Effectively and Efficiently Supporting Visual Big Data Analytics over Big Sequential Data: An Innovative Data Science Approach

Alfredo Cuzzocrea[1,2(✉)], Majid Abbasi Sisara[1,3], Carson K. Leung[4], Yan Wen[4], and Fan Jiang[5]

[1] iDEA Lab, University of Calabria, Rende, Italy
{alfredo.cuzzocrea,majid.abbasi}@unical.it
[2] LORIA, University of Lorraine, Nancy, France
[3] Department of Engineering, University of Trieste, Trieste, Italy
[4] Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada
kleung@cs.umanitoba.ca
[5] Department of Computer Science, University of Northern British Columbia, Prince George, BC, Canada
Fang.Jiang@unbc.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 willingness of many government, researchers, and organizations are led by the initiates of open data to share their data and make them publicly accessible. Healthcare, disease, and epidemiological data, such as privacy-preserving statistics on patients who suffered from epidemic diseases such as Coronavirus disease 2019 (COVID-19), are examples of open big data. Analyzing these open big data can be for social good. For instance, people get a better understanding of the disease by analyzing and mining the disease statistics, which may inspire them to take part in preventing, detecting, controlling and combating the disease. Having a pictorial representation further enhances the understanding of the data and corresponding results for analysis and mining because a picture is worth a thousand words. Hence, in this paper, we present a visual data science solution for the visualization and visual analytics of big sequential data. The visualization and visual analytics of sequences of real-life COVID-19 epidemiological data illustrate the ideas. Through our solution, we enable users to visualize the COVID-19 epidemiological data over time. It also allows people to visually analyze the data and discover relationships among popular features associated with the COVID-19 cases. The effectiveness of our visual data science solution in enhancing user experience in the visualization and visual analytics of big sequential data are demonstrated by evaluation of these real-life sequential COVID-19 epidemiological data.

Keywords: Information visualization · Big data · Sequences · Data science · Visual data science · Data mining · Data analytics · Visual analytics · COVID-19

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022, LNCS 13376, pp. 113–125, 2022.
https://doi.org/10.1007/978-3-031-10450-3_9

1 Introduction

Big data is a phenomenon characterized by the ability to easily generate and collect large quantities of valuable data at rapid speeds 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 [1], biomedical/healthcare data and disease reports (e.g., COVID-19 statistics) [2–5], census data [6], financial time series [7–11], music data [12], patent register [13, 14], social networks [15–17], transportation and urban data [18–22], weather data [23], and web data [24, 25].

There is valuable information embedded in those open big data [26–28] which also makes good use of data mining algorithms [29–34], data analytics methods [35–39, 56–58], visual analytics methods [40–43], and machine learning tools [2–44]. Hence, analyzing and mining these big data can be for social good. By analyzing and mining disease statistics, people can learn more about the disease, such as:

- An outbreak of SARS in 2003 was caused by a coronavirus (CoV) that was associated with SARS.
- Swine flu, which was caused by influenza A virus subtype H1N1 (A/H1N1) and led to a pandemic from 2009 to mid-2010.
- A MERS-CoV caused Middle East respiratory syndrome (MERS). During the period of 2012–2018, the disease was reported from the Middle East (e.g., Saudi Arabia), while in 2015, it was reported from South Korea.
- Zika virus disease, which was primarily transmitted by the bite of an infected mosquito. An outbreak was reported in Brazil during 2015–2016.
- SARS-CoV-2 caused coronavirus disease 2019 (COVID-19). In early 2020, this outbreak was reported to become a global pandemic, and it 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.

We know well that "a picture is worth a thousand words". It is desirable to represent the discovered information and knowledge visually to make it easier to comprehend. 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 presented a big data visualization and visual analytics tool for visualizing and analyzing COVID-19 epidemiological data was presented by us [45]. By analyzing cumulative COVID-19 statistics, it visualizes the most common features (e.g., frequently used transmission methods, hospitalization status, clinical outcomes) associated with the majority of COVID-19 cases.

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, temporal changes may be not revealed in these characteristics during this period. There is a need for a tool or solution to visualize and analyze COVID-19 epidemiological data on a temporal basis. 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,

a data science solution for visualization and visual analytics of temporal data—i.e., a COVID-19 sequence—in the current IV 2021 paper is designed and developed by us.

We designed and developed our data science solution in such a way that notably it visualizes and visually analyses not only 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, social scientists are helped by visualization and visual analytics of temporal employment data 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 major contribution to this paper is the design and development of a data science solution for visual analytics of big sequential data. Our solution visualizes temporal trends in the sequence. The findings (e.g., discovered information and knowledge) from the visual analytics on the temporal sequential data are also visualized by it. The tool reveals, for example, popular features associated with the data (e.g., popular transmission methods, hospitalization status, and clinical outcomes at certain times) and how these characteristics change over time. To illustrate our key ideas, we apply our solution to real- life COVID-19 sequences. Our solution applies to a real-life COVID-19 sequences for evaluating it.

The remainder of this paper follows the following format. The next section discusses background and related works. Our data science solution for visualization and visual analytics of big sequential data are described in Sect. 3. The results of the evaluation based on Canadian COVID-19 epidemiological data are presented in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 Background Analysis

As for application to visualization and visual analytics of COVID-19 data, they have developed many visualizers and dashboards since its declaration as a pandemic. In this regard, the following dashboards are noteworthy: (a) the COVID-19 dashboard created by the World Health Organization (WHO); (b) the COVID-19 dashboard developed by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University; and (c) the COVID-19 dashboard created by the European Center for Disease Prevention and Control (ECDC). They provide summary for global COVID-19 situations. Moreover, visualizers and dash- boards are provided by local governments (e.g., Government of Canada4) and media (e.g., TV5, 6, Wikipedia7) for local COVID-19 situations. A feature that these visualizers and dashboards all have in common is that they display the number 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, we embed additional information and knowledge in the data and yet to be discovered.

To address this issue, we developed a big data visualization and visual analytics tool for identifying infrequent patterns in COVID-19 statistics. In terms of related works on visualizing frequent patterns, Jentner and Keim [46] surveyed several visualization techniques for frequent patterns. we can generalize these techniques into four categories:

- Frequent patterns are best represented visually through lattice representation [53]. With it, frequent patterns are represented as nodes in a lattice (aka concept hierarchy). Edges connect immediate supersets and subsets of a frequent pattern.
- Using pixels as the representation, a frequent k-itemset (a pattern consisting of k items) is represented by a pixel [47].
- Linear visualization, in which frequent patterns are represented linearly. For example, a frequent k-itemset in a polyline is represented by *FIsViz* [48] that connects k nodes in a 2-dimensional space. Rather than represent frequent patterns in an orthogonal diagram (i.e., a wiring-type diagram), *FpVAT* [42] represents them as frequent patterns.
- Tree visualization, in which frequent patterns are represented according to a tree hierarchy. For example, frequent patterns with a side-view of the pyramid are shown by *PyramidViz* [49], in which puts short patterns on the bottom of the pyramid and puts longer related patterns (e.g., extensions of short patterns) on the top. Likewise, *FpMapViz* [50], which shows frequent patterns by overlaying related patterns over shorter patterns, shows frequent patterns in the foreground and background.

Similar to the aforementioned tree visualization, our big data visualization and visual analytics tool [45] for visualizing frequent patterns from the cumulative COVID-19 statistics also follows a hierarchy so that patterns can be connected to their extensions. We can consider it as showing frequent patterns with a top-view. This scheme, however, instead of placing short patterns in the background and overlaying long related patterns in the foreground, places short patterns in the inner ring near the center and long related patterns in the outer ring (although they do not overlap or overlay with the inner ring). Immediate extensions of a pattern are put just outside (but touching) the sector representing the pattern. More specifically, we represent the frequent patterns and their related 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 Canada8 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. Figure 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 the white ring sector represent it) 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.

Expanding the frequent singleton pattern, we find a frequent 2-itemset (i.e., non-singleton pattern) {*domestic acquisition*, *non-hospitalized*}:52.41%. This reveals that, among those who domestically acquired the disease from community exposures, a large fraction ($52.41\%/82.35\% \approx 0.64$) of them did not required hospitalization. This is shown in Fig. 2 that 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. In addition, the figure reveals that, while most of them did not require hospitalization, a significant proportion ($25.83\%/82.35\% \approx$

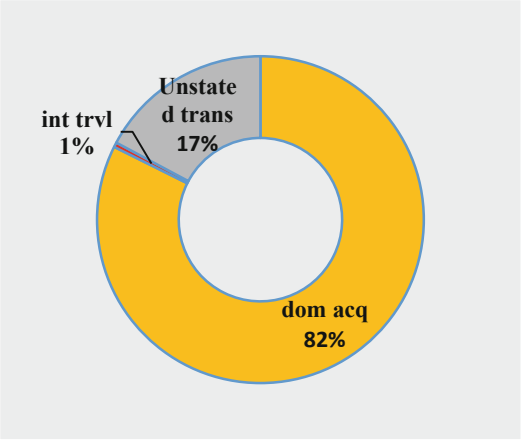


Fig. 1. Visualization of transmission methods.

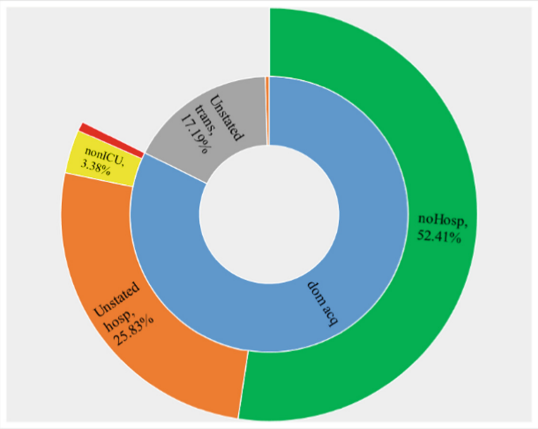


Fig. 2. Visualization of hospital status among those who domestically acquired COVID-19 via community exposures.

0.31) were not listed in hospitals (represented by the gray rings). Smaller fractions of them were admitted to the hospital. Specifically, it hospitalized 3.38%/82.35% $\approx$ 0.04 of them but did not require to admit to the intensive care unit (ICU), as the small yellow ring sector represents it; it admitted 0.73%/82.35% $\approx$ 0.01 of them to the ICU, as the tiny red ring sector represents it.

Among those who domestically acquired but not hospitalized, a significant number (50.87%/52.41% $\approx$ 0.97) of them recovered. In this direction, a frequent 3-itemset {*domestic acquisition*, *not hospitalized*, *recovered*}:50.87% indicates that, among those who domestically acquired but not hospitalized, a significant number (50.87%/52.41% $\approx$ 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 remaining two tiny fractions (i.e., $0.91\%/52.41\% \approx 0.02$ and $0.63\%/52.41\% \approx 0.01$) belong to those without stated clinical outcome and those deceased, as represented by grey and black ring sectors respectively is revealed by figure.

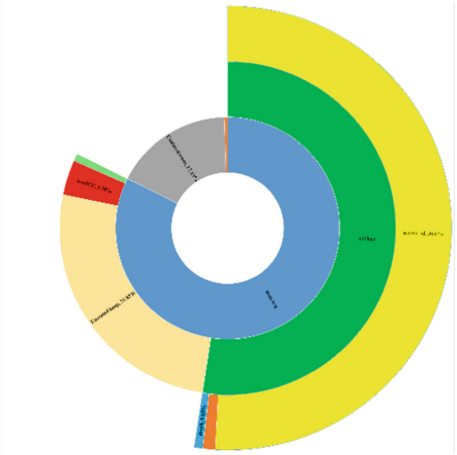


Fig. 3. Visualization of clinical outcomes among those who domestically acquired COVID-19 via community exposures but did not require hospitalization.

A visual data science solution for showing temporal changes and visualizing big sequential data has been designed, developed, and implemented in response to the observation that visualization of frequent patterns may not reveal temporal changes. Jenter and Keim [46] also surveyed several visualization techniques for sequential patterns. Individual representation [51] is included, flow diagram visualization [52], aggregated pattern visualization [53], visual representation with pattern placement strategies [54], and episode visualization [55].

3 An Innovative Data Science Approach for Supporting Visual Big Data Analytics Over Big Sequential Data Effectively and Efficiently

We develop a visual data science solution to explore and visualize temporal changes. It first collects and integrates data from a wide variety of rich data sources. Then the data is preprocessed and a temporal hierarchy to generalize the temporal data is built by it. A fine granularity of data can be collected depending on the application. Having the temporal hierarchy enables us to pick an appreciate level of data granularity. For instance, although it can collect dynamic streaming data at a rapid rate (e.g., per second, minute), may lead to more meaningful and interesting results may be led by analyzing data aggregated at a coarser level (e.g., hourly daily). COVID-19 statistics, for example, are usually updated every day. 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, they are appeared appropriate by analyzing them on a weekly basis.

In addition to choosing the appropriate level of temporal hierarchy, the next key step is to aggregate frequency counts at this level (for example, summing the frequencies over this period of time). The resulting frequent patterns help reveal frequently observed characteristics at a time instance. By repeating the mining procedure over all temporal points, then similarities and differences among frequent patterns discovered at these temporal points are compared and contrasted by us. Using COVID-19 data, we aggregate their daily counts of multiple features associated with the data on a weekly basis to obtain 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. Contrary to that, we focus here on visualizing patterns that occur in sequences. 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).

Pie charts and sunburst diagrams (i.e., doughnut charts) are tempting ways to visualize temporal changes over compositions of features. 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 all information for analysis of temporal changes is captured by the stack of sunburst diagrams, it may be challenges to view and comprehend the details for each diagram, let alone discovering their temporal changes.

The visual data science solution represents and visualizes a feature's temporal composition by using stacked columns to make the composition easy to understand. 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 total number of records should be matched the sum of frequencies of each distinct value. If there are no singleton patterns on the k features, then the sum of the frequencies of the k features, since each one comes from a distinct domain, should be equal to the total number of records that exist. 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 $(2+1)\times(3+1)=12$ combinations of these two features should be matched the total number of records. As a result, the height of the stacked column indicates the frequency of all records at the particular time.

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. Absolute frequency at time t is given by The height of the entire stacked column. In some cases, the height of the entire stacked column or segments of it may reveal the uptrend or downtrend.

In addition, our solution provides users with an alternative representation, in which compositions are represented by 100% stack column. By doing so, we can easily observe the relative frequency (i.e., percentage composition) of different values of features. It is also possible to observe changes in relative frequency.

4 Experimental Assessment and Analysis

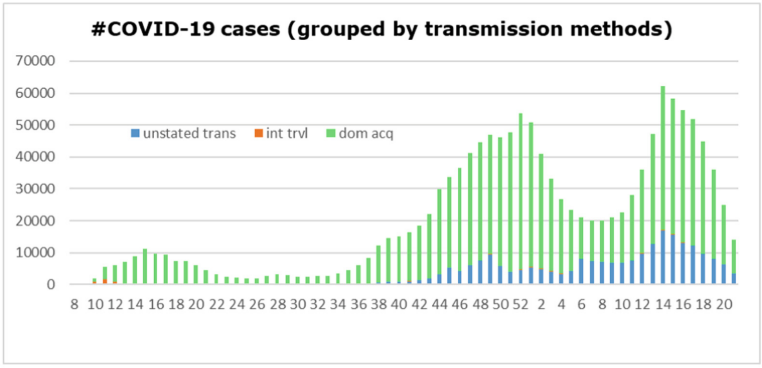
To evaluate our visual data science solution for visualizing and visually analyzing big sequential data, we applied it to the same real-life Canadian COVID-19 epidemiological dataset collected from Public Health Agency of Canada (PHAC) and Statistics Canada for the period from 2020 to May 29, 2021 (i.e.,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

For example, Fig. 4(*a*) shows sequences of stacked columns, in which each column represents the composition of two stated transmission methods. The number of cases for each week is clearly indicated by heights of each column. In each column, the (absolute) heights of the segments representing the domestically acquired cases, foreign-exposed cases, and cases without indication of the method of transmission are shown. 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). As the general pattern shows, the three waves of COVID-19 in Canada oscillate up and down (as well as peak and valley).

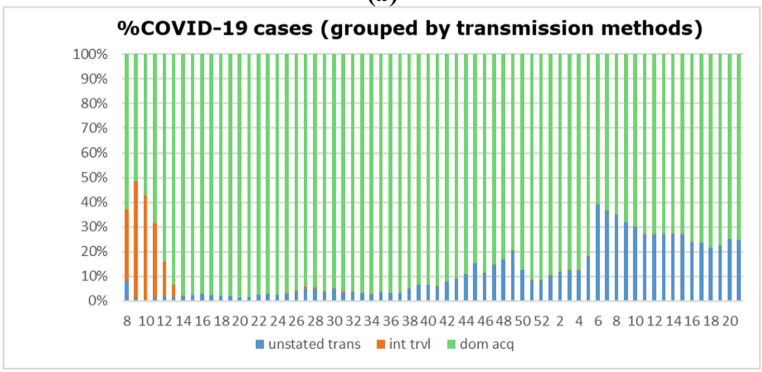
To clearly show the relative percentage of these three transmission methods (including NULL/unstated ones), our visual data science solution provides users with a representation in 100% stacked columns. See Fig. 4(*b*), the significant percentages of cases exposed via international travel—e.g., close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

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, the relative percentages of stated transmission methods of cases is shown by Fig.4(*c*).

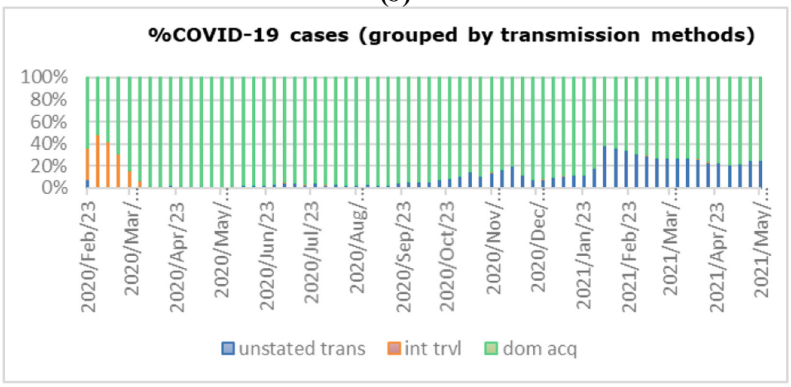
Data Science solution reveals the frequencies of four "*domestic acquisition*", *-combinations of hospital status within the exposure group based on the most frequent transmission method (i.e., domestic acquisition). Figure 5(*a*) shows the frequencies of these hospital statuses—namely, not hospitalized, non-ICU hospitalization, ICU hospitalization, and unstated hospital status—by green, yellow, red and white segments of stacked columns. That majority of these domestically acquired cases were not hospitalized is revealed by the figure. As shown in Fig. 5(*b*) in Week 9 of 2020, almost 20%



(*a*)



(*b*)



(*c*)

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).

of cases acquired domestically led to hospitalization, and close to 10% to ICU admission. Since the second wave, the situation has become stable with less than 10% of domestically acquired cases required hospitalization.

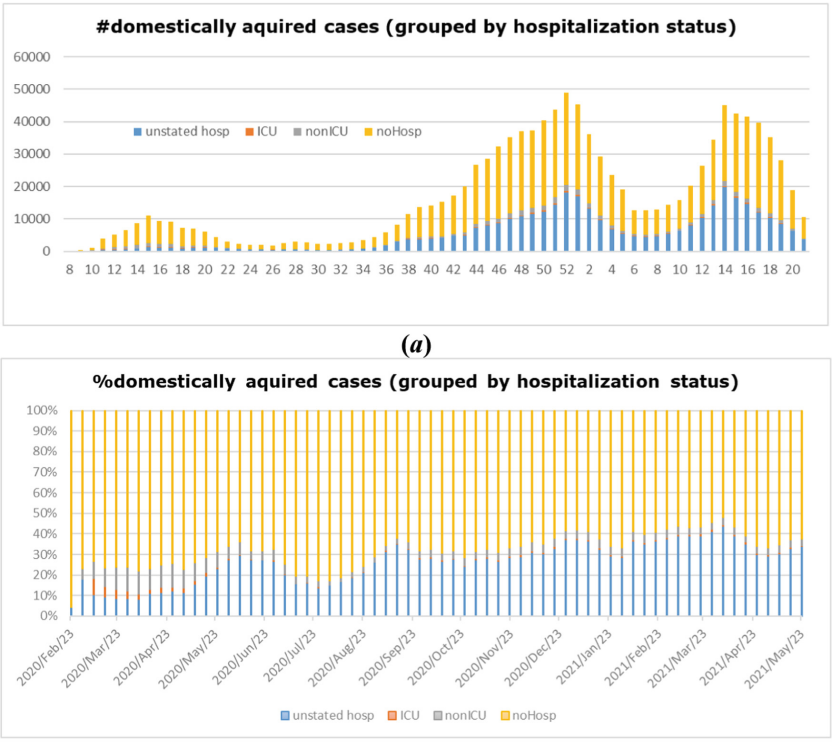


Fig. 5. Our visual data science solution showing (*a*) absolute frequency and (*b*) relative percentage of different hospital statuses among those domestically acquired COVID-19 cases.

Along this direction, the frequencies of the three ⟨*domestic acquisition*, *not hospitalized*, *clinical outcome*⟩-combinations also are visualized by our data science solution. Domestically acquired cases that were not hospitalized for longer than one week were more likely to be recovered than those that were.

5 Conclusions and Future Work

In this paper, we presented a visual data science solution for the visualization and visual analytics of big sequential data. The ideas through its applications to real-life COVID- 19 epidemiological data are illustrated by us. In this approach, we represent feature values in stacked columns, which makes temporal comparisons easier. 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, visibility, interpretability and explainability of our visual data science solution in visualization and visual analytics of big sequential data are further enhanced by us.

References

1. Anderson-Grégoire, I.M., et al.: A big data science solution for analytics on moving objects. AINA **2**, 133–145 (2021)
2. Diallo, A.H., 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 (2019)
3. Hamdi, S., et al.: Intra and inter relationships between biomedical signals: a VAR model analysis. IV 2019, Part I, pp. 411–416 (2019)
4. Pellecchia, M.T., et al.: Identifying correlations among biomedical data through information retrieval techniques. IV 2019, Part I, pp. 269–274 (2019)
5. Shang, S., et al.: Spatial data science of COVID-19 data. IEEE HPCC- SmartCity-DSS 2020, pp. 1370–1375 (2020)
6. Choy, C.M., et al.: Natural sciences meet social sciences: census data analytics for detecting home language shifts. IMCOM 2021, pp. 1–8 (2021)
7. Chanda, A.K., et al.: A new framework for mining weighted periodic patterns in time series databases. ESWA **79**, 207–224 (2017)
8. Jonker, D., et al.: Industry-driven visual analytics for understanding financial timeseries models. IV 2019, Part I, pp. 210–215 (2019)
9. Luong, N.N.T., et al.: A visual interactive analytics interface for complex event processing and machine learning processing of financial market data. IV 2020, pp. 189–194 (2020)
10. Morris, K.J., 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 (2018)
11. Prokofieva, M.: Visualization of financial data in teaching financial accounting, IV 2020, pp. 674–678 (2020)
12. Barkwell, K.E., et al.: Big data visualisation and visual analytics for music data mining. IV 2018, pp. 235–240 (2018)
13. Lee, W., et al.: Reducing noises for recall-oriented patent retrieval. IEEE BDCloud 2014, pp. 579–586 (2014)
14. Leung, C.K., et al.: Information technology-based patent retrieval model. Springer Handbook of Science and Technology Indicators, pp. 859–874 (2019)
15. Huang, M.L., 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 (2020)
16. Jiang, F., et al.: Finding popular friends in social networks. CGC 2012, pp. 501–508 (2012)
17. Singh, S.P., Leung, C.K.: A theoretical approach for discovery of friends from directed social graphs. IEEE/ACM ASONAM 2020, pp. 697–701 (2020)
18. Audu, A.-R.A., Cuzzocrea, A., Leung, C.K., MacLeod, K.A., Ohin, N.I., Pulgar-Vidal, N.C.: An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city. In: Barolli, L., Hussain, F.K., Ikeda, M. (eds.) CISIS 2019. AISC, vol. 993, pp. 224–236. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-22354-0_21
19. Balbin, P.P.F., et al.: Predictive analytics on open big data for supporting smart transportation services. Procedia Comput. Sci. **176**, 3009–3018 (2020)
20. Leung, C.K., et al.: Effective classification of ground transportation modes for urban data mining in smart cities. DaWaK 2018, pp. 83–97 (2018)

21. Leung, C.K., et al.: Urban analytics of big transportation data for supporting smart cities. DaWaK 2019, pp. 24–33 (2019)
22. Shawket, I.M., El khateeb, S.: Redefining urban public space's characters after COVID-19: empirical study on Egyptian residential spaces. IV 2020, pp. 614–619 (2020)
23. Cox, T.S., et al.: An accurate model for hurricane trajectory prediction. IEEE COMPSAC 2018, vol. 2, pp. 534–539 (2018)
24. Leung, C.K., et al.: Explainable machine learning and mining of influential patterns from sparse web. IEEE/WIC/ACM WI-IAT 2020, pp. 829–836 (2020)
25. Singh, S.P., et al.: Analytics of similar-sounding names from the web with phonetic based clustering. IEEE/WIC/ACM WI-IAT 2020, pp. 580–585 (2020)
26. Dierckens, K.E., et al.: A data science and engineering solution for fast k-means clustering of big data. IEEE TrustCom-BigDataSE-ICESS 2017, pp. 925–932 (2017)
27. Leung, C.K., Jiang, F.: A data science solution for mining interesting patterns from uncertain big data. IEEE BDCloud 2014, pp. 235–242 (2014)
28. Muñoz-Lago, P., et al.: Visualising the structure of 18th century operas: a multidisciplinary data science approach. IV 2020, pp. 530–536 (2020)
29. Alam, M.T., et al.: Mining frequent patterns from hypergraph databases. PAKDD 2021, Part II, pp. 3–15 (2021)
30. Fariha, A., et al.: Mining frequent patterns from human interactions in meetings using directed acyclic graphs. PAKDD 2013, Part I, pp. 38–49 (2013)
31. Leung, C.K.: Big data analysis and mining. Encyclop. Inf. Sci. Technol. **4e**, 338–348 (2018)
32. Leung, C.K.: Uncertain frequent pattern mining. Frequent Pattern Mining, pp. 417–453 (2014)
33. Roy, K.K., et al.: Mining sequential patterns in uncertain databases using hierarchical index structure. PAKDD 2021, Part II, pp. 29–41 (2021)
34. von Richthofen, A., et al.: Urban mining: visualizing the availability of construction materials for re-use in future cities. IV 2017, pp. 306–311 (2017)
35. Casalino, G., et al.: Incremental and adaptive fuzzy clustering for virtual learning environments data analysis. IV 2020, pp. 382–387 (2020)
36. Huang, M.L., et al.: Stroke data analysis through a HVN visual mining platform. IV 2019, Part II, pp. 1–6 (2019)
37. Jiang, F., Leung, C.K.: A data analytic algorithm for managing, querying, and processing uncertain big data in cloud environments. Algorithms **8**(4), 1175–1194 (2015)
38. W. Lee, et al. (eds.): Big Data Analyses, Services, and Smart Data (2021)
39. Leung, C.K., Jiang, F.: Big data analytics of social networks for the discovery of "following" patterns. DaWaK 2015, pp. 123–135 (2015)
40. Afonso, A.P., et al.: RoseTrajVis: visual analytics of trajectories with rose diagrams. IV 2020, pp. 378–384 (2020)
41. Kaupp, L., et al.: An Industry 4.0-ready visual analytics model for context-aware diagnosis in smart manufacturing. IV 2020, pp. 350–359 (2020)
42. Leung, C.K., Carmichael, C.L.: FpVAT: A visual analytic tool for supporting frequent pattern mining. ACM SIGKDD Explor. **11**(2), 39–48 (2009)
43. Maçãs, C., et al.: VaBank: visual analytics for banking transactions. IV 2020, pp. 336–343 (2020)
44. Ahn, S., 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 (2019)
45. Leung, C.K., et al.: Big data visualization and visual analytics of COVID- 19 data. IV 2020, pp. 415–420 (2020)
46. Jentner, W., Keim, D.A.: Visualization and visual analytic techniques for patterns. High-Utility Pattern Mining, pp. 303–337 (2019)

47. Munzner, T., 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
48. Leung, C.K., et al.: FIsViz: a frequent itemset visualizer. PAKDD 2008, pp. 644–652 (2008)
49. Leung, C.K., et al.: PyramidViz: visual analytics and big data visualization of frequent patterns. IEEE DASC-PICom-DataCom- CyberSciTech 2016, pp. 913–916 (2016)
50. Leung, C.K., et al.: FpMapViz: a space-filling visualization for frequent patterns. IEEE ICDM 2011 Workshops, pp. 804–811 (2011)
51. Cappers, B.C.M., van Wijk, J.J.: Exploring multivariate event sequences using rules, aggregations, and selections. IEEE TVCG **24**(1), 532–541 (2018)
52. Zhao, J., et al.: MatrixWave: visual comparison of event sequence data. ACM CHI 2015, pp. 259–268 (2015)
53. Chen, Y., et al.: Sequence synopsis: optimize visual summary of temporal event data. IEEE TVCG **24**(1), 45–55 (2018)
54. Stolper, C.D., et al.: Progressive visual analytics: user-driven visual exploration of in-progress analytics. IEEE TVCG **20**(12), 1653–1662 (2014)
55. Jentner, W., et al.: Feature alignment for the analysis of verbatim text transcripts. EuroVis 2017 Workshop on EuroVA, pp. 13– 18 (2017)
56. Cuzzocrea, A., et al.: Fragmenting very large XML data warehouses via K-means clustering algorithm. Int. J. Bus. Intell. Data Min. **4**(3/4), 301–328 (2009)
57. Ceci, M., Cuzzocrea, A., Malerba, D.: Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering. J. Intell. Inf. Syst. **44**(3), 309–333 (2013). https://doi.org/10.1007/s10844-013-0268-1
58. Bellatreche, L., et al.: F&A: a methodology for effectively and efficiently designing parallel relational data warehouses on heterogenous database clusters. DaWak 2010, pp. 89–104 (2010)