precise, some others can be uncertain (e.g., unstated transmission methods) partially due to fast dissemination of the information or privacy-preservation of individual cases.

Although there are some existing works [54] on the epidemiological data, they mostly focused on showing the numbers of confirmed cases and mortality. While the numbers of confirmed cases and mortality are important in revealing the severity of the disease at a specific geographic location, there are other important knowledge that can be discovered from the epidemiological data for revealing additional information associated with the disease (e.g., what are common transmission methods, set of symptoms, etc. among patients in different geographic locations?).

In this paper, we design and develop a data science system that conducts spatial data science of textual-based COVID-19 epidemiological data (rather than images). Instead of projecting the spread of the disease, our system aims to discover common characteristics (beyond just the numbers of confirmed cases and mortality) among COVID-19 cases in a certain geographic location, and compares them with those in other geographic locations.

Our *key contributions* of this paper include our design and development of a data science system that conducts spatial data

DOI 10.1109/HPCC-SmartCity-DSS50907.2020.00177

science of textual-based COVID-19 epidemiological data. With our spatial hierarchy, important information at different spatial granularity can be captured. Moreover, our system discovers frequently co-occurring characteristics (e.g., common sets of symptoms) of COVID-19 cases, compares and contrast among different geographic locations. Taking into account population differences among different geographic locations, we consider both absolute frequency and relative frequency (relative to per thousand inhabitants or percentages of the total populations) when discovering frequently co-occurring characteristics. Furthermore, although the system is designed and developed for spatial analytics of big epidemiological data, it would be applicable to spatial analytics of other big data in many real-life applications and services.

The remainder of this paper is organized as follows. Next section presents our data science system for spatial data analytics. Section III shows evaluation results, and Section IV draws the conclusions.

## II. Our Data Science System for Spatial Data Analytics

In this section, we describe our data science system for spatial data analytics of COVID-19 epidemiological data.

### *A. Data Collection and Integration*

Big COVID-19 epidemiological data can be of a wide variety (e.g., different types of data). They are usually generated and collected from various data sources.

As a concrete example, in Canada, health care is a responsibility of provincial governments. So, Canadian COVID-19 epidemiological data are gathered from each province (or territory), and provincial data are obtained from *health authorities* (which are also known as *health regions*) within the province. For instance, in the province of Manitoba, COVID-19 data can be gathered from Winnipeg Regional Health Authority (WRHA) and four other regional health authorities (RHAs)[1]. In terms of data types, COVID-19 epidemiological data usually contain:

- administrative information, which includes (a) an unique privacy-preserving identifier for each case, (b) its location, and (c) episode day (i.e., symptom onset day or its closest day).
- case details, which include (a) gender, (b) age, and (c) occupation of the cases.
- symptom-related data, which include additional information for the case who is not asymptomatic (i.e., symptomatic case) such as (a) onset day of symptoms, and (b) a collection of symptoms (e.g., cough, fever, chills, sore throat, runny nose, shortness of breath, nausea, headache, weakness, pain, irritability, diarrhea, and other symptoms).
- clinical course and outcomes, which include (a) hospital status—such as hospitalized in the intensive care unit (ICU), non-ICU hospitalized, and not hospitalized—as well as (b) clinical outcomes (e.g., recovery or death).
- exposures, which include transmission methods.

### *B. Data Preprocessing*

After collecting and integrating data from heterogeneous sources, we observe that there are some missing, unstated or unknown information (i.e., NULL values). Given the nature of these COVID-19 cases, it is not unusual to have NULL values because some values may not be available or recorded at the moments for timely reporting of cases. For some other attributes related to case details (e.g., personal information like gender, age), patients may prefer not to report it due the privacy concerns. As there are many cases with NULL values for some attributes, ignoring them may lead to inaccurate or incomplete analysis of the data. Instead, our system keeps all these cases for analysis.

For some attributes (e.g., date), it would be too specific for the analysis. Moreover, delays in testing or reporting (especially, due to weekends) are not uncommon. Hence, it would also be logical to group days into a 7-day interval---i.e., a week. For example, all days within the week of January 19-25 inclusive are considered as Week 3. Side-benefits of such grouping include:

- Summing the frequency of cases over a week (cf. a single day) increases the chance of having sufficient frequency for being discovered as a frequent pattern and getting statistically significant mining results.
- Generalizing the cases help preserve the privacy of the individuals while maintaining the utility for knowledge discovery.

Similarly, for some attributes (e.g., age, occupation), it would be logical to group similar values into a mega-value (say, ages can be binned into age groups). For example:

- grouping ages to age groups (e.g., ≤ 19 years old, 20-29 years old, ..., 70-79 years old, ≥ 80 years old);
- generalizing occupation of the cases to some generalized key occupation groups—say, (a) health care workers, (b) school or daycare workers, (c) long-term care residents, and (d) others;
- generalizing specific transmission methods to some generalized key transmission methods—say, (a) community exposures, (b) travel exposures, and (c) others.

### *C. Spatial Hierarchy*

Recall from Section II-A, COVID-19 epidemiological data can be collected from a wide variety of data sources such as local health authorities. These local data can then be combined and/or aggregated to meta-data at a more generalized granularity level. For instance, we group local data obtained from various facilities (e.g., health centers, hospitals) within a regional health authority (RHA), and then combine and aggregate these data to form the provincial COVID-19 epidemiological data. Along this direction, we then form the data for a national region by

[1] https://www.gov.mb.ca/health/rha/

combining data from some similar provinces (e.g., from Prairie Provinces). Afterwards, we can obtain data for a country, and then a continent, by moving up the spatial hierarchy as shown in Fig. 1.

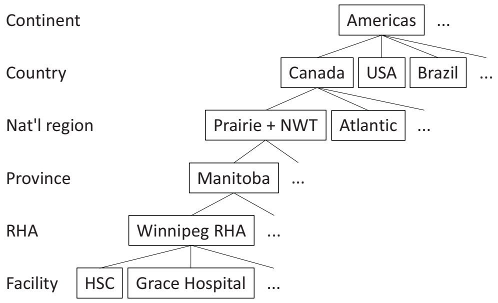

Fig. 1. Spatial hierarchy

With our spatial hierarchy, users can mine frequent patterns and compare contrast patterns among different units at the spatial granularity level of their interest (instead of many repetitive comparisons among data obtained numerous local health authorities). Moreover, users can start conducting spatial data analytics at higher granularity (to avoid distraction) to get an insight or overview. They can then drill in to more detailed data at some specific lower granularity of their interest.

### *D. Frequent and Contrast Pattern Mining*

To discover frequently co-occurring characteristics of COVID-19 cases, we apply frequent patterns to COVID-19 epidemiological data for each geographic location at a certain spatial granularity level in the hierarchy. As data for each location is disjoint, our system can mine each of these disjoint data set independently in parallel.

Partially due to the timely reporting of cases, symptoms were unstated for many cases (i.e., many NULL values for symptoms). As such, the frequency of the symptoms may be lower than values for some other attributes (e.g., domestic acquisition as a transmission method). However, it is scientifically important to know which symptoms—among more than 12 different symptoms—co-occurred more frequently than others. As such, our system provides users with flexible to express their preference or interests. For example, the users can express their interest in finding frequent patterns containing at least one symptoms. As another example, the users can also express their interest in finding frequent patterns consisting of only symptoms.

In addition to finding frequent patterns from each geographic location, our system also compares and contrasts the *ranking* of the discovered patterns among different geographic locations. Moreover, observing that population for each geographic location may vary. Hence, it is logical to take into account the population for that location for comparison. Hence, in addition to reporting the *absolute frequency*, our system also reports the *percentages relative to* (a) *every thousand inhabitants* in the locations, (b) *the population* of the locations and/or (c) the *number of cases* reported for the locations.

## III. Evaluation

### *A. A Case Study on Real-Life COVID-19 Data*

#### *1) Data Collection, Integration and Preprocessing*

To evaluate and demonstrate the usefulness of our data science system, we tested it with COVID-19 epidemiological data from rich data sources like World Health Organization[2] [54], Manitoba Government[3], Wikipedia[4], and Statistics Canada [55, 56]. The last dataset was collected and integrated from provincial and territorial public health authorities by the Public Health Agency of Canada (PHAC). We preprocess data and generalize some attributes to obtain a dataset with the following attributes:

1. A unique privacy-preserving identifier for each case
2. A geographic region/location
3. Episode week (or onset week of symptoms): From Week 3 (i.e., week of January 12-18, 2020) to now
4. Gender
5. Age group: ≤ 19, 20s, 30s, 40s, 50s, 60s, 70s, and ≥ 80s.
6. Occupation group, including:
    a) health care worker,
    b) school or daycare worker (or attendee),
    c) long-term care resident, and
    d) other occupation.
7. Asymptomatic: Yes and No
8. Set of 13 symptoms, including cough, fever, chills, sore throat, runny nose, shortness of breath, nausea, headache, weakness, pain, irritability, diarrhea, and other symptoms.
9. Hospital status, including:
    a) hospitalized in the ICU,
    b) hospitalized but not in the ICU, and
    c) not hospitalized.
10. Transmission method, including:
    a) community exposures, and
    b) travel exposures.
11. Clinical outcome: Recovered and death
12. Recovery week

As of November 12, 2020, the dataset has captured 209,811 COVID-19 cases in Canada. Among them, 190,108 cases with stated episode week. Moreover, although the

[2] https://www.who.int/publications/m/item/weekly-epidemiological-update---15-december-2020

[3] https://news.gov.mb.ca/news/index.html?item=49817&posted=2020-11-15

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

first Canadian case occurred in Week 3, there were not more than two new daily cases for following few weeks. To preserve privacy of these early cases and to cumulate statistically significant mass for analysis, cases from Weeks 3-8 were grouped into (Episode) Week 8 (February 23-29) with 107 cases. From Week 9 onward, the data reflect their reported episode weeks.

### *2) Spatial Hierarchy*

Once the data are preprocessed, our system analyzes and mines data from each geographic location. For instance, as of November 15, 2020, at the regional health authority (RHA) level, the top-3 Manitoban RHAs with the highest number of new daily COVID-19 cases are Winnipeg, Southern and Prairie Mountain RHAs—with 266, 136 and 34 new cases, respectively. Moving up the spatial hierarchy (by combining and aggregating local RHA data), the top-3 provinces are Ontario, Quebec and Alberta—with 1248, 1211 and 991 new daily cases, respectively—among the 13 Canadian provinces and territories. Manitoba is ranked the fifth, with 266+136+34+30+28 = 494 new daily cases.

To avoid distraction and numerous comparison among these 13 provincial and territorial locations, we can also generalize these locations into five national regions. As such, the top-3 Canadian regions are (a) Prairies (consisting of Alberta, Manitoba and Saskatchewan) + NW Territories, (b) Ontario + Nunavut, and (c) Quebec—with 991+494+181+0 = 1666, 1248+10 = 1348, and 1211 new daily cases as of November 15, respectively. Along this direction, top-3 countries in the Americas are USA, Brazil and Argentina—with 181066, 29070 and 11859 new daily cases. Canada is ranked the sixth, with 4741 new daily cases. For completeness, global number of new daily COVID-19 cases is 594000 over all continents.

### *3) Absolute and Relative Frequencies*

Observing that population is not evenly distributed among all geographic locations and locations with lager population may have higher chances of having larger *absolute* numbers of COVID-19 cases, we also present their *relative* figures (e.g., cases per certain number of inhabitants, percentage of population has contracted COVID-19). For example, by incorporating population [57] in the five national regions in Canada, our system their absolute and relative frequencies as of November 12 in Table I.

TABLE I. ABSOLUTE AND RELATIVE CUMULATIVE COVID-19 CASES IN FIVE NATIONAL REGIONS

| **Nat'l region** | **Cum #cases** | | **%cases wrt pop'n** |
|---|---|---|---|
| | ***Absolute#*** | ***per 1M pop'n*** | |
| Quebec | 73,190 | 8,534.5 | 0.853% |
| Ontario + Nunavut | 80,393 | 5,442.1 | 0.544% |
| Prairies + NWT | 37,910 | 5,392.1 | 0.539% |
| BC + Yukon | 16,494 | 3,179.2 | 0.318% |
| Atlantic | 1,824 | 747.2 | 0.075% |
| **Canada** | **209,811** | **5,520.2** | **0.552%** |
| **Worldwide** | **53,766,728** | **6,887.6** | **0.689%** |

Observed from the table, in terms of absolute numbers of cumulative COVID-19 cases, there are more cumulative COVID-19 cases in the national region of (Ontario + Nunavut) than in Quebec. However, in terms of relative numbers, situations in Quebec are more serious (with 0.853% of its population have contracted COVID-19) than Ontario + Nunavut. Such an infection rate in Quebec is higher than the national and global averages (of 0.552% and 0.689%, respectively).

### *4) Frequent Pattern Mining*

In addition to analyzing the number of cases, our system also mine and analyze 12 aforementioned attributes. We observe the following from the Prairies + NW Territories:

- Frequent singleton pattern {not hospitalized}:29208 reveals that 29,208 cases did not needed to be hospitalized. These account for (a) 94.9% of COVID-19 cases with *known* hospitalization status, and (b) 77.0% of all cases (with known and unknown hospitalization status), in this national region.
- {domestic acquisition}: 28617 reveals that 28,617 cases were transmitted via community exposure.
- {recovered}:27167 reveals that 27,167 patients have recovered. These account for an encouraging percentage of 98.3% of patients with *known* clinical outcomes, and 71.7% of all patients, in this region.
- Frequent non-singleton pattern {domestic acquisition, not hospitalized}:27287 reveals that, among the 28,617 cases were transmitted via community exposure, a majority of them (i.e., 27,287 ≈ 95.4%) did not require hospitalization.
- {domestic acquisition, not hospitalized, recovered}: 21795 reveals that, among the 28,617 domestically acquired cases not requiring hospitalization, most of them (i.e., 21,795 ≈ 76.2%) recovered.

As users have flexibility to express their interest or preference (say, finding frequent pattern consisting of only symptoms), our system then incorporates user preference into mining frequent patterns satisfying the user preference. For instance, it finds the following patterns from the same region:

- Frequent patterns {cough}:14431, {headache}:11050, {pain}:9005, {sore throat}:8773 and {fever}:7648 reveal these common symptoms with their absolute frequencies.
- Non-frequent pattern {cough, headache}:6108 reveals the number of cases having the two symptoms together.
- {cough, recovered}:10739 reveals that, among the 14,431 cases with cough, most of them (i.e., 10,739 ≈ 74.4%) recovered.

### *5) Contrast Pattern Mining*

Our data science system applies a similar procedure to other geographic locations to (a) discover frequent patterns from these locations and (b) compare the patterns among different locations. The following are some observations that worth mentioning:

- With 1,704 recovered cases, Atlantic Provinces have a much higher recovery rate (of ~93.4% of all cases in this region) than other four national regions (e.g., a recovery rate of 71.7% in the Prairies + NWT).

- In Quebec, as well as Ontario + Nunavut, occupations of cases are classified into (a) health care workers, (b) others, or (c) unstated. An additional class label of (d) "long-term care residents" is available for cases in BC + Yukon. Another class label—namely, (e) "school/ daycare workers"—is available for cases in the remaining two national regions.
- In Ontario + Nunavut, only 43 cases experienced cough and 9 cases did not. The remaining 80,341 cases did not report any information regarding cough. Similarly, for a majority of cases, their symptoms are unreported.
- Moreover, among all 80,322 cases in Ontario + Nunavut with known occupations, 72,136 cases (i.e., 89.8%) were not health care workers.
- With 15,422 non-hospitalized cases, BC + Yukon have a much higher non-hospitalization rate (~93.5%) than other four regions (e.g., 77.0% in the Prairies + NWT).

### B. *Functionality Check with Related Works*

After demonstrating the features and usefulness of our data science system in conducting spatial data analytics on real-life COVID-19 data, let us evaluate its functionality when compared with related works. First, most of the related works are observed to report mainly the numbers of cases and deaths. They do not provide privacy-preserving details and epidemiological characteristics of those COVID-19 cases, which are provided by our system. Second, for those related works that provide overall data distribution of cases, they are mostly confined to single dimensions/attributes. In contrast, our system provides multi-dimensional information such as relationships among attributes in the form of frequent patterns.

## IV. Conclusions

In this paper, we presented a system for spatial data science on big COVID-19 epidemological data. Our data science system generalizes some attributes for effective analysis. Moreover, it provides users with flexibility of (a) including or excluding these unstated/NULL values and (b) expressing their preference (e.g., "must include symptoms") in mining of frequent patterns. With our spatial hierarchy, the system discovers frequent patterns and contrast patterns at different spatial granularity levels. Evaluation results show the practicality of our system in providing rich knowledge about characteristics of COVID-19 cases. This helps researchers, epidemiologists and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, control and combat the disease. As ongoing and future work, we transfer knowledge learned from the current work to temporal analytics of other big data in many real-life applications and services. We also explore the incorporation of visual analytics [58] with our data science system to conduct visual analytics of spatial big data.

## Acknowledgment

This work is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), as well as the University of Manitoba.

## References

[1] A. Kobusinska, et al., "Emerging trends, issues and challenges in Internet of Things, big data and cloud computing," FGCS 87, 2018, pp. 416-419.

[2] C.K. Leung, "Big data analysis and mining," in Encyclopedia of Information Science and Technology, 4e, 2018, pp. 338-348.

[3] X. Lu, et al., "An integrated high-performance transport solution for big data transfer over wide-area networks," in IEEE HPCC/SmartCity/DSS 2018, pp. 1661-1668.

[4] F. Jiang, C.K. Leung, "A data analytic algorithm for managing, querying, and processing uncertain big data in cloud environments," Algorithms 8(4), 2015, pp. 1175-1194.

[5] C.K. Leung, "Uncertain frequent pattern mining," in Frequent Pattern Mining, 2014, pp. 417-453.

[6] C.K. Leung, et al., "Fast algorithms for frequent itemset mining from uncertain data," in IEEE ICDM 2014, pp. 893-898.

[7] X. Li , et al., "Parallel k-dominant skyline queries over uncertain data streams with capability index," in IEEE HPCC/SmartCity/DSS 2019, pp. 1556-1563

[8] H. Zhou, et al., "NISU: a novel index structure on uncertain data in large-scale publish/subscribe systems," in IEEE HPCC/SmartCity/DSS 2019, pp. 1205-1211.

[9] A. Groulx, C. McGregor, "A social media tax data warehouse to manage the underground economy," in IEEE HPCC/SmartCity/DSS 2018, pp. 1599-1606.

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

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

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

[13] 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.

[14] C.K. Leung, et al., "Urban analytics of big transportation data for supporting smart cities," in DaWaK 2019, pp. 24-33.

[15] X. Lu, et al., "An integrated high-performance transport solution for big data transfer over wide-area networks," in IEEE HPCC/SmartCity/DSS 2018, pp. 1661-1668.

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

[17] C.K. Leung, et al., "Predictive analytics on genomic data with high-performance computing," in IEEE BIBM 2020, pp. 2187-2194.

[18] T. Pawliszak, et al., "Operon-based approach for the inference of rRNA and tRNA evolutionary histories in bacteria," BMC Genomics 21 (Supplement 2), 2020, pp. 252:1-252:14.

[19] O.A. Sarumi, et al., "Spark-based data analytics of sequence motifs in large omics data," Procedia Computer Science 126, 2018, pp. 596-605.

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

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

[22] 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, pp. 106-116.

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

[24] C.K. Leung, F. Jiang, "A data science solution for mining interesting patterns from uncertain big data," in IEEE BDCloud 2014, pp. 235-242.

[25] M. Mahyoub, et al., "Effective use of data science toward early prediction of alzheimer's disease," in IEEE HPCC/SmartCity/DSS 2018, pp. 1455-1461.

[26] P. Braun, et al., “Pattern mining from big IoT data with fog computing: models, issues, and research perspectives,” in IEEE/ACM CCGrid 2019, pp. 854-891.

[27] A. Fariha, et al., “Mining frequent patterns from human interactions in meetings using directed acyclic graphs,” in PAKDD 2013, Part I, pp. 38-49.

[28] C.K. Leung, C.L. Carmichael: FpViz: a visualizer for frequent pattern mining. ACM KDD-VAKD 2009, pp. 30-39.

[29] J. Liu, et al., “Efficient mining of extraordinary patterns by pruning and predicting,” ESWA 125, 2019, pp. 55-68.

[30] W. Sun, et al., “An efficient hash-tree-based algorithm in mining sequential patterns with topology constraint,” in IEEE HPCC/SmartCity/DSS 2019, pp. 2782-2789.

[31] S. Ahn, et al., “A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments,” in FUZZ-IEEE 2019, pp. 1259-1264.

[32] J. de Guia, et al., “DeepGx: deep learning using gene expression for cancer classification,” in IEEE/ACM ASONAM 2019, pp. 913-920.

[33] C.K. Leung, et al., “A machine learning approach for stock price prediction,” in IDEAS 2014, pp. 274-277.

[34] 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,” in IEEE ICMLA 2018, pp. 1486-1491.

[35] B.K. Adhikari, et al., “Statistical analysis for detection of sensitive data using Hadoop clusters,” in IEEE HPCC/SmartCity/DSS 2019, pp. 2373-2378.

[36] C.K. Leung, “Mathematical model for propagation of influence in a social network,” in Encyclopedia of Social Network Analysis and Mining, 2e, 2018, pp. 1261-1269.

[37] W. Lee, et al., “Reducing noises for recall-oriented patent retrieval,” in IEEE BDCloud 2014, pp. 579-586.

[38] C.K. Leung, et al., “Information technology-based patent retrieval model,” in Springer Handbook of Science and Technology Indicators, 2019, pp. 859-874.

[39] P. Braun, et al., “MapReduce-based complex big data analytics over uncertain and imprecise social networks,” in DaWaK 2017, pp. 130-145.

[40] R.C. Camara, et al., “Fuzzy logic-based data analytics on predicting the effect of hurricanes on the stock market,” in FUZZ-IEEE 2018, pp. 576-583.

[41] D. Deng, et al., “An innovative framework for supporting cognitive-based big data analytics for frequent pattern mining,” in IEEE ICCC 2018, pp. 49-56.

[42] K. Hoang, et al., “Cognitive and predictive analytics on big open data,” in ICCC 2020, pp. 88-104.

[43] C.K. Leung, F. Jiang, “Big data analytics of social networks for the discovery of "following" patterns,” in DaWaK 2015, pp. 123-135.

[44] M. Mai, et al., “Big data analytics of Twitter data and its application for physician assistants: who is talking about your profession in Twitter?” in Data Management and Analysis, 2020, pp. 17-32.

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

[46] C.K. Leung, et al., “Visual analytics of social networks: mining and visualizing co-authorship networks,” in HCII-FAC 2011, pp. 335-345.

[47] 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, pp. 846-850.

[48] A.A. Ardakani, et al., “Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks,” Comp. Bio. Med. 121, 2020, pp. 103795:1-103795:9.

[49] B. Robson, “COVID-19 coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed achilles' heel conserved region to minimize probability of escape mutations and drug resistance,” Comp. Bio. Med. 121, 2020, pp. 103749:1-103749:28.

[50] A.A. Amini, et al., “Editorial special issue on "AI-driven informatics, sensing, imaging and big data analytics for fighting the COVID-19 pandemic". IEEE JBHI 24(10), 2020, pp. 2731-2732.

[51] Q. Liu, et al., “A two-dimensional sparse matrix profile DenseNet for COVID-19 diagnosis using chest CT images,” IEEE Access 8, 2020, pp. 213718-213728.

[52] D. Shen, et al., “Guest editorial: special issue on imaging-based diagnosis of COVID-19,” IEEE TMI 39(8), 2020, pp. 2569-2571.

[53] 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, pp. 106617:1-106617:9.

[54] World Health Organization, WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int/

[55] Public Health Agency of Canada, “Detailed preliminary information on confirmed cases of COVID-19 (revised),” Statistics Canada Table 13-10-0781-01. doi:10.25318/1310078101-eng

[56] Public Health Agency of Canada, “Preliminary dataset on confirmed cases of COVID-19,” Statistics Canada Table 13-26-0003. https://www150.statcan.gc.ca/n1/en/catalogue/13260003

[57] Statistics Canada, “Population estimates, quarterly,” Table 17-10-0009-01. doi:10.25318/1710000901-eng

[58] C.K. Leung, C.L. Carmichael, “FpVAT: a visual analytic tool for supporting frequent pattern mining,” ACM SIGKDD Explorations 11(2), 2009, pp. 39-48.