Heterogeneous Model Integration for Multi-source Urban Infrastructure Data

DESHENG ZHANG, Rutgers University, USA
JUANJUAN ZHAO, Shenzhen Institutes of Advanced Technology, China
FAN ZHANG, Shenzhen Institutes of Advanced Technology, China
TIAN HE, University of Minnesota, USA
HAENJU LEE, Daegu Gyeongbuk Institute of Science and Technology, Republic of Korea
SANG H. SON, Daegu Gyeongbuk Institute of Science and Technology, Republic of Korea

Data-driven modeling usually suffers from data sparsity, especially for large-scale modeling for urban phenomena based on single-source urban-infrastructure data under fine-grained spatial-temporal contexts. To address this challenge, we motivate, design and implement UrbanCPS, a cyber-physical system with heterogeneous model integration, based on extremely-large multi-source infrastructures in the Chinese city Shenzhen, involving 42 thousand vehicles, 10 million residents, and 16 million smartcards. Based on temporal, spatial and contextual contexts, we formulate an optimization problem about how to optimally integrate models based on highly-diverse datasets, under three practical issues, i.e., heterogeneity of models, input data sparsity or unknown ground truth. We further propose a real-world application called Speedometer, inferring real-time traffic speeds in urban areas. The evaluation results show that compared to a state-of-the-art system, Speedometer increases the inference accuracy by 29% on average.

Categories and Subject Descriptors: H.4 [Information System Application]: Miscellaneous
General Terms: Algorithms, Model, Experimentation, Application
Additional Key Words and Phrases: Cyber-Physical System, Model Integration

ACM Reference Format:
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION
The recent advance of urban infrastructures increases our ability to collect, analyze and utilize big infrastructure data to improve urban phenomenon modeling [Zheng et al. 2014]. Numerous data-driven models have been proposed based on these infras-
structure data to capture urban dynamics [Aslam et al. 2012] [Shang et al. 2014] [Yuan et al. 2011a]. However, although each infrastructure produces abundant data, almost all resultant models suffer from data sparsity [Zheng et al. 2014]. This is because it is almost impossible to collect complete data about a particular phenomenon under fine-grained spatial-temporal contexts. For example, traffic speeds can be modeled by GPS data from taxicabs [Aslam et al. 2012], but under fine-grained spatial-temporal contexts, such a speed model suffers from data sparsity. As shown by our empirical analysis on the Chinese city Shenzhen, given a middle-length time slot of five minutes during 24 hour a day, 57% of its 110-thousand road segments on average do not have any taxicab, which leads to data sparsity.

In this work, we argue that with increasing updates of urban infrastructures, one urban phenomenon can be separately modeled by many heterogeneous infrastructure datasets. For example, a traffic speed can be directly modeled by vehicle GPS data and loop detector data [Aslam et al. 2012], or indirectly modeled by cellphone and transportation smartcard data [Isaacman et al. 2012]. Integrating these relevant yet heterogeneous models can provide complementary predictive powers by combining the expertise of heterogeneous infrastructures, which is used to address data sparsity issues about single infrastructures. Although many effective models have been proposed based on infrastructure data, they are typically based on single-source data, e.g., taxicab GPS [Aslam et al. 2012], cellphone data [Isaacman et al. 2012], bus data [Bhattacharya et al. 2013], and subway data [Lathia and Capra 2011]. Due to various technical and logistical reasons, little work, if any, has been done to integrate single-source heterogeneous models into a unified multi-source model based on large-scale infrastructure data (TB level data) to address practical issues, e.g., sparse data, for real-world applications. We provide a detailed survey of existing work in Section 6.

To this end, we motivate and design UrbanCPS, a CPS system with a generic heterogeneous-model integration based on extremely-large infrastructure data. In UrbanCPS, we implement five heterogeneous models based on a 14 thousand taxicab network, a 15 thousand truck network, a 13 thousand bus network, a 10 million user cellular network, and an automatic fare-collection system with 17 thousand smartcard readers and 16 million smartcards in Shenzhen. With these five highly-diverse heterogeneous models, we propose a model-integration technique to address their data sparsity, e.g., integrating traffic-speed models based on vehicles data and urban-density models based on cellphone data. However, we face three challenges as follows.

(1) Among all heterogeneous models, some models are only indirectly relevant to a particular phenomenon of interest, e.g., an urban-density model is only indirectly relevant to traffic speeds. Thus, it is challenging to effectively integrate directly-relevant models with indirectly-relevant models due to their heterogeneity.

(2) Indirectly-relevant models normally cannot output a measurement about phenomena of interest directly. Thus, even with complementary knowledge from indirectly-relevant models, it is a non-trivial problem to solve data sparsity for directly-relevant models.

(3) During a model integration, different models have different weights under different temporal, spatial and contextual conditions, and the optimal weights are usually obtained by regression with the ground truth. But the ground truth of urban scale phenomena is almost impossible or really expensive to be obtained.

A unique combination of the above three challenges makes our work significantly different from the previous model integration, where integrated models are often homogeneous and based on complete data with known ground truth. The key contributions of the paper are as follows:
We propose the first generic CPS system UrbanCPS with heterogeneous model integration based on metropolitan-scale data. To our knowledge, the integrated models have by far the highest standard for urban modeling in two aspects: (i) modeling based on the most complete infrastructure data including cellular, taxicab, bus, subway and truck data for the same city, and (ii) modeling based on the largest residential and spatial coverage (i.e., 95% of 11 million permanent residents and 93% of 110 thousand road segments in Shenzhen). The sample data are given in [Sample Data 2015].

We theoretically formulate an optimization problem to integrate heterogeneous models. We propose a technique to dynamically measure heterogeneous-model similarity on phenomena of interest under different temporal, spatial and contextual conditions to address three practical issues as follows: (i) how to integrate indirectly-relevant heterogeneous models; (ii) how to use an integrated model to address data sparsity; (iii) how to assign weights to different models without a regression process based on the ground truth. In particular, we design a technique based on context-aware tensor decomposition to integrate multiple models with data sparsity.

We design and implement a real-world application called Speedometer, which infers real-time traffic speeds in urban areas based on an integration of five models built upon taxicab, bus, truck, cellphone, and smartcard-reader networks. We test UrbanCPS based on a comprehensive evaluation with 1 TB real-world data in Shenzhen. The evaluation results show that compared to a current system, UrbanCPS increases the inference accuracy by 29% on average.

We organize the paper as follows. Section 2 gives our motivation. Section 3 presents the UrbanCPS. Section 4 describes our model integration based on Bayesian model averaging and tensor decomposition. Section 5 validates UrbanCPS with a real-world application, followed by the related work and the conclusion in Sections 6 and 7.

2. MOTIVATION

To show our motivation, we compare two traffic-speed models built upon large-scale empirical data we collected in Shenzhen. The first model is called SZ-Taxi [Transport Commission of Shenzhen Municipality 2014], which is a real-world system deployed and maintained by Shenzhen Transport Committee to infer real-time traffic speeds based on taxicab GPS data in Shenzhen. The second model is called TSE [Shang et al. 2014], which is a state-of-the-art traffic model in the research community based on vehicle GPS data. We feed our bus and truck GPS data to TSE and obtain two models called TSE-Bus and TSE-Truck, respectively. The details are given in Section 5.2. As in Figure 1, we compare three models based on taxicab, bus and truck data to the ground truth on a major road segment in Shenzhen called Shahe Road in 5-min slots during a regular Monday.

The ground truth is obtained by loop detectors, which are deployed in limited intersections of a city to obtain the real-time average traffic speeds. Loop detectors are mostly managed by city transportation agencies. Due to costs and deployment efforts, most cities including Shenzhen only install these detectors on major intersections or road segments instead of urban-scale deployment. The details about loop detectors are given in the evaluation section. Note that although different kinds of vehicles have different speeds on the same road segment, e.g., a bus may have a different speed from a passenger car [Garg et al. 2014b], we focus on developing an average speed model for generic traffic, similar to other state-of-the-art models [Transport Commission of Shenzhen Municipality 2014] [Shang et al. 2014].

In general, all three models have data sparsity issues, i.e., among a total of 288 5-min slots, SZ-Taxi, TSE-Bus and TSE-Truck have data on 87, 49 and 39 slots, i.e., 30%,
17% and 14%, respectively. If the data are all complete for all three models, we should have 24 points for every model, i.e., a total of 72 points, for every red box covering a 2 hour period, but we have much fewer than 72 points as shown in Figure 1. (i) SZ-Taxi has a major data sparsity issue during the early morning when no taxicabs are on this road segment. Further, it typically overestimates the speed in the nighttime since taxicab drivers typically drive much faster than regular drivers in the nighttime when passengers are few, but it underestimates the speed in the daytime due to frequent stopping for pickups and dropoffs as well as long-time waiting for passengers. (ii) TSE-Bus has sparse data in the nighttime when the bus service is not available, and in some regular daytime. Further, it underestimates the speed in the non-rush hour due to frequent stops, but it overestimates the speed in the rush hour because of dedicated fast traffic lines for bus only. (iii) TSE-Truck has sparse data in the morning and evening rush hour, because trucks are forbidden to use several major roads during the rush hour to relief traffic congestion. Even for the time period where trucks are allowed, it still has such an issue. Also, it usually underestimates the speed during other time due to the speed limit of trucks. Note that this road segment was selected as one of ten major road segments in Shenzhen, but we still face major data sparsity issues, which are much worse on other small road segments where taxicabs, buses or trucks are much fewer as shown in Section 3.2.

A seemingly promising solution is to integrate these three models to address data sparsity issues from a homogenous complimentary view. However, such a straightforward homogenous-model integration may still face data sparsity issues due to their inherent homogeneity, e.g., all three models have incomplete data in common slots in the red boxes. In this work, we address this challenge by introducing other heterogeneous models (e.g., urban-density models) based on different datasets (e.g., cellphone data) under the observation that the traffic speed is correlated with urban density in same spatial-temporal contexts [Cox 2015] as shown by Figure 2 where we plot the density and traffic speed on a road segment in Shenzhen on a regular Monday. We clearly found that when the traffic density goes up, the traffic speed goes down. It motivates us to combine density models with speeds models to infer traffic speeds. In fact, in the civil engineering community, such a phenomenon is called the fundamental diagram of traffic flow [Wikipedia]. There has been some previous work to empirically quantify this fundamental diagram [Sen et al. 2013a], but in a small scale with only traffic data. In contrast, our work is to integrate models driven by vehicle GPS data, cellphone data, and smartcard data.
However, how to combine these heterogeneous models for the same objective is challenging. In this work, we propose an integration technique in a reference implementation of an extremely-large CPS system, which presented as follows.

3. URBAN CYBER PHYSICAL SYSTEM

Broadly, a CPS can be considered as a system of systems. Therefore, in this work, we consider a set of urban infrastructure systems (e.g., cellular, taxicab, bus, subway and truck networks) as a Urban Cyber Physical System (UrbanCPS) from a broad perspective: any device in urban infrastructures is considered as a pervasive sensor in Urban CPS, if it generates data that can be used to build a model to describe phenomena of interest. Built upon an integration of models based on multiple data sources, UrbanCPS provides unseen urban dynamics under extremely fine-grained spatio-temporal resolutions to support real-world applications, which cannot be achieved by any model from single data source in isolation, e.g., a monolithic infrastructure.

In Figure 3, we outline UrbanCPS with four components, i.e., Data Collection, Model Generation, Model Integration and Model Utilization. These four components span the whole data-processing chain in UrbanCPS.

As in Figure 3, we provide a road map for the rest of paper as follows. (i) In Section 3.1, we first introduce the data collection where we individually collect multiple-source data from urban infrastructures of Shenzhen. (ii) In Section 3.2, we generate various heterogeneous models based on collected single-source data. (iii) In Section 4, we effectively combine these heterogeneous models by our model integration based on their similarity measurement.
similarity and domain knowledge. (iv) In Section 5, to close the control loop, we propose an application to estimate real-time traffic speeds based on integrated models and other supporting data, e.g., map data and urban partition data. We envision that urban residents would use this application to find efficient routes, which in turn provides feedback to urban infrastructures. As a result, with the highlights on extremely-large data collection and highly-generic heterogeneous model integration, UrbanCPS builds an architectural bridge between multiple domain-independent urban infrastructures and real-world knowledge output tailored by applications.

3.1. Data Collection

In our project, we have been collaborating with several service providers and the Shenzhen Transport Committee (hereafter STC) for the real-time access of urban infrastructures. In Figure 3, we consider five kinds of devices in this version of implementation, which detects urban dynamics from complimentary perspectives.

— **Cellphones** are used to detect cellphone users’ locations at cell tower levels based on call detail records. We utilize cellphone data through two major operators in Shenzhen with more than 10 million users. The cellphone data give 220 million locations per day.

— **Smartcard Readers** are used to detect locations of a total of 16 million smartcards used to pay bus and subway fares. These readers capture more than 10 million rides and 6 million passengers per day. We study reader data from STC, which accesses real-time data feeds of a company that operates the smartcard business.

— **Buses** are used to detect real-time traffic and bus passengers’ locations by cross-referencing data of onboard smartcard readers for fare payments. We study bus data through STC to which bus companies upload their bus status in real time, accounting for all 13 thousand buses generating 2 GPS records/min.

— **Taxicabs** are used to detect real-time traffic and taxicab passengers’ locations based on taxicab status (i.e., GPS and occupancy). We study taxicab data through STC to which taxicab companies upload their taxicab status in real time, accounting for all 14 thousand taxicabs generating 2 GPS records/min.

— **Trucks** are used to detect real-time traffic by logging real-time GPS locations of a fleet of 15-thousand freight trucks, which travel within Shenzhen and around nearby cities. We study this truck network through a freight company that installs GPS devices on all these trunks for daily managements. Every truck uploads its real-time GPS location and driving speed back to the company server every 15s on average, which then are routed to our server.

Since our paper concentrates on system aspects, we briefly introduce our data related issues due to space limitation. We establish a secure and reliable transmission mechanism, which feeds our server the above data collected by STC and service providers with a wired connection.

As in Figure 4, we have been storing a large amount of data to generate single-source models. Their spatial granularity is given in Figure 5 where commercial vehicles, i.e., trucks, buses, regular and electric taxis, generate data at road segment levels but bus smartcards, subway smartcards and cellphones generate data at station levels. Such big data require significant efforts for the daily management. We utilize a 34 TB Hadoop Distributed File System (HDFS) on a cluster consisting of 11 nodes, each of which is equipped with 32 cores and 32 GB RAM. For daily management and processing, we use the MapReduce based Pig and Hive. Due to the extremely large size of our data, we have been finding several kinds of errant data, e.g., missing data, duplicated data and data with logical errors, and thus we have been conducting a detailed clean-
Heterogeneous Model Integration for Multi-source Urban Infrastructure Data

<table>
<thead>
<tr>
<th></th>
<th>Taxicab Dataset</th>
<th>Bus Dataset</th>
<th>Freight Truck Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td>2012/1/1</td>
<td>2013/1/1</td>
<td>2013/9/11</td>
</tr>
<tr>
<td># of Taxis</td>
<td>14,453</td>
<td>13,032</td>
<td>15,001</td>
</tr>
<tr>
<td>Size</td>
<td>1.7 TB</td>
<td>720 GB</td>
<td>1.2 TB</td>
</tr>
<tr>
<td># of Records</td>
<td>22 billion</td>
<td>9 billion</td>
<td>16 billion</td>
</tr>
<tr>
<td>Format</td>
<td></td>
<td>Format</td>
<td>Format</td>
</tr>
<tr>
<td>Plate ID</td>
<td></td>
<td>Date&amp;Time</td>
<td></td>
</tr>
<tr>
<td>Status</td>
<td></td>
<td>Stop ID</td>
<td>Plate ID</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPS&amp;Speed</td>
<td>GPS&amp;Speed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stop ID</td>
<td>Plate ID</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stop ID</td>
<td>Plate ID</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPS&amp;Speed</td>
<td>GPS&amp;Speed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Cellphone Dataset</th>
<th>Smartcard Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td>2013/10/1</td>
<td>2011/7/1</td>
</tr>
<tr>
<td># of Users</td>
<td>10,432,246</td>
<td>16,000,000</td>
</tr>
<tr>
<td>Size</td>
<td>1 TB</td>
<td>600 GB</td>
</tr>
<tr>
<td># of Records</td>
<td>19 billion</td>
<td>6 billion</td>
</tr>
<tr>
<td>Format</td>
<td>Format</td>
<td>Format</td>
</tr>
<tr>
<td>SIM ID</td>
<td>Date&amp;Time</td>
<td>Card ID</td>
</tr>
<tr>
<td>Cell Tower ID</td>
<td>Activities</td>
<td>Device ID</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Station ID</td>
</tr>
</tbody>
</table>

Fig. 4. Datasets from Model Generation

Fig. 5. Data Granularity

ing process to filter out errant data on a daily basis. We protect the privacy of residents by anonymizing all data and presenting models in aggregation. In short, our endeavor of consolidating the above data enables extremely large-scale fine-grained urban phenomenon rendering based on existing single-source models, which is unprecedented in terms of both quantity and quality shown as follows.

3.2. Model Generation
Fellow researchers have proposed many effective single-source models [Zheng et al. 2014], so we restrain ourselves from developing new models. Instead, we directly use our data to generate single-source models based on existing methods.

ACM Transactions on Cyber-physical Systems, Vol. V, No. N, Article A, Publication date: January YYYY.
3.2.1. Model Summary. We implement two kinds of models based on the data collected in UrbanCPS. (i) Speed Models: including $M_T$, $M_B$, $M_F$, which use GPS data from Taxicab, Bus and Freight truck networks individually to estimate real-time traffic speeds. They are implemented similarly according to a state-of-the-art speed model TSE, which uses historical and real-time vehicle data as well as contexts (e.g., physical features of roads) for a collaborative filtering [Shang et al. 2014]. In addition, we consider all Vehicles as a single fleet and feed its data to TSE to obtain a new model $M_V$. (ii) Density Models: including $M_C$ and $M_S$, which use the Cellphone and Smartcard data to estimate real-time urban density (i.e., count of residents). $M_C$ is based on a population density model that predicts future CDR records based on the previous CDR records to indicate the density [Isaacman et al. 2012]. $M_S$ is based on a Gaussian process-based predictive model that uses contexts, e.g., time of day and day of week, to infer transit passenger density [Bhattacharya et al. 2013]. We provide a summary of these models in Table I based on their results in one day. During one day, based on the GPS uploading speeds and traveling patterns, $M_T$, $M_B$, $M_F$, and $M_V$ cover 87%, 59%, 45%, and 93% of all 110 thousand road segments in Shenzhen. During one day, $M_C$ covers 95% of 11 million residents and produces their locations as one of 17,859 cell towers when they use their phones. $M_S$ covers 55% of all residents and produces their locations as one of 10,442 transit stations when they use their smartcards.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Spatial Resolution</th>
<th>Temporal Resolution</th>
<th>Resident Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_T$</td>
<td>87% of Roads</td>
<td>30s</td>
<td>N/A</td>
</tr>
<tr>
<td>$M_B$</td>
<td>59% of Roads</td>
<td>30s</td>
<td>N/A</td>
</tr>
<tr>
<td>$M_F$</td>
<td>45% of Roads</td>
<td>15s</td>
<td>N/A</td>
</tr>
<tr>
<td>$M_V$</td>
<td>93% of Roads</td>
<td>7.5s</td>
<td>N/A</td>
</tr>
<tr>
<td>$M_C$</td>
<td>17,859 Towers</td>
<td>Various</td>
<td>95%</td>
</tr>
<tr>
<td>$M_S$</td>
<td>10,442 Stations</td>
<td>Various</td>
<td>55%</td>
</tr>
</tbody>
</table>

3.2.2. Data Sparsity in Fine Granularity. Although all these models have comprehensive daily data, real-world applications typically require knowledge under fine-grained spatial-temporal contexts [Aslam et al. 2012] [Shang et al. 2014] [Yuan et al. 2011a] where all these models experience data sparsity issues.

Based on the historical data, we pick the first weekday after a national holiday, and in this particular day, all these infrastructure systems generate the biggest data in terms of volumes compared to other days.

We show the percentage of segments where speeds can be captured by speed models in 5-min slots in Figure 6. We found that these models capture a low percentage of segments under 5-min slots, e.g., even for $M_V$ based on all vehicle data, we only have 49% of road segments on average with vehicles, which leads to data sparsity.

Similarly, we show the number of residents captured by $M_C$ and $M_S$ in Figure 7 where the result for $M_S$ is shown by a factor of 10 in order to show the fluctuation. We found that these two density models also have data sparsity issues due to high total population in Shenzhen, e.g., among 11 million permanent residents, $M_C$ can only capture 1 million of them at most during a 5-min slot around 15:00, accounting for only 9% of all residents. $M_C$ can only capture 80 thousand of them at most during a 5-min slot of the morning rush hour, accounting for only 0.7% of all residents.
3.2.3. Opportunity for Model Integration. In this work, we found that although all these models have data sparsity issues, $M^C$ and $M^S$ have more complete data than others, e.g., for every 5-min slot in both $M^C$ and $M^S$, we have density data at cell tower and transit station levels. Therefore, by resetting their spatial granularity to road segment levels (i.e., the details are given in Section 5.2), density models $M^C$ and $M^S$ are capable of providing complimentary knowledge for speed models $M^T$, $M^B$, and $M^F$, which have severe data sparsity issues on road segment levels, e.g., if a speed model does not have GPS data about a road segment during a time slot, we infer missing GPS data based on historical GPS data and the data from road segments with similar urban density, shown by our model integration as follows.

4. MODEL INTEGRATION

We introduce our integration technique by combining models directly or indirectly relevant to phenomena of interest (hereafter direct and indirect models for conciseness). In this work, we simply identify a model as a direct model to an urban phenomenon, if it is based on the data with direct measurements of this phenomenon, e.g., a model based on taxicab data is a direct model for the phenomenon of traffic speeds, because taxicab data have direct measurements of speeds. But a model based on cellphone data is only an indirect model for speeds because it does not have direct measurements on speeds. As discussed before, we also need these indirect models in our integration, because they often provide complimentary knowledge to address data sparsity issues of direct models. Note that direct and indirect models are different from classic supervised and
unsupervised models in data mining, which are both direct models in our context since they are based on data with direct measurements for phenomena of interest.

4.1. Problem Formulation

Let \( x_{t,s} \) be an urban phenomenon we want to characterize associated with a temporal context \( t \) and a spatial context \( s \), and let \( y \) be a class label, where \( x_{t,s} \) and \( y \) are selected from a phenomenon space \( X \) and a label space \( Y \). Based on \( K \) different data sources in various urban infrastructures, we have a set of \( K \) models, i.e., from \( M^1 \) to \( M^K \), and each of them is independently formulated based on a corresponding data source. For example, in our later application, \( x_{t,s} \) is a traffic speed on a road segment \( s \) during a time period \( t \); \( y \) is a label of 20km/h; \( M^1 \) is a model based on taxicab data and assigns a particular label \( y \) to \( x_{t,s} \).

Formally, based on the Bayesian model averaging approach, we have the probability distribution for \( y \) as follows.

\[
P(y|x_{t,s}) = \sum_{k=1}^{K} P(y|M^k, x_{t,s}) \times P(M^k|x_{t,s}),
\]

where \( P(y|M^k, x_{t,s}) \) is the prediction made by \( M^k \) regarding to \( x_{t,s} \); \( P(M^k|x_{t,s}) \) is considered as a model weight for a particular model \( M^k \) given a particular urban phenomenon \( x_{t,s} \) under with a temporal context \( t \) and a spatial context \( s \).

To integrate different models in small-scale systems, Eq.(1) can be directly used. In particular, \( P(y|M^k, x_{t,s}) \) can be accurately obtained by a direct model \( M^k \) directly-relevant to the phenomenon of interest \( x_{t,s} \), based on the complete data. Further, the ground truth of conditional probability \( P(y = y_i|z_{t,s}) \) can also be measured and then used by a regression process to obtain the optimal weight \( P(M^k|x_{t,s}) \) for a model \( M^k \) given \( x_{t,s} \). However, to integrate models in our UrbanCPS with Eq.(1), we face three challenges to directly obtain the two factors, i.e., \( P(y|M^k, x_{t,s}) \) and \( P(M^k|x_{t,s}) \).

First, the models in our UrbanCPS are mostly heterogeneous and based on the data generated by service providers primarily for their own benefits, and thus these models may be only indirectly relevant to the phenomenon of interest. For example, a model based on cellphone data can be used to directly infer cellphone usage and thus urban density. But this model cannot be directly used to infer a traffic speed, though they are somehow related, because normally the higher residential density, the lower the traffic speeds for road segments without bus routes or during the time periods without bus services. As a result, given an indirect model \( M^k \) to the phenomenon \( x_{t,s} \), \( P(y|M^k, x_{t,s}) \) in Eq.(1) is unknown.

Second, due to large-scale phenomena of interest, the data in UrbanCPS are typically quite sparse. For example, the model based on bus GPS data cannot infer traffic speeds for road segments without bus routes or during the time periods without bus services. As a result, even for a direct model \( M^k \) for the phenomenon \( x_{t,s} \), \( P(y|M^k, x_{t,s}) \) in Eq.(1) may still be unknown.

Third, due to technical issues and high costs for direct measurements on urban phenomena, the ground truth for certain phenomena is typically unknown. Without the ground truth, we cannot use a regression process to obtain the optimal weights for all models during integration. Thus, even with known \( P(y|M^k, x_{t,s}) \) based on a direct model with complete data, \( P(M^k|x_{t,s}) \) in Eq.(1) may still be unknown.

A combination of these three challenges provides us a unique design space for our model integration compared to the existing work. As follows, we first show how to solve this problem optimally if we are given all direct models with both the complete data and the ground truth, and then we relax these three assumptions individually to address the three challenges.
4.2. Optimal Solution

Suppose the label space $Y$ is mapped into discrete labels $\{y_1, ..., y_{|Y|}\}$ where $|Y|$ is the number of labels. Let $H_{t,s}$ be a $|Y| \times K$ matrix where $H_{j}^{k} = P(y = y_j | M^k, x_{t,s})$ is the $kj$ entry, and thus it represents all predictions made for $x_{t,s}$ from all $K$ models. Let $w_{t,s}$ be a $K \times 1$ weight vector where $w_{t,s}^k = P(M^k | x_{t,s})$, and thus it represents weights of all $K$ models. As a result, a $|Y| \times 1$ vector $Hw_{t,s}$ is the output of our model integration for $x_{t,s}$, which gives a probability distribution of $x_{t,s}$ on a label space $Y$ of $\{y_1, ..., y_{|Y|}\}$. With this output, we aim to minimize the distance from this output to the true conditional probability (given by the ground truth), which is represented by a $|Y| \times 1$ vector $f_{t,s}$ where $f_j = P(y = y_j | x_{t,s})$. Therefore, based on a straightforward squared error loss without regularization, the key objective of our model integration is to find an optimal weight vector $w^*_{t,s}$ that minimizes the distance between the true $f_{t,s}$ and our output $Hw_{t,s}$ as follows.

$$w^*_{t,s} = \arg \min_{w_{t,s}} (f_{t,s} - Hw_{t,s})^T(f_{t,s} - Hw_{t,s}).$$

The optimal solution of this function can be directly obtained by a least-square linear regression.

However, as discussed before, this optimal solution has three impractical assumptions (i.e., all directly-relevant models, complete data and known ground truth), which leads to two issues. First, an element in $H_{t,s}$, e.g., $H_{j}^{k} = P(y = y_j | M^k, x_{t,s})$, is not always available for an indirect model $M^k$ or a direct model $M^k$ based on sparse data. Second, the true conditional distribution $f_{t,s}$ is mostly unknown due to the unknown ground truth. As in following three subsections, we relax these three assumptions one by one and discuss the issues of (i) how to obtain $P(y|M^k, x_{t,s})$ for an indirect model, (ii) how to obtain $P(y|M^k, x_{t,s})$ for a direct model based on sparse data, and (iii) how to infer the weights without the ground truth, respectively.

4.3. Indirect Models

In our UrbanCPS, various models are built based on the collected data, and some of these may not be directly relevant to the urban phenomenon we try to characterize. But we still need the models based on these indirectly-relevant data, because their diversity can provide additional, more often complimentary, knowledge helping us to solve issues of models directly related. Suppose we have a set of urban phenomena associated with different real-world temporal and spatial contexts $X = \{x_{1,t_1,s_1}, x_{1,t_2,s_2}, x_{2,t_1,s_1}, x_{2,t_2,s_2}\}$, and aim to characterize them into a label space $Y = \{y_1, y_2, y_3\}$. In our later application, $x_{1,t_1,s_1}$ is the average traffic speed on a road segment $s_1$ during time period $t_1$, which can be assigned with a label of $y_1=10\text{km/h}$. Suppose among all $K$ models, the models from $M^1$ to $M^d$ are direct models, and the models from $M^{d+1}$ to $M^K$ are indirect models. For a direct model $M^p \in \{M^1, ..., M^d\}$, $P(y|M^p, x_{t,s})$ is directly obtained; but for an indirect model $M^q \in \{M^{d+1}, ..., M^K\}$, $P(y|M^q, x_{t,s})$ is typically unknown. The main objective of the following is to infer $P(y|M^q, x_{t,s})$ for an indirect model $M^q$. The key idea of our method is to use the internal similarity between an indirect model $M^q$ and all direct models to infer $P(y|M^q, x_{t,s})$ for $M^q$ for a particular temporal spatial combination. However, the internal similarity between models is difficult to be directly quantified, so we introduce a process of categorizing all elements in the phenomenon space $X$ by individual models as follows.

4.3.1. Categorizing. Based on a direct model $M^p$, we directly categorize all elements in $X = \{x_{1,t_1,s_1}, x_{1,t_2,s_2}, x_{2,t_1,s_1}, x_{2,t_2,s_2}\}$ into $|M^p|$ categories, and each of category is associated with a unique label in $Y$. Thus, for a direct model $M^p$, $|M^p| = |Y|$. Similarly, based on an indirect model $M^q$, we also categorize all elements in $X$ into $|M^q|$ categories by a
Therefore, we transfer the problem from the model level $P(y|M^k, x_t, s)$ to the category level $P(y|c^k_i, x_t, s)$, because the comparison between categories is easier to quantify.
Given $x_{t,s} \in c_i^p$ where $c_i^p$ belongs to a direct model $M^p$,
\[ P(y = y_i|x_{t,s}) = \begin{cases} 1 & \text{if } l = i \\ 0 & \text{if } l \neq i \end{cases}. \tag{3} \]
Note that for simplicity we assume that there are no errors during categorizing, i.e., given $x_{t,s} \in c_i^p$, it is always assigned to $y_i$. But if $P(y = y_i|x_{t,s})$ follows an empirical distribution instead of as in Eq. (3), our method still works with a straightforward probabilistic method.

Given $x_{t,s} \in c_i^q$ where $c_i^q$ belongs to an indirect model $M^q$, however, $P(y = y_i|x_{t,s})$ is unknown. Thus, the key question we have now is how to infer $P(y|c_i^q, x_{t,s})$ for a category $c_i^q$ belonging to an indirect model $M^q$. As follows, we solve this issue by exploring similarity between categories from direct and indirect models.

4.3.2. Similarity Measurement. Basically, the rationale behind the similarity measurement is that given a category $c_i^q$ from an indirect model $M^q$ and a category $c_j^q$ from an indirect model $M^q$, the closer $c_i^q$ is to $c_j^q$, the more likely that the members in $c_i^q$ have the same label with the members in $c_j^q$. Essentially, we transfer the expertise from direct models to indirect models by comparing their similarities on category levels.

Formally, for $P(y|c_i^q, x_{t,s})$ where the category $c_i^q$ belonging to an indirect model $M^q$, we have
\[ P(y = y_i|c_i^q, x_{t,s}) = \frac{\sum_{j=1}^d S(c_i^q, c_j^q)}{\sum_{i=1}^{[Y]} \sum_{j=1}^d S(c_i^q, c_j^q)}, \tag{4} \]
where $S(c_i^q, c_j^q)$ is the similarity between two categories $c_i^q$ and $c_j^q$. Therefore, the numerator is the sum of similarity between a category $c_i^q$ and all categories with a particular label $y_i$ from all direct models (i.e., from $M^1$ to $M^d$); the denominator is the sum of similarity between a category $c_i^q$ and all categories with all labels (i.e., from $y_1$ to $y_{[Y]}$) from all direct models (i.e., from $M^1$ to $M^d$).

To quantify similarity between two categories, we use a similarity vector $c_i^k$ to represent the membership of elements in $X$ for a category $c_i^k$. For example, as in Table II, we have $c_1^1 = \{1, 1, 0, 0\}$ indicating the first and second elements in $X$, i.e., $x_{t,s_1}$ and $x_{t,s_2}$, belong to $c_1^1$. With similarity vectors, we calculate $S(c_i^q, c_j^q)$ by Jaccard index.
\[ S(c_i^q, c_j^q) = \frac{|c_i^q \cap c_j^q|}{|c_i^q \cup c_j^q|}. \]
For example, in Table II, $S(c_1^1, c_2^1) = \frac{0}{4}$, and $S(c_1^1, c_2^2) = \frac{1}{4}$. By changing $y_i$ from $y_1$ to $y_{[Y]}$ in Eq. (4), we have the distribution of $P(y|c_i^q, x_{t,s})$.

4.3.3. Summary. In short, based on $P(y|c_i^p, x_{t,s})$ in Eq. (3) for a category $c_i^p$ from a direct model $M^p$ where $p \in [1, d]$ and $P(y|c_i^q, x_{t,s})$ in Eq. (4) for a category $c_i^q$ from an indirect model $M^q$ where $q \in [d+1, K]$, we have $P(y|c_i^k, x_{t,s})$ for any category from both either a direct model $M^p$ or an indirect model $M^q$. As a result, we have $P(y|c_i^k, x_{t,s})$ for all models where $k \in [1, K]$ in Eq. (2), which addressed the challenge of integrating heterogeneous direct models and indirect models.

4.4. Models based on Sparse Data
In this subsection, for models with sparse data, we formulate a tensor decomposition problem to infer real-time urban phenomenon $x_{t,s}$ on road segment $s$ during time $t$. Note that we use traffic speeds as a concrete example of urban phenomena because our tensor decomposition needs specific contexts.
4.4.1. Tensor Construction. We design a three-dimensional tensor $A \in \mathbb{R}^{N \times K \times M}$.

- A speed dimension indicates traffic speed labels: $[y_1, \ldots, y_{|Y|}]$.
- A time slot dimension indicates specific time windows (e.g., one hour window from 5PM to 6PM): $[t_1, \ldots, t_{|T|}]$.
- A spatial unit dimension indicates specific spatial units (e.g., an urban region): $[s_1, \ldots, s_{|S|}]$.
- An entry $A(y, s, t)$ indicates the traffic speed label $y$ for a road segment $s$ during a time slot $t$.

With our data, we fill this tensor $A$, and then obtain all traffic speed labels under a specific spatiotemporal partition. However, a key challenge is that the tensor $A$ is sparse because for road segments without any commercial vehicles during a time window, their corresponding entries are empty due to lacking GPS data.

A typical approach to address this challenge is to use a technique called tensor decomposition. As in Fig. 8, we have a tensor with three dimensions indicating traffic speed, road segments, and time slots. An entry denotes a tuple [speed, location, time]. But this tensor is sparse due to insufficient commercial vehicles. Based on the classic Tucker decomposition model [Kolda and Bader 2009], we decompose $A$ into a core tensor $I$ along with three matrices, $Y \in \mathbb{R}^{|Y| \times d^y}$, $S \in \mathbb{R}^{|S| \times d^s}$, and $T \in \mathbb{R}^{|T| \times d^t}$. $Y$, $S$, and $T$ infer correlations between traffic speeds, road segments, and time slots, respectively. $d^y$, $d^s$ and $d^t$ are the number of latent factors.

We use the following objective function to optimize the decomposition.

$$
||A - I \times Y \times S \times T||^2 + \lambda(||I||^2 + ||Y||^2 + ||S||^2 + ||T||^2)
$$

where the first term is for the measurement of decomposition errors and the second term is a regularization function to avoid over-fitting of modeling. $|| \cdot ||^2$ denotes the $l_2$ norm and $\lambda$ is the parameter to control the regularization function's contribution. By minimizing the above objective function, we obtain the optimized $I$, $Y$, $S$, and $T$ with the sparse tensor $A$, which is given by commercial GPS data. As a result, we use $I \times Y \times S \times T = A'$ to approximate $A$ where $\times$ represents the tensor-matrix multiplication.

However, a key challenge for the above method is that $A$ is over sparse, especially under fine spatiotemporal partitions (small road segment levels under one minute time slots). Therefore it leads to poor performance of the decomposition. We address this...
issue by proposing a technique to use historical traffic data to establish correlated contexts that improve the performance of the decomposition.

4.4.2. Context Extraction. To provide additional information for the decomposition, we use the historical commercial GPS data to extract three contexts, i.e., resident density, speed temporal patterns, and speed spatial patterns. We use three matrices to denote these three contexts as in Fig. 9.

— Resident Densities are given by a matrix $B$ where a row denotes a road segment; a column denotes a time slot; an entry denotes the average active resident count obtained by CDR data and smartcard data in this spatial unit for this time slot over a period of historical time.

— Speed Spatial Patterns are given by a matrix $C$ where a row denotes a road segment; a column denotes a speed label; an entry denotes the probability of this speed label on this road segment given a period of historical time.

— Speed Temporal Patterns are given by a matrix $D$ where a row denotes a time slot; a column denotes a speed label; an entry denotes the probability of this speed label during this time slot given a period of historical time.

All the matrices $B$, $C$, and $D$ can be obtained by a set of historical commercial GPS data.

[Fig. 9. Context Matrix Factorization]
4.4.3 Context-based Tensor Decomposition. Based on the three extracted context matrices, we present a joint tensor decomposition. In particular, we design the objective function as follows.

\[
\min_{\mathcal{I}, \mathcal{Y}, \mathcal{S}, \mathcal{T}} L(\mathcal{I}, \mathcal{Y}, \mathcal{S}, \mathcal{T}) = ||\mathcal{A} - \mathcal{I} \times \mathcal{Y} \times \mathcal{S} \times \mathcal{T}||^2 \\
+ \lambda_1 ||\mathcal{B} - \mathcal{S} \times \mathcal{T}||^2 + \lambda_2 ||\mathcal{C} - \mathcal{S} \times \mathcal{Y}||^2 + \lambda_3 ||\mathcal{D} - \mathcal{T}^T \times \mathcal{Y}||^2 \\
+ \lambda_4 (||\mathcal{I}||^2 + ||\mathcal{Y}||^2 + ||\mathcal{S}||^2 + ||\mathcal{T}||^2).
\]

(5)

where the first term is to measure the error of decomposing \(\mathcal{A}\); the second, third, and forth terms are to measure the error of factorizing matrix \(\mathcal{B}, \mathcal{C}, \) and \(\mathcal{D}, \) respectively; the last term is to avoid over-fitting of the decomposition. In our setting, \(d^y = d^s = d^t\). \(\lambda_1, \lambda_2, \lambda_3, \) and \(\lambda_4\) are preset parameters to indicate term weights. We normalized all values to \([0, 1]\) for the decomposition.

In this objective function, \(\mathcal{A}\) and \(\mathcal{B}\) share \(\mathcal{S}\) and \(\mathcal{T}\); \(\mathcal{A}\) and \(\mathcal{C}\) share \(\mathcal{S}\) and \(\mathcal{Y}\); \(\mathcal{A}\) and \(\mathcal{D}\) share \(\mathcal{Y}\) and \(\mathcal{T}\). Since \(\mathcal{B}, \mathcal{C}, \) and \(\mathcal{D}\) are not sparse, they lead to accurate \(\mathcal{S}, \mathcal{T}\) and \(\mathcal{Y},\) which increases the performance of decomposing \(\mathcal{A}\). As a result, the historical resident densities and traffic speed patterns are transferred into the decomposition of \(\mathcal{A}\), which leads to an accurate tensor decomposition.

Because this objective function does not have a closed-form solution to find the global optimal \(\mathcal{I}, \mathcal{Y}, \mathcal{S}, \) and \(\mathcal{T},\) we use an element-wise optimization algorithm as a numeric method [Karatzoglou et al. 2010] to obtain a local optimal solution. Finally, after we obtain \(\mathcal{I}, \mathcal{Y}, \mathcal{S}, \) and \(\mathcal{T},\) we use \(\mathcal{I} \times \mathcal{Y} \times \mathcal{S} \times \mathcal{T} = \mathcal{A}'\) to address the challenge of modeling based on sparse data.

Note that this method addresses data sparsity for direct models by assuming the data are complete for at least one indirect model, e.g., a density model. If we have missing data for all models, we have to use traditional methods, e.g., weighted averaging, to infer missing data based on historical data.

4.5 Weighting Models without Ground Truth

In this subsection, we address the issues of assigning a weight to a model for the integration without ground truth. Normally, the closer a model \(M^k\) is to the majority of all models, the higher weight it should be assigned with. Therefore, based on the similarity between different models, we assign the weight of a model \(M^k\) for a particular combination of a temporal context \(t\) and a spatial context \(s\) as follows.

\[
P(M^k|x_{t,s}) = w_{t,s}^k = \frac{\sum_{j=1, j\neq k}^K S(M^k, M^j)}{\sum_{i=1}^K \sum_{j=1, j\neq i}^K S(M^i, M^j)},
\]

where the numerator is the sum of the similarity between \(M^k\) and all models; the denominator is the sum of the similarity among all models. In this work, we define the similarity \(S(M^k, M^j)\) between two models \(M^k\) and \(M^j\) as follows.

\[
S(M^k, M^j) = \frac{\sum_{u=1}^{M^k} \sum_{v=1}^{M^j} S(e^k_u, e^j_v)}{|M^k| \cdot |M^j|},
\]

where we use the similarity at category levels to indicate the similarity at model levels.

Note that existing work usually weights each model globally, but our method assigns weights to each model according to a unique phenomenon \(x\) under a unique temporal-spatial combination \(t \cdot s\), which is used to identify variations in the model performance for different real-world contexts. It usually does not exist one weighting scheme that is globally optimal for any phenomenon under all temporal-spatial contexts. Thus, the
urban phenomenon under different temporal and spatial contexts may favor different models. Thus, the weighting scheme based on temporal-spatial contexts is better than the global weighting scheme in terms of prediction accuracy.

4.6. Summary
Based on the problem formulation in the first subsection, we obtain the optimal solution for model weights, which minimizes the distance between the true conditional distribution and the output of our integration. Then in the following three subsections, we relax the three key assumptions in the optimal solution one by one towards a practical model integration. Essentially, the key idea we have been using is to compare internal similarity of effects of different models on a set of given urban phenomena. Then, we transfer predictive powers of indirect models with complete data to direct models with sparse data. The rationale is that the more similar two models are, the more likely they would make the same prediction about an urban phenomenon. Finally, the similarity is used as an indication of a model's weight, by assuming the majority of the models are correct, and thus the closer a model is to other models, the higher weight it carries.

5. APPLICATION: SPEEDOMETER
In this section, we present an application called Speedometer to test the performance of our model integration based on the data we collected in Shenzhen.

5.1. Application Background
The real-time traffic speed in urban regions is an important phenomenon for both residents and transportation authority. An accurate inference about traffic speeds on road segment levels under fine-grained time slots improves many urban applications, e.g., more efficient automobile navigation. A direct yet trivial solution is to install speed detectors, such as loop detectors and traffic cameras as shown in Figure 10, in every road segment.

![Fig. 10. Loop Detectors and Traffic Cameras](image-url)

However, this solution would involve tremendous costs, so these static sensors are only installed in major segments for most cities. To achieve a speed inference for all segments, vehicle GPS data from commercial vehicles, such as taxicab, are utilized to
produce several models to infer traffic speeds [Aslam et al. 2012]. Also, several systems also infer traffic speeds based on participatory sensing [Zheng et al. 2014]. But these models typically are based on single-source homogenous data and are ineffective when data are sparse in fine-grained contexts.

To address this issue, we propose Speedometer, which infers real-time traffic speeds on segment levels based on an integration of five models, i.e., $M_T$, $M_B$, $M_F$, $M_C$ and $M_S$, as proposed in Section 3.2. $M_T$, $M_B$ and $M_F$ are speed models based on Taxi, Bus and Freight truck data; whereas $M_C$ and $M_S$ are density models based on cell-phone and smartcard data. Thus, $M_T$, $M_B$ and $M_F$ individually map a traffic speed $x_{t,s}$ on a segment $s$ during a period $t$ into a label space $Y$ to indicate a traffic speed. $M_C$ and $M_S$ individually infer an urban density into another label space to indicate a density under the same contexts. Based on domain knowledge, $M_T$, $M_B$ and $M_F$ are direct models to speeds, and $M_C$ and $M_S$ are indirect models. Thus, Speedometer effectively integrates them to produce accurate speed inferences based on our integration. For different applications, Speedometer infers traffic speeds on both segment and region levels by aggregating segments with the minimum time slot of 5 mins. Figure 11 gives a visualization on average speeds inferred by Speedometer from 6PM to 7PM in 496 Shenzhen regions where a warmer color indicates a slower speed.

Note that based on our model, another related application is to find the representative intersections to deploy loop sensors to capture traffic dynamics without other data sources. This application can be formulated as an optimization problem to deploy the minimal number of sensors with a guaranteed coverage rate. However, such an application is very hard to deploy and evaluate, so in this paper, we combine the existing infrastructure and commercial vehicle network to predict the traffic speed, which can be evaluated by the data we already have access to.

![Traffic Speeds across Urban Regions](image)

Fig. 11. Traffic Speeds across Urban Regions

### 5.2. Application Evaluation

We compare Speedometer with one real-world system and one state-of-the-art model. The **SZ-Taxi** System: Shenzhen government has a pilot program called TravelIndex to infer congestion levels on road segments for the convenience of its residents, which shows inferred traffic speeds in real time based on GPS data from all taxicabs in
Shenzhen [Transport Commission of Shenzhen Municipality 2014]. SZ-Taxi serves as a single-source model suitable for the situation where the multi-source data are not available. The TSE Model: TSE uses real-time and historical vehicle GPS data and contexts (e.g., physical features of roads) to infer traffic speed with a collaborative filtering [Shang et al. 2014]. For a fair comparison, we aggregate GPS data from taxicabs, buses, and trucks to feed TSE. TSE serves as a naive multi-source approach for the situation where multiple heterogeneous data sources are available, but the integration is at data levels. Differently, Speedometer uses five models, i.e., \( M_T \), \( M_B \), \( M_F \), \( M_C \) and \( M_S \) for an integration at model levels. We reset \( M_C \) and \( M_S \) to the same spatial granularity with \( M_T \), \( M_B \) and \( M_F \). In particular, \( M_C \) and \( M_S \) give the urban density at cell towers and transit station levels, which can be redistributed to road segment levels based on coverage areas of particular cell towers or transit stations. We assign numbers of residents inferred by \( M_C \) and \( M_S \) within a coverage area to all segments in this area. The number of residents assigned to a segments is proportional to the segment length. Further, we use DBSCAN to obtain categories for the similarity measurement. Finally, we investigate the impact of different contexts in our tensor decomposition by adjusting three model parameters, i.e., \( \lambda_1 \), \( \lambda_2 \), and \( \lambda_3 \), which control contributions of different contexts in our tensor decomposition with Eq.(5). The default setting is \( \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \frac{1}{4} \) where we consider all contexts and the regularization term equally.

We utilize 91 days of datasets from all infrastructures in Figure 4. We use a cross-validation approach to divide the data into two subsets: the testing set as streaming data, including the data for one particular day; and the historical set as historical data, including the data for the remaining of 90 days. For a particular day, if we use 10-min slots, at the end of the first slot, i.e., 12:10AM, we use models to infer the speed for the slot from 12:00AM to 12:10AM, based on both the “real-time” data from 12:00AM to 12:10AM in the testing set and all historical data in the historical set. We move the data in the testing set forward for 90 days, leading to 91 experiments. The average results were reported.

We test the models with Mean Average Percent Error (MAPE) as

\[
\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{T_i - \hat{T}_i}{\hat{T}_i} \right|
\]

where \( n \) is the total number of temporal-spatial combinations we tested. We test all models on 18 road segments under 10 min slots, which leads to \( 24 \times 60 \times 18 = 2592 \) combinations for a one-day evaluation. \( T_i \) is the traffic speed inferred by a model under a temporal-spatial combination \( i \); \( \hat{T}_i \) is the ground truth of the traffic speed under a temporal-spatial combination \( i \). An accurate model yields a small MAPE, and vice versa. We test models on these specific road segments because we have access to the ground truth of traffic speeds on these road segments. This ground truth is obtained by loop detectors in Shenzhen road networks, which are inductive loops installed in selected major road segments, and can detect metal and thus accurately detect vehicle speeds. Figure 12 gives the ground truth of traffic speeds about four road segments in Shenzhen.

We first compare all models to show results on four particular road segments and the average result on all road segments. Then, we study impacts of inference slot lengths. Further, we investigate the impact of historical data sizes on the running time and the accuracy of Speedometer to show its feasibility and robustness for the real-time inferences. Finally, we present an evaluation summary.

5.2.1. Accuracy on Road Segments

Figures 13, 14, 15 and 16 plot the MAPE under 10-min slots for four major road segments (i.e., Nantou, Tongle, Fulong and Shennan) in Shenzhen urban area. The first three road segments are in uptown, and the last road segment is in downtown. In general, Speedometer outperforms TSE, which out-

ACM Transactions on Cyber-physical Systems, Vol. V, No. N, Article A, Publication date: January YYYY.
performs SZ-Taxi. This is because SZ-Taxi only considers taxicabs to infer the speeds, which leads to high MAPE, e.g., the early morning in Nantou as in Figure 13. Though TSE uses all data from commercial vehicles, it does not consider other indirect density models. Thus, when the GPS data are not available during certain temporal-spatial combinations, its MAPE is high, e.g., the early morning in Tongle as in Figure 14. For road segments where vehicles are abundant, these three models have the similar MAPE, e.g., the early morning in Fulong as in Figure 15. In general, the performance gain between Speedometer and others is lower during the daytime and the road segments in downtown, e.g., Shennan in Figure 16. This is because taxicabs and other commercial vehicles are abundant and thus quite representative in downtown during the daytime, so all models have better performance.

Figure 17 gives the average MAPE for all road segments under 10-min slots during 24 hours. The MAPE of all three models are typically higher than the MAPE we found...
in Figures 13, 14, 15 and 16. This is because the traffic speed may change dramatically between road segments, and some remote road segments with few vehicles uploading GPS data lead to higher MAPE. But the relative performance between the three models is similar. Speedometer outperforms TSE by 24% on average, and the performance gains are more obvious in the regular daytime, which may result from the consideration of density models. Speedometer outperforms SZ-Taxi by 29%, resulting from its integration of the multiple models.

5.2.2. Impact of Slot Lengths. Figure 18 plots the MAPE of all models with different slot lengths with a default value of 10 mins. The MAPE of all models reduces with an increase in the lengths of the time slots, because in a longer slot we accumulate more data about vehicles, and the traffic speed becomes more stable. Speedometer outperforms TSE and SZ-Taxi significantly if the slot is shorter than 30 mins, which results from the consideration of density models. But when the slot becomes longer than one hour, all models have similar performance, because in such a long slot, all models have enough data for an accurate inference about relatively stable speeds.

5.2.3. Impact of Historical Data. In this subsection, we study the impact of historical data on model accuracies and running times by comparing Speedometer to TSE with a default value of 13 weeks. We did not consider SZ-Taxi, since running times for this model are unknown. Normally, the more the historical data, the more accurate
the models, the lower the MAPE error and the longer the running time. Figures 19 and 20 plot running times and MAPE on different lengths of historical data in terms of weeks. Speedometer has 18% longer running time, which in turn leads to an 29% lower MAPE. This is because Speedometer has to perform its integration involving heterogeneous models, which takes time to calculate the model similarity.

5.2.4. Evaluation Summary. We have the following observations: (i) The inference accuracy is highly dependent on both locations and times as shown by Figures from 13 and 17. On average, all models have better performance in more dense area during the daytime, due to the abundance of the data to feed models. (ii) The length of slots has a significant impact on the performance of all models as in Figure 18. It is intuitive that a longer slot has lower error rates, yet it also reduces the practicality for real-time applications. (iii) As in Figures 19 and 20, the model integration takes a longer running time especially when the historical dataset is big, but it increases the accuracy. A good tradeoff between accuracy and running times has to be designed based on domain knowledge and user preferences. (iv) Looking across different factors, we found that slot lengths have the largest impact, and then locations and times, and finally historical data sizes.

6. RELATED WORK
Three types of work are related to our UrbanCPS: (i) models based on single-source urban infrastructure data, (ii) theoretical ensemble of multiple models, and (iii) data mining based on traffic flow theory.

6.1. Models based on Single-Source Data
Numerous novel models and systems have been proposed based on various urban infrastructure data to improve urban efficiency. We focus on the work closely related to models based on vehicular GPS, cellphone and transit data. Based on GPS data, many models and systems are proposed to benefit various urban residents: estimating traffic volumes or speeds for regular drivers [Aslam et al. 2012]; assisting regular drivers to improve their driving performance [Yuan et al. 2011a]; detecting anomalous taxicab trips to discover driver fraud for taxicab operators [Zhang et al. 2011]; estimating cellphone users’ travel range [Isaacman et al. 2011]; querying expected duration and fare of a planed taxi trip for taxicab passengers [Balan et al. 2011]; inferring gas consumption at road segment levels [Shang et al. 2014]; enabling us to better understand region
functions of cities [Yuan et al. 2012]; discovering temporal and spatial causal interactions to provide timely and efficient services in certain areas with disequilibrium [Liu et al. 2011] [Huang and Powell 2012]; allowing taxicab passengers to query the expected duration and fare of a planned trip based on previous trips and query real-time taxicab availability to make informed transportation choices [Wu et al. 2012]; recommending optimal pickup locations or routes [Ge et al. ] [Yuan et al. 2011b]; learning the dynamics of arterial traffic from probe data [Hofleitner et al. 2012a].

Further, many methods have been proposed for the study of human density and mobility based on cellphone CDR data, e.g., identifying cellphone users’ important locations [Isaacman et al. 2010]; modeling how cellphone users move [Isaacman et al. 2012]; predicting where cellphone users will travel next [Dufková et al. 2009]. Finally, transit GPS data are another important source for research in human density and mobility, e.g., identifying passenger locations based on data from taxicabs [Ganti et al. 2012], buses [Bhattacharya et al. 2013], and subways [Lathia and Capra 2011].

To our knowledge, we are the first to store such a large multi-source dataset, and then build models based on single-source sparse data, and finally systemically integrate these models from a complimentary standpoint. Obviously, the key difference of our work is that our model integration is built upon these models based on single-source data, and then effectively integrates them for better performance.

6.2. Theoretical Ensemble of Multiple Models

Our integration approach is inspired by several studies in the data mining community proposed to theoretically combine different models to improve their performance [Li et al. 2014] [Xie et al. 2014] [Gao et al. 2009] [Gao et al. 2008]. However, these studies are mostly under perfect conditions, e.g., the models are based on the complete data and directly-relevant data [Xie et al. 2014]. Differently, our work is focused on models based on the imperfect data, e.g., sparse and indirectly-relevant data. Further, semi-supervised learning also addresses issues related to imperfect data, e.g., the unlabeled data, but the models in these work are mostly based on same domain knowledge, e.g., similar weather data from different websites [Li et al. 2014] or similar email data from different users [Gao et al. 2009]. In contrast, our approach is to combine much more diverse models, i.e., speed models and density models, based on various urban infrastructure data. In addition, most studies on model integration in the data mining community are based on small-scale data, so their computation is often complex for better performance [Xie et al. 2014], e.g., computing inverse matrices and conducting non-linear programming, which is undesirable for real-time applications based on large-scale urban infrastructure data. Differently, the similarity measurement in our model integration is optimized for computation efficiency, which makes our work suitable for real-time applications.

Our work combining different data sources in urban systems is conceptually similar to the sensor fusion [Crowley and Demazeau 1993]. But the key difference is that for the sensor fusion community, the data used are collected for their models and almost all data are labeled data with directly relevant measurement, and essentially they are homogeneous data. But for our data, they are collected for the billing and management purposes, e.g., vehicle GPS, smartcard data and cellphone data, so some of them are only indirectly related to our model, and so our data are heterogeneous data. The heterogeneity of our data makes our modeling process significantly different from the sensor fusion.

6.3. Data Mining based on Traffic Flow Theory

There have been many studies to estimate and predict macroscopic traffic states (i.e., flow, density, and speed) at a very fine spatio-temporal scale while utilizing the power
of traffic models. Table III systematically compares those studies with UrbanCPS in terms of the used traffic model, data sources, final estimation states, online vs offline estimators, spatial and temporal scales.

[Deng et al. 2013] adopt the Newell-type traffic model to explain a perturbation in traffic flow. They use heterogeneous data sources, including loop detector counts, AVI Bluetooth travel time readings, and GPS samples, to estimate macroscopic traffic states on a freeway segment. They focus on the offline traffic state estimation using the Kalman filters (KFs) to construct a generalized least square estimator. [Nantes et al. 2016] use the first order Lighthill-Whiteham-Richards (LWR) model to explain shock waves that propagate upstream of the intersections in urban contexts. They build up a real-time (i.e., online) traffic prediction model employing the ensemble KFs using data from multiple sources incrementally, whenever they become available. [Work et al. 2010] estimate traffic states based on the LWR model using a Monte Carlo based ensemble KFs. [Hofleitner et al. 2012b] estimate traffic states in arterial networks using sparsely observed probe vehicles. They construct a dynamic bayesian network (DBN) to learn traffic dynamics from historical data and to perform real-time estimation with streaming data. [Herrera et al. 2010] perform a field experiment to show that a 2-3 % market penetration of cell phones is enough to provide accurate measurements of the speed of the traffic flow. They also address concerns, including communication load, device energy consumption, and privacy, existing in collecting GPS data by proposing an appropriate sampling strategy. [Xing et al. 2013] solve an information-theoretic sensor network design problem to minimize total travel time uncertainties. Based on a KF structure, uncertainties are quantified considering several error sources in the travel time estimation process. The above studies are built for lane-based traffic in developed countries. However, in developing regions, heterogeneous vehicles are driven on the same road (i.e., unlaned traffic). [Garg et al. 2014a] propose a smartphone sensor based system to categorize vehicles into four categories: two-wheeler bikes, three-wheeler auto-rickshaws, four-wheeler cars, and public transport like buses for develop-
oping regions. [Sen et al. 2013b] estimate traffic density and speed in unlaned traffic using image processing tools.

Those studies integrate multiple heterogeneous traffic data to build a single prediction model, and they are based on small-scale data (e.g., location data between upstream boundary and downstream boundary). In contrast, UrbanCPS integrates multiple heterogeneous models based on multi-source and large-scale sparse data.

7. CONCLUSION

In this work, we design and implement UrbanCPS to effectively integrate heterogeneous models based on multi-source infrastructure data. Our endeavors offer a few valuable insights which we hope will allow fellow researchers to utilize our system for not only model integration but also real-world applications. Specifically, these insights are that (i) heterogeneous models based on different urban infrastructure data provide different yet complimentary view for the same urban phenomenon, and thus an effective integration of them would boost the model performance; (ii) for many urban phenomena, indirectly-relevant models are often powerful to address the issue of directly-relevant models, e.g., sparse data, but we need an effective method to integrate them with direct-relevant models; (iii) though difficult to be obtained, the ground truth data about urban phenomena are vital for both model designs and evaluations. (iv) while it is challenging to integrate heterogeneous models, it is more challenging to negotiate with service providers for large-scale infrastructure data to feed models.

REFERENCES


Raghu Ganti, Mudhakar Srivatsa, Anand Ranganathan, and Jiawei Han. 2012. Inferring Human Mobility Patterns from TaxiCab Traces (UbiComp ’13).

Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. 2008. Knowledge Transfer via Multiple Model Local Structure Mapping. In ACM KDD’08.

Jing Gao, Wei Fan, Yizhou Sun, and Jiawei Han. 2009. Heterogeneous Source Consensus Learning via Decision Propagation and Negotiation. In ACM KDD’09.


Yong Ge, Hui Xiong, Alexander Tuzhilin, Keli Xiao, and Marco Gruteser. An energy-efficient mobile recommender system (KDD ’10).


Yan Huang and Jason W. Powell. 2012. Detecting regions of disequilibrium in taxi services under uncertainty (SIGSPATIAL ’12).


Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation. In ACM SIGMOD’14.


Sihong Xie, Jing Gao, Wei Fan, Deepak Turaga, and Philip S. Yu. 2014. Class-Distribution Regularized Consensus Maximization for Alleviating Overfitting in Model Combination. In ACM KDD’14.

ACM Transactions on Cyber-physical Systems, Vol. V, No. N, Article A, Publication date: January YYYY.


Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. 2011a. Driving with knowledge from the physical world (KDD ’11).


