



The SIGSPATIAL Special

**Newsletter of the Association for Computing Machinery
Special Interest Group on Spatial Information**

Volume 11 Number 2 July 2019

The SIGSPATIAL Special

The SIGSPATIAL Special is the newsletter of the Association for Computing Machinery (ACM) Special Interest Group on Spatial Information (SIGSPATIAL).

ACM SIGSPATIAL addresses issues related to the acquisition, management, and processing of spatially-related information with a focus on algorithmic, geometric, and visual considerations. The scope includes, but is not limited to, geographic information systems.

Current Elected ACM SIGSPATIAL officers are:

- Chair, Cyrus Shahabi, University of Southern California
- Past Chair, Mohamed Mokbel, University of Minnesota
- Vice-Chair, Goce Trajcevski, Iowa State University
- Secretary, Egemen Tanin, University of Melbourne
- Treasurer, John Krumm, Microsoft Research

Current Appointed ACM SIGSPATIAL officers are:

- Newsletter Editor, Andreas Züfle, George Mason University
- Webmaster, Chrysovalantis (Chrys) Anastasiou, University of Southern California

For more details and membership information for ACM SIGSPATIAL as well as for accessing the newsletters please visit <http://www.sigspatial.org>.

The SIGSPATIAL Special serves the community by publishing short contributions such as SIGSPATIAL conferences' highlights, calls and announcements for conferences and journals that are of interest to the community, as well as short technical notes on current topics. The newsletter has three issues every year, i.e., March, July, and November. For more detailed information regarding the newsletter or suggestions please contact the editor via email at azufle@gmu.edu.

Notice to contributing authors to The SIGSPATIAL Special: By submitting your article for distribution in this publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights:

- to publish in print on condition of acceptance by the editor,
- to digitize and post your article in the electronic version of this publication,
- to include the article in the ACM Digital Library,
- to allow users to copy and distribute the article for noncommercial, educational or research purposes.

However, as a contributing author, you retain copyright to your article and ACM will make every effort to refer requests for commercial use directly to you.

Notice to the readers: Opinions expressed in articles and letters are those of the author(s) and do not necessarily express the opinions of the ACM, SIGSPATIAL or the newsletter.

The SIGSPATIAL Special (ISSN 1946-7729) Volume 11, Number 2, July 2019.

Table of Contents

	Page
Message from the Editor <i>Andreas Züfle</i>	1
<u>Section 1: Issue on SIGSPATIAL Visions and Challenges</u>	
Introduction to this Special Issue <i>Andreas Züfle</i>	2
GeoAI at ACM SIGSPATIAL: Progress, Challenges, and Future Directions <i>Yingjie Hu, Song Gao, Dalton Lunga, Wenwen Li, Shawn Newsam, Budhendra Bhaduri</i>	5
Batman or the Joker? The Powerful Urban Computing and its Ethics Issues <i>Kaiqun Fu, Abdulaziz Alhamadani, Taoran Ji, Chang-Tien Lu</i>	16
Quantum Spatial Computing <i>Martin Werner</i>	26
<u>Section 2: Tools and Datasets</u>	
UCR-STAR: The UCR Spatio-temporal Active Repository <i>Saheli Ghosh Tin Vu Mehrad Amin Eskandari Ahmed Eldawy</i>	34

Message from the Editor

Andreas Züfle

Department of Geography and GeoInformation Science, George Mason University, USA

Email: azufle@gmu.edu

Dear SIGSPATIAL Community,

The newsletter serves the community by publishing short contributions such as SIGSPATIAL conferences' highlights, calls and announcements for conferences and journals that are of interest to the community, as well as short technical notes on current topics.

The first section of this July 2019 issue features three technical reports highlighting potential new future research directions and challenges. These directions include challenges in GeoAI, ethical issues in urban computing, and the vision of quantum spatial computing. The second section contains a technical report of a system of broad interest to spatial data scientists in the SIGSPATIAL community. It describes UCR-STAR, a repository that not only currently stores more than a hundred spatial and spatio-temporal datasets but that allows to visually explore these datasets prior to downloading them.

You can download all Special issues from:

<http://www.sigspatial.org/publications/newsletter/>

I want to sincerely thank all authors for their generous contributions of time and effort that made this issue possible. I hope that you will find the newsletters interesting and informative and that you will enjoy this issue.

Yours sincerely,

Andreas Züfle

SIGSPATIAL Newsletter Editor



The SIGSPATIAL Special

Section 1:

SIGSPATIAL Visions and Challenges

ACM SIGSPATIAL
<http://www.sigspatial.org>

Visions and Challenges in GeoAI, Ethics, and Spatial Quantum Computing

Andreas Züfle

Department of Geography and GeoInformation Science, George Mason University

Email: azufle@gmu.edu

Geographic information systems are changing rapidly as new technologies, such as advances in artificial intelligence and quantum computing, proliferate. Towards gaining a glimpse of possible new directions for the next decades, it is paramount to disseminate and discuss new research directions, challenges, and visions, that may be of broad interest to the SIGSPATIAL community. In the last few years, our community has excelled at providing a forum for such visions and challenges in dedicated *Vision and Challenge* tracks, leading to many visionary research directions [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. A goal of this newsletter is to continue and complement these visions. Three vision and challenge topics are covered in this special issue:

1. the first article is contributed by the organizers of GeoAI, the most successful ACM SIGSPATIAL Workshop of the last years, having 40+ registrations each year, and having 50+ people in the workshop room at a time. The authors survey and summarize the directions of GeoAI publications of the last three years, and provide a list of open research directions for this rapidly advancing field;
2. in the second article, Fu et al. discuss ethical issues of mining urban spatial data. They identify ethical vulnerabilities from three primary research directions of urban computing: urban safety analysis, urban transportation analysis, and social media analysis for urban events;
3. in the third article, Martin Werner introduces central aspects of quantum algorithms and explores future directions on how quantum algorithms can be used to leverage spatial and spatio-temporal algorithms. This article describes example algorithms, such as map coloring and dynamic time warping, which may benefit from advances in quantum computing.

In addition to these vision and challenge papers, this issue also proudly presents UCR-STAR, a spatio-temporal data repository that allows to browse and visualize a plethora of publicly available spatio-temporal data sets, ranging from point and polygon data on OpenStreetMap [22] to commonly used trajectory data sets such as TDrive [23] and GeoLife [24]. Such a repository is particularly useful for PhD students searching for data sets. It allows them to simultaneously explore and visualize many public data sets, and to pick the best for their research needs.

I hope the readers will enjoy this issue and find it useful in their research work. I'd also like to call upon readers to send me suggestions for news that they would like to appear in the next issues of this newsletter. If you have exciting news, visions, and challenges that you would benefit the SIGSPATIAL community, and that you would like to disseminate, please reach out to me! Finally, I want to cordially thank the authors for their excellent contributions to this issue.

References

- [1] Y. Xie, S. Shekhar, R. Feiock, and J. Knight, “Revolutionizing tree management via intelligent spatial techniques,” in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 71–74, ACM, 2019.
- [2] H. S. Al-Olimat, V. L. Shalin, K. Thirunarayan, and J. P. Sain, “Towards geocoding spatial expressions (vision paper),” in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 75–78, ACM, 2019.
- [3] R. A. Alghamdi, A. Magdy, and M. F. Mokbel, “Towards a unified framework for event detection applications,” in *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, pp. 210–213, ACM, 2019.
- [4] J. Xu, H. Lu, and R. H. Güting, “Understanding human mobility: A multi-modal and intelligent moving objects database,” in *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, pp. 222–225, ACM, 2019.
- [5] H. Kavak, J.-S. Kim, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle, “Location-based social simulation,” in *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, pp. 218–221, ACM, 2019.
- [6] G. Giannopoulos and M. Meimaris, “Learning domain driven and semantically enriched embeddings for poi classification,” in *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, pp. 214–217, ACM, 2019.
- [7] T. Dasu, Y. Kanza, and D. Srivastava, “Geofences in the sky: herding drones with blockchains and 5g,” in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 73–76, ACM, 2018.
- [8] A. Degbelo and C. Kray, “Intelligent geovisualizations for open government data (vision paper),” in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 77–80, ACM, 2018.
- [9] S. Schmoll and M. Schubert, “Vision paper: reinforcement learning in smart spatio-temporal environments,” in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 81–84, ACM, 2018.
- [10] O. Wolfson, “Understanding the human brain via its spatio-temporal properties (vision paper),” in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 85–88, ACM, 2018.
- [11] H. Aly, M. Youssef, and A. Agrawala, “Towards ubiquitous accessibility digital maps for smart cities,” in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 8, ACM, 2017.
- [12] T. Dasu, Y. Kanza, and D. Srivastava, “Geotagging ip packets for location-aware software-defined networking in the presence of virtual network functions,” in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 9, ACM, 2017.
- [13] K. Ramamohanarao, J. Qi, E. Tanin, and S. Motallebi, “From how to where: Traffic optimization in the era of automated vehicles,” in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 10, ACM, 2017.

- [14] T. C. van Dijk and A. Wolff, “Algorithmically-guided user interaction,” in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 11, ACM, 2017.
- [15] V. Zakhary, C. Sahin, T. Georgiou, and A. El Abbadi, “Locborg: Hiding social media user location while maintaining online persona,” in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 12, ACM, 2017.
- [16] K. Janowicz and G. McKenzie, “How “alternative” are alternative facts? measuring statement coherence via spatial analysis,” in *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2017.
- [17] O. Gkountouna, D. Pfoser, C. Wenk, and A. Züfle, “A unified framework to predict movement,” in *International Symposium on Spatial and Temporal Databases*, pp. 393–397, Springer, 2017.
- [18] C. Jonathan and M. F. Mokbel, “Towards a unified spatial crowdsourcing platform,” in *International Symposium on Spatial and Temporal Databases*, pp. 379–383, Springer, 2017.
- [19] M. Sarwat and A. Nandi, “On designing a geoviz-aware database system-challenges and opportunities,” in *International Symposium on Spatial and Temporal Databases*, pp. 384–387, Springer, 2017.
- [20] K. A. Schmid, A. Züfle, D. Pfoser, A. Crooks, A. Croitoru, and A. Stefanidis, “Predicting the evolution of narratives in social media,” in *International Symposium on Spatial and Temporal Databases*, pp. 388–392, Springer, 2017.
- [21] Z. Li, “Semantic understanding of spatial trajectories,” in *International Symposium on Spatial and Temporal Databases*, pp. 398–401, Springer, 2017.
- [22] M. Haklay and P. Weber, “Openstreetmap: User-generated street maps,” *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.
- [23] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, “T-drive: driving directions based on taxi trajectories,” in *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, pp. 99–108, ACM, 2010.
- [24] Y. Zheng, X. Xie, W.-Y. Ma, *et al.*, “Geolife: A collaborative social networking service among user, location and trajectory,” *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

GeoAI at ACM SIGSPATIAL: Progress, Challenges, and Future Directions

Yingjie Hu¹, Song Gao², Dalton Lunga³, Wenwen Li⁴, Shawn Newsam⁵, Budhendra Bhaduri³

¹University at Buffalo, USA

²University of Wisconsin, Madison, USA

³Oak Ridge National Laboratory, USA

⁴Arizona State University, USA

⁵University of California, Merced, USA

Abstract

Geospatial artificial intelligence (GeoAI) is an interdisciplinary field that has received tremendous attention from both academia and industry in recent years. This article reviews the series of GeoAI workshops held at the Association for Computing Machinery (ACM) International Conference on Advances in Geographic Information Systems (SIGSPATIAL) since 2017. These workshops have provided researchers a forum to present GeoAI advances covering a wide range of topics, such as geospatial image processing, transportation modeling, public health, and digital humanities. We provide a summary of these topics and the research articles presented at the 2017, 2018, and 2019 GeoAI workshops. We conclude with a list of open research directions for this rapidly advancing field.

1 Introduction

While the field of artificial intelligence (AI) was born in the 1950s, it has recently experienced a tremendous resurgence and is currently drawing significant attention from academia, industry, news media, and the general public. Through machine learning and in particular deep learning, AI has provided novel solutions to a variety of challenging problems ranging from computer vision to natural language processing, achieving near human-level performance. The impact of deep learning has reached many application domains, and geography is no exception. Remote sensing and geospatial image processing [57] is one area in geography and geographic information science (GIScience) that has been quick to embrace artificial intelligence techniques. For example, deep learning techniques have been adopted and further improved for tasks such as hyperspectral image analysis [2] and high-resolution satellite image interpretation [56]. Beyond remote sensing, researchers have also utilized deep learning techniques to extract information from other sources of geospatial images, such as Google Street View [35] and scanned historical maps [8]. In natural language processing (NLP), deep learning models, such as deep recurrent neural networks, have been employed to improve the accuracy of place name extraction from textual data [23, 43]. Other neural network based NLP techniques, such as word embeddings, have been employed to help quantify changes in stereotypes and attitudes toward women and ethnic minorities over a 100 year study period in the United States [10]. And, there are many other examples of research that integrates geography and AI, such as extracting building footprints using convolutional neural networks (CNNs) [50, 54], deep semantic segmentation for automated driving [32], vehicle trajectory prediction [48], indoor navigation [17], gazetteer conflation [31], and spatial epidemics [44]. The integration of geography and AI has given rise to the new and exciting interdisciplinary field of geospatial artificial intelligence (GeoAI).

Three factors in particular are contributing to the emergence of GeoAI as a field. First, the growing availability of large amounts of geospatial data continue to enable the training of increasingly complex AI models. Large datasets have long existed in the field of geography, with examples including remote sensing images, national-level census data, road networks, and land use and land cover (LULC) data. However, a variety of newer geospatial datasets, such as location-based social media and high resolution GPS trajectory data, have become available in the past two decades. In addition, companies such as Yelp and Uber have started to share their data (such as Yelp points of interest (POI) and Uber vehicle trajectories). These large and diverse geospatial datasets capture diverse aspects of natural and social environments, and enable the training of AI models to address a rich variety of problems.

The second factor contributing to GeoAI is that novel AI models and other computational methods are starting to be developed specifically for geographic problems. For example, Li et al. [18] extend a Faster-RCNN (Faster Region-based CNN) model to support the automatic detection of natural terrain features from remote sensing imagery. The model addresses a number of unique geospatial challenges, such as the ambiguous boundary of natural features in comparison to man-made features such as buildings and roads. Marcos et al. [27] propose a neural network model called RotEqNet for land cover mapping. This model encodes rotation equivariance in a CNN, and can address the challenge of recognizing rotated versions of the same object in remote sensing images. It is worth noting that geographic location and geographic information systems (GIS) play critical roles in developing GeoAI models since these models often integrate heterogeneous data from different sources. Geographic location and GIS are essential to establishing links between the data (e.g., linking streets to nearby residential areas and green spaces).

The third factor is that the increasing availability of high-performance computing hardware makes it possible to efficiently train GeoAI models with big geo data. GIS, as a special type of computer and information system, has already been integrated with supercomputing infrastructure (e.g., CyberGIS) [42]. Training GeoAI models on supercomputing infrastructure can significantly reduce the often lengthy training times, and enable broader hyperparameter and architecture search to identify the optimal models. Major companies such as ESRI and Microsoft have begun offering new computing resources (e.g., GeoAI Data Science Virtual Machine) with the goal of bringing AI, geospatial analysis, and high-performance computing together. Accelerated inference for deployment with AI models is now achievable via production ready system including the use of the Apache Spark computing framework [22].

It is in this context that a series of GeoAI workshops have been organized at ACM SIGSPATIAL, the premier conference at the intersection of geospatial data analysis and computer science. GeoAI 2017 and GeoAI 2018 took place in Los Angeles and Seattle respectively [26, 11], and, at time of writing, GeoAI 2019 has confirmed the list of accepted papers and will take place in Chicago. These GeoAI workshops bring together GIScientists, computer scientists, engineers, entrepreneurs, and decision makers from academia, industry, and government to discuss the latest trends, successes, challenges, and opportunities in this interdisciplinary field of GeoAI. The 2017 and 2018 GeoAI workshops were some of the more popular workshops at the ACM SIGSPATIAL conference based on conference reported statistics, and we expect the 2019 workshop to be similarly well attended. The papers accepted and presented at the workshops have covered a wide range of GeoAI research topics. In this article, we systematically review and summarize these papers. We then propose future GeoAI research directions.

2 Research Topics at the GeoAI Workshops

In this section, we review the research topics and problems covered at the three GeoAI workshops. We first review the set of papers presented at the 2017, 2018, and 2019 workshops. We then provide a summary of the topics covered.

2.1 GeoAI'17

ACM SIGSPATIAL 2017 marked the first gathering of the GeoAI workshop series and took place in Los Angeles. This workshop received a total of 14 submissions, and after a rigorous review process, 8 papers were accepted and presented at the workshop. This workshop also featured two keynotes: one was an academic keynote by Dr. Shawn Newsam (University of California, Merced) on “Geographic knowledge discovery using ground-level images and videos”, and the other was an industry keynote from Dr. Saikat Basu (Facebook) on “Using AI to help generate roads for OpenStreetMap”. Among the accepted 8 papers, there were two major research topics: geospatial image processing and transportation modeling. On the topic of geospatial image processing, Li et al. [19] used the Faster-RCNN model to support the automatic detection of terrain features, such as craters, from remote sensing images. Law et al. [16] applied a CNN to classify street frontages using both Google Street View images and 3D models produced by ESRI's City Engine. Collins et al. [5] developed a deep CNN model to enhance low resolution multispectral satellite imagery where no corresponding high resolution images are available. Duan et al. [8] proposed an algorithm to automatically align vector data to scanned historical maps to generate large datasets for training a CNN to recognize geographic features in such maps.

On the topic of transportation modeling and analysis, Kulkarni and Garbinato [14] explored the effectiveness of RNNs for generating synthetic traffic trajectory datasets. Murphy et al. [28] proposed an image-based classification method that uses CNNs to classify the noise level in GPS trajectories with the goal of improving the accuracy of distance estimation for ride-sharing applications. Within the general scope of transportation and navigation, Li et al. [17] developed a method that combines deep learning based object recognition and second-order hidden Markov models to detect indoor landmarks. In addition to the papers which focused on the two main themes, Majic et al. [25] developed an unsupervised approach to identify equivalent OpenStreetMap keys based on their co-occurring patterns and the geometries of the features annotated by the keys.

A special issue on “GeoAI: Artificial Intelligence Techniques for Geographic Knowledge Discovery” was organized in the *International Journal of Geographical Information Science* [12], which complemented the GeoAI'17 Workshop. This special issue attracted full paper submissions from not only the GeoAI 2017 workshop but also interested scholars from the community.

2.2 GeoAI'18

Due to the popularity and success of the first GeoAI workshop, the second workshop, GeoAI 2018, was held at ACM SIGSPATIAL in Seattle. This workshop received 19 submissions, and 10 papers were accepted after review. Dr. Rangan Sukumar (Cray Inc.) gave an industry keynote on “The AI Journey in Geospatial Discovery: Navigating Shapes, Sizes and Spaces of Data”, and Dr. Bruno Martins (University of Lisbon) gave an academic keynote on “GeoAI Applications in the Spatial Humanities”. Many of the accepted papers continued the two major research topics from GeoAI 2017, namely geospatial image processing and transportation modeling. On the topic of geospatial image processing, Xu et al. [47] reviewed the use of computer vision and CNN models for locating aerial images collected by unmanned aerial vehicles (UAVs). Sun et al. [38] used a CNN model, a modified U-Net, to combine satellite imagery and GPS data for road extraction. And, Srivastava et al. [36] trained a CNN model for classifying the multiple functions of buildings from Google Street View images in the city of Amsterdam.

On the topic of transportation modeling and analysis, Van Hinsbergh et al. [41] combined GPS trajectory with in-car signal data, and used a location extraction technique, Gradient-based Visit Extractor, to extract interesting locations of a trajectory, such as drop-off, parked, and pick-up. Pourebrahim et al. [30] examined the potential of adding Twitter data to neural network and gravity models to enhance commuter trip prediction.

Two papers at GeoAI 2018 specifically focused on methodology. Swan et al. [39] explored the negative impact of noise in training data on trained deep neural network models, and found that models trained with small amounts of noise had little change in precision but considerable increases in recall; however, as noise

levels continued increasing, both precision and recall decreased. Aydin et al. [1] proposed an unsupervised and consensus-based regionalization algorithm, SKATER-CON, to create spatially contiguous regions, which addresses the chaining problem of computing minimum spanning trees by using random spanning trees and outperformed two state-of-the-art regionalization methods, SKATER and ARISEL.

A number of other research topics were also covered at GeoAI 2018. Xi et al. [44] proposed a deep residual network for influenza prediction by integrating the spatial-temporal properties of influenza at an intra-urban scale, and applied their model to a dataset in Shenzhen, China. Chow [4] discussed the possibility and challenges of integrating AI and agent-based models (ABMs) to simulate and estimate the head count of a moving crowd. Elgarroussi et al. [9] developed Aconcagua which is a spatio-temporal framework that can perform emotion analysis on tweets and monitor the change of positive and negative emotions over time and space.

2.3 GeoAI'19

The third GeoAI workshop will be held at ACM SIGSPATIAL in Chicago in November 2019. This workshop received 25 submissions, and 17 papers were accepted after review. Dr. Xin Chen (HERE Technologies) will give an industry keynote on “HD Live Map for Automated Driving: An AI Approach”, and Dr. Raju Vatsavai (North Carolina State University) will give an academic keynote on “Geospatial AI for Monitoring Crops to Nuclear Proliferation Using Global Earth Observations”. Many papers continue the two major research topics from GeoAI 2017 and GeoAI 2018, geospatial image processing and transportation modeling. On the topic of geospatial image processing, Chen et al. [3] propose a ChangeNet to identify relevant pixelwise changes in time-varying images taken at the same location and utilize conditional generative adversarial networks (GANs) to improve classification results. Dorji et al. [7] present a machine learning approach to estimate the median income levels of sub-districts in Thailand using nighttime satellite images, population density, road density, and distance to major metropolitan areas. Peng et al. [29] propose a residual patch similarity CNN (ResPSNet) to map urban flood hazard zones using bi-temporal high resolution (3 meters) pre- and post-flooding multispectral surface reflectance satellite imagery. Law and Neira [15] propose a convolutional principle component analysis (ConvPCA), which is applied to both street level and street network images to find a set of uncorrelated and ordered visual latent components to predict urban characteristics such as street quality classification and street enclosure regression tasks. Liang and Newsam [21] explore deep learning regression approaches to estimate the spatial resolution of very high-resolution overhead imagery and show that a stacked auto-encoder (SAE) frontend outperforms a standard CNN feature extractor. Xin and Adler [45] use a convolutional long short-term memory (Convolutional-LSTM) neural network model for grass identification in multi-temporal Sentinel-2 images. Tavakkol et al. [40] introduce the Kartta Labs open-data project aimed at unrendering and organizing the world’s historical maps, and support user queries for the corresponding vector-based geospatial content as well as recreating the historical maps in various cartographic styles.

On the topic of transportation modeling, Yin et al. [53] propose a web-based analytic platform that leverages deep learning techniques in computer vision for vehicle plate number identification and illegal parking detection. Xing et al. [46] from Didi utilize street images captured by driving vehicle recorders (DVR) and develop a traffic sign discovery driven system for various types of traffic rule automatic update such as no left/right/U turn, no parking, speed limit, etc. Several works use GPS trajectories to gain insights. J. Krumm and K. Krumm [13] show how to use mobility traces from 1.8 million users to infer the number of businesses and residences in each 250m * 250m grid cell of the Seattle metropolitan area. Yin et al. [52] introduce a multi-scale graph CNN for road intersection detection from ride-hailing service GPS trajectories in Singapore. Mai et al. [24] learn the probability of pick-up and drop-off locations and times from taxi GPS trajectories and develop a spatial-temporal intelligent recommendation system to improve the expected net revenue of the electric-taxi.

Beyond geospatial image processing and transportation modeling, Soliman and Terstriep [34] develop Keras Spatial, a Python package for preprocessing and augmenting geospatial data. Yuan and Crooks [55] use a CNN to extract user opinions in natural language from more than 3 million user-contributed Yelp restaurant reviews.

They discover homogeneity among cities regarding the average proportions of aspects (i.e., terms and categories of semantics) in restaurant reviews. Snyder et al. [33] combine a deep neural network (Deepgeo) with a social media analytics and reporting toolkit (SMART) to infer the city-level geolocation of tweets and create real-time visual analytics. Finally, there are two vision papers focusing on the spatiotemporal intelligence of human behaviors and decision making. Li and Huang [20] outline a novel spatial-temporal imitation learning (STIL) framework that defines, investigates, and addresses the emerging research challenges of analyzing and learning human decision-making strategies from human-generated spatial-temporal data. And, Yang and Jankowska [51] share thoughts on how health behaviors can be contextualized in space and time through the use of GeoAI just-in-time adaptive interventions models, and highlight challenges such as on-the-fly feature engineering and spatiotemporal contextual uncertainty that need to be addressed in future work.

2.4 Summary

Table 1 groups the papers from the three GeoAI workshops based on their research topics. As can be seen, the majority of papers focus on geospatial image processing and transportation modeling. The popularity of these two topics can be attributed to the many advancements made by deep learning in image processing and the widespread interest of the ACM SIGSPATIAL community in transportation modeling. Other topics addressed include digital humanities, cartography, public health, disaster response, and social media analysis. Several papers specifically focus on methodological research, while almost all papers adapt and improve existing AI methods to target geographic problems. Interesting applications and visions, such as integrating AI and agent-based models (ABMs) for crowd analysis, have also been discussed.

3 Future Research Directions for GeoAI

Many global challenges are yet to benefit from GeoAI. With the ever growing collections of geospatial data we anticipate new GeoAI technological advances to emerge. However, near-human performance with AI seems to hinge on the availability ground-truth data for training. More opportunities could be opened up by taking advantage of the vast amount of unlabeled data for training unsupervised AI based methods. Furthermore, domain problems need to be fully understood and problem owners and AI experts from academia and industry need to be engaged during the design of the GeoAI solutions. Below, we list and organize possible future research directions for GeoAI into four categories.

Research directions and questions motivated by the uniqueness of geospatial problems and phenomena:

- Building spatiotemporally explicit models. In order to better understand the complex geospatial contexts and geographical process on the ground, it is crucial to employ spatiotemporally explicit models and evaluate the results by integrating both human intelligence and machine intelligence evaluations [49].
- Enhancing model generalizability in the context of geography. Geographical data sets are always collected from certain regions. How can we ensure that the GeoAI models trained using data from one geographic area can generalize to other geographic areas?
- Accommodating uncertainty in geospatial problems and datasets. Deep learning methods have traditionally not been designed with data uncertainty in mind yet uncertainty is a fundamental concept in geography. Can geography actually contribute to deep learning more broadly by developing methods to imbue the models with uncertainty analysis?

Research directions and questions motivated by the wealth of geospatial data:

- Developing new and open geospatial data infrastructures. ImageNet [6] played a key role in revolutionizing the field of computer vision. Future GeoAI applications could benefit from similar platforms by

Table 1: Papers presented in the three GeoAI workshops.

Research Topic	Related Papers in GeoAI Workshops
Geospatial image processing	<ul style="list-style-type: none"> - Li, W. et al. (2017) [19] - Law, S. et al. (2017) [16] - Collins, C.B. et al. (2017) [5] - Duan, W. et al. (2017) [8] - Xu, Y. et al. (2018) [47] - Sun, T. et al. (2018) [38] - Srivastava, S. et al. (2018) [35] - Chen et al. (2019) [3] - Dorji et al. (2019) [7] - Law and Neira (2019) [15] - Liang and Newsam (2019) [21] - Xin and Adler (2019) [45]
Transportation modeling and analysis	<ul style="list-style-type: none"> - Kulkarni, V. and Garbinato, B., (2017) [14] - Murphy, J. et al. (2017) [28] - Li, Q. et al. (2017) [17] - Sun, T. et al. (2018) [38] - Van Hinsbergh, J. et al (2018) [41] - Pourebrahim, N. et al (2018) [30] - Yin et al. (2019) [53] - J. Krumm and K. Krumm (2019) [13] - Xing et al. (2019) [46] - Yin et al. (2019) [52] - Mai et al. (2019) [24]
Digital humanities	<ul style="list-style-type: none"> - Duan, W. et al. (2017) [8] - Tavakkol et al. (2019) [40]
Public health	<ul style="list-style-type: none"> - Xi, G. et al. (2018) [44] - Yang and Jankowska (2019) [51]
Disaster response	<ul style="list-style-type: none"> - Peng et al. (2019) [29]
Social media and geo-text analysis	<ul style="list-style-type: none"> - Pourebrahim, N. et al (2018) [30] - Elgarroussi, K., et al. (2018) [9] - Yuan and Crooks (2019) [55] - Snyder et al. (2019) [33]
Methods and techniques	<ul style="list-style-type: none"> - Swan, B. et al. (2018) [39] - Aydin, O. et al. (2018) [1] - Soliman and Terstriep (2019) [34]
Novel applications and visions	<ul style="list-style-type: none"> - Majic, I. et al. (2017) [25] - Chow, T. E. (2018) [4] - Li and Huang (2019) [20]

investigating open and indexable rich geospatial data archives. A great example is the BigEarthNet platform [37] which has been demonstrated to be significantly larger than the existing archives in remote sensing and has been used as a diverse training source in the context of deep learning.

- Fusing multi-source geospatial data for knowledge discovery. The fusion of diverse geospatial datasets at different spatiotemporal resolutions through feature engineering and deep learning can enable novel geographic knowledge. How can machine learning help automate, streamline, or assist geospatial data integration?

Research directions and questions motivated by the need of labeled training data:

- Building domain datasets via novel techniques. As most GeoAI models involve supervised learning, the availability of large, benchmark datasets becomes essential to promote research across geospatial communities and to validate the generalizability of the research results. Can we leverage certain novel techniques to facilitate the development of domain datasets? For example, is there a role for generative adversarial networks (GANs) to play in augmenting GeoAI training data?
- Reducing the need of labeled data through self-supervised learning. Self-supervised learning methods have demonstrated their capability in other domains with their potentially unlimited capacity to uncover patterns in unlabeled data. Could they offer better scalability and reduce the need for labeled training data in the geography domain?

Research directions and questions motivated by the need of reliable and explainable models:

- Building robust and reliable GeoAI models. How can we ensure that future GeoAI systems are robust and reliable, and how do we evaluate them at scale?
- Building explainable GeoAI models. Most AI learning systems remain a black box. While these systems have demonstrated good performance in object inspection and classification, it is important to understand their learning and decision making process when applied to a variety of geospatial problems in both the physical and social sciences domains.

The research directions and questions discussed above are not exhaustive, and there exist many other important directions to be explored. We look forward to seeing exciting new research to be shared and published in the ACM SIGSPATIAL GeoAI workshops and other venues in the coming years.

References

- [1] O. Aydin, M. V. Janikas, R. Assunção, and T.-H. Lee. SKATER-CON: Unsupervised regionalization via stochastic tree partitioning within a consensus framework using random spanning trees. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 33–42. ACM, 2018.
- [2] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2094–2107, 2014.
- [3] Y. Chen, X. Ouyang, and G. Agam. ChangeNet: Learning to detect changes in satellite images. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 24–31. ACM, 2019.

- [4] T. E. Chow. When GeoAI meets the crowd. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI'18, pages 52–53, New York, NY, USA, 2018. ACM.
- [5] C. B. Collins, J. M. Beck, S. M. Bridges, J. A. Rushing, and S. J. Graves. Deep learning for multisensor image resolution enhancement. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 37–44. ACM, 2017.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [7] U. J. Dorji, A. Plangprasopchok, N. Surasvadi, and C. Siripanpornchana. A machine learning approach to estimate median income levels of sub-districts in Thailand using satellite and geospatial data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 11–14. ACM, 2019.
- [8] W. Duan, Y.-Y. Chiang, C. A. Knoblock, V. Jain, D. Feldman, J. H. Uhl, and S. Leyk. Automatic alignment of geographic features in contemporary vector data and historical maps. In *Proceedings of the 1st workshop on artificial intelligence and deep learning for geographic knowledge discovery*, pages 45–54. ACM, 2017.
- [9] K. Elgarroussi, S. Wang, R. Banerjee, and C. F. Eick. Aconcagua: A novel spatiotemporal emotion change analysis framework. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 54–61. ACM, 2018.
- [10] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.
- [11] Y. Hu, S. Gao, S. Newsam, and D. Lunga. GeoAI 2018 workshop report the 2nd acm sigspatial international workshop on GeoAI: AI for geographic knowledge discovery seattle, wa, usa-november 6, 2018. *SIGSPATIAL special*, 10(3):16–16, 2019.
- [12] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri. GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, pages 1–13, 2020.
- [13] J. Krumm and K. Krumm. Land use inference from mobility traces. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 1–4. ACM, 2019.
- [14] V. Kulkarni and B. Garbinato. Generating synthetic mobility traffic using rnns. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 1–4. ACM, 2017.
- [15] S. Law and D. M. N. Alvarez. An unsupervised approach to geographical knowledge discovery using street level and street network images. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 56–65. ACM, 2019.
- [16] S. Law, Y. Shen, and C. Seresinhe. An application of convolutional neural network in street image classification: The case study of london. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 5–9. ACM, 2017.

- [17] Q. Li, J. Zhu, T. Liu, J. Garibaldi, Q. Li, and G. Qiu. Visual landmark sequence-based indoor localization. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 14–23. ACM, 2017.
- [18] W. Li and C.-Y. Hsu. Automated terrain feature identification from remote sensing imagery: a deep learning approach. *International Journal of Geographical Information Science*, pages 1–24, 2018.
- [19] W. Li, B. Zhou, C.-Y. Hsu, Y. Li, and F. Ren. Recognizing terrain features on terrestrial surface using a deep learning model: An example with crater detection. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 33–36. ACM, 2017.
- [20] Y. Li and W. Huang. Imitation learning from human-generated spatial-temporal data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 9–10. ACM, 2019.
- [21] H. Liang and S. Newsam. Estimating the spatial resolution of very high-resolution overhead imagery. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 77–80. ACM, 2019.
- [22] D. Lunga, J. Gerrand, H. L. Yang, C. Layton, and R. Stewart. Apache spark accelerated deep learning inference for large scale satellite image analytics, 2019.
- [23] A. Magge, D. Weissenbacher, A. Sarker, M. Scotch, and G. Gonzalez-Hernandez. Deep neural networks and distant supervision for geographic location mention extraction. *Bioinformatics*, 34(13):i565–i573, 2018.
- [24] K. Mai, W. Tu, Q. Li, H. Ye, T. Zhao, and Y. Zhang. STIETR: Spatial-temporal intelligent e-taxi recommendation system using gps trajectories. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 5–8. ACM, 2019.
- [25] I. Majic, S. Winter, and M. Tomko. Finding equivalent keys in openstreetmap: semantic similarity computation based on extensional definitions. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 24–32. ACM, 2017.
- [26] H. Mao, Y. Hu, B. Kar, S. Gao, and G. McKenzie. Geoai 2017 workshop report: the 1st acm sigspatial international workshop on geoai:@ ai and deep learning for geographic knowledge discovery: Redondo beach, ca, usa-november 7, 2016. *SIGSPATIAL Special*, 9(3):25–25, 2018.
- [27] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. *ISPRS journal of photogrammetry and remote sensing*, 145:96–107, 2018.
- [28] J. Murphy, Y. Pao, and A. Haque. Image-based classification of gps noise level using convolutional neural networks for accurate distance estimation. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 10–13. ACM, 2017.
- [29] B. Peng, X. Liu, Z. Meng, and Q. Huang. Urban flood mapping with residual patch similarity learning. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 40–47. ACM, 2019.
- [30] N. Pourebrahim, S. Sultana, J.-C. Thill, and S. Mohanty. Enhancing trip distribution prediction with twitter data: comparison of neural network and gravity models. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 5–8. ACM, 2018.

- [31] R. Santos, P. Murrieta-Flores, P. Calado, and B. Martins. Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, 32(2):324–348, 2018.
- [32] M. Siam, S. Elkerdawy, M. Jagersand, and S. Yogamani. Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017.
- [33] L. Snyder, M. Karimzadeh, R. Chen, and D. Ebert. City-level geolocation of tweets for real-time visual analytics. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 85–88. ACM, 2019.
- [34] A. Soliman and J. Terstriep. Keras Spatial: Extending deep learning frameworks for preprocessing and on-the-fly augmentation of geospatial data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 69–76. ACM, 2019.
- [35] S. Srivastava, J. E. Vargas Muñoz, S. Lobry, and D. Tuia. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science*, pages 1–20, 2018.
- [36] S. Srivastava, J. E. Vargas-Muñoz, D. Swinkels, and D. Tuia. Multilabel building functions classification from ground pictures using convolutional neural networks. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on AI for geographic knowledge discovery*, pages 43–46. ACM, 2018.
- [37] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *CoRR*, abs/1902.06148, 2019.
- [38] T. Sun, Z. Di, and Y. Wang. Combining satellite imagery and gps data for road extraction. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 29–32. ACM, 2018.
- [39] B. Swan, M. Laverdiere, and H. L. Yang. How good is good enough?: Quantifying the effects of training set quality. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 47–51. ACM, 2018.
- [40] S. Tavakkol, Y.-Y. Chiang, T. Waters, F. Han, K. Prasad, and R. Kiveris. Kartta labs: Unrendering historical maps. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 48–51. ACM, 2019.
- [41] J. Van Hinsbergh, N. Griffiths, P. Taylor, A. Thomason, Z. Xu, and A. Mouzakitis. Vehicle point of interest detection using in-car data. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 1–4. ACM, 2018.
- [42] S. Wang. A cybergis framework for the synthesis of cyberinfrastructure, gis, and spatial analysis. *Annals of the Association of American Geographers*, 100(3):535–557, 2010.
- [43] X. Wang, C. Ma, H. Zheng, C. Liu, P. Xie, L. Li, and L. Si. Dm_nlp at semeval-2018 task 12: A pipeline system for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 917–923, Stroudsburg, PA, USA, 2019. ACL.
- [44] G. Xi, L. Yin, Y. Li, and S. Mei. A deep residual network integrating spatial-temporal properties to predict influenza trends at an intra-urban scale. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 19–28. ACM, 2018.

- [45] Y. Xin and P. R. Adler. Mapping miscanthus using multi-temporal convolutional neural network and google earth engine. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 81–84. ACM, 2019.
- [46] T. Xing, Y. Gu, Z. Song, Z. Wang, Y. Meng, N. Ma, P. Xu, R. Hu, and H. Chai. A traffic sign discovery driven system for traffic rule updating. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 52–55. ACM, 2019.
- [47] Y. Xu, L. Pan, C. Du, J. Li, N. Jing, and J. Wu. Vision-based uavs aerial image localization: A survey. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 9–18. ACM, 2018.
- [48] Y. Xu, Z. Piao, and S. Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5275–5284, 2018.
- [49] B. Yan, K. Janowicz, G. Mai, and S. Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 35. ACM, 2017.
- [50] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2600–2614, Aug 2018.
- [51] J.-A. Yang and M. Jankowska. Contextualizing space and time for geoai jitais (just-in-time adaptive interventions). In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 66–68. ACM, 2019.
- [52] Y. Yin, A. Sunderrajan, X. Huang, J. Varadarajan, G. Wang, D. Sahrawat, Y. Zhang, R. Zimmermann, and S.-K. Ng. Multi-scale graph convolutional network for intersection detection from gps trajectories. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 36–39. ACM, 2019.
- [53] Z. Yin, H. Xiong, X. Zhou, D. Goldberg, D. Bennett, and C. Zhang. A deep learning based illegal parking detection platform. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 32–35. ACM, 2019.
- [54] J. Yuan. Learning building extraction in aerial scenes with convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2793–2798, 2017.
- [55] X. Yuan and A. Crooks. Assessing the placeness of locations through user-contributed content. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 15–23. ACM, 2019.
- [56] F. Zhang, B. Du, and L. Zhang. Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1793–1802, 2015.
- [57] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

Batman or the Joker? The Powerful Urban Computing and its Ethics Issues

Kaiqun Fu, Abdulaziz Alhamadani, Taoran Ji, Chang-Tien Lu
Department of Computer Science, Virginia Tech, Falls Church, USA
E-mail: {fukaiqun, hamdani, jtr, ctlu}@vt.edu

Abstract

The exponential growth of the urban data generated by urban sensors, government reports, and crowdsourcing services endorses the rapid development of urban computing and spatial data mining technologies. Easier accessibility to such enormous urban data may be a double-bladed sword. On the one hand, urban data can be applied to solve a wide range of practical issues such as urban safety analysis and urban event detection. On the other hand, ethical issues such as biasedly polluted urban data, problematic algorithms, and unprotected privacy may cause moral disaster not only for the research fields but also for the society. This paper seeks to identify ethical vulnerabilities from three primary research directions of urban computing: urban safety analysis, urban transportation analysis, and social media analysis for urban events. Visions for future improvements in the perspective of ethics are addressed.

1 Introduction

Ethics issues in data mining such as privacy and anonymity protection, algorithmic biases, and surveillances have been raised by some of the researchers in the recent decade [22, 14, 3]. With the ubiquitous deployment of the urban sensors and rapid growth of crowdsourcing technologies, urban computing and spatial data mining techniques begin to thrive in detecting and analyzing urban events. An enormous amount of urban data is generated from multiple sources such as traffic sensors, air quality meters, various government agency reports, and event social media as crowdsourcing. Such a rise of “big urban data” has boosted the development of the urban computing techniques in several important directions such as urban safety inference, intelligent transportation systems, and social media analysis for urban events. However, the ethical issues in urban computing and spatial data mining fields do not receive enough attention as the exponential growth of the big urban data.

This paper is dedicated to providing discussions of ethics for several research directions of urban computing in detail. The paper is structured into three directions:

- *Urban Safety Analysis.* Safety and security-related events detection and prediction are important research topics in urban computing and spatial data mining. With the rapid deployment of urban sensors, more accessible government reports, and social media platforms, there are more innovative approaches and applications to tackle urban safety analysis problems. However, the importance of ethical issues is rarely addressed by the newly proposed methods. We summarize the state-of-the-art works in urban safety analysis, point out the ethical issues in the existing practices, and provide potential improvements.
- *Urban Transportation Prediction and Analysis.* To provide and maintain benign mobility within urban areas is one of the most fundamental requirements of Intelligent Transportation Systems (ITS). The massive mounted traffic sensors nowadays are generating Gigabytes of data per hour. With effective spatial data

mining techniques and urban computing algorithms, various ITS related topics such as traffic forecasting, incident and disruption detection, and incident impact analysis have been proposed. However, such transportation-related urban data is mostly generated by government agencies, which leads to unscrupulous concerns such as intruding commuters' privacy and discriminatory surveillance. We summarize the state-of-the-art works in urban transportation analysis, point out the ethical issues in the existing works, and provide potential improvements.

- *Social Media Analysis for Urban Events.* Social media and location-based services with user-posted content have generated a staggering amount of information that has potential applications in various areas such as urban event detection and incident analysis. The omnipresence of social media and location data is capable of representing people's behavior, attitudes, feelings, and relationships. These features can be ethically problematic, even where such data exist in the public domain. Unfortunately, such issues are rarely addressed in the research fields of urban computing and spatial data mining. We summarize the state-of-the-art works in urban events detection from social media, point out the ethical issues in the existing works, and provide potential improvements.

2 Ethics of Urban Computing and Urban Safety

With the ubiquitous deployment of the urban sensors and rapid growth of crowdsourcing technologies, urban computing and spatial data mining techniques begin to thrive in detecting and analyzing urban events. Among all categories of urban events, safety and security-related events should be treated as one of the most important ones without a doubt. The urban computing community has addressed important problems such as urban safety and crime prediction [6, 26, 9], safe route recommendations [33, 12], and threats detection [21]. However, the convenience and accessibility of such abundant urban and spatial data generated by the urban sensors, end-users, and city administrators put a spotlight on unethical issues such as biased datasets, biased algorithms, biased results, and compromised privacy. Such problems are rarely addressed by the researchers in the urban safety analysis fields. In this section, we summarize some of the pioneering research works in the urban safety analysis field, address the potential ethical issues, and then provide our visions on how to tackle and improve or mitigate the current research status of the ethical issues in the urban computing and spatial data mining fields.

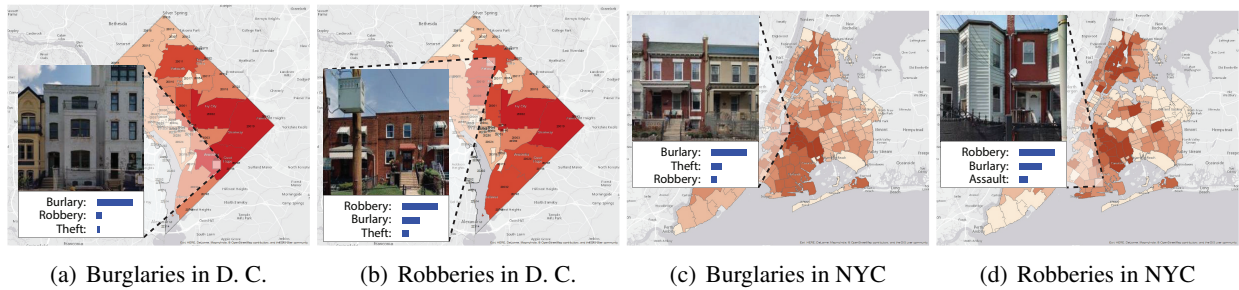


Figure 1: Spatial Distributions of Different Categories of Crime

2.1 Current Works for Urban Safety Analysis

Urban Crime Perception with Convolutional Neural Network. Nadai et al. [6] studied the relationship between a neighborhood's appearance of safety and its levels of human activity in two main Italian cities. By combining the recognized safety scores predicted using a convolutional neural network trained on Google Street View images with mobile phone data, they found that there is a significant positive and negative correlation

depending on gender and age demographics. Urban recommendations could be given to any specific city if the data were available. Figure 1 shows the correlation between the crime distribution and the physical appearances of the city. In another study to explore the connection between urban perception and crime inferences, Liu et al. [26] present a unified framework to learn to quantify safety attributes of physical urban environments using crowd-sourced street-view photos without human annotations. A large-scale urban image dataset is collected in multiple major cities. Safety scores from the government’s criminal records are collected as objective safety indicators. A deep convolutional neural network is proposed to parameterize the instance-level scoring function. Figure 2 shows the structure of the proposed model. The method is capable of localizing interesting images and image regions for each place.

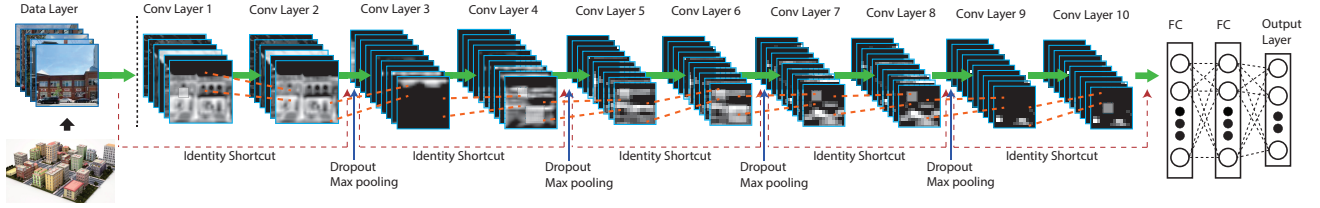


Figure 2: Representation of the Convolutional Neural Network Structure for Learning Urban Crime

Safe Route Recommendation with Crowdsourcing. Shah et al. [33] propose a travel route recommendation system that suggests safe travel routes in real-time by incorporating social media data resources and points of interest review summarization techniques. The system consists of an efficient route recommendation service that considers safety and user interest factors, a transportation-related tweets retriever with high accuracy, and a text summarization module that provides summaries of location-based Twitter data and Yelp reviews to enhance route recommendation service.

Threats Identification from News and Social Media. Airports are a prime target for terrorist organizations, drug traffickers, smugglers, and other nefarious groups. Homeland security professionals must rely on measures of the attractiveness of an airport as a target for attacks. Khandpur et al. [21] present an open-source indicators approach, using urban data sources such as social media and news articles, to conduct a relative threat assessment, i.e., estimating if one airport is under greater threat than another.

2.2 Ethics Issues for Urban Safety Analysis

For crime prediction and safety, analysis works in urban computing, predictive policing and similar urban safety AI models are subjects to problematical algorithms. The problematical algorithms can be generated in several ways:

1) Biased data sources. Existing works involve safety analysis [6, 12] utilize crime reports from the police departments. Such reports from the authorities may inevitably be biasedly polluted during recording, for example, based on stereotypes towards certain groups of people. The review “Policing Predictive Policing” revealed that the crime data is not just biased, it is “notoriously incomplete”. There are certain crimes which were under-reported to authorities such as sexual assault, domestic violence, and fraud. Not only that, some communities, frustrated with current policing practices, simply decline to report crimes [8]. The review, also, reported that half of the crimes with victims go unreported. With such incomplete data, it is not fair to employ PredPol in areas that do not have enough information or poor data. Another case where data can be biased in a different way happened when the GPS location data of potholes in Boston was collected by the smartphone app StreetBump to help patch the potholes in Boston. The idea was great but there was a major problem. Crawford reported that people in lower-income groups and elder residents groups are less likely to have smartphones, where smartphone access can be as low as 16%. This indicates that there is a significant amount of the population have not reported the potholes, therefore there is a crucial part missing from the data [4].

2) Biased Data produces biased results (garbage in, garbage out): Deep learning algorithms learn from the existing data, and when bias exists in the data, discrimination will exist in the results. The flaw in the data will allow these algorithms to “inherit the prejudices of prior decisions makers” [1]. For example, the primary goal of predictive policing is to inform better personnel deployment, then areas that are “less hot” as per the software would most definitely see less police deployment and hence less crime intervention by law enforcement in general. However, this effect would be seen in cases when the police department relied solely on the predictive policing tool over its network of police informants and other sources of ground-level reconnaissance [8]. The intuitive solution to mitigate discrimination in AI models is to equally represent everyone. However, it appears to be another problem. Nordling [30] points out that information from certain population groups usually is not representative because of missing data. Therefore, this lack of diversity is likely to result in biased algorithms. Another intuitive solution is “simply to throw more data at the AI” to expose the AI to wrong cases and correct the errors, another solution is to “change the first network’s inputs” [15]. However, such solutions are not yet enough to fix those errors but at least the errors are recognized.

Vision: An effort mentioned by Kamiran when there is bias in data because data is not accessible, missing or underrepresented is to compensate some of the bias by oversampling underrepresented communities [19]. Further, one of the several solutions for the problem of the discriminatory algorithm that Barocas et al. [1] mentioned is to educate the employees (researchers) to rectify the problem if they understand the causes of the problem, although it is hard to identify the problems sometimes because the resulting discrimination is unintentional. There has been a lot of research to solve the bias in data and bias outcomes. Yet, the problem still persists which requires more effort in future work.

3 Transportation Predictions with Urban Data Sources

Detecting and analyzing transportation-related incidents and forecasting transportation status is critical to the success of the research fields of Intelligent Transportation Systems (ITS) and smart cities. Predicting traffic on urban traffic networks using spatiotemporal models has become a popular research area in the past decade [34]. Transportation-related incident detection, analysis of transit systems [17] and road networks [10] have also gained increasing attention in recent years. Various previous works on traffic prediction and incident analysis apply traffic data sources from traffic sensor providers such as INRIX¹ and Regional Integrated Transportation Information System² (RITIS). However, most of the existing research works ignore ethical issues while utilizing the traffic data generated by urban traffic sensors. Ethical issues such as surveillance on urban activities with such an enormous amount of sensor-generated data and privacy issues on conducting experiments on big data that is generated based on people’s daily mobility should be brought under the spotlight by the researchers in the research areas of intelligent transportation systems and smart city.

3.1 Current Works for Transportation Predictions

Transit Service Disruption Detection from Social Media. Transit agencies are seeking to move beyond traditional customer questionnaires and manual service inspections to leveraging open source indicators like social media for detecting emerging transit events. Inspired by the multi-task learning framework, Ji et al. [17] propose the Metro Disruption Detection Model (*MDDM*), which captures the semantic similarity between transit lines in the Twitter space. The *MDDM* model novel constraints on feature semantic similarity exploiting prior knowledge about the spatial connectivity and shared tracks of the metro network. Figure 3 shows disruption events for 2015 on the Orange, Silver, and Blue lines operated by the Washington Metropolitan Transit Authority. Disruption 1, disruption 2, and disruption 3 occurred on the Orange line. Disruption 4 and disruption 5 occurred on the

¹INRIX: <http://inrix.com/>

²Regional Integrated Transportation Information System: <https://ritis.org/>

Silver line. Disruption 6, disruption 7 and disruption 8 occurred on the Blue line. *MDDM* model successfully detects disruptions 1, 4 and 6.

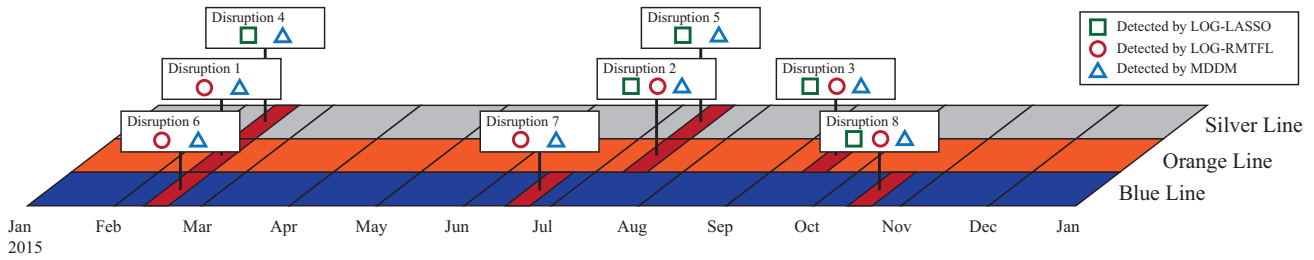


Figure 3: A timeline of metro disruptions on the Orange, Blue, and Silver metro lines in 2015. Events along these spatially interconnected lines often co-occur

Traffic Incident Detection Analysis with Social Media Summarization. Fu et al. [11] propose a social media-based traffic status monitoring system (*Steds*). The system is initiated by a transportation-related keyword generation process. Then an association rules based iterative query expansion algorithm is applied to extract real-time transportation-related tweets for incident management purposes. The feasibility of summarizing the redundant tweets to generate concise and comprehensible textual contents is confirmed.

Traffic prediction in a bike-sharing system. Li et al. [25] propose a hierarchical prediction model to predict the number of bikes that will be rented from/returned to each station cluster in a future period so that reallocation can be executed in advance. A bipartite clustering algorithm is proposed to cluster bike stations into groups, formulating a two-level hierarchy of stations. The total number of bikes that will be rented in a city is predicted by a Gradient Boosting Regression Tree (GBRT). A multi-similarity-based inference model is proposed to predict the rent proportion across clusters, and the inter-cluster transition, based on the number of bikes rented from/ returned to each cluster, can be easily inferred.

3.2 Ethics Issues for Transportation Predictions

Various types of urban sensors are deployed for recording transportation-related information such as vehicle travel speed, road occupancy, and volumes. The mobility of the commuters is monitored by multiple types of techniques such as loops, microwave, acoustic, video, and crowdsourcing such as Twitter and Waze. Collecting seemingly anonymous data that appear hard to identify an individual may in fact be problematic in many ways.

1) The risk of compromising the privacy of an individual. In the research [23], the MIT group combined two anonymized datasets of people in Singapore, one of mobile phone logs and the other of transit trip, each including “geolocation stamps” with time and place of each point. They succeeded to match up 17% of the users in one week and 11 weeks to get 95%. if GPS data from smartphones were added, it took less than a week to reach that number. The group acted as ethical or ‘white hat’ hackers to prove that someone acting in bad faith could do the same and compromise the privacy of many people.

2) Governments’ surveillance on urban mobility. Surveillance is infringing upon personal privacy. The times we live in, there’s a necessity for some surveillance in public spaces to maintain order, but as soon as we set into private spaces, it becomes an issue. The level of government surveillance varies from one government to another depending on the level of freedom practiced by the system. Most governments justify their unnecessary surveillance with words such as “National Threats”, “Security Stabilization”, or “Fighting Crime”. This will lead everyone (citizens) to feel that it is their duty to allow governments to collect their data even when it is not needed. As a result, the collected data from a government surveillance will have different impacts. Data can be collected from anywhere, GPS locations, car license plate scanners, phone calls, social media, closed-circuit television CCTV, and so on. Data can be used in “positive” and “negative” ways. This website presents what

happens when we are negatively being monitored by the government³. Surveillance in public spaces must be done without violating privacy. Some of the previous works in transportation analysis utilize GPS based data sources such as user check-in records and driver's travel paths. These works are vulnerable to violating privacy protection, which should be further addressed in the urban computing and spatial data mining fields.

Vision: The intelligent transportation system community can strengthen its ethics awareness by considering several additional aspects: 1) differential privacy to protect users' privacy, 2) users' awareness to share mobility data, 3) transparency between urban data collectors and users, and 4) comprehensive legislation. One of the advanced methods to preserve the privacy of users' information is implementing Differential privacy which was introduced by Dwork [7]. The method overlaps two areas: data analytics and statistics. Differential privacy uses techniques such as noise injection to keep the data of individual users completely private, and guarantees that the outcome prevents enabling anyone to learn anything about an individual. Meanwhile, it allows the researcher to query from the data and obtain good research results. By maintaining the balance between the implemented model and differential privacy, the privacy of an individual can be protected [18]. Users should have full awareness of the accessibility and availability of the mobility data generated by themselves. Therefore, the user should realize the potential risks of sharing their mobility data. Furthermore, the transparency between the data vendors and the users is not balanced: the vendors or institutions which collect and share data about people know a lot about the users but the users do not know as much about the data collectors. For that, users have the right to know how their data are used and have the right to withdraw from any experiment if asked. Finally, in the future developments of the urban computing industry, more comprehensive legislations should be proposed. Tighter regulations for collecting and sharing data should be established to prevent users' privacy from infringed such as General Data Protection Regulation known as GDPR which was applied in Europe.

4 Social Media based Event Detection and Analysis

Location-based service embedded social media sources such as Twitter and Foursquare have become popular data sources as surrogates for monitoring and detecting events. Targeted domains such as crime, election, and social unrest require the creation of algorithms capable of detecting events pertinent to these domains. Due to the unstructured language, short-length messages, dynamics, and heterogeneity typical of social media streams, it is technically difficult and labor-intensive to develop and maintain supervised learning systems. Recent studies in the research fields of event detection and analysis have applied both supervised and unsupervised learning methods to better modeling and forecasting the spatiotemporal events from social media data sources [36, 37, 20]. However, while utilizing the social media-based data sources, some of the ethical issues, such as built-in biases in dataset and algorithms, have not been paid with enough attention nowadays. In this section, we review some work on location-based social media analysis, state, and propose solutions to the ethical issues arising in the research fields of spatial data mining and social media analysis.

4.1 Current Works for Event Detection from Social Media

Spatial Event Forecasting in Social Media. Social media has become a significant surrogate for spatial event forecasting. The accuracy and discernibility of a spatial event forecasting model are two key concerns, which respectively determine how accurate and how detailed the model's predictions could be. Zhao et al. [37] propose a multi-resolution spatial event forecasting (*MREF*) model that concurrently addresses all the above challenges by formulating prediction tasks for different locations with different spatial resolutions, allowing the heterogeneous relationships among the tasks to be characterized.

Cyber Attack Detection using Social Media. Social media can also be viewed as a sensor into various societal events such as disease outbreaks, protests, and elections. Khandpur et al. [20] describe the use of social

³<https://theyarewatching.org/>

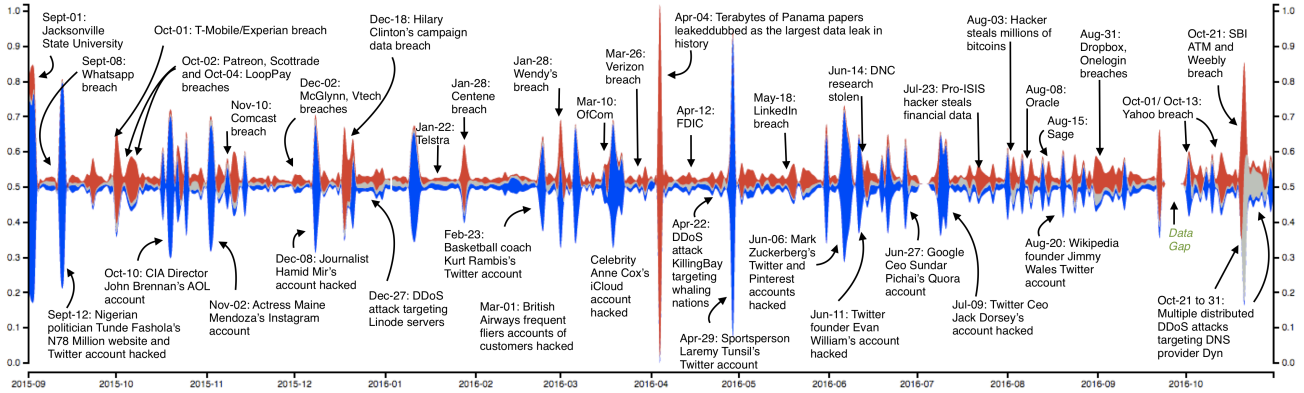


Figure 4: Streamgraph showing normalized volume of tweets (September 2015 through October 2016) tagged with data breach (red), DDoS activity (grey) and account hijacking (blue) types of cyber-security events

media as a crowdsourced sensor to gain insight into ongoing cyber-attacks. The proposed social media-based approach detects a broad range of cyber-attacks in a weakly supervised manner using just a small set of seed event triggers and requires no training or labeled samples. A new query expansion strategy based on convolution kernels and dependency parses helps model semantic structure and aids in identifying key event characteristics. Figure 4 shows the wide range of events that the proposed system is able to detect. Notice the clear bursts in Twitter activity that the query expansion algorithm is able to detect.

Real-Time Detection of Traffic Accidents from Twitter. D'Andrea et al. [5] propose a real-time monitoring system for traffic event detection from Twitter stream analysis. The system fetches tweets from Twitter according to several search criteria, processes tweets, by applying text mining techniques, and finally performs the classification of tweets. The proposed SVM-based system aims to assign the appropriate class label to each tweet, as related to a traffic event or not. The traffic accident detection system is employed for real-time monitoring of several areas of the road networks, allowing for detection of traffic events almost in real-time, often before online traffic news sites.

4.2 Ethics Issues for Event Detection from Social Media

For the work on social media based event detection and analysis, several concerns may contribute to the ethical issues: **1) the built-in biases in social media data.** Bias in social media can happen when a sample is collected in such a way that some members of the intended population are not equally represented in the sample, resulting in a non-random sample. If such an error occurs in sampling, the results of the research could be mistakenly attributed to the study phenomenon instead of the sampling method [2]. In addition, due to the way humans generate their opinions, it is argued that misinformation in social media data can be produced in several ways, purposefully or accidentally. In a study experimented on Facebook's News Feed [16, 24], the group demonstrated how negative or positive posts disseminate by reducing positive or negative expressions. One case is that emotional contagion can play a massive role in generating biases in the social media platform, and this is demonstrated by the 2016 US elections when emotional contagion is used to foster negative viewpoints against specific politicians [35]. If such biases exist in Social media data of Urban planning projects, discriminatory results may occur.

2) the consequences of flaws in data being manipulated by the users. Relying on data collected from users in social media in order to use the model to recommend safer routes or detect threats is vulnerable to manipulation by malicious users. Recommendation systems, a similar system to the mentioned works, have been under attack since their appearance. An attack is carried out when fake users or certain legitimate malicious

users feed content to the system in a certain way that makes the system frequently favors a recommendation over what would the system should result [29, 13]. Leveraging human’s tendency to use their peer’s opinions to make their own decisions, large companies tend to use fake accounts and bots on social media for product promotion, posting positive comments, and silencing critics [28].

3) Social media privacy concerns. Social media is a beneficial source for real-time and historical data which makes it a great candidate to be surveilled positively and negatively. Social media has helped law enforcement to detect human trafficking activities, solve murder cases and other criminal activities [27]. In a negative way, in some governments where censoring is more strict than others. However, those governments do not censor many social media sources. Social media is where people tend to speak out about their local problems expecting solutions from the government. Yet, the government has escalated their efforts to monitor and suppress [32]. In a recent report published by the Brennan Center, Social Media Monitoring, there is proof that Department of Homeland Security DHS is exploiting personal data mined from social media to surveil protestors, religious and ethnic minorities [31]. The surveillance has expanded from positive (national security) to discriminatory surveillance.

Vision: While using social media data, researchers do need to be very careful and watchful for these issues, i.e., ensuring that the data’s source and gathering methodologies mitigate bias, the cleaned data did not incidentally impose bias, and the data is used in a way that does not use or impose any bias. Monitored social media chatter could be used to inform law enforcement personnel of large potentially violent or non-violent protest events or other sociopolitical gatherings that might require a police presence to prevent untoward incidents. Such alerts, if provided to police departments with enough lead time, will allow for better planning of deployment logistics ensuring the safety of the officers and the civilians they are deployed to protect. EMBERS is once such an anticipatory intelligence system that forecasts population-level events in multiple countries in Latin America and provides not only the counts of the number of protests but also other details pertaining to the date, time and location of the potential protest event.

5 Conclusion

The rapid growth of the urban data for urban computing and spatial data mining methods raises new challenges along with new opportunities for various application fields such as urban safety analysis, intelligent transportation systems, and event detection. However, unfortunately, the ethical issues while obtaining, utilizing, and inferring from such enormous urban data are rarely addressed by the current researchers in urban computing communities. This paper reviews the most popular research branches of urban computing and spatial data mining, including urban safety analysis, intelligent transportation systems, and event detection. Ethical vulnerabilities for the existing urban computing and spatial data mining works such as predictive policing, biased data sources, compromised privacy, surveillance, and discrimination are revealed by this paper. Promising potential solutions and prospective visions towards such identified ethical vulnerabilities are directed. We offer these discussions of ethics in urban computing with the hope that they contribute highlight attention from the urban computing communities.

References

- [1] S. Barocas and A. D. Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [2] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.
- [3] J. Cranshaw. Whose city of tomorrow is it?: on urban computing, utopianism, and ethics. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 17. ACM, 2013.

- [4] K. Crawford. The hidden biases in big data. *Harvard Business Review*, 1:1, 2013.
- [5] E. D’Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni. Real-time detection of traffic from twitter stream analysis. *IEEE transactions on intelligent transportation systems*, 16(4):2269–2283, 2015.
- [6] M. De Nadai, R. L. Vieriu, G. Zen, S. Dragicevic, N. Naik, M. Caraviello, C. A. Hidalgo, N. Sebe, and B. Lepri. Are safer looking neighborhoods more lively?: A multimodal investigation into urban life. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1127–1135. ACM, 2016.
- [7] C. Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.
- [8] A. G. Ferguson. Policing predictive policing. *Wash. UL Rev.*, 94:1109, 2016.
- [9] K. Fu, Z. Chen, and C.-T. Lu. Streetnet: preference learning with convolutional neural network on urban crime perception. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 269–278. ACM, 2018.
- [10] K. Fu, T. Ji, L. Zhao, and C.-T. Lu. Titan: A spatiotemporal feature learning framework for traffic incident duration prediction. In *Proceedings of the 27nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2019.
- [11] K. Fu, C.-T. Lu, R. Nune, and J. X. Tao. Steds: Social media based transportation event detection with text summarization. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1952–1957. IEEE, 2015.
- [12] K. Fu, Y.-C. Lu, and C.-T. Lu. Treads: A safe route recommender using social media mining and text summarization. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 557–560. ACM, 2014.
- [13] I. Gunes, C. Kaleli, A. Bilge, and H. Polat. Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review*, 42(4):767–799, 2014.
- [14] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126. ACM, 2016.
- [15] D. Heaven. Why deep-learning ais are so easy to fool., 2019.
- [16] M. Ienca and E. Vayena. Cambridge analytica and online manipulation. *Scientific American*, 30, 2018.
- [17] T. Ji, K. Fu, N. Self, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for transit service disruption detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 634–641. IEEE, 2018.
- [18] Z. Ji, Z. C. Lipton, and C. Elkan. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*, 2014.
- [19] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.
- [20] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan. Crowdsourcing cybersecurity: Cyber attack detection using social media. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1049–1057. ACM, 2017.

- [21] R. P. Khandpur, T. Ji, Y. Ning, L. Zhao, C.-T. Lu, E. R. Smith, C. Adams, and N. Ramakrishnan. Determining relative airport threats from news and social media. In *Twenty-Ninth IAAI Conference*, 2017.
- [22] J. M. Kleinberg. Challenges in mining social network data: processes, privacy, and paradoxes. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 4–5. ACM, 2007.
- [23] D. Kondor, B. Hashemian, Y.-A. de Montjoye, and C. Ratti. Towards matching user mobility traces in large-scale datasets. *IEEE Transactions on Big Data*, 2018.
- [24] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.
- [25] Y. Li, Y. Zheng, H. Zhang, and L. Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 33. ACM, 2015.
- [26] X. Liu, Q. Chen, L. Zhu, Y. Xu, and L. Lin. Place-centric visual urban perception with deep multi-instance regression. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 19–27. ACM, 2017.
- [27] A. Mateescu, D. Brunton, A. Rosenblat, D. Patton, Z. Gold, and D. Boyd. Social media surveillance and law enforcement. *Data Civ Rights*, 27:2015–2027, 2015.
- [28] J. McGregor. Reddit is being manipulated by big financial services companies. *Forbes Review*, 1:2, 2017.
- [29] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology (TOIT)*, 7(4):23, 2007.
- [30] L. Nordling. Mind the gap, 2019.
- [31] F. Patel, R. Levinson-Waldman, S. DenUyl, and R. Koreh. Social media monitoring. 2019.
- [32] B. Qin, D. Strömberg, and Y. Wu. Why does china allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives*, 31(1):117–40, 2017.
- [33] S. Shah, F. Bao, C.-T. Lu, and I.-R. Chen. Crowdsafe: crowd sourcing of crime incidents and safe routing on mobile devices. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 521–524. ACM, 2011.
- [34] Y.-J. Wu, F. Chen, C.-T. Lu, and S. Yang. Urban traffic flow prediction using a spatio-temporal random effects model. *Journal of Intelligent Transportation Systems*, 20(3):282–293, 2016.
- [35] U. Yaqub, S. Chun, V. Atluri, and J. Vaidya. Sentiment based analysis of tweets during the us presidential elections. In *Proceedings of the 18th annual international conference on digital government research*, pages 1–10. ACM, 2017.
- [36] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS one*, 9(10):e110206, 2014.
- [37] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-resolution spatial event forecasting in social media. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 689–698. IEEE, 2016.

Quantum Spatial Computing

Martin Werner

Institute for Applied Computer Science & Research Institute CODE,
Bundeswehr University Munich, Germany
martin.werner@unibw.de

Abstract

Quantum computing is expected to create a major shift across the whole computing industry given that it can solve a certain set of operations that used to be very hard in comparably short time. In this article, the author intends to introduce central aspects of quantum algorithms to the GIS community and explores some probable directions of research related to quantum computing for spatial algorithms. The author hopes that this paper enables researchers of our community to place an informed decision whether and when it is worth looking at this exciting new area of algorithm research and computing while staying on an abstraction level high enough for concise presentation.

1 Introduction

In the last decade, quite a few companies are having success with building first real-world quantum computers. Among these are D-Wave systems [6], Google [10], and IBM [12]. However, these machines are currently very small and their usage implies lots of noise. Nevertheless, quantum computers are nowadays existing and there is a race of growing them to a useful size. The interesting aspect of quantum computing is that it can provide exponential speedups in certain situations. For example, it is possible to factor integer numbers in polynomial time using Shor’s algorithm on a quantum gate computer or it is possible to solve certain NP-hard problems in polynomial time.

In this paper, we explore how quantum computing could affect the spatial computing domain in a very broad sense and as the community might not have been exposed too much with basic concepts of quantum computing, the paper will first briefly recall the computational models that quantum computers operate in, because these form a meeting point between the engineers and physicians actually trying to build computers and the computer scientists looking for algorithms that can exploit quantum computing peculiarities.

2 Quantum Computing

Quantum computing is an idea in which quantum effects are being used to model information. Similarly to digital computers, the atomic unit of information is a bit, called qubit in the quantum computing domain. Similarly to the situation of a digital computer, a qubit can typically be set to two different values (e.g., true and false) and if it is read, it will realize a binary value (true or false). Internally, however, it can be in a superposition state where it actually takes “both values at the same time”. If a quantum bit is in perfect superposition and you read it, you will get a zero and a one with equal probability. When introducing quantum gate computers in a later section, we will understand much more about qubits.

From an application perspective, the simplest form of quantum computing is adiabatic quantum computing or quantum annealing. A quantum annealer is a specialized quantum computers that is built to solve a certain type of optimization problem using qubits that are in superposition and entangled with each other according to a problem specification. As we want to avoid explaining the physics and engineering of building quantum annealers, we will describe the computational model only, which is the model of Quadratic Unconstrained Binary Optimization (QUBO). This model is equivalent to the (physics-inspired) Ising model, you can jump back and forth with simple algebraic modifications.

A QUBO is given as a quadratic form $x^t Ax$ where $x \in \mathbb{B}^n$ is a vector of binary values and A is a matrix of real numbers. The QUBO task is to find the vector $x \in \mathbb{B}^n$ that minimizes this quadratic form given that x is binary. As a rough idea of how a physical implementation works, let us just say that it is based on the energy minimization principle of nature. That is, the matrix is somehow encoded into some hardware and then the hardware is cooled down near to the absolute zero point of temperature and then the bits can be read that minimize the QUBO because the solution to the QUBO minimizes the energy of the physical construction used. This is essentially how the D-Wave quantum annealer works. That is, from an application perspective (e.g., as a spatial computing researcher), you will have to encode your problem into the QUBO formulation, let the quantum annealer solve your QUBO and expect the results to be very noisy or, in other words, your algorithm should be able to deal with good yet suboptimal solutions to the given QUBO. Usually, quantum computers are used over the Internet, because operating them is a very difficult task (e.g., cooling down as much as possible, minimizing all sorts of noise influence to the machine, etc.). Hence, even more hands-on, you would prepare a QUBO matrix, send it to the annealer and get a vector back that might be good in terms of minimizing the QUBO form.

The second and more advanced form of quantum computing is a quantum gate computer. To (informally) explain a quantum gate computer, we need to go back to the concept of a qubit. In fact, a qubit is represented as the tensor product of two one-dimensional complex vector spaces. It is customary to introduce two symbols $|1\rangle$ and $|0\rangle$ to represent two basis elements of this tensor product. That is, $|0\rangle$ can be identified with the vector $(1, 0)$ and $|1\rangle$ with $(0, 1)$. Then a qubit is given by two complex numbers α and β forming a linear combination

$$\alpha|0\rangle + \beta|1\rangle \text{ constrained by } |\alpha|^2 + |\beta|^2 = 1.$$

This allows for a geometric interpretation as the topological space of a qubit is homeomorphic to the surface of a sphere in \mathbb{R}^3 known as the Bloch sphere. As quantum gate computers come with a preferred basis which provides the link to boolean values (e.g., a quantum $|1\rangle$ represents true, a qubit $|0\rangle$ represents false, all other qubits represent true and false at the same time with varying probability), quantum gates are given as 2×2 matrices with complex entries. These matrices represent rotations, thus, need to be unitarian. This is an interesting aspect as this implies that quantum gate computers are reversible computers, e.g., each operation you can do can be undone as well.

Here are some widely-used gates, namely the N gate which provides negation, the Hadamard gate H which brings basis elements into “perfect superposition”. The V and Z gate operate only on the $|1\rangle$ part leaving $|0\rangle$ unchanged, where $V_\pi = Z$.

$$N = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, V_\theta = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix}, Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

It is customary to look at the Hadamard gate once: Operating H on $|0\rangle$ gives

$$H|0\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$$

The magnitudes of α and β now give the probability that a measurement of the qubit returns true ($|\alpha|^2$) or false ($|\beta|^2$). With this information, it is easy to see that measuring $H|0\rangle$ returns true and false with a probability

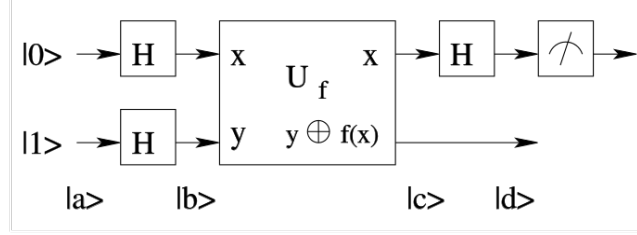


Figure 1: Two-qubit implementation of quantum algorithm of Deutsch.

of 50% each. Given that there are infinitely many unitarian matrices that could be made a quantum gate in a physical quantum computer implementation, it is worth noting that there are universal sets of gates with which any unitarian matrix can be produced up to some given error. One such set is given by H , N_C (the application of some generic procedure on N), and $V_{\frac{\pi}{4}}$. For details, the reader is referred to the excellent tutorial [19].

Now, these quantum gate computers usually bind together a few qubit into a quantum register and the gates operate on register level. The first non-trivial algorithm is given by the following quantum algorithm which is able to decide whether two things are equal by only reading once. This is a powerful concept and actually motivates the Shor algorithm for factorizing large integer numbers.

The basic process of applying a quantum gate computer to a problem is now to first set a certain quantum register to a known base-state (e.g., combinations of the base states $|0\rangle$ and $|1\rangle$ per qubit). Then, some operations bring some of the qubits in superposition such that the following operations can operate on them. Finally, the qubits should have been modified by the algorithm that sampling from them reveals the answer to the given problem.

One widely-used trick here, valid for any boolean function f , is the construction of U_f , which given a Boolean function

$$f : \mathbb{B}^m \rightarrow \mathbb{B}^n$$

provides a unitarian map

$$\hat{f} : \mathbb{B}^m \times \mathbb{B}^n \rightarrow \mathbb{B}^m \times \mathbb{B}^n$$

by sending $(x, y) \rightarrow (x, y \oplus (fx))$ ¹.

This is, for example, used in the most simple non-trivial quantum algorithm known as the algorithm of Deutsch [9]. This can be used to decide, whether a given one-dimensional Boolean function $f : \mathbb{B} \rightarrow \mathbb{B}$ is constant, e.g., $f(0) = f(1)$, or balanced, e.g., $f(0) \neq f(1)$. In classical computing, it is possible to do this by reading both $f(0)$ and $f(1)$, but not faster. For a quantum computer, however, it is possible to do so faster.

To understand this algorithm, one can just compute it operation by operation following the diagram in Figure 1². While you are doing so, you realize that f is evaluated exactly once but using a qubit in superposition and that the algorithm returns the needed information about f , which would be classically impossible. Assuming that the evaluation of f takes, say, one year, this brings runtime from two years in the classical model down to one year using a quantum computer.

This is possibly the simplest situation in which one can see how and that Quantum algorithms can be significantly faster for traditional problems formulated in Boolean functions and concludes our short trip of quantum gate computers.

¹In this context, \oplus is used for addition modulo 2

²The image has been taken from Wikipedia, cf. <https://de.wikipedia.org/wiki/Deutsch-Jozsa-Algorithmus>

3 Quantum Computing Examples

This section provides a few examples and resources on existing quantum algorithms with a tie to spatial computing. However, it is not intended to be complete or to introduce any of these solutions, it shall rather provide a concise starting point for interested readers to explore the field.

3.1 Logical Networks of Light Switches

Adiabatic quantum computing is often introduced with the light switch game. Given a network of n light switches connected with some logics to a single lamp, find the setting in which the light is turned on without knowing the connectivity network of the light switches. In fact, this is a hard problem which could practically only be solved with a brute force approach or additional information. A brute force approach, however, needs to evaluate 2^n possible states of the n switches and becomes intractable even for moderate n .

As a good source for learning this problem, we refer the user to the more general class of circuit satisfiability problems [18] and to the D-Wave introduction to their quantum computer [5].

3.2 Map Coloring

The map coloring problem is a nice problem for understanding how an adiabatic quantum computer works in a spatial setting and it is also given a core example in the D-Wave manual, because it is conceptually easy yet shows the limitations of the D-Wave hardware (namely that there aren't couplers between all qubits) and how to circumvent this by "duplicating" qubits constraining them to be equal. The central question when doing spatial computing with quantum computers will be what a bit represents. Traditionally, spatial computing takes place a lot in a geometry domain with floating point representations of geometry. This domain is not well-suited for adiabatic quantum computing, because it is not easy to find what a qubit should represent. Topology in contrast provides many problems that are of combinatorial nature and, therefore, very well-suited for adiabatic quantum computing.

The map coloring problem aims to assign colors to a planar map such that no two neighbors have the same color. In the mathematics, this problem has been widely discussed, because it remained unclear and is considered the first real mathematical problem of relevance that has been solved using a computer. In 1852, Francis Guthrie formulated the conjecture that any given planar map can be coloured with four colors while creating a map of England. After that, a sequence of attempts in proving this conjecture has been established where most solutions survived a decade, but finally were proven wrong. During this time, however, an upper bound with five colors was established by Heawood. Heinrich Heesch proposed a computer proof which has never been implemented due to limited machine power, however, it was refined a few times bringing down the set of solutions which need manual or computer checking to 1936 (Appel and Haken, 1976) leading to a publication with 400-page appendix in 1989. Finally, this problem was formally proven in 2005 using the computer-assisted proof environment Coq by Benjamin Werner (still not closing the debate in mathematics, because hardliners accept proofs only if they are fully transparent in the sense that they are accessible to the mind of humans).

How can this be implemented in a quantum computer? As said, the pressing question is how to map the problem to qubits. And in this case, it can be done as follows: Each country in the map is assigned a color represented as a one-hot encoding of qubits. That is, we assign a set of k qubits to each country in the map and constrain them such that exactly one of those qubits is one at a time. Then, the topology of neighbors is encoded as constraints, essentially, that their vectors must not be the same. This can be done in the QUBO framework and it can be mapped to the D-WAVE hardware if the number of qubits suffices. There is a D-WAVE white paper with lots of details on this problem [7].

3.3 Dynamic Time Warping

Dynamic Time Warping is another very nice example showing how geometric problems can be supported by quantum computers. Dynamic time warping is a distance definition which has found many applications. It has been applied to varying application areas including speech recognition [20], handwriting recognition and signature verification [4], in computer vision [1], in shape retrieval [16], in computational biology and medicine [15], in pattern recognition [2], and recently similarity search for spatial trajectories [21].

Dynamic Time Warping is based on extending the idea of a point-wise Euclidean distance of time series by allowing the “matching” to change according to dynamic programming rules. In fact, the dynamic time warping distance of trajectories (and time series) is defined to be the sum of the distances of points of both trajectories, where the points are matched based on minimizing over all matchings. A matching, in this context, is an assignment of points of one trajectory with points of the other trajectories. For dynamic time warping, a matching always starts by matching the first two points of the trajectory. Then, it can either go one step in one of the trajectories or one step in both. Ultimately, however, it has to reach the last vertex of both trajectories. This is a natural dynamic programming problem and can be formulated as follows [22]:

$$d_{\text{DTW}}(a_{1\dots n}, b_{1\dots m}) = \begin{cases} 0 & \text{if } a \text{ and } b \text{ are empty} \\ \infty & \text{if only one of } a \text{ and } b \text{ is empty} \\ d(a_n, b_m) + \min \begin{cases} d_{\text{DTW}}(a_{1\dots n-1}, b_{1\dots m-1}) \\ d_{\text{DTW}}(a_{1\dots n-1}, b_{1\dots m}) \\ d_{\text{DTW}}(a_{1\dots n}, b_{1\dots m-1}) \end{cases} & \text{otherwise} \end{cases}$$

This distance can be computed in $O(mn)$ time [8]. It is usually composed of first computing the full distance matrix and, then, searching for the shortest path in this distance matrix going right, up, or both at the same time.

We will now simplify to dynamic time warping problems given with equal-length trajectories of n points each. As a quadratic optimization problem, DTW can be written using the distance matrix and a set of constraints only. Concretely, consider the set $\mathcal{A}_{m,n}$ of monotonous alignment matrices. These are binary matrices containing a path from the upper-left corner (1,1) to the lower-right corner (m-1,n-1) containing only of moves \rightarrow , \searrow , and \downarrow . The size of this set is known as the Delannoy number, cp. sequence A001850 in the OEIS. With this set in place, the DTW distance is given as

$$d_{\text{DTW}}(x, y) = \min_{A \in \mathcal{A}_{n,n}} \langle A, \Delta(x, y) \rangle, \quad (1)$$

where $\langle -, - \rangle$ is the Frobenius inner product of matrices.

But how can we map this to a quantum computer? One approach is to write the expression using binary vectors $x \in \mathbb{B}^{n^2}$ obtained by flattening the path matrices exploiting the fact that the Frobenius product is the sum of the entries of the Hadamard product of the two given matrices.

$$d_{\text{DTW}}(x, y) = \min_{x = \text{flatten}(A) \text{ for } A \in \mathcal{A}_{n,n}} x^t D x$$

This is now clearly a quadratic optimization problem, but how can we get rid of the constraint? It is possible to represent certain constraints as matrices as well. Such a matrix C would, for example, take values

$$x^t C x = \begin{cases} 0 & \text{if } \text{mat}(x) \in \mathcal{A}_{n,n} \\ 1 & \text{else,} \end{cases}$$

where mat denotes the operation turning the vector x back into a square path matrix. Assuming such a matrix C exists, DTW could be solved by solving

$$\min_x x^t D x + \lambda x^t C x = \min x^t (D + \lambda C) x \quad (2)$$

If λ is chosen large enough (e.g., $2n \max D$), then each time the constraint matrix would fire, the value in Eq. 2 would be impossible to become smaller than the worst value one could generate with the distance matrix part. If, however, the constraint is fulfilled, we are left with the situation of DTW in Eq. 1.

This concludes our meta-algorithm for DTW. The art of programming adiabatic quantum computers is now to come up with constraints that are good enough to guarantee an optimal solution. Note that for the case of DTW, a matrix C with the properties above cannot exist, because $x = 0 \in \mathbb{B}^{n^2}$ is clearly not a path matrix, yet $x^t A x$ is zero. Therefore, and this is the heart of QUBO-based quantum computing, we have to relax from the case of ideal constraints and come up with alternatives that allow for concluding the final result or parts thereof.

3.4 Quantum Gate Computing Strategy

Quantum gate computing allows for solving a larger set of problems as opposed to the adiabatic case, however, it is not as accessible. However, quantum gate computers can solve many of the NP-hard combinatorial problems as well and go beyond that. A nice overview of existing algorithms for quantum gate computers is maintained online in the Quantum Algorithm Zoo [13].

If you want to exploit the power of a quantum gate computer in spatial computing, you have to design what each bit will represent and at the same time, you have to prepare an intuition of what a superposition of these bits can bring to you and how quantum gates will encode the constraints of the solution to your problem at hand. This domain is still in its infancy as it is not very easy to come up with intuitions like that. Therefore, for spatial computing, I expect quantum gate computers to be used more often in an indirect way: *If you want to apply a quantum gate computer to spatial computing and don't have an innovative intuition on how to exploit and modify bits in superposition, you will still be able to rely on the solution capabilities of quantum computers for known quantum algorithms.*

For a quantum computer, this includes the solution of traveling salesman type problems as well as QUBOs. That is, one common way of “programming” a quantum gate computer might be by decomposing the spatial problem into a set or sequence of hard problems for which a quantum algorithm has already been proposed.

There are plenty of examples of how to formulate spatial optimizations in form of traveling salesman type problems and it is left to the reader to exercise with this philosophy.

3.5 Quantum Machine Learning

As quantum computers are very fast in solving certain types of optimization, they are a candidate for many algorithms in the machine learning domain that internally rely on optimization. One point in time where this journey starts might be seen in the paper [3]. A major milestone in the sequel is Harrow, Hassidim, and Lloyd’s quantum algorithm for solving linear systems [11]. Lately, Kerenidis and Prakash’s algorithm for recommender systems provided exponential speedup over any known classical algorithm [14]. Most interestingly, this algorithm has given the intuition to a classical algorithm with an equivalent asymptotic runtime complexity [17]. This is a nice story for finishing this short introduction to quantum computing as it makes clear that understanding quantum algorithms can directly feedback to the field of classical algorithms and is tightly related to a family of algorithms nowadays known as randomized algorithms, which are, as well, underrepresented in the spatial computing community of these days.

4 Research Need

Spatial computing was mostly inspired by the geometry of the things we discuss. It is quite typical that this happens in a continuous representation of space using floating point numbers though there are a few approaches that try to solve spatial computing problems in an integer domain (graph rounding, fixed-precision arithmetics, binarization). In the last decades, the topology of things is being used more and more in spatial computing

and methods from topology quite naturally generate combinatorial problems. The map coloring problem is a typical example where the topology relations of neighboring polygons is representing the spatial aspect and the resulting optimization problem is combinatorial.

Given that quantum computers will remain small for a certain time, the spatial computing community should look into how certain problems can be solved in a combinatorial or at least in an integer-rounded fashion in order to make efficient use of individual qubits.

In addition, the spatial computing community was used to solve their problems on their own by extending and adopting solutions to similar non-spatial problems. The author is convinced that there are quite a few spatial computing problems that can be solved by using the quantum computer as an oracle solving a certain set of hard problems. In this setting, quantum algorithm research takes the form of finding problem transformations and problem reformulations such that a given problem of interest is expressed as a sequence or low-complexity algorithm formulating how the problem should be solved under the assumption that a set of hard problems has a quantum computing algorithm of small time- and space complexity.

5 Conclusion

We are living in a time in which the first commercial quantum computers reach the market. At the same time, there has been a lot of research in the area of quantum computing which motivated to build a quantum computer. The community is centered around proving “quantum supremacy” in many senses, namely, that the expensive research on quantum computers and the expensive operation of those is needed, because there are problems intractable for classical computers that become solvable. In consequence, research that uses quantum computers on problems that are not intractable today is in its infancy and an exciting area for further research. As an example, the dynamic time warping distance has been heavily researched in the classical computing model but is typically used in a restricted setting where the amount of warping is limited in order to make it both indexable and fast. With a quantum computer and certain developments that are beyond the scope of this paper, we might hope to remove this typical restriction and, thereby, allow dynamic time warping to be used without such warping restriction in a scale-free manner.

With this paper, the author hopes to have motivated a few spatial computing researchers to consider quantum computing in future research by looking for instances of hard problems that you are currently avoiding with tricks but quantum computer could directly solve as well as for problems with a natural structure of binary optimization.

References

- [1] A. Almog, A. Levi, and A. M. Bruckstein. Spatial de-interlacing using dynamic time warping. In *IEEE International Conference on Image Processing*, volume 2, pages 1006–1010. IEEE, 2005.
- [2] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.
- [3] N. H. Bshouty and J. C. Jackson. Learning dnf over the uniform distribution using a quantum example oracle. *SIAM Journal on Computing*, 28(3):1136–1153, 1998.
- [4] W.-D. Chang and J. Shin. Modified dynamic time warping for stroke-based on-line signature verification. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 724–728. IEEE, 2007.
- [5] Quantum computing primer. <https://www.dwavesys.com/tutorials/background-reading-series/quantum-computing-primer>.

- [6] D-WAVE Systems, 2019. <https://www.dwavesys.com/>.
- [7] E. D. Dahl. Programming with d-wave: Map coloring problem, 2013.
- [8] K. Deng, K. Xie, K. Zheng, and X. Zhou. Trajectory indexing and retrieval. In *Computing with spatial trajectories*, pages 35–60. Springer, 2011.
- [9] D. Deutsch. Quantum theory, the church–turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 400(1818):97–117, 1985.
- [10] Google AI Quantum, 2019. <https://ai.google/research/teams/applied-science/quantum/>.
- [11] A. W. Harrow, A. Hassidim, and S. Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15), 2009.
- [12] IBM Q, 2019. <https://www.ibm.com/quantum-computing/>.
- [13] S. Jordan. Quantum Algorithm Zoo, 2019. <https://quantumalgorithmzoo.org/>.
- [14] I. Kerenidis and A. Prakash. Quantum recommendation systems. *arXiv preprint arXiv:1603.08675*, 2016.
- [15] B. Legrand, C. Chang, S. Ong, S.-Y. Neo, and N. Palanisamy. Chromosome classification using dynamic time warping. *Pattern Recognition Letters*, 29(3):215–222, 2008.
- [16] A. Marzal, V. Palazón, and G. Peris. Contour-based shape retrieval using dynamic time warping. In *Current Topics in Artificial Intelligence*, pages 190–199. Springer, 2005.
- [17] E. Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 217–228. ACM, 2019.
- [18] A. S. Tannenbaum. NP-hard problems. 2014.
- [19] B. Valiron. Quantum computation: a tutorial. *New Generation Computing*, 30(4):271–296, 2012.
- [20] V. Velichko and N. Zagoruyko. Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2(3):223–234, 1970.
- [21] H. Wang, H. Su, K. Zheng, S. Sadiq, and X. Zhou. An effectiveness study on trajectory similarity measures. In *Proceedings of the Twenty-Fourth Australasian Database Conference-Volume 137*, pages 13–22. Australian Computer Society, Inc., 2013.
- [22] B.-K. Yi, H. Jagadish, and C. Faloutsos. Efficient retrieval of similar time sequences under time warping. In *Proceedings of the 14th International Conference on Data Engineering*, pages 201–208. IEEE, 1998.



The SIGSPATIAL Special

Section 2: Tools and Datasets

ACM SIGSPATIAL
<http://www.sigspatial.org>

UCR-STAR: The UCR Spatio-temporal Active Repository

<https://star.cs.ucr.edu/>

Saheli Ghosh Tin Vu Mehrad Amin Eskandari Ahmed Eldawy

Department of Computer Science and Engineering, University of California, Riverside, USA
{sghos006,tvu032,mamin021,eldawy}@ucr.edu

Abstract

This article describes the UCR Spatio-temporal Active Repository (UCR-STAR). UCR-STAR is a visual catalog for big spatial datasets. Rather than a boring tabular listing of datasets, it provides an interactive map interface that allows users to explore these datasets to assess their coverage, quality, and distribution. This article describes both the functionality of UCR-STAR as well as the underlying system architecture. We believe that this article can help the research community by explaining how to realize research ideas into a workable product.

1 Introduction

Recently, there has been a tremendous growth in spatial data collection from various sources such as satellites, IoT sensors, smartphones, autonomous cars, and others. At the same time, there is a move for open data led by governments, non-profit organizations, and industry which makes hundreds of thousands of datasets publicly available. For example, Data.gov [4], which is maintained by the US government, hosts more than 140,000 datasets that are tagged as *geospatial*. Similarly, other governments, non-government organizations (NGOs), and companies keep releasing data on the web as part of the open data movement [3, 5, 13, 6, 11]. To browse these datasets, existing data repositories provide a plain listing of the datasets with references on how to access and download them, e.g., see Figure 1(a). Users will either have to *guess* what the dataset really contains or download these datasets, figure out how to import them into their favorite tool, before they can interact with the data.

To break from this old interface, this article describes UCR-STAR which provides an interactive web-based interface that allows users to interact with the datasets to assess their coverage, quality, and distribution before even downloading them. Figure 1(b) shows a screenshot of how UCR-STAR looks like. The majority of the screen is devoted to the map-based visualization of the dataset while a smaller part is used to list the datasets and its details. The map portion provides the standard map interactions such as pan and zoom. When a dataset is selected on the left, the map is updated in a fraction of a second to visualize the selected dataset no matter how big it is. Users can also search for datasets using a keyword search or advanced search based on the size or type of the dataset.

UCR-STAR is currently hosting more than 100 datasets with a total size of nearly one terabyte. We welcome requests to add additional datasets to the archive to further facilitate the access to these datasets and help the research community. The rest of this article gives more details about the architecture of UCR-STAR and the research behind it.

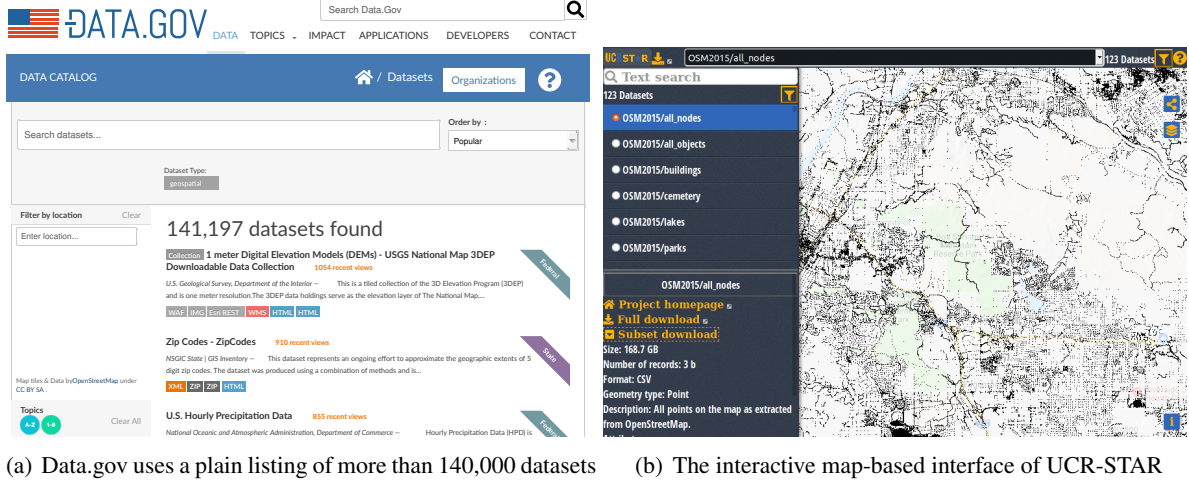


Figure 1: A contrast between the plain static interface of Data.gov and the interactive interface of UCR-STAR

2 UCR-STAR Architecture

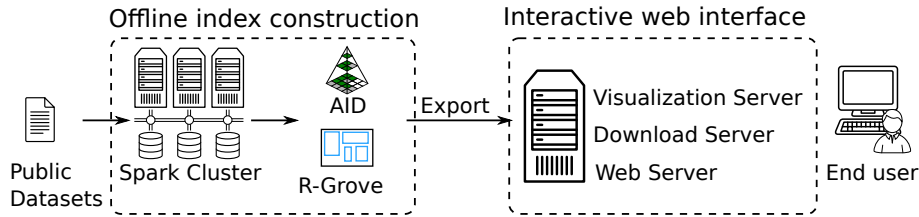


Figure 2: Overview of the system architecture of UCR-STAR

Figure 2 gives an architectural overview of UCR-STAR. The system consists of two major components, an offline index construction component and an interactive web-based interface. The offline index construction runs on a Spark cluster. It takes a publicly available dataset as input and produces two indexes, R-Grove and AID. The R-Grove index [12] is a spatial index for big spatial data that allows efficient retrieval of any rectangular query range. The AID [8] index is a light-weight adaptive visualization index that enables multilevel visualization of very large datasets. These two indexes are built on an in-house Spark cluster to scale to very large datasets. For example, a 100 GB dataset usually takes less than an hour to construct both indexes on a 12-node Spark cluster.

After the two indexes are constructed on Spark, they are then exported from the cluster into a single machine that hosts the second interactive web server. That single machine hosts three logical servers, a visualization server, a download server, and an Apache web server. The visualization server uses both indexes, AID and R-Grove, to serve *image tiles*; these are small images that are combined together to provide the map visualization. The download server provides the *subset download* features which allow users to download a part of the dataset in any format. The web server handles all requests to static files, e.g., HTML pages and images, and forwards tile and download requests to the corresponding servers. The next sections provide more details on how these two components work.

3 Offline Index Construction

The goal of the offline index construction process is to construct the R-Grove spatial index and the AID visualization index. Both are constructed on a Spark cluster to scale to big datasets.

3.1 Spatial Index Construction (R-Grove)

To be able to quickly access the data for very large datasets, we build a spatial index which is used by both the visualization and the download servers. R-Grove[12] is a novel partitioning and indexing method for big spatial data. R-Grove is an adaptation of the R-Tree family to big data. It inherits the quality index characteristics of R-Trees while balancing the load for distributed query processing. R-Grove is mainly a partitioning method that partitions the input dataset into 128 MB blocks. This is followed by constructing traditional local indexes for each partition. We use the RR*-tree [2] index for the local indexing step. The constructed index is stored as separate file, one for each partition, in the distributed file system accompanied with a small *master file* that stores the metadata of the partitions, e.g., the minimum bounding rectangle and the size of each one. Figure 3 shows an example of how the R-Grove global index partitions a 7.1 GB roads dataset into 60 partitions.

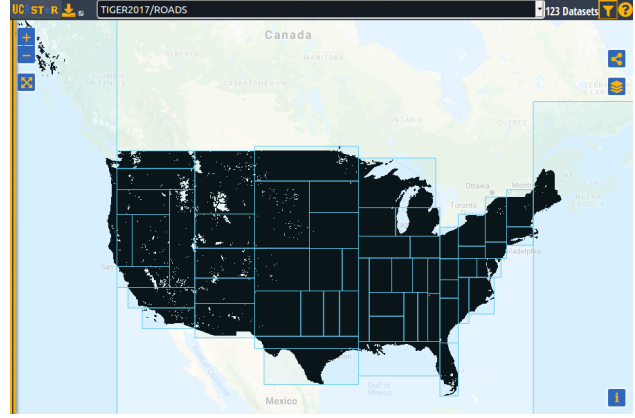


Figure 3: R-Grove index plot of a 7.1 GB roads dataset

3.2 Visualization Index Construction (AID)

To provide an interactive map visualization for the dataset, we use the standard tile-based pyramid structures illustrated in Figure 4. This structure is used by most web maps where the top tile covers the entire world and each deeper level multiplies the number of tiles by four. A standard pyramid of 20 levels will contain up to 366 billion tiles which is impractical to construct and maintain for hundreds of datasets. However, we make an observation that not all tiles are equal from the query processing perspective. For example, a tile representing the heavily populated New York City will have way more records than a tile representing the scarcely populated Palms Spring. In this case, it makes more sense to materialize the first tile as an image tile while the second one can be generated on demand.

In UCR-STAR, we use the Adaptive-Image-Data (AID) index [8] which builds on this observation to classify the tiles into four classes based on a threshold parameter (θ). The threshold θ represents the largest size that can be visualized on demand. Any tile that contains more θ records needs to be pre-generated and materialized to disk. Based on this threshold, AID defines the following four classes of tiles.

1. **Image tile** covers a large amount of data ($> \theta$) and is very expensive to visualize. This tile is pre-generated and materialized as an image.
2. **Data tile** has a small amount of data ($\leq \theta$) while its parent is an image tile. This tile can be retrieved and processed on-the-fly from the corresponding R-Grove index. If no spatial index is available, the records contained in this tile are stored in a *data file*, e.g., a CSV or GeoJSON file.
3. **Shallow tile** covers a small amount of data ($\leq \theta$) and its parent tile is not an image tile. This means that it is covered by an existing data tile and can also be generated on-demand.
4. **Empty tile** does not contain any records (zero records) and does not need to be stored. A prime example is a water area in a road network dataset.

The AID index that we use only contains the image tiles while all other tiles are generated upon request using the corresponding R-Grove index. The index construction process starts by building a histogram of the input data which is used to classify the tiles based on their sizes as explained above. After that, the image construction process uses the histogram to locate and generate only the image tiles.

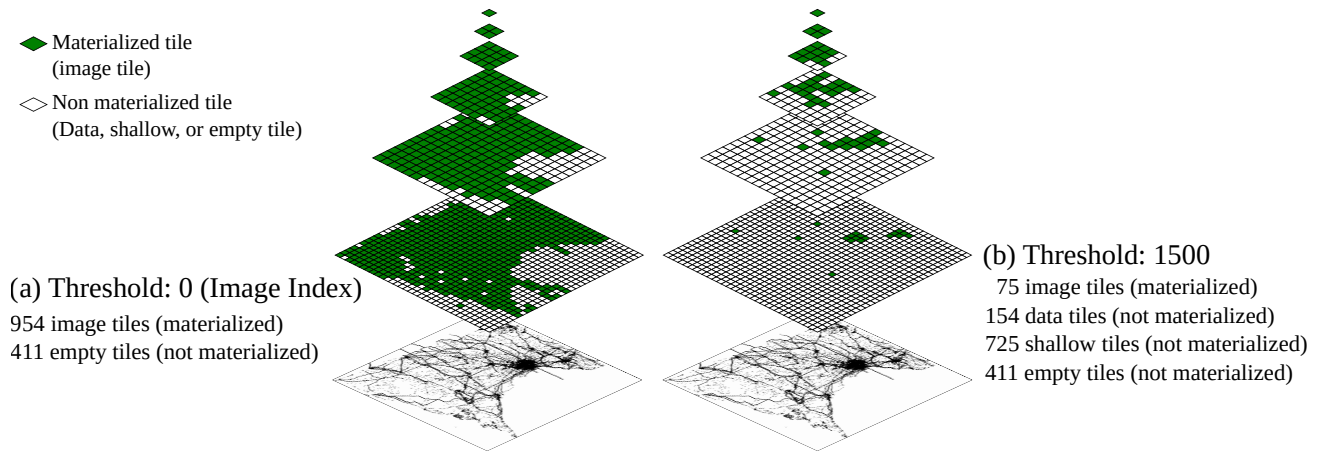


Figure 4: Examples of adaptive indexing with different threshold values

Figure 4 provides an example of the AID index, where (a) has a threshold 0, which means that all non-empty tiles are pregenerated and materialized as images, whereas in (b), where the threshold is 1500, only tiles having more than 1500 records are computed as image tiles while the rest are either data or shallow tiles and are not materialized. This not only reduces index size but also reduces the index construction time.

4 Interactive Web Interface

After the two indexes, R-Grove and AID, are constructed, they are exported from the distributed file system to a single machine that serves as the back-end server. This single physical machine hosts three server processes, visualization server, download server, and web server, described below.

4.1 Visualization Server

The visualization server serves a single query called GETTILE. This query requests a tile identified by a dataset ID, and a tile ID (z, x, y) , where the dataset ID identifies which dataset is visualized and the tile ID represents the zoom level (z) and tile location in that level (x, y) . The result is always a PNG image of size 256×256 pixels that the browser displays. As the user interacts with the maps, the browser sends a series of GETTILE queries to retrieve all the tiles that are visible on the screen and stitches them together to display the visible part of the map. The main challenge in this query is that not all the tiles are readily available as a PNG image on the server. Rather, some of them need to be generated on demand upon user request and they have to be served within 500 milliseconds to maintain the interactivity.

To answer the GETTILE query, the visualization server locates the AID index that corresponds to the dataset ID. First, it searches for a pre-generated PNG image by locating the file `'tile-z-x-y.png'`. If the file is found, then it is returned immediately. If no such file is available, then a tile needs to be generated on-the-fly. In this case, the corresponding R-Grove index is located. Then, through simple math, the minimum bounding rectangle (MBR) of the tile is calculated and a simple range query is executed on the R-Grove index. The resulting records are then visualized by simply scanning them. AID ensures that when a tile has to be generated on-demand, it is always small enough to be generated in less than 500 milliseconds.

To further improve the performance, the server caches all generated tiles in case they are requested again by other users. The server is configured to hold up to 100,000 tiles in the cache in PNG format which consumes an average of 100 MB of memory. Since its release, UCR-STAR served more than 820,000 tile requests with 27%, 48%, and 25%, static, on-demand, and cached tiles, respectively

4.2 Download Server

The download server empowers the users by allowing them to download a subset of the data based on an arbitrary rectangular range. While this feature is not new, UCR-STAR handles it in a unique way that was not done before. In short, it provides immediate download to any range of arbitrary size. To understand how this is different, we compare UCR-STAR to existing web sites that provide similar functionality.

- **GeoFabrick [7]** provides extracts from OpenStreetMap for prespecified ranges, e.g., countries or states. On the positive side, it provides immediate download to the files, but on the negative side it does not allow users to select arbitrary ranges and there is an overhead on the server on keeping all these files.
- **TAREEG [1]** follows a different approach that allows users to choose any arbitrary range and then it extracts the corresponding data and makes it available for download. However, it has a huge drawback in that it does not provide immediate access to the downloaded files. It asks the user to enter an email address and it sends the download link to that email after the file is ready. The extracted files are kept on the server for a month before they get deleted.
- **OpenStreetMap [10]** allows users to export an arbitrary rectangular range as an OSM file with immediate download. However, it only works for extremely small ranges that can be served on-the-fly. All requests for large regions, e.g., an entire city, are rejected.

UCR-STAR provides the best of all these options. It allows the user to choose any arbitrary range and the download starts immediately no matter how big the file is; even for hundreds of gigabytes of data. The trick is to utilize the HTTP *chunking* feature which allows the response to be sent in chunks rather than in one piece. When the user request is received, the download server searches the R-Grove index to locate the partitions that match the query. As soon as the first partition is found, the server processes it and starts sending the data back to the client which makes the download start immediately. As the user downloads the data, the server keeps reading the next partitions and sends the data until the entire request is satisfied. This approach has many advantages over the traditional approach of serving static files including:

1. The download starts immediately regardless of the size.
2. The server does not need to store any additional files.
3. If the user cancels the download, the server immediately frees up the resources and does not continue the extraction.
4. The server can handle hundreds of concurrent download requests as each request is very light on resources.
5. The server can provide many file formats for download as the files are generated on-the-fly.

The only drawback is that the browser does not know the total size of the download. This means that there is no progress bar for the download. It also means that the download has to always restart from the beginning in case of a network failure. Since the release of this feature, UCR-STAR served 189 download requests with a total size of 87 GB. Note that the requests to download the full data are redirected to third party servers and are not tracked by UCR-STAR.

4.3 Web Server

As noted above, UCR-STAR uses two Java servers to host visualization and downloads. These servers could be running or hosted on one or multiple machines with arbitrary ports. On the other hand, an Apache Web server is available on the standard HTTP (80) and HTTPS (443) ports. Therefore, the Apache web server received all

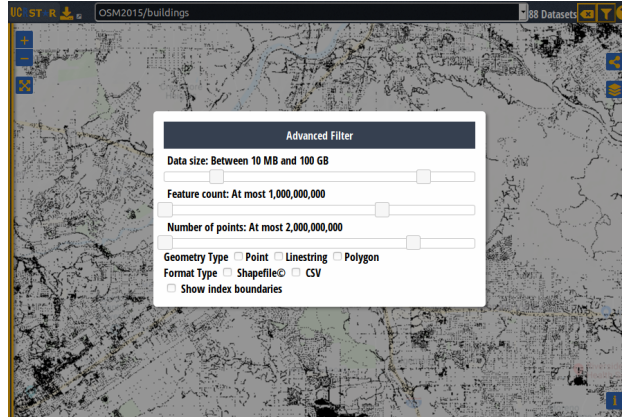


Figure 5: The search feature runs completely on the client side

visualization and download requests and it needs to forward them to the corresponding visualization or download server. In order to address this problem, we use the virtual host feature on Apache Web Server. In particular, based on the path after the domain, the request for visualizing a specific dataset is forwarded to the corresponding Java server. Similarly, a different path on the server is forwarded to the download server. This configuration allows one public domain to act as a single server for an array of services that can be hosted on single or multiple machines. This configuration also allows the web server to handle all static resources efficiently while the Java server can focus on the dynamic part of the application.

4.4 Front-end

This part briefly describes the set of HTML and JavaScript pages that are loaded into the end-user browser.

Map interface: We use OpenLayers [9] for displaying the map. It contains a base layer of either OpenStreetMap, Google Maps, or Google Satellite Imagery. On top of it, we put a tiled layer (XYZ Layer) that displays the dataset that is selected by the user. The XYZ layer is customized to send all the requests to the visualization server which hosts the tiles.

Search feature: The search functionality (Figure 5) is implemented entirely in JavaScript and is executed locally in the browser. Since the server currently hosts only a few hundred datasets, the browser can efficiently implement the search feature by scanning all these datasets for each query. This also relieves the server from addressing these additional search requests. A nice advantage to the users is that it adds some level of privacy as the server does not track the datasets that users search for.

Index display: One of the interesting features in UCR-STAR is the ability to show the underlying indexes on the datasets in a short amount of time regardless of the size of the dataset (Figure 3). To accomplish this, the front-end requests the *master file* of the selected index in GeoJSON format. The retrieved GeoJSON file is then used as a source for a vector layer that is added on top of the base and data layers. This layer is typically very small as it has one entry for each 128 MB partition. OpenLayers can efficiently handle a few thousand features in GeoJSON which is enough for all the datasets that are currently served by UCR-STAR. Since the client has the entire GeoJSON file, we allow the user to interact with the displayed index by clicking the partitions to show more details about them, e.g., total size and number of features in the partition.

Embed visualizations: Another interesting feature of UCR-STAR is the ability of users to *embed* a dataset visualization on another website without the need to install any additional software. Simply, UCR-STAR provides an HTML code snippet that users can copy/paste into their websites to get a fully interactive map display of a specific dataset similar to the one on the UCR-STAR website. This is a great option to share the datasets and provide richer engagement in the community. To implement the embed feature, UCR-STAR hosts a special JavaScript file that is designed to automatically initialize a map with the configured location and dataset visualization on top of it. The JavaScript file itself is static, but the HTML code snippet contains information about the geographical location, the zoom level to initialize the map, and the ID of the dataset to display on top of it. Once the map is loaded, OpenLayers handles all map interactions as usual.

Acknowledgments

This work is supported in part by the National Science Foundation (NSF) under grants IIS-1838222 and CNS-1924694 and by the USDA National Institute of Food and Agriculture, AFRI award number A1521.

References

- [1] L. Alarabi, A. Eldawy, R. Alghamdi, and M. F. Mokbel. TAREEG: a mapreduce-based system for extracting spatial data from openstreetmap. In *Proceedings of the ACM Symposium on Advances in Geographic Information Systems, ACM SIGSPATIAL*, pages 83–92, Dallas/Fort Worth, TX, Nov. 2014.
- [2] N. Beckmann and B. Seeger. A Revised R*-tree in Comparison with Related Index Structures. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD*, pages 799–812, Providence, RI, July 2009.
- [3] Big Ten Academic Alliance Geoportal. <https://geo.btaa.org/>.
- [4] Data.gov. The home of the U.S. Governments open data. <https://www.data.gov/>.
- [5] Data.gov.uk. <https://www.data.gov.uk/>.
- [6] Los Angeles - Open Data Portal, 2019. <https://data.lacity.org/>.
- [7] GeoFabrik Homepage, 2019. <https://www.geofabrik.de/>.
- [8] S. Ghosh, A. Eldawy, and S. Jais. AID: An Adaptive Image Data Index for Interactive Multilevel Visualization. In *Proceedings of the IEEE International Conference on Data Engineering, IEEE ICDE*, Macau, China, Apr. 2019.
- [9] OpenLayers API: A high-performance, feature-packed library for all your mapping needs, 2019. <https://openlayers.org/>.
- [10] OpenStreetMap, 2019. <https://www.openstreetmap.org/>.
- [11] United Nations Open Data - A World of Information, 2019. <http://data.un.org/>.
- [12] T. Vu and A. Eldawy. R-Grove: Growing a Family of R-trees in the Big-data Forest. In *Proceedings of the ACM Symposium on Advances in Geographic Information Systems, ACM SIGSPATIAL*, pages 532–535, Seattle, WA, Nov. 2018.
- [13] Yahoo! Webscope Datasets, 2018. <https://webscope.sandbox.yahoo.com/>.

join today!

SIGSPATIAL & ACM

www.sigspatial.org

www.acm.org

The **ACM Special Interest Group on Spatial Information (SIGSPATIAL)** addresses issues related to the acquisition, management, and processing of spatially-related information with a focus on algorithmic, geometric, and visual considerations. The scope includes, but is not limited to, geographic information systems (GIS).

The **Association for Computing Machinery (ACM)** is an educational and scientific computing society which works to advance computing as a science and a profession. Benefits include subscriptions to *Communications of the ACM*, *MemberNet*, *TechNews* and *CareerNews*, full and unlimited access to online courses and books, discounts on conferences and the option to subscribe to the ACM Digital Library.

- ☐ SIGSPATIAL (ACM Member) \$ 15
- ☐ SIGSPATIAL (ACM Student Member & Non-ACM Student Member) \$ 6
- ☐ SIGSPATIAL (Non-ACM Member) \$ 15
- ☐ ACM Professional Membership (\$99) & SIGSPATIAL (\$15) \$114
- ☐ ACM Professional Membership (\$99) & SIGSPATIAL (\$15) & ACM Digital Library (\$99) \$213
- ☐ ACM Student Membership (\$19) & SIGSPATIAL (\$6) \$ 25

payment information

Name _____

ACM Member # _____

Mailing Address _____

City/State/Province _____

ZIP/Postal Code/Country _____

Email _____

Mobile Phone _____

Fax _____

Credit Card Type: ☐ AMEX ☐ VISA ☐ MC

Credit Card # _____

Exp. Date _____

Signature _____

Make check or money order payable to ACM, Inc

ACM accepts U.S. dollars or equivalent in foreign currency. Prices include surface delivery charge. Expedited Air Service, which is a partial air freight delivery service, is available outside North America. Contact ACM for more information.

Mailing List Restriction

ACM occasionally makes its mailing list available to computer-related organizations, educational institutions and sister societies. All email addresses remain strictly confidential. Check one of the following if you wish to restrict the use of your name:

- ☐ ACM announcements only
- ☐ ACM and other sister society announcements
- ☐ ACM subscription and renewal notices only

Questions? Contact:

ACM Headquarters
2 Penn Plaza, Suite 701
New York, NY 10121-0701
voice: 212-626-0500
fax: 212-944-1318
email: acmhelp@acm.org

Remit to:

ACM
General Post Office
P.O. Box 30777
New York, NY 10087-0777

SIGAPP



Association for
Computing Machinery

www.acm.org/joinsigs

Advancing Computing as a Science & Profession



The SIGSPATIAL Special

ACM SIGSPATIAL
<http://www.sigspatial.org>