# PHD Research Proposal

Title:

# High-Dimensional Spatiotemporal Trajectories Analytics in Smart Sustainable Cities

Applicant: Ruidong Xue

Domestic Supervisor: Professor Jiajin Le

Domestic Institution: School of Computer Science and Technology,
Donghua University

Potential Supervisor: Dr Weiren (William) Yu

Potential Institution: School of Engineering and Applied Science,
Aston University

# Abstract

The most timely and straightforward detection method for monitoring the state of city systems is to deploy large scale sensors. Various data collected from sensors often contain potential knowledge. The knowledge can help engineers and administrators to make early warnings and better decisions. With the increasing scale and complexity of cyber physical systems, sensor data streams pose significant challenges in data streams analysis in real time. Interesting patterns and anomalous phenomena need be effectively characterised and analysed using an automated and robust approach. Motivated by this, my research aims to devise novel techniques that can efficiently manage and analyze high-dimensional spatiotemporal trajectories generated from a diverse set of sensors to enable future cities to be more intelligent and sustainable.

**Key words:** sensors, smart city, data mining, spatiotemporal trajectories

## Contents

# 1 Introduction

Cities are our future. One salient feature of city living in modern time is the fragmentation and changing community from communities of locality to communities of interest [1]. With the advance of micro-electronics technology, dozens of sensors have been installed in urban space. These sensors are used to record the state of the whole system. The timely feedback from sensors to the status of subsystems facilitates fast maintenance and repair of the whole system, and improves the service life of city systems.

With the rapid development of Information and Communication Technology (ICT), modern cities will be the epicenter for various Internet of Things (IoT) applications. Latency, real-time response, and micro batch updating are the main characteristics of data stream processing. Therefore, tradition methods for static data processing are not appropriate to tackle data streams generated from modern cities.

Furthermore, smart cities need to timely find fault, understand the reasons, and thus can reduce costs and improve efficiency. The structure and functions of the city system are usually very complex and involve a very large number of subsystems, which requires large-scale sensor deployment. Multiple sensors record different indicators of city system. If these indicators can be integrated together for management and analysis, we will be able to get more potential knowledge from the city.

My research will focus on the management of large scale spatiotemporal trajectory data obtained from sensors, using the relevant knowledge of interdisciplinary, methods from computer science, and applied mathematics.

# 2 Literature Review

In recent years, many researchers from data mining areas focus on managing long high-dimensional spatiotemporal trajectories generated from a diverse set of sensors. In general, sensor data processing includes data collection, cleaning, data management, knowledge discovery and mining. I will summarize the existing main research problems which arise in the context of sensor data processing as follows:

### 2.1 Data Collection and Cleaning

The data obtained from sensors is often erroneous. One method for data cleaning is to mark a raw sensor value as an outlier if the raw sensor value deviates significantly from the inferred sensor value. Another important approach for data cleaning is known as declarative data cleaning [2, 3, 4]. In this approach, the user registers SQL-like queries that define constraints over the sensor values. Sensor values are marked as outliers when these constraints are violated. In addition to these methods, I also learn many other data cleaning approaches which are relevant to this project. For example, Extensible Sensor stream Processing (ESP) can successfully alleviate both missed and unreliable readings in sensor data [5]. Compared to the simple moving average (SMA), the proposed weighted moving average (WMA) algorithm can effectively clean data and offer quick response time [6].

### 2.2 Data Compression

The large volumes of collected data pose significant challenges for collected data. Sometimes, the volume of the data is so large that it may be impractical to store the entire raw data, and it may be desirable to either compress or drop portions of the data. The application of baseline-dependent window functions can reduce both the overall level of far sidelobe confusion noise, and lessens the impact of A-team sources in sidelobes [7].

### 2.3 Data Denoising

Most algorithms have not yet attained a desirable level of applicability. All show an outstanding performance when the image model corresponds to the algorithm assumptions, but fail in general and create artifacts or remove image fine structures [8]. The denoising performance of all considered methods are compared in this project and I will propose an algorithm which will be optimal under a generic statistical image model.

### 2.4 One-pass Streaming Algorithm

In computing, a one-pass algorithm reads its input exactly once, in order, without unbounded buffering. A one-pass algorithm generally requires O(n) time and less than O(n) storage (typically O(1)), where n is the size of the input. There are two algorithms use analytical rules to adjust the model extracted from the stream instead of recomputing the entire solution on the augmented datasets. We show that these algorithms are more accurate than the current extension of CVMs [9].

### 2.5 In-network Processing

In-network query processing refers to the complete or partial evaluation of database queries at the edges of a network, rather than in a centralized database server [10]. The information-driven sensor collaboration has several advantages: tracking is distributed, and the network is energy-efficient, activated only on a when-needed basis [11].

### 2.6 Frequent Pattern Mining

In the case of data streams, one may wish to find frequent itemsets either over a sliding window or the entire data stream [12, 13]. In the case of data streams, the problem of frequent pattern mining can be studied under several models: Entire Data Stream Model and Damped Window Model. The first method Sketches are often used to determine heavy-hitters in data streams, and therefore, an extension of the methodology to the problem of finding frequent patterns is natural. The advantage of the second model is when a new transaction arrives, the overall effect of such an approach is to create an exponential decay function on the arrivals in the data stream.

## 3 Research Objectives and Potential Challenges

The purpose of my PHD research is to propose scalable and cost-effective ways which will be able to enhance urban infrastructures and ecosystems, thereby enabling future cities to be better connected and more sustainable.

Because the large scale diverse sensors record data every second or minute, and this project needs algorithms that use a small amount of time and memory resources, and that are able to adapt to changes and not to stop learning. How to deal with data stream accurately in real time is the main challenge for the projects in the near future. In this project, data arrives at high speed, and algorithms process it with very strict constraints of space and time. Furthermore, these algorithms should be distributed.

According to my master final project in Mid Sweden University(self-pay), I will apply my improved presentation and discretization method to deal with large scale and high dimension raw data set to get normalized data set. Furthermore, interdisciplinary and combining methods will be applied to propose innovative analysis techniques.

Finally, based on my initial analysis model, I will propose efficient and innovative analysis techniques strategies to support managing and analyzing high-dimensional spatiotemporal trajectories which from large scale diverse sensors.

## 4 Feasibility to Meet Expected Goals

Given my strong data mining and analytics background in both Donghua

University and Mid Sweden University (self-pay), I will make significant scientific contributions on both theoretical break through which can overcome current challenges and establish effective high-dimensional spatiotemporal trajectories management system, as well as the challenging technical problems involved in my PhD studies. Computer science plays a considerable role in our daily life. Hence, this PHD project not only has strong associations with my master period research field, but also is closely related to my previous publication.

Moreover, I have double master degree from Mid Sweden University (self-pay) and Donghua University. During my master study, I have been actively involved in my research projects that are related to data mining algorithms and my publication, such as a data mining project of Bosch Company and a data mining project of hospital.

In addition, I major in computer science and technology in both undergraduate and postgraduate study. Thus, I am good at database systems, data mining, machine learning, and algorithm design and implementation. Moreover, I have proficient programming experience, such as C++, Python, Java, and SQL.

Finally, based on my published paper *Sensor time series association rule discovery based on modified discretization method* [14] which focus on the large scale diverse sensors, I will be able to continue my research on high-dimensional spatiotemporal trajectories analytics in smart sustainable cities.

## 5 Methodology & Tasks

My research approaches are interdisciplinary, combining methods from computer science, and applied mathematics to improve how scalable spatiotemporal trajectories are managed for smart sustainable cities maintenance to enhance citizen well-being.

My methods are broken down to the following four tasks. For each task, the research methods are summarized as follows:

**Task 1) Data Collection and Cleaning.** a) Take good advantage of abundant UK research resources in the field of Big Data to obtain real-world datasets. b) Propose efficient and automatic methods to collect and filtrate spatiotemporal trajectories data. c) Apply discretization method to obtain normalized data sets.

**Task 2) Scalable High-dimensional Spatiotemporal Trajectories Management.** a) Exploit new technologies to spot motifs (repeated patterns) in long high-dimensional spatiotemporal trajectories. b) Define an intuitive DTW (Dynamic Time Warping) like distance to assess the similarity between high-dimensional spatiotemporal trajectories. c) Propose optimization techniques to substantially accelerate the measure of similarity between two spatiotemporal trajectories that may vary in dimension and length.

**Task 3) Efficient Algorithm Design for Frequent Pattern Mining.** a) Design one-pass streaming algorithm to reduce the time complexity. b) Apply efficient in-network processing methods to for distributed and efficient data query. c) Propose new frequent pattern mining method to reduce the search cost and improve algorithm.

**Task 4) Experimental Evaluation.** Conduct performance evaluation to evaluate the performance of my proposed techniques. The efficiency will be evaluated in terms of CPU overhead, I/O cost, memory usage, scalability, and index size.

**Significance**. My proposed research will bring innovative analysis techniques (such as clustering, matching, filtering and visualization) to high-dimensional spatiotemporal trajectories management, including motif discovery, data characteristics understanding, hypotheses validation, and private data publication. State failure detection will provide early warning to our future city.

## 6 Time Schedule

Based on my research tasks in Section 5, I illustrate the time schedule of my four-year PHD research in Table 1, and visualize my timeline in Figure 1 via Gantt chart.

**Table1 Time Schedule**

| Tasks | Period |
|---|---|
| T1. Data Collection and Normalization | 01/09/2017-01/03/2018 |
| T2. Scalable High-dimensional Spatiotemporal Trajectories Management | 01/11/2017-01/12/2020 |
| T3. Efficient Algorithm Design for Frequent Pattern Mining | 01/06/2018-01/09/2020 |
| T4. Experimental Evaluation | 01/12/2018-01/09/2021 |

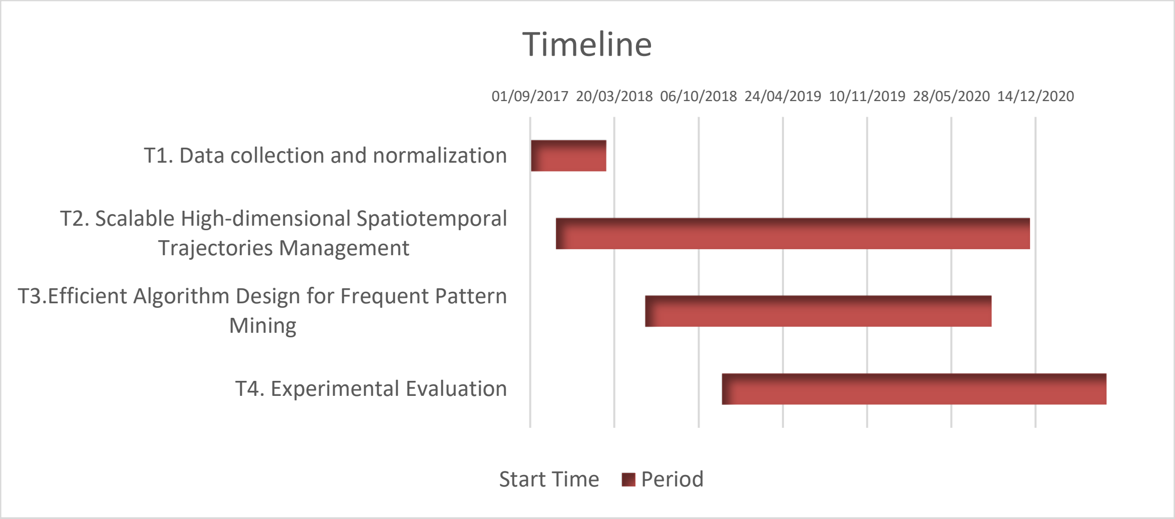

**Figure 1: My Research Schedule visualized in Gantt chart**

# References

[1] M Höjer, J Wangel. Smart sustainable cities: definition and challenges [M]. ICT Innovations for Sustainability. Springer International Publishing, 2015: 333-349.

[2] S R Jeffery, M Garofalakis, M J Franklin. Adaptive cleaning for RFID data streams. The 32nd International Conference on Very Large Data Bases. VLDB Endowment, 2006: 163-174.

[3] C Mayfield, J Neville, S Prabhakar. ERACER: a database approach for statistical inference and data cleaning. The ACM SIGMOD International Conference on Management of Data. 2010: 75-86.

[4] J Rao, S Doraiswamy, H Thakkar, et al. A deferred cleansing method for RFID data analytics. The 32nd International Conference on Very Large Data Bases. VLDB Endowment, 2006: 175-186.

[5] S R Jeffery, G Alonso, M J Franklin, et al. Declarative support for sensor data cleaning. The International Conference on Pervasive Computing. Springer Berlin Heidelberg, 2006: 83-100.

[6] Y Zhuang, L Chen, X S Wang, et al. A weighted moving average-based approach for cleaning sensor data. The 27th International Conference on Distributed Computing Systems, ICDCS, 2007: 38-38.

[7] M T Atemkeng, O M Smirnov, C Tasse, et al. Using baseline-dependent window functions for data compression and field-of-interest shaping in radio interferometry[J]. Monthly Notices of the Royal Astronomical Society, 2016, 462.

[8] A Buades, B Coll, J M Morel. A Review of Image Denoising Algorithms, with a New One[J]. Siam Journal on Multiscale Modeling & Simulation, 2005, 4(2):490--530.

[9] R Anculef, H Allende, et al. Two one-pass algorithms for data stream classification using approximate MEBs. The International Conference on Adaptive and

Natural Computing Algorithms. Springer-Verlag, 2011:363-372.

[10] S Madden. In-Network Query Processing [M]. Springer US, 2009.

[11] J Liu, J Reich, F Zhao. Collaborative In-Network Processing for Target Tracking [J]. EURASIP Journal on Advances in Signal Processing, 2003, 2003(4):1-14.

[12] C Giannella, J Han, J Pei, et al. Mining frequent patterns in data streams at multiple time granularities [J]. Next generation data mining, 2003, 212: 191-212.

[13] R Jin, G Agrawal. An algorithm for in-core frequent itemset mining on streaming data. The International Conference on Data Mining (ICDM). 2005.

[14] R Xue, T Zhang, D Chen, et al. Sensor time series association rule discovery based on modified discretization method. The 2016 IEEE International Conference on Computer Communication and the Internet (ICCCI),.2016: 196-202.