2021 25th International Conference Information Visualisation (IV) | 978-1-6654-3827-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IV53921.2021.00044

A Visual Data Science Solution for Visualization and Visual Analytics of Big Sequential Data

Carson K. Leung[1] (✉), Yan Wen[1], Chenru Zhao[1], Hao Zheng[1], Fan Jiang[2], Alfredo Cuzzocrea[3]
[1] *Department of Computer Science, University of Manitoba*, Winnipeg, MB, Canada
[2] *Department of Computer Science, University of Northern British Columbia* (*UNBC*), Prince George, BC, Canada
[3] *iDEA Lab, University of Calabria*, Rende, Italy
✉ kleung@cs.umanitoba.ca

***Abstract*—In the current era of big data, huge volumes of valuable data have been generated and collected at a rapid velocity from a wide variety of rich data sources. In recent years, the initiates of open data also led to the willingness of many government, researchers, and organizations to share their data and make them publicly accessible. An example of open big data is healthcare, disease and epidemiological data such as privacy-preserving statistics on patients who suffered from epidemic diseases like the coronavirus disease 2019 (COVID-19). Analyzing these open big data can be for social good. For instance, analyzing and mining the disease statistics helps people to get a better understanding of the disease, which may inspire them to take part in preventing, detecting, controlling and combating the disease. As "a picture is worth a thousand words", having the pictorial representation further enhances people's understanding of the data and the corresponding results for the analysis and mining. Hence, in this paper, we present a visual data science solution for the visualization and visual analytics of big sequential data. We illustrate the ideas through the visualization and visual analytics of sequences of real-life COVID-19 epidemiological data. Our solution enables people to visualize COVID-19 epidemiological data and their temporal trends. It also allows people to visually analyze the data and discover relationships among popular features associated with the COVID-19 cases. Evaluation of these real-life sequential COVID-19 epidemiological data demonstrates the effectiveness of our visual data science solution in enhancing user experience in the visualization and visual analytics of big sequential data.**

***Keywords*—information visualization, big data, sequences, data science, visual data science, data mining, data analytics, visual analytics, COVID-19**

I. Introduction

In the current era of big data [1, 2], huge volumes of valuable data can be easily generated and collected at a rapid velocity from a wide variety of rich data sources. In recent years, the initiates of open data also led to the willingness of many government, researchers, and organizations to share their data and make them publicly accessible. Examples of open big data include biodiversity data [3], biomedical/healthcare data and disease reports (e.g., COVID-19 statistics) [4-7], census data [8], financial time series [9-13], music data [14], patent register [15, 16], social networks [17-19], transportation and urban data [20-24], weather data [25], and web data [26, 27].

Embedded in these open big data are useful information and valuable knowledge that can be discovered by *data science* [28-30]—which make good uses of data mining algorithms [31-36], data analytics methods [37-41], visual analytics techniques [42-45], machine learning tools [46-48] and/or mathematical and statistical models [49, 50]. Hence, analyzing and mining these big data can be for social good. For instance, analyzing and mining the disease statistics helps people to get a better understanding of the disease such as:

- severe acute respiratory syndrome (SARS), which was caused by a SARS-associated coronavirus (CoV) and led to an outbreak in 2003.
- Swine flu, which was caused by influenza A virus subtype H1N1 (A/H1N1) and led to a pandemic from 2009 to mid-2010.
- Middle East respiratory syndrome (MERS), which was caused by a MERS-CoV. The disease was reported from places like Middle East (e.g., Saudi Arabia) between 2012-2018 and South Korea in 2015.
- Zika virus disease, which was primarily transmitted by the bite of an infected mosquito. An outbreak was reported in Brazil during 2015-2016.
- coronavirus disease 2019 (COVID-19), which was caused by SARS-CoV-2. This was reported to break out in 2019, became a global pandemic in early 2020, and is still prevailing in 2021.

A better understanding of the disease may inspire people to take part in preventing, detecting, controlling and combating the disease.

It is well known that "a picture is worth a thousand words". For ease of comprehension of the discovered information and knowledge, a pictorial representation of the discovered information and knowledge is desirable. This calls for an important branch of data science—*visual data science*, which makes good use of visualization and visual analytics techniques.

In IV 2020, we [51] presented a big data visualization and visual analytics tool for visualizing and analyzing COVID-19 epidemiological data. It analyzes cumulative COVID-19 statistics and visualizes popular features (e.g., popular transmission methods, hospitalization status, clinical outcomes)

978-1-6654-3827-8/21/$31.00 ©2021 IEEE
DOI 10.1109/IV53921.2021.00044

associated with a majority of COVID-19 cases from these cumulative statistics.

As COVID-19 has spanned more than a year, the cumulative statistics reveal numerical summary of some characteristics of COVID-19 cases for the entire period. However, it may not reveal temporal changes in these characteristics during this period. This calls for a tool or solution for visualizing and analyzing *temporal aspects* of COVID-19 epidemiological data. This motivates our current work. As a non-trivial but logical extension to our big data visualization and visual analytics tool from IV 2020 work, we design and development of a data science solution for visualization and visual analytics of temporal data—i.e., a COVID-19 sequence—in the current IV 2021 paper.

It is important to note that our data science solution is designed and developed in such a way that it visualizes and visually analyzes, not only temporal COVID-19 data, but also other big sequences. For example, visualization and visual analytics of financial time series or stock prices helps financial analysts to get a better understanding of the trends (e.g., uptrends, downtrends) of a stock. Similarly, visualization and visual analytics of temporal employment data helps social scientists to get a better understanding of social and/or economic situations and impacts of any irregular events or shocks (e.g., COVID-19 pandemic).

Our *key contribution* of this paper is our design and development of a data science solution for visualization and visual analytics of big sequential data. Our solution visualizes temporal trends in the sequence. It also visualizes the findings (e.g., discovered information and knowledge) from the visual analytics on the temporal sequential data. For instance, it reveals popular features associated with the data (e.g., popular transmission methods, hospitalization status and clinical outcomes at certain periods) and how these features change over time. To illustrate our key ideas, we apply our solution to real-life COVID-19 sequences. We also evaluation our solution by applying it to a real-life COVID-19 sequences.

The remainder of this paper is organized as follows. The next section discusses background and related works. Section III describes our data science solution for visualization and visual analytics of big sequential data. Section IV shows evaluation results on real-life Canadian COVID-19 epidemiological data. Finally, conclusions are drawn in Section V.

II. Background and Related Works

As for application to visualization and visual analytics of COVID-19 data, many visualizers and dashboards have been developed since its declaration as a pandemic. Notable ones include (a) World Health Organization (WHO) COVID-19 Dashboard[1], (b) COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)[2], and (c) COVID-19 dashboard by European Center for Disease Prevention and Control (ECDC)[3]. They provide summary for global COVID-19 situations. Moreover, local governments (e.g., Government of Canada[4]) and media (e.g., TV[5, 6], Wikipedia[7]) also provides visualizers and dashboards for local COVID-19 situations. One commonality among these visualizers and dashboards is that they focus on the total numbers of new/confirmed cases and deaths, as well as their cumulative totals. They serve the purpose of fast dissemination of these crucial numbers related to COVID-19 cases. However, there is additional information and knowledge that are embedded in the data and yet to be discovered.

In response, we presented a big data visualization and visual analytics tool for visualizing *frequent patterns* from the cumulative COVID-19 statistics. In terms of related works on visualizing frequent patterns, Jentner and Keim [52] surveyed several visualization techniques for frequent patterns. These techniques can be broadly generalized into four categories:

- Lattice representation, which is the most intuitive representation of frequent patterns [53]. With it, frequent patterns are represented as nodes in a *lattice* (aka *concept hierarchy*). Immediate supersets and subsets of a frequent pattern are connected by edges.
- Pixel-based visualization, in which multiple frequent *k*-itemsets (i.e., patterns containing *k* items) of the same length *k* are represented by a pixel [54].
- Linear visualization, in which frequent patterns are represented linearly. For example, FIsViz [55] represents a frequent *k*-itemset in a polyline that connects *k* nodes in a 2-dimensional space. To avoid bending and crossing-over of polylines, FpVAT [44] represents frequent patterns in a wiring-type diagram (i.e., an orthogonal graph).
- Tree visualization, in which frequent patterns are represented according to a tree hierarchy. For example, PyramidViz [56] shows frequent patterns with a *side-view* of the pyramid, in which short patterns are put on the bottom of the pyramid and longer related patterns (e.g., extensions of short patterns) are put on the top. As another example, FpMapViz [57] shows frequent patterns with a *top-view*, in which short patterns are put in the background and longer related patterns are overlay in the foreground.

Similar to the aforementioned tree visualization, our big data visualization and visual analytics tool [51] for visualizing *frequent patterns* from the cumulative COVID-19 statistics also follows a hierarchy so that patterns can be connected to their

[1] https://covid19.who.int/

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

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

[4] https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html

[5] https://newsinteractives.cbc.ca/coronavirustracker/

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

extensions. It can be considered as showing frequent patterns with a top-view. However, instead of putting short patterns in the background and longer related patterns are overlay in the foreground, it put short patterns in the inner ring near the center and longer related patterns in the outer ring (but do not overlay or overlap with the inner ring). Immediate extensions of a pattern are put just outside (but touching) the sector representing the pattern. More specifically, frequent patterns and their related patterns are represented in a pie chart or a sunburst diagram (i.e., a doughnut chart).

To elaborate, when mining and analyzing a Canadian COVID-19 epidemiological dataset collected from Public Health Agency of Canada (PHAC) and Statistics Canada[8] for the period from 2020 to May 29, 2021 (i.e., Week 21 of 2021), our big data visualization and visual analytics tool visualizes transmission methods of 1,368,422 COVID-19 cases in Canada. Fig. 1 shows a frequent 1-itemset (i.e., a singleton pattern) {domestic acquisition}:82.35% and its patterns {unstated transmission method}:17.19% and {international travel}: 0.46%. These patterns reveal that 82.35% (as represented by the white ring sector) of these cases acquired the disease domestically via community exposures, 17.19% (as represented by the grey ring sector) were without any stated transmission methods, and the remaining 0.46% (as represented by the tiny light-blue ring sector) were exposed to the disease via international travel.

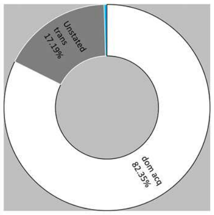

Fig. 1. Visualization of transmission methods.

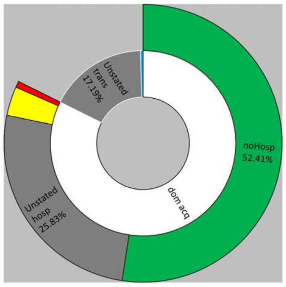

Fig. 2. Visualization of hospital status among those who domestically acquired COVID-19 via community exposures.

Extending the frequent singleton pattern, we discover a frequent 2-itemset (i.e., non-singleton pattern) {domestic acquisition, not hospitalized}:52.41%. This reveals that, among those who domestically acquired the disease from community exposures, a large fraction (52.41%/82.35% ≈ 0.64) of them did not required hospitalization. Fig. 2 shows how our big data visualization and visual analytics tool represents this frequent 2-itemset by putting a green ring sector outside of (but touching) the white ring sector representing the domestic acquisition. The figure also reveals that, while a majority did not require hospitalization, still a noticeable fraction (25.83%/82.35% ≈ 0.31) of them were without stated hospital status (as represented by the grey ring sector). Smaller fractions of them were admitted to the hospital. Specifically, 3.38%/82.35% ≈ 0.04 of them were hospitalized but did not require to admit to the intensive care unit (ICU), as represented by the small yellow ring sector; 0.73%/82.35% ≈ 0.01 of them were admitted to the ICU, as represented by the tiny red ring sector.

Along this direction, a frequent 3-itemset {domestic acquisition, not hospitalized, recovered}:50.87% reveals that, among those who domestically acquired but not hospitalized, a significantly large fraction (50.87%/52.41% ≈ 0.97) of them recovered. This pattern is represented by a golden ring sector, which was put outside of (but touching) the green ring sector representing the domestically acquired but not hospitalized cases, in Fig. 3. The figure also reveals that the remaining two tiny fractions (i.e., 0.91%/52.41% ≈ 0.02 and 0.63%/52.41% ≈ 0.01) belong to those without stated clinical outcome and those deceased, as represented by grey and black ring sectors respectively.

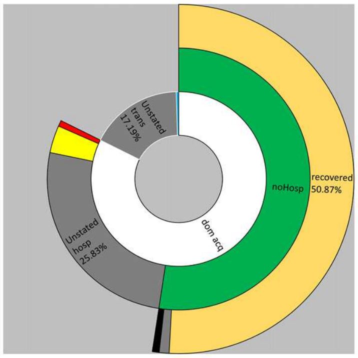

Fig. 3. Visualization of clincial outcomes among those who domestically acquired COVID-19 via community exposures but did not require hospitalization.

Observed that visualization of frequent patterns may not reveal temporal changes, we design and develop a visual data science solution to show temporal changes and visual big sequential data. Jentner and Keim [52] also surveyed several visualization techniques for sequential patterns. These include individual representation [58], flow diagram visualization [59], aggregated pattern visualization [60], visual representation with pattern placement strategies [61], and episode visualization [62].

III. Our Visual Data Science Solution

To explore and visualize temporal changes, we design and develop a visual data science solution. It first collects and integrates data from a wide variety of rich data sources. It then preprocesses the data and builds a temporal hierarchy to generalize the temporal data. Depending on the application, collected data can be of a fine granularity. Having the temporal hierarchy enables us to pick an appreciate level of data granularity. For instance, although dynamic streaming data can

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

be collected at a rapid rate (e.g., per second, minute), analyzing data aggregated at a coarser level (e.g., hourly daily) may lead to more meaningful and interesting results. As a concrete example, COVID-19 statistics is usually updated on a daily basis. We observed that, analyzing the data on a yearly basis may miss some details, but analyzing them on a daily basis may lead to a huge solution space and may be sensitive to unnecessary fluctuation (e.g., delay in testing or reporting cases due to weekends). Consequently, analyzing them on a weekly basis appear to be appropriate.

After selecting an appropriate level of temporal hierarchy, the next key step is to mine frequent patterns at this level by aggregating frequency counts (e.g., summing frequencies of data over this temporal unit). The resulting frequent patterns help reveal frequently observed characteristics at a time instance. By repeating the mining procedure over all temporal points, we then compare and contrast similarities and differences among frequent patterns discovered at these temporal points. As a concrete example, we mine sequences of COVID-19 data on a weekly basis by aggregating their daily counts of various features associated with the data to form the corresponding weekly counts. Then, we discover frequent patterns revealing characteristics (e.g., transmission method, hospital status, clinical outcome) of COVID-19 in a particular week (e.g., Week 8 of 2020).

In terms of visualization for sequential data, many related works focus on visualizing collections of individual sequences. In contrast, in this paper, we focus on visualizing sequences of collections of patterns. For example, instead of analyzing and visualizing the trend of each individual stock, our visual data science solution focuses on visualizing temporal changes in the composition of stocks. As another example, for sequences of COVID-19 data, our solution focuses on visualizing temporal changes in the composition of some features (e.g., raise or drop in the number of domestically infected cases among all transmission methods).

To visualize temporal changes over compositions of features, it is tempting to stacking all pie charts or outward sunburst diagrams (i.e., doughnut charts). Given an outward sunburst diagram for a temporal point, one could repeat the mining and visualization process to generate multiple sunburst diagrams (with one diagram for each temporal point). While the stack of sunburst diagrams capture all information for analysis of temporal changes, it may be challenges to view and comprehend the details for each diagram, let alone discovering their temporal changes.

Instead, for comprehensible view, our visual data science solution represents and visualizes the composition of a feature at a temporal point by a stacked column. We observe that, when visualizing the composition of a (categorical) feature, *each record takes on a single value* (*including NULL*) *for the feature*. Hence, for singleton patterns on a feature, the sum of frequencies of each distinct value should match the total number of records. For non-singleton patterns on *k* features, as each feature comes from a domain, the sum of frequencies of each distinct combination of values for the *k* features should again match the total number of records. As a concreate example, with 2 transmission methods (i.e., domestic acquisition and international travel) and unstated transmission method (i.e., NULL) for the feature "transmission method", the sum of frequencies of these 2+1 = 3 feature values should match the total number of records. With 3 hospital statuses (i.e., ICU, non-ICU hospitalized, not hospitalized) and NULL for an additional feature "hospital status", the sum of frequencies (2+1) × (3+1) = 12 combinations of these two features should match the total number of records. This explains why the height of the stacked column would give the frequency of all records at the temporal point.

For easy comparison of compositions of features over *n* temporal points, our visual data science solution represents these *n* compositions with *n* stacked column arranged according to their temporal order. The height of the entire stacked column gives absolute frequency at time *t*. Changes in the height of the entire stacked column and/or of segments of the column reveal the uptrends or downtrends.

In addition, our solution provides users with an alternative representation, in which compositions are represented by 100% stack column. By doing so, the relative frequency (i.e., percentage composition) of different values of features can be easily observed. Changes in relative frequency can thus be easily observed too.

IV. Evaluation

To evaluate our visual data science solution for visualizing and visually analyzing big sequential data, we applied it to the same real-life Canadian COVID-19 epidemiological dataset collected from Public Health Agency of Canada (PHAC) and Statistics Canada for the period from 2020 to May 29, 2021 (i.e., Week 21 of 2021) mentioned in Section II. Our solution represents compositions of features associated with these 1,368,422 Canadian COVID-19 cases in stacked column or 100% stacked column charts. Each column represents the composition of features in a week.

For example, Fig. 4(a) shows sequences of stacked columns, in which each column represents the composition of two stated transmission methods. Heights of each column clearly indicates the number of cases for each week. The heights of white, light blue, and grey segments of each column indicates the (absolute) frequencies of domestically acquired cases, cases exposed through international travel, and cases without stated transmission methods. In earlier weeks, observable numbers of cases were exposed via travel. In later/recent weeks, more cases whose transmission methods were unknown (and probably still under investigation). The general shape shows the ups and downs (as well as peaks and valleys) of the three waves of COVID-19 in Canada.

To clearly show the relative percentage of these three transmission methods (including NULL/unstated ones), our visual data science solution provides users with a representation in 100% stacked columns. See Fig. 4(b), from which users can easily observe the significant percentages of cases exposed via international travel—e.g., close to 50% of cases infected in Week 9 (i.e., March 01-07) of 2020. The numbers became insignificant starting Week 14 (April 05-11) of 2020 partially due to international travel restriction.

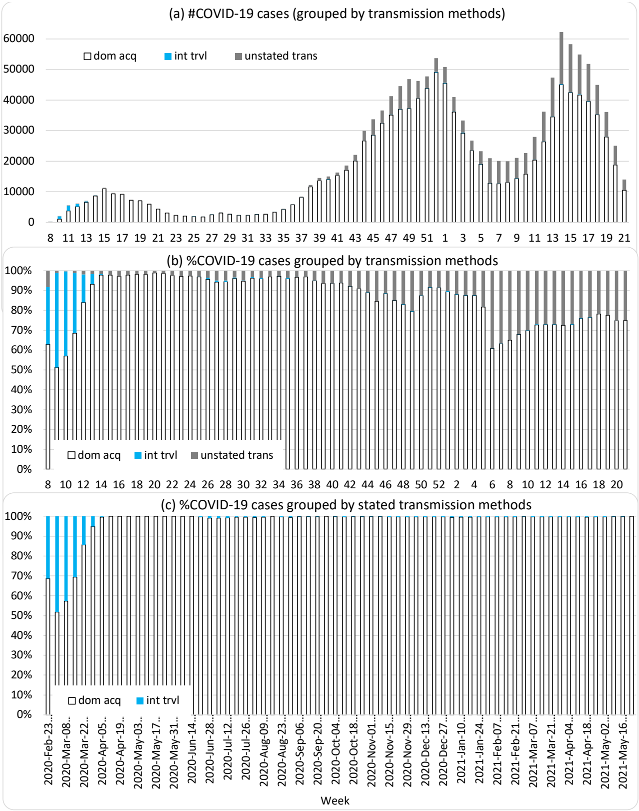

Fig. 4. Our visual data science solution showing (a) absolute frequency and (b) relative percentage of different transmission methods, as well as (c) relative percentage of *stated* transmission methods for Canadian COVID-19 cases from Week 8 of 2020 to Week 21 of 2021 (i.e., Feb. 23, 2020 to May 29, 2021).

For user convenience, our visual data science solution also provides users with options to include or exclude NULL values. As an instance of excluding NULL values, Fig. 4(c) shows the relative percentages of *stated* transmission methods of cases.

To let user explore the composition of different hospital statuses among those exposed via the most frequent transmission method (i.e., domestic acquisition), our data science solution shows the frequencies of the four ⟨"domestic acquisition", *hospital status*⟩-combinations. Fig. 5(a) shows the frequencies of these hospital statuses—namely, not hospitalized, non-ICU hospitalization, ICU hospitalization, and unstated hospital status—by green, yellow, red and white segments of stacked columns. The figure reveals that majority of these domestically acquired cases were not hospitalized. Fig. 5(b) reveals that, during the first wave, close to 20% of domestically acquired cases were hospitalized and close to 10% of them were admitted to the ICU in Week 9 of 2020. Since the second wave, the situation has become stable with less than 10% of domestically acquired cases required hospitalization.

Along this direction, our data science solution also visualizes the frequencies of the three ⟨"domestic acquisition", "not hospitalized", *clinical outcome*⟩-combinations. This reveals that a majority of domestically acquired cases that did not hospitalized were recovered.

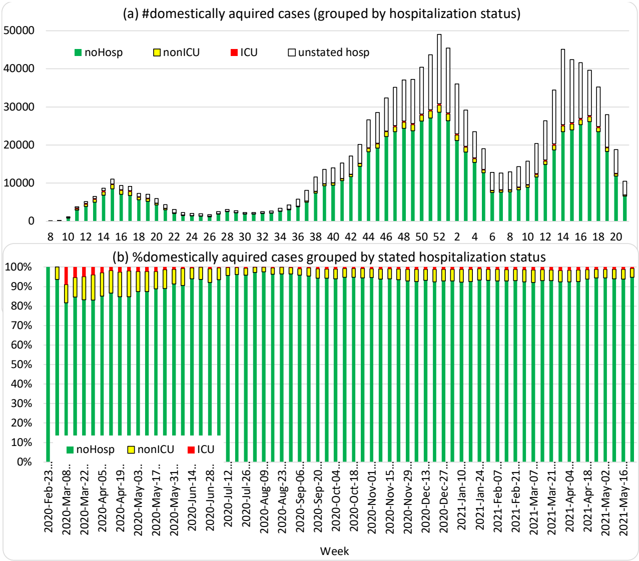

Fig. 5. Our visual data science solution showing (a) absolute frequency and (b) relative percentage of different hospital statuses among those domestically acquired COVID-19 cases.

V. Conclusions

In this paper, we presented a visual data science solution for the visualization and visual analytics of big sequential data. We illustrate the ideas through its applications to real-life COVID-19 epidemiological data. Our solution represents compositions of combinations of feature values in stacked columns, which enables easy comparison in the temporal dimension. Although we evaluated and showed its practicality by using COVID-19 data, it can be applicable to visualization and visual analytics of other big sequential data. As *ongoing and future* work, we further enhance visibility, interpretability and explainability of our visual data science solution in visualization and visual analytics of big sequential data.

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

References

[1] C. Ordonez, et al., "An intelligent visual big data analytics framework for supporting interactive exploration and visualization of big OLAP cubes," IV 2020, pp. 421-427.

[2] A. Perrot, et al., "HeatPipe: high throughput, low latency big data heatmap with Spark streaming," IV 2017, pp. 66-71.

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

[4] A.H. Diallo, et al., "Proportional visualization of genotypes and phenotypes with rainbow boxes: methods and application to sickle cell disease," IV 2019, Part I, pp. 1-6.

[5] S. Hamdi, et al., "Intra and inter relationships between biomedical signals: a VAR model analysis," IV 2019, Part I, pp. 411-416.

[6] M.T. Pellecchia, et al., "Identifying correlations among biomedical data through information retrieval techniques," IV 2019, Part I, pp. 269-274.

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

[8] C.M. Choy, et al., "Natural sciences meet social sciences: census data analytics for detecting home language shifts," IMCOM 2021. DOI: 10.1109/IMCOM51814.2021.9377412

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

[10] D. Jonker, et al., “Industry-driven visual analytics for understanding financial timeseries models,” IV 2019, Part I, pp. 210-215.

[11] N.N.T. Luong, et al., “A visual interactive analytics interface for complex event processing and machine learning processing of financial market data,” IV 2020, pp. 189-194.

[12] 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.

[13] M. Prokofieva, “Visualization of financial data in teaching financial accounting,” IV 2020, pp. 674-678.

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

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

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

[17] M.L. Huang, et al., “Designing infographics/visual icons of social network by referencing to the design concept of ancient oracle bone characters,” IV 2020, pp. 694-699.

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

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

[20] 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.

[21] 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.

[22] C.K. Leung, et al., “Effective classification of ground transportation modes for urban data mining in smart cities,” DaWaK 2018, pp. 83-97.

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

[24] I.M. Shawket, S. El khateeb, “Redefining urban public space's characters after COVID-19; : empirical study on Egyptian residential spaces,” IV 2020, pp. 614-619.

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

[26] 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.

[27] 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.

[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, F. Jiang, “A data science solution for mining interesting patterns from uncertain big data. IEEE BDCloud 2014: 235-242

[30] P. Muñoz-Lago, et al., “Visualising the structure of 18th century operas: a multidisciplinary data science approach,” IV 2020, pp. 530-536.

[31] M.T. Alam, et al., “Mining frequent patterns from hypergraph databases,” PAKDD 2021, Part II, pp. 3-15.

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

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

[34] C.K. Leung, “Uncertain frequent pattern mining,” Frequent Pattern Mining, 2014, pp. 417-453.

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

[36] A. von Richthofen, et al., “Urban mining: visualizing the availability of construction materials for re-use in future cities,” IV 2017, pp. 306-311.

[37] G. Casalino, et al., “Incremental and adaptive fuzzy clustering for virtual learning environments data analysis,” IV 2020, pp. 382-387.

[38] M.L. Huang, et al., “Stroke data analysis through a HVN visual mining platform,” IV 2019, Part II, pp. 1-6.

[39] 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.

[40] W. Lee, et al. (eds.), Big Data Analyses, Services, and Smart Data, 2021.

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

[42] A.P. Afonso, et al., “RoseTrajVis: visual analytics of trajectories with rose diagrams,” IV 2020, pp. 378-384.

[43] L. Kaupp, et al., “An Industry 4.0-ready visual analytics model for context-aware diagnosis in smart manufacturing,” IV 2020, pp. 350-359.

[44] 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.

[45] C. Maçãs, et al., “VaBank: visual analytics for banking transactions,” IV 2020, pp. 336-343.

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

[47] C.K. Leung, et al., “Machine learning and OLAP on big COVID-19 data,” IEEE BigData 2020, pp. 5118-5127.

[48] F.A. Orji, J. Vassileva, “Using machine learning to explore the relation between student engagement and student performance, IV 2020, pp. 480-485.

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

[50] C. Servin, et al., “Adversarial teaching approach to cybersecurity: a mathematical model explains why it works well,” IV 2020, pp. 313-316.

[51] C.K. Leung, et al., “Big data visualization and visual analytics of COVID-19 data,” IV 2020, pp. 415-420.

[52] W. Jentner, D.A. Keim, “Visualization and visual analytic techniques for patterns,” High-Utility Pattern Mining, 2019, pp. 303-337.

[53] G. Bothorel, et al., “Visualization of frequent itemsets with nested circular layout and bundling algorithm,” ISVC 2013, Part II, pp. 396-405.

[54] T. Munzner, et al., Visual mining of power sets with large alphabets. Tech. rep. TR-2005-25, UBC, 2005. https://www.cs.ubc.ca/tr/2005/tr-2005-25

[55] C.K. Leung, et al., “FIsViz: a frequent itemset visualizer,” PAKDD 2008, pp. 644-652.

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

[57] C.K. Leung, et al., “FpMapViz: a space-filling visualization for frequent patterns,” IEEE ICDM 2011 Workshops, pp. 804-811.

[58] B.C.M. Cappers, J.J. van Wijk, “Exploring multivariate event sequences using rules, aggregations, and selections,” IEEE TVCG 24(1), 2018, pp. 532–541.

[59] J. Zhao, et al., “MatrixWave: visual comparison of event sequence data,” ACM CHI 2015, pp. 259–268.

[60] Y. Chen, et al., “Sequence synopsis: optimize visual summary of temporal event data,” IEEE TVCG 24(1), 2018, pp. 45–55.

[61] C.D. Stolper, et al., “Progressive visual analytics: user-driven visual exploration of in-progress analytics,” IEEE TVCG 20(12), 2014, pp. 1653–1662.

[62] W. Jentner, et al., “Feature alignment for the analysis of verbatim text transcripts,” EuroVis 2017 Workshop on EuroVA, pp. 13– 18.