DEPARTMENT: Applications

Spatio-Temporal Urban Data Analysis

A Visual Analytics Perspective

Harish Doraiswamy
New York University

Juliana Freire
New York University

Marcos Lage
Universidade Federal Fluminense

Fabio Miranda
New York University

Claudio Silva
New York University

Editor:
Mike Potel
potel@wildcrest.com

Visual analytics systems can greatly help in the analysis of urban data allowing domain experts from academia and city governments to better understand cities, and thus enable better operations, informed planning and policies. Effectively designing these systems is challenging and requires bringing together methods from different domains. In this paper, we discuss the challenges involved in designing a visual analytics system to interactively explore large spatio-temporal data sets and give an overview of our research that combines visualization and data management to tackle these challenges.

Cities are centers of resource consumption, of economic activity, and of innovation. At the same time, inadequate expansion leads to serious problems, from rapid sprawl and pollution to increased social inequality. Our increasing ability to collect, transmit, and store data, coupled with the growing trend towards openness,[1] creates a unique opportunity that can benefit government, science, residents and industry alike.

Urban data is unique in that it captures the behavior of the different components of a city, namely its residents, existing infrastructure (physical and policies), and the environment. To understand a city, it is important to *explore these components and how they interact over space and time*. Urban data analysis, however, has often been limited to well-defined questions, what Tukey described as *confirmatory data analysis*.[2] The common practice is for domain experts to formulate hypotheses on the basis of theory and anecdotal experience, then for data scientists to select relevant data and carry out analyses, and finally, the domain experts can inspect the results to either disprove or support the hypotheses. Such a batch-oriented analysis pipeline hampers exploration across data sets essential for understanding trends and potential causal mechanisms. The lack of interactivity, along with the recent explosion in data volume and complexity, make it clear that this process cannot scale.

Visualization and visual analytics systems have been successful at enabling users to obtain insight: well-designed visualizations substitute perception for cognition, freeing up limited cognitive/memory resources for higher-level problems.[3] In this paper, with the focus on spatio-temporal data sets, we discuss the various challenges involved in designing visual analytics systems catered to empower domain experts explore large urban data sets, illustrated through an overview of our recent work discussing the research initiated towards the design of such systems to better understand as well as improve cities.

VISUAL ANALYTICS FRAMEWORK FOR URBAN DATA SETS

A visual analytics system requires the integration of techniques from multiple domains, including visualization and data management, as well as a plethora of analytics methods (Figure 1). There are many challenges involved in designing such systems to explore urban data. First, although there has been much work on scaling databases for big data, existing technologies do not meet the interactivity requirements of these systems. It is thus important to bridge the gap between database techniques and visualization systems.[4] Second, a significant number of urban data sets contain one or more spatial (or geometric) components. Formulating complex queries over these data is difficult and evaluating them is computationally expensive, making it difficult to attain the interactive speeds visual analytics demands. Third, manual (exhaustive) exploration of large data sets is not only time consuming, but often becomes impractical. Analytics techniques are needed that help focus the users' attention and guide them during the exploration process. Fourth, the dependency on data scientists to carry out analyses distances the domain experts from the data, limiting their opportunity to generate new hypotheses and explore new directions. Thus, these tools must be usable and within reach for domain experts who often lack computer science expertise.

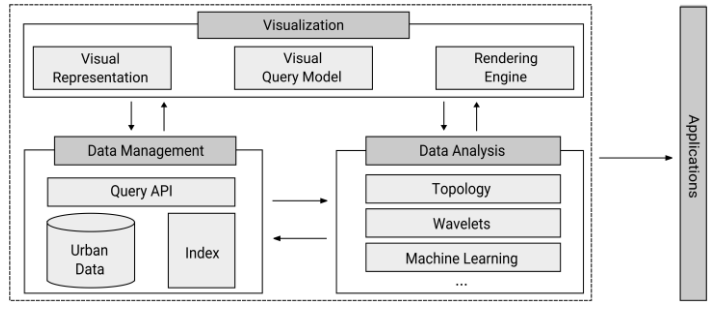

Figure 1. Typical architecture of a visual analytics system consists of three tightly coupled components: visualization, data management and data analysis. The final composition of the components is defined by the end application using such a framework.

In what follows, we briefly discuss the requirements and corresponding challenges involved in designing the components of visual analytics systems. We also give an overview of some of our research that addresses these challenges.

Visualization

Visual analytics systems typically consist of a collection of linked visualization widgets providing users with different perspectives of the data. Traditionally, these widgets use common visualization abstractions such as heat and choropleth maps (see Figure 2), histograms, time series, etc. Exploration is achieved through queries over the data, usually expressed in SQL, to populate the visualizations. However, this approach is not effective when working with data that contain one or more spatial and temporal attributes, a common feature of urban data.

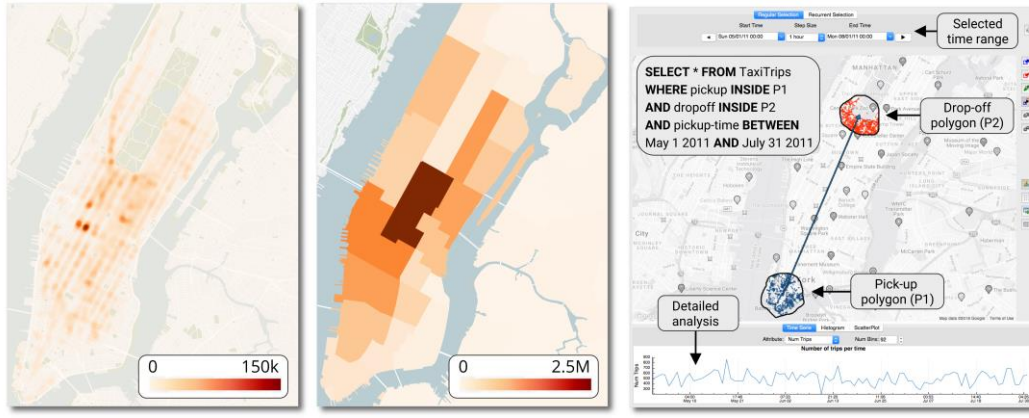

Figure 2. The number of taxi pick-ups for a 1 month period visualized as a heat map (left) and over the neighborhoods of NYC (center). As illustrated (right), the TaxiVis system allows users to specify queries visually: here, a user select trips within a 3 month period that have pickups in Lower Manhattan and dropoffs in the selected region in Midtown. The query results are visualized over the map, as well as a time series showing daily variation of these trips.

For example, consider the well-known New York City's (NYC) Taxi data set (www.nyc.gov/html/tlc/html/about/trip_record_data.shtml). This data captures detailed information about the 500,000 daily trips carried out by yellow cabs in NYC. It consists of two spatial attributes (pickup and dropoff locations), two temporal attributes (pickup and dropoff times), and additional attributes including distance traveled, fare and tip amount. Queries to explore trip patterns (e.g., exploring properties of trips between two locations of interest within a given time frame) over the data are difficult to express using traditional SQL. To address this challenge, we designed a novel visual query model[5] that allows users to visually express complex spatio-temporal queries over *origin-destination* data such as the taxi data. This model encodes constraints corresponding to the components of such data: *spatial, temporal, and attributes*. The model is expressive and supports a wide class of queries, including the query classes for spatio-temporal data defined in Peuquet's triad framework.[6] Figure 2 (right) illustrates one such visual query that returns trips from lower Manhattan to Midtown for a period of three months.

Another challenge arises from the fact that existing visual encodings are often not sufficient for exploring urban data. Considering again the taxi data, given the large number of trips, traditional methods of visualizing trip paths heavily clutter the visualization making it ineffective. To reduce such clutter, we introduced a novel clustering technique that uses vector fields to induce a notion of similarity between trajectories, thus allowing for the definition and representation of clusters.[7] Furthermore, to capture the local mobility dynamics resultant from the collective motion of taxis (or vehicle in general), we also proposed a visual representation for vehicular movement based on vector field visualization techniques.[8]

While the above solutions are for just one class of data set (origin-destination data), similar challenges arise for other data classes as well. For example, public transportation is a key component of any city. However, analyzing the schedule efficiency is a complex process. To ease this process in the analysis of subway efficiency, inspired by the famous *Marey's train schedule*, we designed a visual representation that allows users to easily identify, inspect and compare spatio-temporal patterns between planned and actual subway service.[9]

Thus, when working with urban data sets, it is essential to carefully design both query models as well as visual representations that enable easy and effective visual exploration of these data.

Data Analysis

While visually exploring urban data sets can greatly help in the understanding of the data and their underlying context, to fully understand urban processes, it is important to analyze how the different components of a city, its residents, infrastructure and environment, interact and change

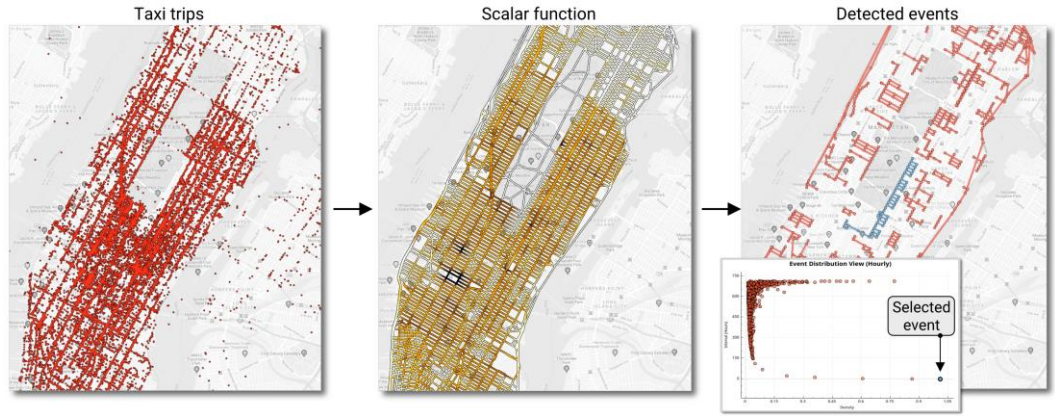

Figure 3. Taxi pickup and dropoff locations (left) are transformed into a scalar function defined over the nodes of the road network of NYC (middle). The topological features of this scalar function are used as the set of potential events. Their properties (visualized using the scatter plot) allow users to select interesting events. The selected event (right) captures the Israeli Day Parade in 2012.

through space and time. This introduces two main problems. As mentioned earlier, manual exploration is often not practical and makes the task of identifying interesting phenomena (or features) from the data difficult (e.g., the taxi data comprises information about over 170 million trips per year). While aggregation and sampling can help overcome this problem, they do so at the cost of occluding small or local patterns in the data. Therefore, mechanisms are needed to guide users during data analysis. Moreover, several phenomena intrinsic to the urban fabric are not described by any single data set. Thus, we need to make collective use of several of these data sets to derive potentially interesting properties of a city.

To help guide users towards interesting patterns and data slices in the taxi data, we designed an automated event-detection algorithm using concepts from computational topology.[10] The events are identified as local topological features (Figure 3) that are "globally interesting" from a topological perspective which is captured by the notion of *topological persistence*. In recent work, Valdivia et al. have explored the use of graph wavelets in combination with pattern recognition/classification to analyze NYC taxi data (Figure 4).[11] By properly handling wavelet coefficients as feature descriptors for classification, such an approach removes the need for specific knowledge of the particularities of the transform and is able to reveal regions and time intervals with similar dynamic, regardless of their spatial and temporal distance.

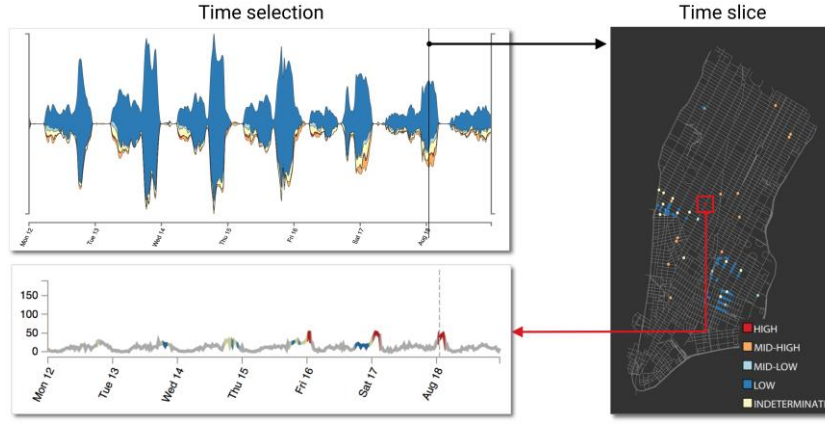

Figure 4. Graph wavelets are used to identify frequency classes based on the temporal behavior of the taxi trip locations (left top). Selecting a time slice allows visualizing the locations belonging to the same frequency class (right). Individual locations (one in a high frequency class is chosen) can be further explored to study their temporal behavior (left bottom).

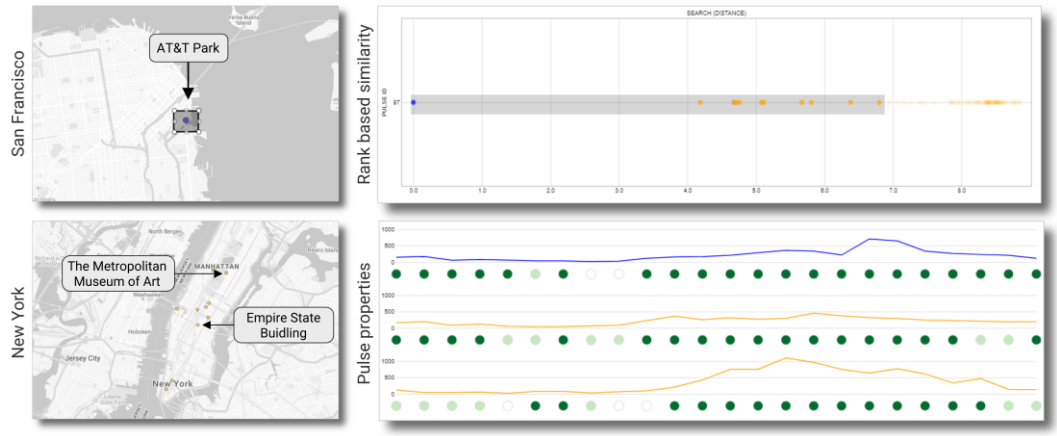

Figure 5. Using locations of photos posted on Flickr.com as a proxy to compare tourist activity patterns in San Francisco (SF) and NYC. The top 10 pulses in NYC similar to the one at AT&T Park in SF are selected. The top two among these correspond to the Empire State Building and the Met museum. In addition to the locations, the properties of the pulses corresponding to these locations are visualized using a custom visual encoding.

Techniques such as the ones mentioned above focus on identifying features or events in the data in the context of a city. However, as mentioned earlier, to *understand the phenomena in a city* it is crucial to support the simultaneous analysis of multiple data sets. Towards this goal, we proposed the concept of *urban pulse* that captures the spatio-temporal dynamics across a city over multiple temporal resolutions.[12] The key idea is to capture both the spatial and temporal variation of the "activity" in a city from the various data sets, which we aim to represent as a collection of *pulses*. In conjunction with a visual interface, these pulses can then be used to not only compare the behavior of different parts of a city, but also compare multiple cities based on the data. Figure 5 shows the web interface of this system and illustrates a scenario of comparing the tourist patterns between San Francisco and NYC.

Furthermore, to provide explanations for the features and event patterns identified from a given data set, we proposed the *data polygamy framework*, that uses concepts from computational topology to identify possible relationships between salient features in data sets.[13]

Data Management

To be effective, visual analytics systems must be *interactive*, requiring sub-second response times.[3] Liu and Heer[3] have shown that even a 500-milliseconds difference can significantly impact visual analysis, reducing interaction and data set coverage during analysis, as well as the rate in which users make observations, draw generalizations and generate hypotheses. However, having been designed for batch queries issued through text-based or terminal interfaces, existing relational database technologies and business intelligence systems used for OLAP (OnLine Analytical Processing) analyses are not suitable for interactive tools.[14]

Thus, an essential component in the design of visual analytics tools for urban data analysis is to enable interactive query performance in the presence of large spatio-temporal data sets. Our initial focus in this aspect has been catered towards specific data types and queries that are common in visual interfaces. To efficiently handle interactive OLAP queries over large time series data, we designed the Time Lattice data structure,[15] that enables sub-second responses to queries over time series having even a billion points while supporting interactive updates of new data.

For spatio-temporal data, we designed approaches that leverage the parallelism provided by GPUs to significantly speedup query processing. STIG[14] (Spatio-Temporal Indexing using GPUs) is an indexing scheme that supports complex spatio-temporal selection queries over large, historical data at interactive rates — attaining over 2 orders of magnitude speedup compared to commercial relational databases. To handle queries that perform spatial aggregation that involve

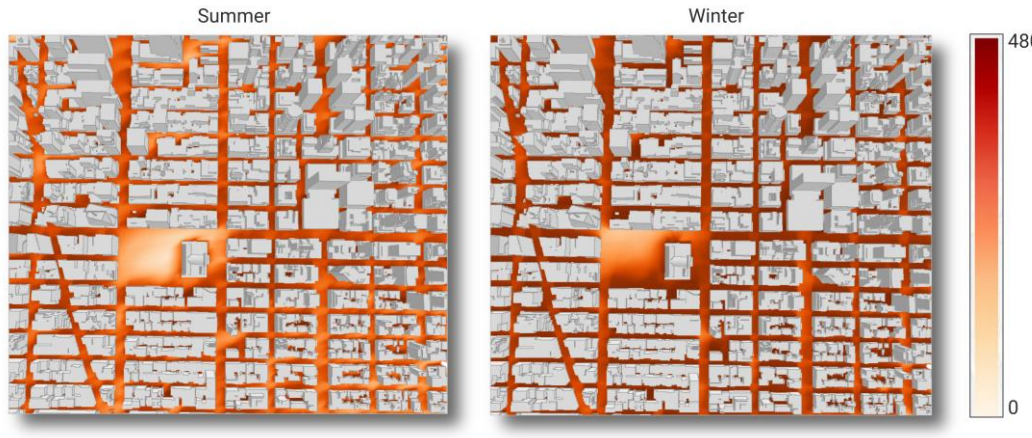

Figure 6. Comparing the average amount of shadows cast per day (in minutes) on the streets of NYC during summer and winter. This analysis reveals that there are several streets that are always in shadow during winter.

aggregating the results of a spatial join between the urban data and a set of spatial regions, we proposed Raster Join,[16] a technique that uses GPU rasterization to efficiently support these queries. It requires only about a second to handle queries comprising of over 850 million points.

Putting it all together

The different techniques mentioned above were proposed as part of visual analytics tools we developed for specific problems faced during the analysis of the different urban data sets. For instance, we built TaxiVis (github.com/ViDA-NYU/TaxiVis), an open-source system that uses the visual query model as well as the STIG index to enable the interactive exploration of the taxi data. The system was deployed at the NYC Department of Transportation. The topology-based approach to guide users, as well as the taxi flow system, were implemented on top of the TaxiVis framework.

Many of the systems we built were designed in collaboration with domain experts. The transportation schedule explorer system, which uses visual representation inspired by *Marey's* schedule, was designed in collaboration with experts from NYC Metropolitan Transportation Authority.[9] Similarly, we designed the Urban Pulse system (github.com/ViDA-NYU/urban-pulse) in collaboration with architects from Kohn Pedersen Fox (KPF), a leading architecture firm in NYC. We also describe in the next section, a more recent system, Urbane, that supports operations that go beyond the traditional 2D-based analysis to better understand and analyze cities, and is also being used by our architect collaborators.

MOVING BEYOND FLATLAND

Existing GIS and visual analytics tools to explore and analyze urban data typically work in 2-dimensions with the map providing the primary context. However, given the rapid verticalization of cities, many phenomena of interest to various stakeholders, including policy makers, urban planners and architects are inherently 3-dimensional. It is therefore necessary for tools to not only support data exploration in 3D but also support complex 3D analysis.

The inclusion of this additional dimension further increases the difficulty in addressing the various challenges involved in designing visual analytics tools. For example, queries would now be over 3D space. While existing databases already struggle to handle queries over 2D space, handling these more complex queries for interactive visualization is almost impossible using current technologies. Similarly, visual encodings that work on 2D urban data (like the taxi data) can no

Figure 7. The sky exposure computes the fraction of the sky blocked by the buildings when viewed from street level (left). Note that this fraction is high in the Financial District, an area that has a dense cluster of tall buildings. The landmark view analysis (right) allows architects and city agencies to test the impact of proposed building designs (yellow tower) on the views to selected landmarks (highlighted in green). Buildings for which the views improve are highlighted in blue, and those negatively affected in red. The Empire State Building, Bank of America Tower, and the Chrysler Building are selected as landmarks for this analysis.

longer be used for data available with a 3D context. For example, noise levels in a neighborhood would vary based on the location as well as altitude, and can be represented as a function defined on a volume. Common 2D visual metaphors such as heat maps cannot be used for such data. While volume rendering is a common way to visualize such data, doing so would result in losing the context of the city. Thus, new visual metaphors are needed for 3D urban data.

The complexity arising from the additional dimension is carried over to the analysis component as well. It further introduces new problems and challenges that go along with it. For instance, city planners want to ensure that any new constructions adhere to the various city regulations. This requires an analysis pipeline that allows for existing buildings to be replaced by potential new constructions, and computes the effect/impact of these changes, allowing users to check their adherence to the regulations.

To tackle these challenges, we initiated the development of Urbane,[17] a 3-dimensional multi-resolution framework that enables a data-driven analysis of cities. In addition to supporting the traditional 2D-based visual analysis, Urbane also enables users to perform what-if analyses by changing the 3D state of a city, such as removing and/or replacing existing buildings with potential new constructions. The framework is extensible, making it possible to compute new metrics and impact measures, as well as help drive simulations performed over a city.

One such example is the city-scale analysis of shadows. Shadows can be both detrimental (e.g., inhibit vegetation growth, reduce solar energy potential, etc.), as well as beneficial (e.g., reduce urban heat island effect created by paved surfaces, etc.). It is therefore important to analyze the impact of shadows so that various stakeholders (e.g., city council, urban designers and developers) can make informed decisions on the verticalization of a city. To allow meaningful analysis of the impact of shadow, we must accumulate shadows over long time periods (e.g., shadows during summer). To support such analysis, we integrated the Urbane framework with a novel technique[18] for the fast accumulation of shadows that allows the stakeholders to analyze the shadow impact of different building designs on public spaces. Figure 6 shows one such example which compares the amount of shadows cast over the streets of NYC during summer and winter. Note that such an analysis at a city scale was not possible using existing tools. Moreover, such an infrastructure can also be used to inform public discourse (e.g., see www.nytimes.com/interactive/2016/12/21/upshot/Mapping-the-Shadows-of-New-York-City.html). The current version of Urbane also supports analyses involving other measures of interest such as how tall towers affect the amount of sky that is exposed, as well as the impact on landmark visibility due to proposed new towers (see Figure 7). Urbane has also been extended to inform new building designs based on view quality[19] (Figure 8).

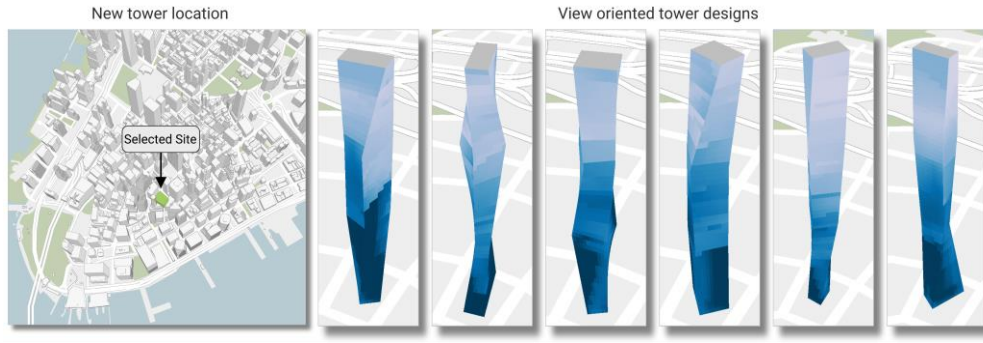

Figure 8. Possible building designs that optimize the view quality for the selected site. Here the view quality is defined based on the ability to view landmarks, water bodies, as well as unobstructed visibility. Regions with a darker shade of blue denote low view quality, while lighter shades denote higher view quality.

CONCLUSIONS

Designing visual analytics systems to handle large urban data requires interdisciplinary groups whose expertise straddle several computer science areas as well as in the application domain. In our experience, these collaborations have led not only to the successful translation of research results into deployed tools, but also to the identification of new fundamental research problems. This paper provides a brief overview of the work we have done in the intersection between visualization and data management. Many challenges still exist, notably around data discovery and data quality[20] where visual analytics can play a significant role.

ACKNOWLEDGEMENTS

This work was supported in part by: the Moore-Sloan Data Science Environment at NYU; NASA; DOE; NSF awards CNS-1229185, CCF-1533564, CNS-1544753, CNS-1730396, and OAC 1640864; CNPq; and FAPERJ. J. Freire and C. T. Silva are partially supported by the DARPA MEMEX and D3M programs. We thank Dr. L. G. Nonato (USP-SC) for Fig. 4.

REFERENCES

1. L. Barbosa, K. Pham, C. T. Silva, M. R. Vieira, and J. Freire, “Structured open urban data: understanding the landscape,” *Big Data*, 2(3):144–154, 2014.
2. J. W. Tukey, “Exploratory Data Analysis,” Pearson, 1977.
3. Z. Liu and J. Heer, “The effects of interactive latency on exploratory visual analysis,” *IEEE TVCG*, 20(12):2122–2131, 2014.
4. R. Chang, J. D. Fekete, J. Freire, and C. E. Scheidegger, “Connecting visualization and data management research,” Dagstuhl Reports, 7(11):46–58, 2017.
5. N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, “Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips,” *IEEE TVCG*, 19(12):2149–2158, 2013.
6. D. Peuquet, “It’s about time: A conceptual framework for the representation of temporal dynamics in geographic information systems,” *Ann. Assoc. Am. Geogr.*, 84(3):441–461, 1994.
7. N. Ferreira, J. T. Klosowski, C. E. Scheidegger, and C. T. Silva, “Vector field kmeans: Clustering trajectories by fitting multiple vector fields,” *CGF*, 32(3):201–210, 2013.
8. J. Poco, H. Doraiswamy, H. T. Vo, J. L. D. Comba, J. Freire, and C. T. Silva, “Exploring traffic dynamics in urban environments using vector-valued functions,” *CGF*, 34(3):161–170, 2015.

9. C. Palomo, Z. Guo, C. T. Silva, and J.Freire, “Visually exploring transportation schedules,” *IEEE TVCG*, 22(1):170–179, 2016.
10. H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. T. Silva, “Using topological analysis to support event-guided exploration in urban data,” *IEEE TVCG*, 20(12):2634–2643, 2014.
11. P. Valdivia, F. Dias, F. Petronetto, C. T. Silva, and L. G. Nonato, “Wavelet-based visualization of time-varying data on graphs,” *Proc. IEEE VAST*, pp. 1–8, 2015.
12. F. Miranda, H. Doraiswamy, M. Lage, K. Zhao, B. Goncalves, L. Wilson, M. Hsieh, and C. T. Silva, “Urban pulse: Capturing the rhythm of cities,” *IEEE TVCG*, 23(1):791–800, 2017.
13. F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire, “Data polygamy: The many-many relationships among urban spatio-temporal data sets,” *Proc. SIGMOD*, pp. 1011–1025. 2016.
14. H. Doraiswamy, H. T. Vo, C. T. Silva, and J. Freire, “A gpu-based index to support interactive spatio-temporal queries over historical data,” *Proc. IEEE ICDE*, pp. 1086–1097, 2016.
15. F. Miranda, M. Lage, H. Doraiswamy, C. Mydlarz, J. Salamon, Y. Lockerman, J. Freire, and C. T. Silva, “Time lattice: A data structure for the interactive visual analysis of large time series,” *CGF*, 37(3):13–22, 2018.
16. E. Tzirita Zacharatou, H. Doraiswamy, A. Ailamaki, C. T. Silva, and J. Freire, “GPU rasterization for real-time spatial aggregation over arbitrary polygons,” *PVLDB*, 11(3):352–365, 2017.
17. N. Ferreira, M. Lage, H. Doraiswamy, H. T. Vo, L. Wilson, H. Werner, M. Park, and C. T. Silva, “Urbane: A 3D framework to support data driven decision making in urban development,” *Proc. IEEE VAST*, pp. 97–104, 2015.
18. F. Miranda, H. Doraiswamy, M. Lage, L. Wilson, M. Hsieh, and C. T. Silva, “Shadow accrual maps: Efficient accumulation of city-scale shadows over time,” *IEEE TVCG*, 2018.
19. H. Doraiswamy, N. Ferreira, M. Lage, H. T. Vo, L. Wilson, H. Werner, M. Park, and C. T. Silva, “Topology-based catalogue exploration framework for identifying view-enhanced tower designs,” *ACM TOG*, 34(6):230, 2015.
20. J. Freire, A. Bessa, F. Chirigati, H. T. Vo, and K. Zhao, “Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips,” *IEEE Data Eng. Bull.*, 39(2): 63-77, 2016.

ABOUT THE AUTHORS

Harish Doraiswamy is a Research Scientist at the Center of Data Science and a Research Assistant Professor at the Dept. of Computer Science and Engineering at NYU. He received his Ph.D. in Computer Science and Engineering from the Indian Institute of Science, Bangalore. He is a member of IEEE. Contact him at harishd@nyu.edu.

Juliana Freire is a Professor of Computer Science and Engineering and Data Science at NYU. She is the lead PI and executive director of the NYU Moore-Sloan Data Science Environment, an ACM Fellow, the elected chair of the ACM SIGMOD, and a council member of the Computing Community Consortium (CCC). Contact her at juliana.freire@nyu.edu.

Marcos Lage is a Professor in the Dept. of Computer Science at UFF and is one of the principal investigators of the Prograf lab. He has a Ph.D. in applied mathematics from PUC-Rio. He is a member of IEEE. Contact him at mlage@ic.uff.br.

Fabio Miranda is a Ph.D. candidate in the Dept. of Computer Science and Engineering at NYU. He received a M.Sc. degree in computer science from PUC-Rio. He is a student member of IEEE. Contact him at fmiranda@nyu.edu.

Claudio Silva is a Professor of Computer Science and Engineering and Data Science with NYU. He is an IEEE Fellow and the elected chair of the IEEE Visualization & Graphics Technical Committee. He has received a number of awards including the IEEE Visualization Technical Achievement Award. Contact him at csilva@nyu.edu.

Contact editor Mike Potel at potel@wildcrest.com.