Research Statement

My research passion lies in discovering real-world problems with authentic data, and resolving these problems by creating solid solutions that transform science and daily life. Over the past seven years, I have broadly investigated urban Big-Data-driven Cyber-Physical Systems, a new information paradigm integrating communication, computation and control, with a special focus on data-intensive aspects in urban infrastructures. This statement summarizes my research impact, vision, contributions, and future agenda.

1 Research Impact

The impact of my research is substantiated by awards, services, applications, course materials, and grants.

Based on my research impact, I was honored with the Chinese Government Award for Outstanding Students as one of only four Computer Science students selected from 250,000 Chinese students in the United States. Further, my dissertation was selected for several awards including Doctoral Dissertation Fellowship Award and Excellent Thesis Travel Award. My master's thesis was also selected for the only Excellent Thesis Award in Computer Science from 46,000 graduate students in Heilongjiang Province of China.

The impact of my research has led to high visibility within the community. I was invited to give several talks at top universities and research institutes, e.g., MIT, New York University, Boston University, Northeastern University, University at Buffalo, Microsoft and IBM Research. I was also invited as a panelist for several panels including an awarding ceremony hosted by the Consulate-General of the P.R. China at University of Wisconsin-Madison. Further, I served as a reviewer for several top journals and conferences including IEEE IoT, TPDS, TMC, INFOCOM and ICDCS.

My research also provides several social applications. It has reduced development costs for real-world applications, which led the company PChomes to potentially commercialize one of my systems, a carpooling service, in the Dallas/Fort Worth area, and also led the Shenzhen Institutes of Advanced Technology, working with the Shenzhen government to use my models for a pilot program to predict bus arrivals under a platform used by 100,000 residents every day. As a part of my research group, I am also working with the Department of Transportation in Washington D.C. to optimize their BikeShare Network under a NSF funded CPS project. Finally, my results led to the first released 20GB heterogeneous urban datasets and are used for urban research by many groups from different disciplines such as Civil Engineering and Urban Planning.

My research impact has also been recognized through course materials. My results have been adopted as graduate-level course materials in several universities such as the University of Utah (CS7943: Networking Seminar, Spring 2015), the Stony Brook University (ESE670: Topics in Electrical Sciences, Fall 2014), and the University at Buffalo (CSE726: Selected Topics in CPS, Fall 2014). They have also been translated and taught overseas such as at Shanghai JiaoTong University and National Cheng Kung University.

Especially in these times of financial turmoil, it is critical for a future faculty member to be able to secure funding based on excellent proposals. Even as a graduate student, I have noticeable experience with grant proposals. I substantially assisted my advisor in creating two proposals to the NSF CSR program and NSF CPS program recently funded for $500,000. All of which are primarily based on my research results. More importantly, I have been actively seeking research funding with proposals written by myself. I was awarded with several grants and fellowships, and the total amount is more than $100,000, which is a major support for a student. These grants prove my ability to secure future funding for my own research group.

The above impact indicates that my research results are favorably perceived by the community and that my research agenda is on a fruitful track.

2 Research Vision

For the first time ever, we have more people living in urban areas than rural areas. To address their emerging challenges in this inevitable urbanization, the White House announces Smart Cities Initiative with $160 million investment in September 14, 2015. Under the scope of this initiative, my work addresses emerging problems in the urban transportation sector by big-data-driven applications with a cyber-physical-systems (CPS) approach. CPS provides a seamless interaction between the cyber and physical worlds, which is the foundation for data-driven transit services such as transit dispatching.

Traditional urban transit services are typically driven by *small data* collected at small scales, from single sources, or in an offline fashion. As a result, they typically employ suboptimal and fixed components such as schedules and routes. My research first questions whether such legacy small-data-driven services are sufficient for transit today. Then, it seeks to identify and make sense of dynamic relationships between urban-scale phenomena (such as demand and supply) to improve urban efficiency with applications driven by *big data*: large-scale multi-source data collected online with high volume, variety and velocity.

In particular, my research on big-data-driven applications is built upon large-scale urban infrastructures. Compared to dedicated sensor and actuator networks, the devices in these already-deployed infrastructures can be utilized without additional costs. My technical vision is that (i) various physical devices in existing urban infrastructures collaboratively sense urban phenomena and send sensing data to peer devices or data centers in the cloud, (ii) and then peer devices or data centers compute and analyze models to describe urban phenomena, and (iii) finally these peer devices or data centers inform or control physical devices, which in turn provide feedback about urban phenomena. I consider this integration of sensing, modeling, control and feedback as the core theme for my research. This core theme motivates me to uniquely utilize the CPS approach to design urban applications with an emphasis on agile model-based control. In these urban applications, I pursue fundamental properties and optimizations such as accelerated data acquisition, model-oriented inference, and application-based control that are rigorously defined for the efficiency they bring to real-world services.

Looking at my research from another direction, it reexamines legacy transit services given the new reality of big data and CPS. Existing transit services driven by small data without CPS control are straightforward for the relatively-stable demand/supply relationship of the past. However, this relationship today experiences various spatiotemporal dynamics, and thus existing services quickly hit bottlenecks that prevent offering satisfactory performance. In contrast, with the emphasis on big-data-driven CPS, I design services based on large-scale multi-source data collected online with feedback-based control, so these services can improve urban transit for an even dynamic demand/supply relationship in the future.

In short, my focus on analyzing and designing big-data-driven applications with a CPS approach is maintained throughout the research process, and thus defines my position as a researcher.

3 Technical Contributions

My research results have led to more than 20 publications, featuring 10 first-author conference papers in very competitive venues (such as MobiCom, SenSys, IPSN, ICCPS, SIGSPATIAL, BigData, ICDCS, and RTSS with 10% to 20% acceptance ratios), three best paper or runner-up awards in ICCPS, CWSN and CyberC, one best-poster award in INFOCOM, one invited conference paper in CTS, and one invited fast-tracked journal article to Proceedings of the IEEE.

In particular, my research bridges cyber-physical systems (CPS) and Big Data in extremely-large-scale urban infrastructures (such as a 10-million-cellphone network and a 16-million-smartcard network in a metropolitan area). More importantly, my results are uniquely built upon these infrastructure data that are typically at least one or two orders of magnitudes bigger than the data from academic experimental systems, enabling

my research to discover novel results that are difficult or impossible to reveal using small data. My research correlates the three phases of data-driven designs (*i.e.*, acquisition, analyses, and application) with three components of CPS (*i.e.*, communication, computation and control) by these three themes:

3.1 Performance-Guaranteed Communication for Data Acquisition

The first theme of my research is to improve data acquisition, which is the key enabler for accurate models and responsive applications. For years, many applications have been based on straightforward yet centralized acquisition such as through cellular networks with extra delay and costs, which motivates me to explore distributed acquisition through neighbor-device-discovery services. This approach is challenging because many devices are mobile and have to autonomously turn their radios on and off or to a different frequency (*i.e.*, duty-cycling) to save energy or maintain other services, leading to extremely-limited time slots for data acquisition, *e.g.*, 1 over 100. My work [1, 2] has pioneered the research of accelerated neighbor discovery. The major effort [1] was published in SenSys'12, the flagship conference in the sensor network community.

In particular, although several discovery protocols have been proposed, they are aimed to save energy or bandwidth, leading to a long delay and thus inaccurate models. Instead of proposing another discovery protocol, my work in SenSys'12 provides a middleware Acc to seamlessly accelerate all existing protocols with guaranteed performance. The key concept I developed is indirect discovery, which uses the discovery capability of neighbor devices for data acquisition, whereas all previous work is based on direct discovery. The data acquisition hardware we designed has been installed in 98 vehicles for a field study. More importantly, I conducted the first and only evaluation of distributed data acquisition in a large-scale mobile network, a 14-thousand-taxicab network, showing that Acc-assisted protocols reduce the delay by 52% while consuming the same energy or bandwidth. This work provides valuable insights on using distributed data acquisition in urban infrastructures for real-time applications.

3.2 Model-Driven Computation for Data Analytics

Based on data acquisition, my second theme is the design of theoretical models to describe urban phenomena, *e.g.*, mobility patterns [3] [4], traffic speed [5], passenger demand [6], and transit supply [7]. Previously, the models for urban phenomena (such as mobility) were based on single-source data for a group of sampled residents (such as cellphone users) by assuming a uniform sampling. However, my research, for the first time, reveals that urban modeling based on single-source data leads to a serious yet quantifiable bias. My most favorably-perceived result on urban modeling is a mobility model mPat [3] based on heterogeneous data sources, which was published in MobiCom'14, the top conference in the mobile computing community.

Specifically, my work spearheaded the concept of considering a set of infrastructures as a heterogeneous CPS where a device is treated as a pervasive sensor generating data to model phenomena of interest. Under this concept, I proposed the first work, called mPat, to model urban mobility with big data, *i.e.*, four kinds of online data with a size of 4 TB including cellphone, taxicab, bus, and subway data, and then optimally integrate them to produce a comprehensive model by exploring spatiotemporal and contextual correlation. Thus, mPat outperforms a state-of-the-art model in accuracy by 31%. Due to the impact of this work, it has been invited to be fast tracked to Proceedings of the IEEE, which is considered the most highly-cited general interest journal in electrical engineering and computer science.

3.3 Demand-Centric Supply Control for Application Design

Based on the model computation, the third theme is focused on the unbalanced urban demand/supply relationship, the key root cause of transit-system inefficiency. It leads to wasted gas or poor passenger experiences

such as long waiting time. My work covers a complete set of the demand/supply relationship under three scenarios: nonsupply [8], oversupply [9], and undersupply [10]. Essentially, my applications control the supply based on the corresponding demand with a CPS approach taking into feedback of the control. A major thread under this theme published in SenSys'13 is a carpooling application called coRide [10], which delivers passengers by sharing taxicabs to increase taxicab supply yet with the same number of taxicabs.

Although taxicab carpooling exists in real world, it is typically in an *ad hoc* fashion. In contrast, I investigate benefits, opportunities, incentive, and routing mechanisms for the on-demand carpooling service coRide. To encourage coRide's adoption, I propose a win-win fare model as incentive for both passengers and drivers to participate. The evaluation results based on 14,000 taxicabs show that compared with the ground truth, coRide reduces 33% of total mileage used; the win-win fare model lowers passenger fares by 49% and simultaneously increases driver profits by 76%. This work is well accepted by the community, and led a company called PChomes to commercialize this system in the Dallas/Fort Worth area of Texas.

4 Future Research Agenda

Strategically, to increase my research impact, my long-term goal is threefold: (i) addressing the most fundamental technical challenges in data-driven CPS, (ii) consolidating more diverse data for data-driven research for the community, and (iii) building large-scale platforms for system implementation and evaluation.

4.1 Addressing Technical Challenges

In short, the fundamental challenge I want to address for CPS is to design application-specific control algorithms based on heterogeneous models from uncertain data. (i) Uncertain Data: due to loose-control sensing in heterogeneous systems, the physical data we have are uncertain, i.e., noisy, sparse, implicit, untimely, and inconsistent; (ii) Heterogeneous Models: based on these uncertain data from different systems, the models we have are heterogeneous in terms of scale, timeliness, granularity, and completeness, especially when we consider the multi-source data-driven models; (iii) Application-specific control algorithms: based on these models, we need some application-specific designs to control physical components of systems to improve their efficiency, which typically requires domain knowledge along with real-world data and platforms for design and evaluation. I use three concrete directions to introduce my future work to address these challenges.

Uncertain-data Inference. Under the traditional data acquisition paradigm such as using a dedicated sensor network, we can explicitly control and configure data acquisition schemes to obtain data with certainty. However, because urban infrastructures are mainly used to provide services instead of collecting data, we have several temporal, spatial, and contextual constraints for data acquisition, which result in data uncertainty issues such as incomplete, skewed, noisy, and implicit data. I plan to investigate how to improve these data with tensor decomposition on the technical level, such as missing-data reconstruction based on spatiotemporally and contextually correlated data, as well as on the policy level, such as providing incentive mechanisms for tightly-controlled crowdsourcing. In particular, we are working with our collaborators in Shenzhen to improve the Shenzhen bus arrival prediction service. The bus system in Shenzhen has more than 14,000 buses, and 10,000 stations with more than a 10 million ridership. The goal is to accurately infer bus arrival time by real-time bus GPS. A preliminary version of the service has been deployed, and has more than 100,000 users. But the key challenge for this service is that the bus GPS data are missing all the time due to communication. To address this challenge, we aim to propose a solution to integrate multi-source data from nearby taxis, smartcards and user app data to increase the accuracy of this service. The technical approach is to use context-aware tensor decomposition to recover missing bus GPS data by minimizing recovering errors.

Heterogeneous-Model Integration. Though multi-source urban data, such as cellphone and taxicab data,

are heterogenous in nature, the models driven by these data are spatiotemporally confined in the same urban area under contextual correlations. The legacy data mining and modeling schemes have been well explored for single-source data-driven or multi-source homogenous-data-driven models, but the similar schemes for correlated multi-source heterogeneous data-driven models are missing. My recent work coMobile [4] provides the first attempt to integrate four models driven by multi-source data, and I plan to advance this direction by deeply integrating these data-driven models to compensate for individual limitations related to scale, timeliness, and granularity. Specifically, we are working with our collaborators in Shenzhen to model real-time energy consumption at road segment level. It is challenging because of the fine spatiotemporal granularity from multiple heterogeneous data sources. Based on our datasets in Shenzhen, we can model electricity and gas consumption of commercial vehicles, but the key challenge is about private vehicles because there is no urban infrastructure that can capture private vehicles at urban scale in real time. We aim to use existing infrastructures, e.g., cellphone, OBD devices, cameras and loop sensors, to infer traffic volumes and speeds, thus inferring the energy consumption. The technical approach is to preform model integration based on semi-supervised multi-view learning to minimize disagreement between heterogeneous models.

Application-specific Control Algorithms. Finally, the last direction is about how to control system components based on real-world applications. In particular, I will focus on supply adjustment in urban transportation systems. The existing robust control algorithms for transportation are mostly based on uncertain demand models developed with historical data. In contrast, I plan to design control algorithms based on both historical data-driven models and real-time system dynamics to rebalance the relationship between demand and supply. In particular, we are working with our collaborators in Department of Transportation in Washington D.C. to design an efficient bike rebalancing algorithm for the D.C. Bike System. Currently the system operators need to use trucks to move bikes heuristically between stations to ensure sufficient bike supply. But the key challenge is the dynamic real-time passenger demand in the D.C. bike system. We aim to propose a uncertain demand and supply model based on real-time data with correlated contexts, e.g., weather, train schedules, and other events. The technical approach is to preform robust model predictive control based on real-time uncertain demand models to balance bike supply among different stations.

4.2 Consolidating Diverse Data

The second agenda is uniquely based on an urban-data repository I have been vertically and horizontally managing and studying for five years. Vertically, I manage and study heterogenous data for a metropolitan area with 14 million population by working with Shenzhen Institutes of Advanced Technology (SIAT), including data from a 42-thousand-vehicle network, a 10-million-user cellphone network, and a 16-million-smartcard network. Horizontally, I mange and study the data from six taxicab networks across three continents including New York City, San Francisco, Rome, Beijing, Shanghai, and Shenzhen. In the future, I plan to further negotiate with service providers based on my research results to obtain more data for research. During the consolidation, we aim to address two technical challenges, i.e., data management and data privacy.

It is crucial yet challenging to effectively manage diverse urban data for the better performance of models and services. I have been performing straightforward yet separate management for all these heterogenous datasets given that the related work for heterogenous urban-data management is extremely limited. But the most data I studied are steaming data from different sources yet correlated in multiple ways. As more heterogenous data are added to the data repository for the same city, I plan to investigate a hybrid index with temporal, spatial, and contextual correlation to improve the efficiency of later analyses.

Consolidating diverse data also introduces privacy and security issues, which have been well studied targeting at single-source data. However, the multi-source data provide new challenges, e.g., a late-night taxicab GPS data and its associated smartcard transaction data for fare may expose a smartcard with a billing address. I plan to further study these issues in a promising direction of differential privacy.

4.3 Building Real-world Large-scale Platforms

The third agenda is to actively build real-world platforms in order to validate our models and to fully unleash the potential of the above multi-source data repository. Based on my research impact, I was fortunate to test my several designs by small-scale real-world trials, such as an inter-regional transit [3]. In the future, I plan to advance my collaboration with Pchomes in Texas to commercialize my carpooling application coRide and with District Department of Transportation in Washington D.C. to improve Capital Bikeshare Program, which serve as a starting point for application deployment in the United States. Further, I have been working with SIAT to use multi-source data to implement a real-world service to predict bus arrival time in Shenzhen. As a pilot program for Intelligent Transportation Initiatives of the Shenzhen government, this service as a smartphone app has been used by 100,000 residents every day. Finally, I am actively looking for other cities to test my multi-source data-driven designs, *e.g.*, I am studying heterogenous datasets including taxicab GPS, fares, and bike trips in New York City, which is my next target city for deployments. With these valuable platforms, I plan to further analyze, model, and implement urban services based on multi-source data to improve urban efficiency.

During my Ph.D research, I am fortunate enough to study one of the largest urban datasets consolidated for academic research. I have explored a wide range of cutting-edge topics and obtained many hands-on skills and experiences by providing leading research on big-data-driven CPS. In the future, I will dedicate my effort to address challenges caused by urbanization with independent research and interdisciplinary collaboration.

References Cited

[1] D. Zhang, T. He, Y. Liu, Y. Gu, F. Ye, R. K. Ganti, and H. Lei, "Acc: Generic On-Demand Accelerations for Neighbor Discovery," in *the 10th ACM Conference on Embedded Networked Sensor Systems (SenSys'12)*.

[2] D. Zhang, T. He, F. Ye, R. K. Ganti, and H. Lei, "EQS: Neighbor Discovery and Rendezvous Maintenance with Extended Quorum System for Mobile Applications," in *the 32nd International Conference on Distributed Computing Systems (ICDCS'12)*.

[3] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales," in *the 20th ACM International Conference on Mobile Computing and Networking (MobiCom'14)*.

[4] D. Zhang, J. Zhao, F. Zhang, and T. He, "coMobile: Real-time Human Mobility Modeling at Urban Scale by Multi-View Learning," in *The ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'15)*.

[5] D. Zhang, J. Zhao, F. Zhang, and T. He, "UrbanCPS: a Cyber-Physical System Based on Multi-source Data with Model Integration," in *the ACM/IEEE 6th International Conference on Cyber-Physical Systems (ICCPS'15)*.

[6] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Dmodel: Online Taxicab Passenger Demand Model from Large Roving Sensor Networks," in *the 3rd IEEE International Congress on Big Data (BIGDATA'14)*.

[7] D. Zhang, T. He, Y. Liu, and J. A. Stankovic, "CallCab: A Unified Recommendation System for Carpooling and Regular Taxicab Services," in *the IEEE International Conference on Big Data (BIGDATA'13)*.

[8] D. Zhang, J. Zhao, F. Zhang, R. Jiang, and T. He, "Feeder: Supporting Last-Mile Transit with Extreme-Scale Infrastructure Data," in *the 14th ACM Conference on Information Processing in Sensor Networks (IPSN '15)*.

[9] D. Zhang and T. He, "pCruise: Reducing Cruising Miles for Taxicab Networks," in *the 33rd IEEE Real-time Systems Symposium (RTSS'12)*.

[10] D. Zhang, Y. Li, F. Zhang, M. Lu, Y. Liu, and T. He, "coRide: Carpool Service with a Win-Win Fare Model for Taxicab Networks," in *the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys'13)*.