Preface

The last decade has seen a tremendous increase in the size of business and personal data collected: data of all forms but particularly data with spatial (geographic) and temporal dimensions. We can find numerous examples of this around us: smartphones equipped with GPS sensors or location-aware applications recording the mobility of their owners, mobile call transactions recorded by spatially-registered cell towers, surveillance cameras monitoring movements of people, cars and other objects, financial systems recording commercial transactions of people shopping, and so on, all with location and time signature. We are flooded with petabytes and exabytes of such data, which, of course, are more commonly known as 'Big Data'.

One of the primary reasons why big data is 'big', is that we now have the capabilities for observing the same phenomena from different vantage points (locations) at different moments of time. More and more -over 98% - computers are embedded in the physical world and increasingly larger proportion of all data now comes inscribed with space and time metadata (Acatec, 2012). In fact, if you take away location and time as the primary axes to represent this data and the data loses much of its semantics and utility. In effect, space and time provide the unifying axes for organizing, visualizing and interpreting multiple streams of heterogeneous big data (Singh et al., 2012). While multiple traditional computational approaches, such as data-mining techniques or machine learning algorithms, can still be applied for analyzing these datasets, there is growing interest and need to define specialized methods for analyzing, visualizing and making sense of ***spatio-temporal*** big data. This, indeed is the main motivation of our book.

In essence the book tries to connect data with its semantics - people, places and things – as without them data holds little value. In doing so, it also associates the big data trends with the emerging field of computational social science. As put forward by Gonzalez et al. (2008) and Lazer et al. (2009), computational social science aims at analyzing and understanding human behavior in general and human mobility in particular based on digital breadcrumbs they leave behind, which constitute what we know today as Big Data. Prior to this new approach, only statistical methods were being applied to data sets resulting from limited number of individuals being

subjected to experiments conducted by sociologists and psychologists. The new era of using Big Data analytics on extensive data sets potentially provides us with bigger opportunities for understanding human behavior and using this understanding for the common good of the mankind. Our book definitely falls along the same lines in that with most of the work presented herein, researchers have either worked with large spatio-temporal datasets that involve human mobility, or their techniques and results can be extended to work with such datasets.

While computational algorithms will continue to be important for interpreting 'big data', we will also see a growing need for human-in-the-loop analytics systems. In fact, one could argue that any mission critical big data decision system should always include a human in the loop. Considering this, one of the aims of this book is to provide a balanced treatment to both the computational algorithms as well as the visualization approaches. Such visualization approaches optimize a critically scarce resource – *human attention* – and will be of pivotal importance in the emerging geo-temporal big data systems. They will also allow for easy consumption of large scale data by a variety of stake holders and aid human decision ability (Andrienko et al., 2010; Keim et al., 2010).

The wide array of contributions in this book, coming from authors who have their academic homes in a wide variety of departments, underscore the breadth of the relevance of this books agenda. While the trends in Big Data continue to influence their traditional homes in the computational departments, they are increasingly relevant and influencing the research in a wide array of fields including, geography, industrial engineering, business informatics, information science, computational social sciences, urban planning, epidemiology, financial engineering and so on. The confluence of such ideas is also shaping multiple aspects of human life including health, wellness, productivity, transportation, sanitation, power, education, shopping, and sustenance (Pentland, 2012; Ratti et al., 2006; Šćepanović et al., 2015; Fotheringham & Rogerson, 2013).

Consequently, the book aims to provide a balance between the foundational concepts as well as application perspectives. As more and more of big data tools and techniques become widely accessible, they are increasingly relevant to experts in multiple disciplines who might be interested in big data analytics, but perhaps from their unique perspectives in urban planning, epidemiology, business analytics, and so on. This emerging confluence of interests and ideas underscores the need for the next generation of researchers as well as practitioners who are well-versed in computational as well as visualization techniques related to big data analytics. This is again how the book an important role in the emerging landscape.

This book highlights the recent developments on theoretical and empirical research in the area of spatio-temporal big data analysis and visualization, while giving specific attention to the emerging applications. The book is intended not only as a

reference book for academicians who are interested in some of the recent empirical research, but also as one that offers insights to practitioners and professionals who would like to take away lessons for future implementations of the techniques covered. We also think that many instructors would like to cover some recent methodologies as well as interesting case results from this book as classroom material in related Big Data analytics courses.

The book is organized primarily into three main parts. Chapters 1 through 4 present approaches for analyzing datasets that have spatio-temporal dimensions, whereas Chapters 5-8 focus more on visualization techniques and tools for such datasets. Chapters 9 and 10 provide modeling and algorithmic approaches for a better understanding of the corresponding data, and the last chapter addresses issues regarding the privacy and security of such datasets when they are stored in the Cloud. We find that it is fitting to end the book with a chapter pertaining to privacy as we expect multiple newer privacy challenges to emerge to the fore as spatio-temporal data about human behavior become widely available for research and analysis (de Montjoye et al., 2015).

In further detail, let us provide a brief overview of each chapter.

The book starts with the analysis of a dataset available from a telecom operator. In Chapter 1, Salman et al. consider call detail records (CDR) of subscribers who reside in the city of Istanbul, Turkey. The dataset includes not only a spatio-temporal stamp for each call record, but also additional demographic attributes associated with the subscriber making each call. The authors provide an analysis of this dataset from a disaster preparedness perspective, hoping to have a better understanding of how people move around the city, which areas they are highly concentrated at what time of the day, and how mobile the disabled subscribers are. The results of this analysis will hopefully assists logistics planners to plan ahead of a disaster for deploying critical resources well ahead of time before such a disaster hits.

A similar dataset is used in Chapter 2 by Altshuler et al., but this time for validating a proposed campaign optimization model. Considering how important it is to efficiently use financial and human resources in reaching out to existing and potential customers through marketing campaigns, the authors propose a mathematical model for calculating optimized campaigns that best decide on the configuration of customer-interacting units and their activities, as well as the geographical areas to operate in. Their results are validated against two datasets, one from a mobile phone operator, the other from a taxi operator company, both with spatio-temporal characteristics, by understanding the mobility patterns of individuals and using this information to decide on the optimal strategies for the marketing campaigns.

Chapter 3 considers yet another "Big Data" type of data set, one that results from the purchase transactions of the customers of a major financial institution. In this case, each transaction represents a purchase made at a specific merchant location

at a specific time, coupled with certain demographic information of the customer who made the purchase. The authors Srivastava et al. analyze this data set from a merchant financial-wellbeing point of view. Specifically, they investigate how the demographic profile of those customers as well as their spatio-temporal spending behavior correlate with the merchant well-being, which is measured in terms of total revenue generated and how consistently this revenue is generated. The authors argue that the findings may complement existing methods used by financial institutions in assessing the financial risks and creditworthiness of their business customers (i.e. merchants), and hence provide a more reliable way of doing so.

The next chapter presents a different vein of analysis, considering crowd-funding projects are geographically distributed. In Chapter 4, Bishop analyzes data on website image impressions, clicks and positions and shows how this data is significantly correlated to geographical factors. Bishop concludes that many interacting variables contribute to the observed phenomena and combining these with other related data sources for the Big Data analysis to be effective. He provides an extensive set of statistical tests to support his conclusions.

With the next set of four chapters, the contributing authors primarily focus on visualization approaches. In Chapter 5, Turk et al. propose an interactive tool for visualizing spatio-temporal data as well as detecting recurring patterns. They consider two sources of data: financial transactions of a bank's customers and location queries of a mobile operator's Friend Finder service (subscribers querying locations of their favorite list of friends, family members, etc.). Their proposed tool presents the analyst user with a heat map showing the density and distribution of location queries or purchase transaction locations, allowing her to filter the data –and hence pivot the visualization– by geography, time, day of week, etc. The user is then left with the task of manually detecting recurring patterns, if any, by examining a series of images and assessing how similar they are. The proposed tool assists this process by calculating and reporting what the authors refer to as structural similarity index. The authors propose that this system be used for not only detecting recurring patterns, but anomalies as well.

The next chapter, Chapter 6, takes the tool in the preceding chapter and the related analysis one step further and considers a three-dimensional visualization of data. The authors Kaya et al., however, look at it from a usability perspective, analyzing how suitable 2D- and 3D-visualizations are when it comes to visualizing spatio-temporal data. Through a series of experiments and interviews with domain experts, they identify factors that do and do not work for the visualization of such datasets to be effective. Their experiments reveal that, in terms of task completion time, none of the tested approaches is significantly better than the other, but the 2D density map visualization leads to more effective appreciation of the images presented in terms of trend detection and making comparisons.

Chapter 7 makes use of a very common form of spatial data visualization: Voronoi diagrams. While the classical Voronoi definition involves drawing catchment areas around point locations (e.g. stores, warehouses, landmark points), Devulapalli et al. consider the *inverse* Voronoi problem, where the goal is to determine center location "weights", given the point locations and desirable areas around them. This inverse problem is applicable to different information visualization contexts, as discussed by the authors in the chapter. Devulapalli et al. propose a method for determining the location weights, using convex optimization theory.

The last chapter that focuses on spatio-temporal data visualization is by Kwan et al. who consider population movement. In Chapter 8, the authors study data captured from mobile phone networks and propose a visualization approach using vectors that represent the direction in which large masses of people move. Their approach essentially involves clustering of spatio-temporal records in a highly efficient manner and associated each cluster with directional vector that represents that major movement activity in that cluster. The authors make note of the fact that other systems that produce massive datasets with spatio-temporal attribute signatures for individual population can similarly be used to visualize population movement.

The next two chapters, Chapter 9 and 10, propose modeling or algorithmic approaches that make use of spatio-temporal data. In Chapter 9, Santi and Ratti consider GPS location stamps collected from a fleet of taxis, and use this dataset to understand the impact of a shared taxi system on the urban environment. Specifically, they consider the modeling of taxis for shared rides, i.e. two or more passengers with different but similar trip origin and destination locations sharing the same taxi for a portion of the ride. A wide-scale use of such services clearly have a positive impact on traffic congestion in a densely populated urban area, by potentially reducing the traffic congestion and allowing individuals to have more pleasant journeys. The authors consider the dense street network of an urban area, cluster the street segments of this network into "cells", from which passenger origin and destination locations are tallied and matched for the purpose of sharing the ride.

Chapter 10 by Chauhan and Kaur proposes an efficient algorithm for retrieving spatial clusters from large location-aware databases. The authors apply this framework to medical databases of images with spatial elements, for retrieving clusters of various shapes and size, using an algorithm for discovering clusters and retrieving statistical information at different levels of data granularity. They suggest that through this efficient algorithm, an improved knowledge discovery process can be applied with more effective outcomes.

The last chapter of our book, Chapter 11, covers a different perspective on location-aware big datasets: the fact that these datasets can be stored on the cloud and the privacy and security issues that might be of concern. Gebremeskel et al. present a thorough review of major Cloud systems, their infrastructural issues in

terms of privacy and security concerns, and how to alleviate such concerns through certain measures that can be implemented. These issues we find extremely critical as spatio-temporal databases, in addition to traditional datasets with demographic attributes, reveal much more private information about their subjects and must hence be properly addressed. Gebremeskel et al. provide a good overview of hints and approaches in doing so.

We are extremely grateful to all contributors who participated in the creation of this book. They bring together a wide array of expertise, which is evident from their research topics as well as geographical and departmental affiliations. This is many ways marks the breadth and scope of this topic and at the same time underscores the unique scope of this book in integrating and assimilating the findings from such diverse traditions of research methods, agendas, and scopes.

All in all, this book is nothing but an invitation to start a conversation with, *you*, our dear reader. As this field grows and broadens in its approaches and application directions, we anticipate a growing interest and curiosity in issues pertaining to big data, visualization, and geo-temporal analytics. We hope the interested readers from academic as well as practitioner communities will appreciate this books offerings and reach out to us with their insights, comments, critiques, and suggestions.

Sincerely,

Burcin Bozkaya
Sabanci University, Turkey

Vivek Singh
Rutgers University, USA & Massachusetts Institute of Technology, USA

REFERENCES

Acatec – The German Academy of Science and Engineering. (2012). Cyber-Physical Systems: Driving Force for Innovations in Mobility, Health, Energy and Production. Author.

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., & Tominski, C. et al. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*, *24*(10), 1577–1600. doi:10.1080/13658816.2010.508043

de Montjoye, Y. A., Radaelli, L., Singh, V. K., & Pentland, A. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539. doi:10.1126/science.1256297 PMID:25635097

Fotheringham, S., & Rogerson, P. (Eds.). (2013). *Spatial analysis and GIS*. CRC Press.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. doi:10.1038/nature06958 PMID:18528393

Keim, D. A. Kohlhammer, Ellis, & Mansmann (Eds.). (2010). Mastering the information age-solving problems with visual analytics. Florian Mansmann.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., & Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science*, *323*(5915), 721. doi:10.1126/science.1167742 PMID:19197046

Pentland. (2012). Society's Nervous System: Building Effective Government, Energy, and Public Health Systems. *IEEE Computer Magazine*.

Ratti, C., Williams, S., Frenchman, D., & Pulselli, R. M. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning. B, Planning & Design*, *33*(5), 727–748. doi:10.1068/b32047

Šćepanović, S., Mishkovski, Hui, Nurminen, & Ylä-Jääski. (2015). Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator. *PLoS ONE*, 0124160. PMID:25897957

Singh, G., & Jain. (2012). Situation Recognition: An Evolving Problem for Heterogeneous Dynamic Big Multimedia Data. In *Proceedings ACM Multimedia*. ACM.