Fairness-aware Spatio-temporal Prediction for Cities

An Yan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Bill Howe, Chair

Carole L. Palmer

Cynthia Chen

Program Authorized to Offer Degree:
Information Science

University of Washington

Abstract

Fairness-aware Spatio-temporal Prediction for Cities

An Yan

Chair of the Supervisory Committee:
Associate Professor Bill Howe
Information School

Equitable prediction is of critical importance to urban applications such as transportation resource allocation and predictive policing. Decisions based on unfair predictions may lead to inequitable service distribution or impose disproportionate impact on underrepresented minorities. Machine learning based spatio-temporal prediction models have been widely adopted in urban settings, but few of them have built fairness into their design. This dissertation is a pioneering work to explore a suite of fairness-aware spatio-temporal prediction methods for cities, including measuring fairness for urban applications (metrics), designing fairness-aware spatio-temporal prediction algorithms (algorithms), learning bias-free data representations (data), and evaluating fairness-aware systems for real-world applications (applications). Specifically, I propose *FairST*, a fairness-aware spatio-temporal prediction model based on 3D convolutional neural network. A key feature of FairST is the integration of fairness regularizers to the model to encourage equitable prediction. I also propose two fairness metrics that measure equity gaps between social groups for urban mobility systems. Experiments on two real-world new mobility datasets demonstrate that FairST is able to close more than 80% of fairness gap while achieving *better* accuracy than state-of-the-art but fairness-oblivious baseline methods. Further experiments show that FairST is able to reduce unfairness for multiple attributes without sacrificing much accuracy. I propose an unsupervised algorithm framework to learn fair, accurate, and reusable (FAR) data

representations, *the EquiTensors*, for heterogeneous and multi-dimensional urban datasets. Experiments with 23 input datasets and 4 real applications suggest that EquiTensors could help mitigate the effects of the sensitive information embodied in the biased data. Meanwhile, applications using EquiTensors outperform models that ignore exogenous features and are competitive with "oracle" models that use hand-selected datasets. EquiTensors can be trained and released by government agencies or trusted data brokers over both public open data and unreleased data. It presents a novel way to allow downstream applications a means of improving accuracy, avoiding data discovery and pre-processing, and limiting their exposure to new sources of discriminatory bias. This dissertation will make methodological contributions to urban data science and machine learning research. The proposed methods will inform the development of fairness assessment measures and bias-removal strategies for stakeholders such as public resource/service distributors and government agencies, allowing for intelligent and responsible decision-making that benefits all citizens.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my chair and advisor, Dr. Bill Howe, whose unreserved support and guidance made this dissertation possible. It has been a great pleasure working closely with him over the past few years. Bill, thank you for all the patience, encouragement, inspiration, and freedom that you gave me to help me grow as an independent researcher.

I would like to express my deep gratitude to Dr. Carole L. Palmer for her extremely generous support over my five and half years at the University of Washington. I have learned so much from her in research, teaching, writing, project management, and many more.

I am very grateful to Dr. Cynthia Chen for the inspiring conversations and very helpful suggestions. I would also thank Dr. Xuegang Ban for his insightful comments on my research and being my GSR.

Special thanks also to Dr. Nic Weber for his helpful advice in research and teaching. Working with him has been an enjoyable experience. I also owe my sincere thanks to Dr. Amy Ko. She advised me on my first first-author publication in Information School.

Additional thanks to my fellow PhD students at the Information School for their friendship and support during my PhD study.

This work is dedicated to my family.

Chapter 1
INTRODUCTION

Predicting dynamic urban activities such as energy consumption, air pollution, public safety, travel time, and traffic flows has become a fundamental task for the public and private sector. For example, mobility system operators such as ride-hailing and bikeshare companies often use accurate demand estimates to guide resource optimization and maximize system utility [146, 85, 9, 115]. Law enforcement agencies increasingly rely on crime prediction to deploy police forces [124].

Many modern prediction systems for cities are based on machine learning and urban data. However, machine learning algorithms can be discriminatory because they can reproduce and magnify the biased signals embedded in the data [186, 81, 6, 26]. Increasing evidence has shown that algorithms may produce unfair predictions for urban applications. Recent studies show that algorithms that distribute app-based mobility services may discriminate against people of color [15, 50]. For example, influenced by Uber's pricing algorithm, neighborhoods with more white people experienced higher service quality [132]. Similar concerns were raised in public safety domain. For example, one study [101] revealed that a widely-used predictive policing tool, PredPol, would reinforce the bias in the police records, resulting in disproportionate policing of minority communities. Moreover, decisions informed by unfair predictions are likely being incorporated into the prediction models as the ground truth, producing a negative feedback loop and reinforcing the structural inequity [26].

This dissertation aims to incorporate fairness into urban prediction systems. The core enabling methodologies are related to two research areas: fairness in machine learning and spatio-temporal prediction for cities. The emerging field of fairness in machine learning seeks to identify biases and remove biases embodied in data or algorithms. Spatio-temporal

prediction research focuses on developing algorithms that can capture the spatial and temporal dynamics of urban activities. However, fair machine learning community invests almost exclusively in methods, which are only tested in cases such as loan distribution, college admission, and advertising so far. There is a dearth of research on designing fairness-aware approaches for spatio-temporal context and urban applications. Furthermore, to the best of our knowledge, few existing work in predicting urban activities considers fairness in their solutions. This dissertation bridges the gap between fairness in machine learning and spatio-temporal prediction for cities. Based on real-world urban data and applications, I intend to explore several challenging yet open questions: How to accurately model the spatio-temporal dynamics of urban activities (utility)? How to measure the fairness of spatio-temporal predictions in various urban applications (metrics)? How to discover and remove discriminatory signals from urban data (data)? How to design fairness-aware machine learning algorithms (algorithms)? The goal of the study is to propose spatio-temporal prediction methods that can achieve fairness and utility at the same time, providing building blocks for cities to create equitable and efficient decision-making systems that benefit all citizens.

1.1 Definitions

Before proceeding to the rest of the dissertation, there are a number of important terms, including *fairness*, *discrimination*, and *bias*, that deserve clarification. These terms may be ambiguous to readers or have controversial meanings. This section first reviews the definitions of fairness and discrimination, and discusses methods to measure fairness in various contexts. It then focuses on the fairness definitions and metrics that this dissertation is grounded in. The final part of this section reviews the definitions of two types of bias that are common in urban predictions and urban data.

1.1.1 Discrimination and Fairness

Discrimination is generally understood as disadvantageous treatment of an individual based on his or her membership in protected group(s) (e.g., race and age) [186, 110, 6]. *Fairness*

is broadly considered as "the absence of discrimination" [76].

There are many ways to measure fairness, depending on the context. For example, Title VII - equal employment opportunity under the United States Civil Rights Act forbids discrimination based on sensitive attributes such as sex and color in employment [64]. To enforce Title VII, the "80% Rule" was advocated by the US Equal Employment Opportunity Commission to detect *disparate impact* in employee selection procedures. The 80% Rule states that if the selection rate for minorities is less than 80% of the rate of non-minorities, the procedure is considered to be discriminatory [36, 6].

In academia, the implementation of the Civil Rights Act in 1964 ushered in a plethora of research on discovering and measuring discrimination and fairness in many fields such as employment [37], education [28], transportation [63, 50], and housing [169], etc. Many studies examine fairness using statistical tests like regression slope test [169] and t-test, which typically check whether the differences in target variables (e.g., salary) between demographic groups are significant [186]. For example, Hughes and MacKenzie [63] investigated the relationships between wait times for UberX and socioeconomic indicators in Seattle using regression methods. The underlying fairness definition of this line of approaches is that the target variables (e.g., hiring decisions, transportation resources) should not be (unjustifiable) associated with the sensitive attributes.

Recently, the use of automatic decision-making systems powered by machine learning and big data has become prevalent across application domains (e.g., loan distribution, online ads delivery). Fairness metrics such as statistical parity [35] and equalized odds [57] have been the mainstream way for measuring fairness in machine learning. Fairness metrics have several advantages: 1) Metrics can indicate the extent of discrimination [11, 78, 186]. 2) It is generally straightforward for machine learning models to incorporate fairness metrics as additional objectives to their optimization process. 3) Metrics don't rely on strict assumptions like independence and random sampling that statistical tests require [186].

We focus on fairness metrics used in machine learning research, as it is most pertinent to the methods developed in this dissertation. We now review the underlying fairness definitions

of different metrics used in machine learning.

There are mainly two popular fairness definitions in machine learning research. For *Group fairness*, it asks for parity or approximate parity of some measures across protected groups [26, 35]. For example, *statistical parity* asks for independence between predicted outcome (e.g., hiring decisions) and sensitive attributes (e.g., gender, race). *Equalized odds* [57], sometimes referred to as *separation* [4], requires equal *true positive rates* and *false positive rates* across two groups. This definition stresses that the model should have equal accuracy across groups instead of performing well only on the majority. *Equal opportunity* is similar to equalized odds [57], but it only requires equal true positive rates. There are many other variants of group fairness metrics such as *calibration* and *predictive parity*. Interested readers can refer to [5, 25] for detailed discussions around each of them. *Individual fairness* says that "any two individuals who are similar with respect to a particular task should be classified similarly" [35]. To enforce individual fairness, "similarity" between a pair of individuals from advantaged and disadvantaged demographic groups respectively has to be defined. Group fairness and individual fairness conflict with each other when there is a large dissimilarity between two demographic groups. For example, when making hiring decisions based on education level, if members of a demographic group A are over-represented in low education and under-represented in high education, enforcing individual fairness would result in a lower hiring rate for members in group A than B, thus violating group fairness [35].

There is no agreed upon definition for fairness so far. **The fairness definition in this dissertation is grounded in the notion of *Group Fairness***. The metrics developed in Chapter 3 are based on the idea of statistical parity, and the metrics in Chapter 4 reflect the spirit of equalized odds and equal opportunity. Group fairness is straightforward and aligns well with equity visions of the government, such as equitable distribution of urban resources across demographic groups in cities [118]. In particular, there is a correspondence between *group fairness* and *vertical equity*, a concept in transportation equity literature. *Vertical equity* is about allocating resources to individuals or groups that differ in income, social class, mobility need, or ability [89, 17]. When there is an uneven distribution of

transport supply across different socioeconomic groups in cities, vertical equity encourages compensating for such inequalities by policies favoring disadvantaged groups[89]. This aligns with group fairness (statistical parity) that the level of transportation supply in a city should be the same across different groups.

1.1.2 Bias

Another term that is frequently used in this dissertation is *bias*. Suresh and Guttag [135] identified six types of bias that may arise from automatic decision-making systems and their input data: historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, and deployment bias. This dissertation is mainly concerned with *historical bias* and *measurement bias*, as they typically arise in urban data. Historical bias refers to "a misalignment between world as it is and the values or objectives to be encoded and propagated in a model" [135]. It is not brought by sampling or measurement, but rather by the world's status, which is shaped by human practices. For example, a road network map can accurately represent the locations and types of road segments in a city, but it reflects the historical transportation infrastructure development strategies of a city, which might not be favorable to disadvantaged communities. Housing prices, points of interest, and traffic flows data may also embody this type of bias. Measurement bias often happens when proxy data is used to approximate the variable of interest, which is not directly observable [135]. For example, the number of trip starts is often used as a proxy for ride-share demand, but the demand of people who cannot use this service (e.g., people who do not have smartphones or internet access) is not captured in the data. Similarly, police incident reports or arrest statistics are often used as a proxy for crime data, but they may reflect human bias rather than the true distribution of crime [143, 135].

1.2 Research Objectives

This study aims to explore a suite of fairness-aware spatio-temporal prediction methods for cities. There are four main research objectives: Modeling the spatio-temporal dynamics of

urban activities (utility), measuring fairness for spatio-temporal settings (metrics), integrating fairness into prediction models (algorithms), and learning fair data representations from heterogeneous urban data (data).

RO1: Modeling the spatio-temporal dynamics of urban activities. Urban activities such as mobility demand usually exhibit complex spatial and temporal patterns, and are influenced by many external factors such as weather and road network [85]. Therefore, the keys to accurate predictions are the modeling of spatio-temporal dynamics and the incorporation of relevant exogenous data. Capturing the spatio-temporal patterns is challenging because of the complex non-linear interactions of multiple variables in space and time. These dynamics are usually beyond the capability of traditional methods such as time series analysis (e.g., ARIMA) or conventional machine learning methods (e.g., Naive Bayes). Another prominent challenge lies in data integration. As open urban data with various dimensions (e.g., temporal, spatial, or spatio-temporal) and formats (e.g., csv, shapefile) becomes increasingly available, how to extract meaningful information from these heterogeneous data sources and integrate them into a single prediction model presents a challenge [185].

This dissertation proposes a generic framework for modeling urban activities based on convolutional neural networks (CNN). We partition a city into a regular grid and slice a study period into equal time intervals, forming a spatio-temporal grid (or tensor). We aggregate urban activities (e.g., criminal incidents) and all exogenous features into the common grid. We use a 3D convolutional neural network (3D CNN) as the core building block in our model to capture spatio-temporal dynamics. To incorporate the exogenous features such as weather and traffic, we adopt a three-stream model architecture that fuses together 1D, 2D, and 3D convolutional layers, respectively. A 1D CNN is used to extract information from 1D temporal features such as city-wide temperature, and a 2D CNN is used to extract information from 2D spatial features such as house prices. This novel architecture is applicable to a wide range of spatio-temporal prediction problems and can accommodate arbitrary number of exogenous features.

RO2: Measuring fairness for spatio-temporal settings. Although the specification

and assessment of fairness metrics is an active research area [52, 64], most of the proposals are inapplicable to spatio-temporal settings. First, the prediction targets in spatio-temporal settings (e.g., ride-hailing demand) are typically continuous whereas many fairness metrics such as statistical parity and equalized odds are designed for discrete classification settings [35, 57]. Second, in spatio-temporal settings, each record is usually a geographic area representing an entire subpopulation, and therefore cannot necessarily be assigned a specific attribute value (e.g. white), but rather a percentage of the subpopulation that has that value (e.g. percentage white). Among the few studies that propose fairness metrics for regression, a categorical sensitive attribute is typically required [18, 11]. Finally, fairness assessment can be application specific. For example, ride-hailing companies may consider examining fairness on a per capita basis as the demand is typically positively associated with zonal population [39]. However, fairness metrics thus far are generic across applications. To address these challenges, this dissertation proposes to adapt a set of exiting fairness metrics for spatio-temporal settings and discusses their variations for specific urban applications.

RO3: Integrating fairness into the prediction model. Machine learning algorithms can replicate and amplify the biases encoded in the training data [26, 19]. Researchers have proposed a variety of remedies that occur at all stages of machine learning pipeline, including pre-processing [71], training [35, 176], and post-processing [38, 57]. Nevertheless, existing studies largely focused on classification settings using conventional machine learning methods and rarely considered spatio-temporal applications driven by neural networks. Moreover, current literature provided limited theoretical discussion or empirical experiments on addressing multiple sensitive attributes scenarios. This dissertation bridges these research gaps by integrating fairness in spatio-temporal predictions using deep neural networks. Specifically, we incorporate fairness metrics as additional constraints (i.e., regularizers) in the loss function of the prediction models. Multiple sensitive attributes can be naturally handled via a weighted sum of fairness regularizers for each attribute in the loss function. We discuss the trade-offs between accuracy and fairness through experiments on real-world datasets.

RO4: Learning fair data representations from heterogeneous open urban data.

The use of relevant exogenous features can improve prediction accuracy [181, 174]. In urban applications, many prediction problems are potentially influenced by a common set of spatio-temporal factors (e.g., weather, housing prices, traffic, road networks). For example, predicting bikeshare demand depends on weather, topography, and traffic [160, 115], but the same data sources are also helpful for predicting citywide crowd flow and accident patterns [181, 174]. Researchers in urban data science often rely on open datasets to produce exogenous features for their predictive analysis or model development [164]. However, open data are often too noisy, incomplete, and biased to use directly in research [171, 67]. In particular, most urban datasets are polluted by systemic socioeconomic and racial discrimination. For example, housing prices reflect historical discriminatory urban development policies [8], public safety data reflects racist policing practices [126], and transportation data reflects biased policies toward wealthy neighborhoods [123]. These sources of bias are propagated into prediction tasks, resulting in unfair predictions [186, 6] and exacerbating structural inequity. Nevertheless, thus far there is no generalized approach for integrating open urban data in support of fair spatio-temporal prediction.

Based on the aforementioned observations, we propose an unsupervised algorithm to learn fair data representations from a wide range of commonly used urban datasets, collected mostly from open data repositories. The learned data representations can be used as exogenous features in many downstream applications, and are fair, accurate, and reusable (FAR principles). Such data representations could potentially help the downstream tasks to achieve better accuracy while limiting their exposure to additional discriminatory signals. Meanwhile, they do not require access to the raw data and the training pipelines of the downstream tasks. The proposed method is based on a convolutional denoising autoencoder that learns an integrated representation from heterogeneous multi-dimensional urban datasets, an adversarial model that learns to detect a sensitive attribute (race, income, etc.) from the learned representation, and a disentangling module that further separates the sensitive attribute from other from other information in the latent space [62, 80, 106]. The fairness approach combines adversarial learning [159, 157, 128] and learning disentangled

representations [96, 29], but adapt them for an unsupervised data integration setting with continuous and distributed sensitive attributes (e.g., percent of high income residents in a region) as opposed to categorical attributes. To enhance the fidelity of the representation, this dissertation explores an adaptive weighting scheme inspired by recent work in multi-task learning [100, 168, 114, 94, 22].

1.3 Contributions

This dissertation makes methodological contributions to urban data science and machine learning research. It is a pioneering work to explore a suite of fairness-aware spatio-temporal prediction methods for cities, focusing on five aspects: utility, metrics, algorithms, data, and applications. The proposed methods will provide fairness assessment measures and bias-removal strategies for stakeholders such as public resource distributors and government agencies, allowing for intelligent and responsible decision-making that benefits all citizens. The main contributions are summarized as follows:

- **A generic framework for accurate spatio-temporal prediction.** We propose a generic spatio-temporal prediction algorithm based on 3D convolution neural network (3D CNN) to model complex spatio-temporal dynamics and integrate exogenous features. We test its effectiveness with two mobility demand prediction cases (Chapter 3) and two public safety cases (Chapter 4). We show that our method outperforms traditional methods and several state-of-art deep learning based methods.

- **Novel fairness metrics.** Drawing on existing work in quantifying fairness in the context of machine learning, we describe two types of fairness metrics for spatio-temporal settings in Chapter 3 and Chapter 4, respectively. The first type assumes that the prediction values should be independent of the sensitive attributes, whereas the second type assumes the prediction errors should be independent of the sensitive attributes. In Chapter 3, we propose two metrics: *region-based fairness gap (RFG)* and *individual-based fairness gap (IFG)* for urban applications. Both metrics measure the gap be-

tween mean per capita prediction values (e.g., taxi demand) across demographic groups (e.g., white and non-white groups) over a certain period of time. The difference lies in that RFG focuses on discrete sensitive attributes while IFG deals with continuous attributes. We show that prediction algorithms without fairness treatment can be discriminatory against underrepresented minorities. Chapter 4 adapts three fairness metrics to spatio-temporal settings: residual difference (RD), positive residual difference (PRD), and negative residual difference (NRD). RD, PRD, and NRD measure the gap between the prediction errors across groups. We show that the use of exogenous features may introduce additional biases to the prediction results.

- **A supervised spatio-temporal prediction algorithm to remove biases during model training.** Based on the aforementioned prediction framework and fairness metrics, Chapter 3 proposes **FairST**, a fairness-aware spatio-temporal prediction framework. It aims to enforce independence between the predictions and the sensitive attributes. FairST integrates RFG or IFG as regularizers into the loss minimization pipelines to encourage fair prediction. It does not require bias-free training data and can handle multiple sensitive attributes at the same time. Experiments on two real-world new mobility datasets show that FairST is able to close more than 80% of fairness gap for a single sensitive attribute and at the same time achieve better accuracy than state-of-the-art but fairness-oblivious baseline methods. Further experiments show that FairST is able to reduce unfairness for multiple attributes, outperforming baselines in both accuracy and fairness.

- **An unsupervised algorithm for fair integrations of heterogeneous urban datasets.** In Chapter 4, we propose an end-to-end model to learn fair data representations, called **EquiTensors**, from a wide range of commonly used urban datasets. EquiTensors aim to improve the downstream prediction accuracy without introducing much extra biases. We introduce FAR (Fairness, Accuracy, and Reusability) principles for evaluating such pre-trained data representations. The proposed algorithm consists

of two main components: an integrative model that summarizes a large number of multi-dimensional urban datasets as a single data representation and a fairness module based on adversarial learning and disentangled representation to remove discriminatory effects from the learned representation. Our experiments show that EquiTensors can help mitigate the effect of sensitive demographic information embodied in the datasets and at the same time help the downstream tasks to achieve prediction accuracy comparable to an 'oracle' networks that trained with hand-selected features.

- **Empirical evaluation of the proposed methods on real-world applications.** We evaluate FairST with two mobility demand prediction cases in Seattle and Austin, respectively (Chapter 3). We produce "EquiTensors" for Seattle, and evaluate their effectiveness on four urban applications in mobility and public safety domains (Chapter 4). Overall, our experiment results suggest that with the proposed methods, it is possible to achieve fairness and utility at the same time.

1.4 Dissertation Organization

The dissertation is structured as follows: Chapter 1 provides an overview of the dissertation. Chapter 2 provides study background and reviews the related literature. Chapter 3 addresses RO1, RO2, and RO3. A fairness-aware supervised spatio-temporal prediction algorithm, called *FairST*, along with two new fairness metrics is introduced. Chapter 4 addresses RO2 and RO4. An unsupervised algorithm, *the core integrative model*, is introduced to learn reusable data representations from heterogeneous open urban datasets. Chapter 4 also describes three fairness metrics adapted to the spatio-temporal setting, and presents the *EquiTensor* architecture on top of the core integrative model, for learning fair representation for heterogeneous urban data. Chapter 5 concludes the dissertation and discusses the limitations and future directions.

Chapter 2
BACKGROUND AND RELATED WORK

This chapter presents study background (Section 2.1) and reviews the related literature (Section 2.2). The background section first provides a broad overview of fairness issues resulting from the increasing adoption of automatic decision systems in the public and private sector. It then focuses on the fairness issues in urban data and applications that motivate the development of fairness-aware methods for cities, and examines examples of discriminatory data-driven applications and biased urban datasets in resource/service allocation and public safety. The third part of the background section discusses the potentials and challenges to the use of open data, which motivate the development of fair integrations of multi-source data to provide safer and easier access to open data, broadening the utility of open data portals. Two of my own studies, one on the usage pattern of open government data and another one on data reuse practices in Earth System Science, are presented as additional context. The literature review section examines data science for cities including spatio-temporal prediction algorithms and learning representations for urban data, and fairness in machine learning including discovering and measuring unfairness and fairness-aware machine learning methods.

2.1 Background

2.1.1 Fairness, Accountability, and Transparency of Automated Decision Systems

Automated decision systems powered by machine learning and big data have been widely employed in many applications including credit scoring, criminal justice, online advertising, employment, etc. [186, 120, 112, 126]. These systems have been hailed as efficient, objective, and accurate alternatives to human decision-makers [4]. However, increasing evidence has

shown that data-driven systems contain biases. For example, Google's image recognition system wrongly identified black users as gorillas [56]. Amazon's same-day delivery services excluded predominantly black neighborhoods in many cities [65].

Researchers pointed out that even if the algorithms themselves are well-intentioned, they can replicate and amplify human biases encoded in the data, resulting in unequal distribution of impact across different demographic groups [186, 81, 6]. This is because machine learning algorithms seek to fit the training data as much as possible to make accurate predictions. The process of learning also "accurately" captures signals of discrimination [26]. In 2017, a study found that an influential language corpus [119] generated by machine learning algorithms accurately reproduced historic biases [19]. The corpus reflects societal stereotypes such as female names are more associated with family while male names are more associated with career. Not only do algorithms pick up discrimination in data, but they also magnify them [26]. This is often due to the fact that minority groups are underrepresented in training data, and that algorithms tend to fit more towards the majority groups, leading to higher error rates for the minorities [26]. One study revealed that a widely-used predictive policing tool, PredPol, would reinforce the bias in the police records, resulting in disproportionate policing of minority communities [101].

One case that has attracted tremendous public attention is the "COMPAS debate". COMPAS is a recidivism prediction tool powered by machine learning [25]. It estimates a risk score to represent the likelihood that a defendant will reoffend based on 137 survey questions. This software has been used to assess more than 1 million defendants [33]. The debate started by ProPublica reporting that COMPAS software discriminates against black defendant [2]. ProPublica's analysis indicated that COMPAS had notably higher *false positive rates* (FPR) for black defendants than for white defendants. Moreover, the likelihood of a recidivating white defendant being identified as low risk is twice as high as that of black defendants, meaning that COMPAS also shows lower *false negative rates* (FNR) for black defendants than white defendants. In confronting the public criticism, Northpointe, the company that developed COMPAS, argued that their software satisfied a fairness criteria called *predictive*

parity [32]. Later, researchers [43, 25] showed that it is not possible to satisfy predictive parity, equal FPR, and equal FNR at the same time.

This dissertation is situated in the broader context of social impact of automated decision systems, focusing on developing fairness approaches for urban application domain.

2.1.2 Equity Issues in Urban Data and Applications

Equity Issues in Data-driven Urban Applications. Cities are increasingly relying on automatic decision systems based on big data and algorithms, but evidence has shown that data-driven applications may discriminate certain groups of people. Two typical domains that have fairness concerns are resource/service allocation and public safety. The former involves mobility resource allocation, public facility placement, and online shopping, etc.[3]. The latter primarily involves predictive policing.

Mobility operators such as ride-hailing and bikeshare companies often rely on demand prediction to guide resource optimization [146, 85]. For example, Uber predicts demand to direct drivers to high-demand areas [9]. However, algorithms that distribute app-based mobility services may discriminate African American riders, resulting in longer waiting times and higher trip cancellation rates [15, 50, 132]. Similarly, evidence suggests that online shopping services provided by tech companies may discriminate against under-represented communities. For example, Amazon's same-day delivery services excluded predominantly black neighborhoods in many cities [65]. Underestimation of resource/service demand for underrepresented minorities may result in insufficient supply to them, which can produce a feedback loop: racial and income disparities are misinterpreted in the model as lack of demand, reinforcing reduced access to services/resources.

In public safety domain, data-driven decision making systems have been adopted in police forces and law enforcement agencies around the world [111]. These systems use historical data to predict crime incidences and identify risk areas in a city. Recently, studies have shown that predictive policing tools may discriminate people of color [101]. For example, a recidivism prediction tool called COMPAS was found to have notably higher *false positive rates* for

black defendants than for white defendants [2, 25]. The use of such systems may result in stigmatizing certain individuals and groups and disproportionately imposing negative impacts on them [111].

This dissertation seeks to alleviate the equity issues by proposing a fairness-aware machine learning algorithm (FairST) for cities. The proposed method is promising in helping the private or public sector to make well-informed decisions and meet equity goals at the same time.

Equity Issues in Urban Data. Discriminatory signals encoded in data is a primary source of bias in data-driven decision making systems [26]. It has become an obstacle for data sharing for both private and public sector [171]. For private data owners, privacy and biases are major concerns of releasing data for public use. Sharing the data as it is might run the risk of propagating biases [171]. Public sector and research communities have stronger motivations to share data than private sector. Open government data can increase government transparency and open scientific data can strengthen scholarly communication and improve research reproducibility [163, 164]. As a result, an overwhelming amount of urban datasets has been released to public domain. Most of these datasets contain location or temporal information (spatio-temporal datasets) and have been widely used to support urban science research and applications [164]. However, some datasets are polluted by systemic bias due to socioeconomic and racial discrimination. For example, housing prices reflect historical discriminatory urban development policies [8] and police incident reports usually reflect *more* about policing practices than criminal activity [126]. The direct use of such datasets in research or real-world applications may lead to biased results and even unintended consequences [111].

This dissertation proposes to remove discriminatory signals from spatio-temporal urban datasets through learning fair data representations (EquiTensors). We propose FAR principles (Fairness, Accuracy, and Re-usability) for evaluating such representations. FAR principles are related to but different from FAIR principles (Findability, Accessibility, Interoperability, and Re-usability) for data sharing [152] and FACT principles (Fairness, Accuracy,

Credibility, and Transparency) for responsible data science [141] in objectives and scopes.

2.1.3 The Reuse of Open Data

This dissertation is also motivated by the observation that the potential of open data has not been fully exploited [171, 55]. For example, one study found out that a majority of open data published by data.gov.uk has never been accessed [14]. Challenges to the use of open data are related to lack of metadata [108], data quality [67] (i.e. completeness, accuracy, and clarity), proprietary formats [108], data portal usability [172], and lack of centralized access point [40]. In particular, it is difficult to identify useful datasets across multiple open data portals; and the cleaning, management, and interpretation of open data is labor-intensive [40]. At the same time, data owners are reluctant to release certain types of information due to privacy concerns and biases in the dataset [171, 151].

This dissertation proposes to learn reusable and fair representations from a large number of urban open datasets to benefit an array of prediction tasks. This approach presents a novel way to allow downstream applications a means of improving accuracy, avoiding data discovery and pre-processing, and limiting their exposure to new sources of discriminatory bias. It also complements "raw" access through open data portals and potentially provides a single point of control for managing data use policies (e.g., for privacy). The rest of this section describes two background studies, focusing on the pattern of open government data usage [164] and data reuse in Earth System Science [163], respectively.

Exploring the use pattern of open government data. Open government data (OGD) refers to government-related data that is free and open for public use [150]. Despite a plethora of existing work on the potential of and intentions to use OGD, we still lack understanding of the actual use of it. Using a text mining approach, this study explores the use pattern of OGD in academic research by answering three research questions: 1) How are OGD used in academic research? 2) What sources of OGD do researchers use? 3) What fields of academic research are using OGD? Part of the reason why we focus on academic research is that the massive scientific publications provide clear and traceable evidence of

OGD usage through explicit citations and mentions; and it is difficult to collect data usage evidence from other users from the private sector. Nevertheless, the results in this study may allow us to peep into the general usage patterns of OGD that go beyond academic research.

Specifically, we developed a protocol to identify, select, and categorize published literature where open government data was explicitly used as a research input. We first developed a list of 302 open government data portals and their URLs based on the resources on data.gov web page. We searched the Scopus, Springer, and IEEE databases using the 302 data portal URLs and retrieved 2486 papers published from January 2009 to July 2017. A final study sample of 1229 papers were obtained after applying our inclusion and exclusion criteria. Our analysis of the study sample revealed several interesting findings:

- There is an upward trend of publications using OGD over time from 2009 to 2016, implying the increasing attention to OGD in scientific research. This may be because of the expanding open data holdings and the increasing awareness of OGD in academia and beyond.

- Chicago is ranked as the sixth most frequent source of OGD among the 302 sources, which include international and country level portals. Open data portals in New York City and Seattle are also among the top 30 sources, suggesting that open urban data is playing an important role in supporting research.

- Some open data portals had one particular type of dataset that had been used most frequently. For example, in the United Kingdom, an "index of deprivation" dataset was used frequently as an ancillary source for Medicine and Public Health research. For Chicago, crime data was the most used source for social, urban, and computer science studies. These popular datasets may be the most authoritative or only source to answer a research question. It is also possible that these datasets can help answer questions that are of great local interest. Interestingly, this may also imply that a vast majority of open data is underused.

- OGD was used by nearly all research fields, including Chemistry and Dentistry that seem less likely to use OGD. This suggests that OGD has exhibited more potential use for research than reflected in exiting literature. Medicine, Environmental Sciences and Social Sciences occupy the top three fields that use OGD in publications. Computer Sciences and Engineering rank No. 4 and No. 6 in the list, suggesting that OGD is contributing to technical innovations.

- OGD was primarily used as data sources for new research: 33.4% used OGD as main source and 19.5% as an auxiliary source. About one third of the papers use OGD for providing context of a study. Other usage types include testing new methods, providing new services or systems, result evaluation, and creating composite datasets.

Overall, our results suggest that OGD is a valuable source for scientific research of a wide range of fields. In particular, urban data portals such as Chicago, New York City, and Seattle have been used heavily to support social, urban, and computer science studies.

Examining Data Reuse in Earth System Science. Open data can promote reproducibility, openness, and innovation in science [121]. Sources of open data used by scientists include federally funded research data centers, disciplinary and institutional repositories, open government data, and private companies, as well as shared data offered by individual researchers (or research labs). Open data produced by one organization or individual can be reused to form new analysis without recollection, therefore accelerating new discovery. Moreover, open data can also be reused to reproduce an existing study, during the process of which the validity and rigor of the original study would be evaluated. This study reports on results from a survey examining how researchers reuse data to support new research and reproducing existing research, using Earth System Science (ESS) as a case study. ESS is data-intensive and highly collaborative, making it an informative site for us to examine cross-disciplinary data practices. The findings of this study provide a baseline on data reuse and reproducibility practices and perspectives in ESS, and will inform investments in research data services for supporting researchers in ESS and other interdisciplinary fields.

We conducted a survey of active ESS researchers from 126 U.S. universities and research centers, representing a wide variety of scientific fields. Over half of the 207 respondents had more than 20 years of research experience. The survey has twenty-four questions covering respondent demographics, experiences and practices reusing data and reproducing published results, views on how to improve the current state of practices, and the greatest challenges particular to ESS. We summarize here some interesting findings related to the context of this dissertation.

- Data reuse practices were strong across the sample, with 73.0% reporting that they always or often use data generated by others, and only 1% indicating they never do. More than seventy percent (70.3%) of respondents use data from federally funded data centers always or often, with lower levels of data use from three other sources — data supplements to published papers (30.8%), government open data portals (26.4%), and researcher websites (20.5%).

- Data was most frequently reused for "new analysis" (87.0%), followed by comparing results (70.4%), providing information (53.7%), and testing or developing new methods of analysis (52.5%). Only 18.5% reported reusing data for reproducing published studies.

- More than 80% of the respondents reported needing research support services. Technical support for tools was rated the highest (59.2%), followed by consulting services on data management and data sharing (54.4%) and assistance with data preprocessing and cleaning (48.1%).

- Challenges to data reuse and reproducibility are related to incentives, documentation, data management, cost, data sharing, methods complexity, and data services. Lack of documentation of data with spatial and temporal context was prioritized by many respondents. Large scale and volume of data, cost and effort, limitations in data services were also strongly represented.

Our results revealed why data are reused and where data are accessed, as well as challenges to data reuse. Data was primarily reused for new research and much less for reproducing existing studies. This is consistent with our previous findings that OGD was primarily used as data sources for new research. Although data reuse were strong among the pool of respondents, practical and technical challenges to data reuse are pronounced. These challenges mainly include lack of documentation, cost and effort, and limited data services. A strong majority acknowledged the need for professional assistance with data processing and documentation, archiving and deposit, as well as the need to offload some labor. This is aligned with observations from previous studies that it is often very time-consuming to collect, process, and manage data [40, 98, 79]. Data owners and research data services could consider investing more in these areas to promote data reuse.

2.2 Related Work

This section discusses algorithms used for spatio-temporal prediction and learning representations from urban data, as well as methods for discovering, measuring, and removing discriminatory biases from machine learning algorithms. For each of these areas, we review representative works, identify research gaps, and propose our solutions.

2.2.1 Data Science for Cities

Spatio-temporal Prediction. Accurate prediction is an important step towards effective resource allocation (e.g., bike rebalancing) and timely responses to emergencies (e.g., fire department dispatches). Early work adopted time series analysis methods such as ARIMA or classical machine learning algorithms such as Gradient Boosting Regression Trees (GBRT) to predict urban events [144, 170, 86]. However, they have limited capability of modeling complex spatio-temporal dynamics. Recently, deep neural networks have become popular for modeling spatio-temporal data due to their performance modeling complex non-linear interactions [181, 146]. Recurrent Neural Networks (RNN) can capture temporal dependencies [42, 156] and Convolutional Neural Networks (CNN) can capture spatial structures [166].

Therefore, researchers use variants of RNNs and CNNs to model spatio-temporal problems [93] such as forecasting city crowd flows [181]. Combinations of CNNs and RNNs were proposed to learn both temporal and spatial dependencies in one network [174]. ConvLSTM adopts a LSTM network structure, but can take image-like data as input, therefore achieving the advantages of CNNs and RNNs [155]. 3D Convolutional Networks were initially used for modeling video data [138], but recently were also used for transportation demand prediction. For example, StepDeep is a network based on 3D convolutions to predict the number of taxi trips leaving and entering a certain region of a city at a certain time. StepDeep achieved better accuracy than other methods including DeepSD [129, 146]. Graph CNNs is suitable for data of graph structures such as road network, so they have gained popularity in predicting traffic forecasting [54, 12, 30].

Our prediction algorithms are based on 3D CNN architecture because it can learn spatial and temporal correlations simultaneously in one network. Our method is similar to StepDeep [129], but StepDeep architecture only incorporated temporal features such weather and is mainly intended for mobility demand prediction. Our method aims to be generic across various urban application domains and can handle arbitrary number of temporal, spatial, and spatio-temporal features. Furthermore, few existing work in modeling urban activities considers fairness in their solutions.

Representation Learning in Urban Applications. Most of the prediction methods are designed for individual urban application. Nevertheless, we observe that these tasks share similar model structures (e.g., CNN based models) and frequently use the same features (e.g., transportation network and weather). Collecting and formatting these features take substantial efforts and training separate models for each task is expensive in time and computation. It is therefore highly desirable that the knowledge (features or representations) learned from data through deep neural networks can be reused for other tasks or shared among multiple related tasks. Representation learning aims to extract useful information from data for better prediction and enable feature reuse [10]. A nascent thread of research focusing on spatio-temporal representation learning holds promises in deriving reusable and

sharable knowledge from data through multitask learning, embedding methods, and unsupervised learning [90, 180, 47, 48, 10, 122, 59]. For example, Space2Vec [105] and LESR [68] are unsupervised deep learning methods for representing spatial points or regions. However, they only considered the spatial domain. Several studies [149, 180, 148, 83] proposed methods for learning representations for spatio-temporal data. For example, Wang et al.[149] learned a temporal-aware representation through deep autoencoder with from GPS trajectories for driving behavior analysis. Nevertheless, most of them targeted at specific application domains, therefore their representations have limited re-usability.

This dissertation proposes to maximize the re-usability of the learned data representations by considering as many as possible available open urban datasets without assuming specific application domains. As such, the learned data representations can potentially benefit a wide range of downstream applications.

2.2.2 Fairness in Machine Learning

The wide application of machine learning algorithms in predictive policing [126], daily life [112, 120], and legal systems [7] has raised concerns that algorithms can discriminate against certain demographic groups. This is because algorithms can replicate and amplify the disparities in data [26]. Research in identifying and removing the biases embodied in algorithms is referred to as fairness in machine learning [97]. This emergent research field has two main objectives: 1) achieving fairness by design [57] and 2) achieving maximum possible utility under fairness constraints. There are currently two main research directions [186]: 1) Discovering and measuring unfairness, and 2) fairness-aware machine learning methods, including correcting discrimination in algorithms and removing biases from data.

Discovering and Measuring Fairness. There are mainly two ways to discover discrimination: statistical tests and fairness metrics [186]. Early research on fairness typically fits a regression model with sensitive attributes as independent variables. Statistical tests are frequently used in domains such as transportation equity and education equity. Recently, many fairness criteria (metrics) have been proposed to measure the degree of discrimination

[11, 78]. These metrics have been the mainstream ways for measuring fairness in machine learning community. The most widely adopted metrics for fair classification are Statistical parity [35], Equalized odds [57], and Equal opportunity [57]. Statistical parity asks for independence between predicted outcome (e.g., hiring decisions) and sensitive attributes (e.g., gender, race) [35]. Equalized odds requires equal mis-classification rates [57], that is, equal *true positive rates* and *false positive rates* across two demographic groups. Equal opportunity [57] is a relaxation of equalized odds, as it only requires equal true positive rates. There is less attention devoted to fairness metrics for regression. Equal means [18] resembles statistical parity in classification. It requires that the mean prediction outcomes across two groups are the same. Balanced residuals resembles equalized odds, requiring equal *mean positive residual* and equal *mean negative residuals* across groups [18]. Equal positive residuals [18, 60, 167] resembles equal opportunity, if we consider overestimation of outcome as beneficial.

Fairness-aware Machine Learning Methods. Fairness-aware machine learning methods aim to make discrimination-free predictions [35, 5]. Researchers have proposed a variety of remedies that occur at all stages of machine learning pipeline, including pre-processing, training, and post-processing [38, 186, 5, 158].

Fairness remedies at pre-processing stage seek to remove bias from data, so that the sanitized data can be used by any fairness-agnostic predictors. This approach is particularly desirable when data owners hope to release the data for public use, and the data is expected to be reused in many scenarios. There are two ways to "debias" data: directly modifying the dataset [71, 107, 73] and learning a fair representation for the dataset [41, 178, 20, 99, 179, 104]. Discrimination correction can be applied during the training process. Fairness can be either encoded as hard constraints [72, 18, 35, 176, 183] or as soft constraints (an additional regularization term) in the loss function [11, 16]. Post-processing methods build fairness into machine learning models by adjusting the prediction outcomes of a predictor [154, 57]. Post-processing methods do not require access to individual-level predictions or training data, so it is suitable for applications that have privacy concerns [57]. The limitation

is that these desirable features of post-processing approach often lead to a significant loss of utility [154, 4, 5]. This dissertation focuses on discrimination correction during training and learning fair representations.

Correcting Discrimination in Algorithms. This family of methods is featured with fairness constraints during the model learning process. They do not assume bias-free training data. Hard constraints have stronger theoretical basis and are usually used in conventional machine learning algorithms. Soft constraints can be applied to deep learning models and usually allow models to achieve multiple, sometimes conflicting objectives at the same time, with some trade-offs among the objectives. Compared to the pre-processing and post-processing approaches, debiasing during training is task-specific and can achieve higher utility [5]. The limitation of this approach is that it requires full access to the training pipeline, which is not always possible in practice [38, 186, 5, 158].

While existing papers primarily deal with conventional machine learning, there are a few studies that use fairness regularizers in deep learning models. For example, Hendricks et al. [16] addresses gender bias in deep learning based image captioning models through the use of regularizers that encourage fairness. However, Hendricks et al.'s setting is different from ours because their sensitive attribute (i.e. gender) is global for an image. In our case, sensitive attributes are distributed across the city: each region (pixel) has its own sensitive attribute value (e.g., percentage of white). Furthermore, the majority of fairness research focuses on classification settings rather than regression settings [78]. Calders et al. proposed using equal means as a fairness metric in linear regression. Fairness was incorporated through constraints in loss functions [18]. Berk et al. developed a series of convex fairness regularizers for linear and logistic regression. They used group fairness and individual fairness analogs in regression settings [11]. Our proposed method, FairST, was inspired by Berk et al.'s work, but the metrics and the formulation of the loss function are novel, as is the spatio-temporal setting.

Learning Fair Data Representations. One way to correct unfairness during the pre-processing phase is to learn a fair representation from the data such that little about the

sensitive attributes can be learned from the representation [46, 178]. In other words, learning a fair representation is to encode the original data into another feature space so that 1) non-sensitive information is preserved as much as possible and 2) information about the sensitive attributes is removed [178]. Fair data representations are promising in relieving the tensions in transparency and fairness. Consider the case when data owners hope to release their data but are reluctant to do so due to fairness concerns, this approach allows them to release a fair version of the original dataset while preserving fidelity. Compared to directly modifying the training data, this solution has several advantages. First, it does not hand-tune the dataset, so it is more generic and scalable. It strives to preserve the maximum non-sensitive information through optimization, which is based on theories rather than modifying the data according to some ad-hoc rules.

There are mainly two ways to achieve fair presentation: unsupervised learning [178, 127, 99] and adversarial learning [159, 157, 128, 145]. For example, Madras et al. [104] proposed a method called LAFTR. Based on an encoder-decoder structure, LAFTR learns a representation Z that predicts a supervised target and reconstructs the input. Meanwhile an adversary attempts to predict the sensitive information from Z. Most of the existing studies focus on classification tasks, and learning fair representations in spatio-temporal settings is absent from the literature. Our method is similar to LAFTR [104], the main difference is that our sensitive attributes are spatially continuous (e.g., a map of income level), whereas LAFTR uses binary sensitive attributes (e.g., gender). Our method is also related to work on conditional image generation where the task is to generate different version of an input image (e.g., a face) by varying a attribute (e.g., gender). For example, Lample et al. uses adversarial learning to learn latent representations for images that are invariant to an attribute such as gender [80]. Our method is different from theirs in that their manipulable attributes are discrete and global for an image, whereas our sensitive attributes are distributed at "pixel" level.

2.3 Summary

Abundant evidence suggests that data-driven urban applications in resource/service allocation and public safety can be discriminatory, but there is a dearth of research on fairness-aware prediction algorithms designed for cities so far. Spatio-temporal datasets such as house prices and transportation demand may contain bias, which can propagate through analytic pipelines and lead to biased research results or decisions. Currently there is a lack of research on how to remove discriminatory signals from spatio-temporal datasets. Many datasets supporting urban applications and research come from open data repositories, but these data sources are often too noisy, incomplete, and biased to use directly. Moreover, it is often difficult to discover, collect, and pre-process datasets from across multiple open data repositories. Improved data services and new ways of data release are needed to promote the reuse of open data.

Deep neural networks are powerful in addressing two key challenges in predicting urban activities: capturing rich spatio-temporal correlations and extracting information from heterogeneous urban data. Recent research in modeling urban activities suggests that models based on ConvLSTM or 3D CNN, generally yield better prediction accuracy than RNN or CNN based models [129, 174]. Meanwhile, it is generally agreed that the use of exogenous features such as weather and road networks helps prediction accuracy [184]. While various prediction frameworks have been proposed for urban predictions, the generalizability of them outside their original application domains is unclear. Spatio-temporal representation learning can extract knowledge from multi-source urban data and generate reusable features across downstream tasks. Nevertheless, existing work in this arena largely focuses on specific application domains, therefore the resulting data representations tend to have limited reusability.

There are two main streams of research in fair machine learning: discovery and measurement of unfairness, and fairness-aware methods [186, 4]. Various ways of measuring fairness have been proposed, but none of them is designed for spatio-temporal settings. Fairness-

aware corrections throughout all stages of machine learning have been developed. Most of them aim to achieve group fairness for classification. While a majority of the remedies were tested in conventional machine learning algorithms, learning fair data representation and enforcing fairness as regularizers during model training are well-suited for deep learning models. The former does not require access to raw data and the training pipelines of downstream tasks, and the latter shows promises in satisfying fairness constraints while maintaining high utility.

Overall, this dissertation differs from existing work in objectives and approaches. First, few existing work in modeling urban activities considers fairness in their solutions. FairST builds on the state of the art 3D CNN and is generic across tasks. It incorporates fair regularizers to guide the model to learn equitable and accurate predictions. Second, we propose to learn EquiTensors, reusable and fair data representations. EquiTensors are task-agnostic representations trained from a large number of open urban datasets in an unsupervised way, while most of the representation learning in urban settings focuses on specific application domains and only considered limited number of datasets. Third, EquiTensors aim to remove discriminatory signals from spatio-temporal data, which is absent from the existing literature. Finally, EquiTensor presents a novel way to share open data, avoiding data discovery and pre-processing, limiting their exposure to new sources of discriminatory bias, and broadening the utility of open data repositories.

Chapter 3

FAIRNESS-AWARE SPATIO-TEMPORAL PREDICTION

In this chapter, we develop two new fairness metrics and a fairness-aware spatio-temporal prediction algorithm [1]. While we use new mobility applications as a case study, our methods can be extended to other spatio-temporal scenarios with fairness concerns such as crime incidence prediction.

3.1 Introduction

New mobility services such as car-sharing, bike-sharing, and ride-hailing have been deployed in many cities as affordable and on-demand transportation options for citizens. For example, dockless bike share systems have been introduced in many cities in China and the United States. They are docking-station-free and GPS-tracked. Using a mobile app, users can locate and pick up a bike closest to them, and park the bike anywhere they want [115, 82, 156]. Ride-hailing companies such as Uber and Lyft connect drivers to riders through mobile phone apps. Today, they are providing over 12 million trips per day worldwide [15, 21].

Supply and demand in new mobility systems are often unbalanced due to complex and dynamic factors such as traffic conditions and weather. Accurate and high-resolution demand estimates are therefore important to guide resource optimization and maximize system utility [146, 85]. For example, ride-hailing companies predict demand to direct drivers to high-demand areas [9]. Similarly, bikeshare operators use trucks to *rebalance* bikes from low-demand to high-demand areas based on demand estimation [115].

Beyond accuracy, incorporating equity into demand prediction is crucial for delivering a transportation system that benefits all citizens, particularly for historically underrepresented

[1] These contributions were presented in the following publications [160, 161].

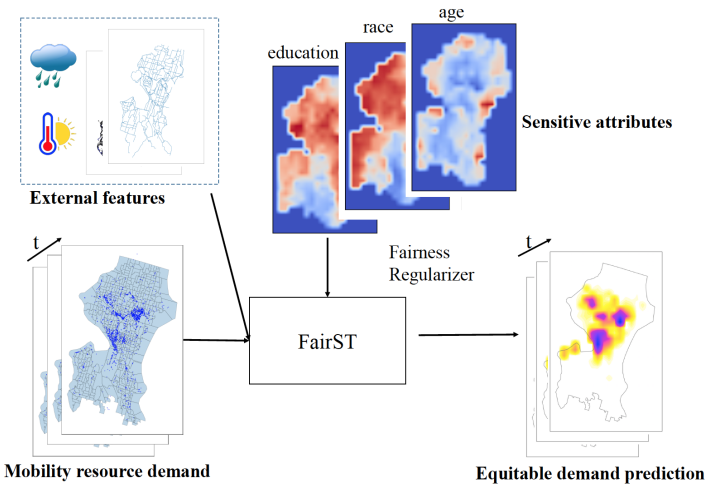

Figure 3.1: FairST is a deep learning based demand prediction model for new mobility systems. It not only models the spatial-temporal dynamics of mobility system, but also makes equitable predictions by incorporating a fairness regularizer that encourages equal prediction between groups defined by, for example, race, age, or education level.

groups. An individual's access to resources allocated or predicted by algorithms should not be dependent on sensitive attributes such as race and age. However, recent studies show that algorithms that distribute app-based mobility services may discriminate against people of color [15, 50]. For example, influenced by Uber's pricing algorithm, neighborhoods with more white people experienced higher service quality [132]. Compared to traditional transportation modes, new mobility services may lead to greater inequity. For example, people without smart phones are not able to access the services. Underestimation of mobility resource demand for these groups may result in insufficient supply to these areas, which can produce a feedback loop: racial and income disparities are misinterpreted in the model as lack of demand, reinforcing reduced access to services.

In this chapter, we incorporate fairness in a demand prediction framework for new mobility systems, while acknowledging broader applications to any fair allocation of resources in space and time. To achieve our goal, the proposed approach addresses three challenges: ac-

curate modeling of the spatial-temporal dynamics of the new mobility system, defining novel fairness metrics suitable for this task, and effective integration of fairness into the prediction model.

modeling the spatial-temporal dynamics of mobility resource demand. Resource demand exhibits complex spatial and temporal patterns, and is influenced by many external factors such as weather and road network [85]. In systems such as ride-hailing and dockless bikeshare, the demand is continuous over space. Thus an initial challenge is to properly discretize continuous demand and model the spatial dependencies among neighbouring regions.

We address this challenge by partitioning the city into a regular grid and aggregating demand into time intervals. We use a 3D convolutional neural network (3D CNN) as the core building block in our model to capture spatial-temporal dynamics. To incorporate the exogenous features such as weather and traffic that can influence demand, we adopt a three-stream model architecture that fuses together 1D, 2D, and 3D convolutional layers, respectively. A 1D CNN is used to extract information from 1D temporal features such as city-wide temperature or rainfall, and a 2D CNN is used to extract information from 2D spatial features such as the location of bike lanes.

Designing fairness metrics for mobility resource demand. Although the specification and assessment of fairness metrics is an active research area [52, 64], most of the proposals are inapplicable in spatial-temporal settings. First, the prediction target in our setting (e.g., ride-hailing demand) is continuous whereas many fairness metrics such as statistical parity and equalized odds are designed for discrete classification settings [57]. Second, in mobility systems, each record is usually a geographic area representing an entire subpopulation, and therefore cannot necessarily be assigned a specific attribute value (e.g. white), but rather a percentage of the subpopulation that has that value (e.g. percentage white). Among the few studies that propose fairness metrics for regression, a categorical sensitive attribute is typically required [18, 11]. These methods cannot be directly applied to our problem unless we discretize our sensitive attributes. Finally, a fairness metric for mobility resource demand

prediction should consider the overall population distribution. The transportation literature suggests that mobility resource demand is positively associated with zonal population [39], so fairness should be examined on a per capita basis.

To address these challenges, we interpret fairness in demand prediction as the requirement that individuals of different groups have access to a similar amount of the resource in demand. We propose two fairness metrics: *region-based fairness gap (RFG)* and *individual-based fairness gap (IFG)*. Both assess the gap between mean per capita demand across groups over a period of time. However, RFG assumes that a distinct label is assigned to the entire region. For instance, a neighborhood with a majority white population may be assigned the label "white." IFG instead is assigned a distribution based on demographics rather than a single label.

Integrating fairness into the prediction model. Fairness can be incorporated into a prediction model during data preprocessing [71], model training [35, 176], or postprocessing [57]. During model training, fairness can be either encoded as a hard constraint or as additional terms in the loss function [11]. We propose two possible terms, corresponding to RFG and IFG. To the best of our knowledge, our work is the first to incorporate fairness in a spatial-temporal urban mobility setting using deep neural networks.

To this end, we introduce FairST, a **Fair**ness-aware **S**patial-**T**emporal model that accounts for dynamics of mobility resource demand and enforces fairness through regularizers (Figure 3.1). FairST can be naturally extended to other scenarios that involve spatial-temporal modeling and have fairness concerns such as crime incidence prediction. We summarize our main contributions as follows:

- We propose a new mobility resource demand prediction algorithm based on 3D convolution neural network (3D CNN) to model the temporal and spatial dependencies. The proposed algorithm adopts a three-stream architecture to integrate exogenous features with various dimensions.

- We propose two fairness metrics: *region-based fairness gap (RFG)* and *individual-based*

fairness gap (IFG) for urban mobility. Both metrics measure the gap between mean per capita demand across groups over a certain period of time. The difference lies in that RFG focuses on discrete sensitive attributes while IFG deals with continuous attributes.

- We design and implement two fairness regularizers for deep networks in spatial-temporal settings, region-based fairness and individual-based fairness based on RFG and IFG. They are integrated into the loss minimization pipelines to encourage fair prediction.

- We evaluate our method using two real-world datasets. Our experiments demonstrate that our method effectively closes the fairness gaps while achieving better accuracy than state-of-the-art fairness-oblivious models.

3.2 Related Work

Equity in New Mobility Systems. A number of researchers have studied equity in bike sharing systems. Ursaki and Aultman-Hall [140] found that there are significant differences in race, education level, and income of population inside and outside bikeshare service areas in four U.S. cities. Other studies also indicate that in North America, advantaged groups have more access to docked bikeshare than disadvantaged groups [61]. In examining access equity of dockless bikes in Seattle, Mooney et al.[115] found that more college-educated and higher-income residents have access to more bikes, and that bike demand is high correlated with rebalancing destinations. Overall, current literature suggests that disparities exist in the access of bikeshare systems. The equity of ride-hailing services is less clear. Although some studies found that service quality is not necessarily associated with the income or minority fraction of pickup locations [63, 147], the findings from some other studies suggest that ride-hailing companies provide poor services to low-income neighborhoods [132]. Moreover, several studies [50, 15] found that ride-hailing drivers discriminate against African American riders, resulting in longer waiting times and higher trip cancellation rates.

Existing studies focus mostly on assessing equity based on the outcomes of deployed systems, we argue that approaches for preventing unequal resource distribution or dynamically correcting unfairness are lacking.

Spatial-temporal Prediction. Accurate demand prediction is an essential step towards effective resource allocation (e.g., bike rebalancing and ride dispatch) strategies. Early work adopted time series analysis methods such as ARIMA or classical machine learning algorithms such as Gradient Boosting Regression Trees (GBRT) to predict mobility resource demand [144, 170, 86]. Recently, deep neural networks have become popular for modeling spatial-temporal data due to their performance modeling complex non-linear interactions [181, 146]. Recurrent Neural Networks (RNN) can capture temporal dependencies [42, 156] and Convolutional Neural Networks (CNN) can capture spatial structures [166]. Therefore, researchers use variants of RNNs and CNNs to model spatial-temporal problems [93] such as forecasting city crowd flows [181]. Combinations of CNNs and RNNs were proposed to learn both temporal and spatial dependencies in one network [174]. ConvLSTM adopts a LSTM network structure, but incorporated convolutional operators in replace of fully-connected nodes, therefore achieving the advantages of CNNs and RNNs [155]. 3D Convolutional Networks were initially used for modeling video data [138], but recently were also used for transportation demand prediction. StepDeep is a network based on 3D convolutions to predict the number of taxi trips leaving and entering a certain region of a city at a certain time. StepDeep achieved better accuracy than other methods including DeepSD [129, 146].

Few existing work in modeling urban resource demand considers fairness or equity in their solutions. FairST builds on the state of the art 3D CNN approaches and incorporates fair regularizers to guide the model to learn equitable spatial-temporal prediction.

Fairness in Machine Learning. Studies on fairness in machine learning focus on identifying and removing bias in the outcome variable with respect to some sensitive group (e.g., race, gender, income) [64]. Although many competing definitions of fairness have been proposed, most involve the idea that the predicted outcomes should be statistically independent from a given sensitive attribute [35]. Several fairness metrics have been proposed

for classification settings. Individual fairness, in contrast to group fairness, captures the idea that similar individuals should be treated similarly [35]. Group fairness is better aligned with most legal and practical definitions, arguing that members of a disadvantaged group should receive similar treatment to an advantaged group, by experiencing similar predicted outcomes [41]. Equalized odds requires equal mis-classification rates across groups [57]. Based on these concepts, researchers have proposed fairness-aware remedies that occur at all stages of the machine learning pipeline [11].

The majority of fairness research focuses on classification settings rather than regression settings [78]. Metrics for classification involve discrete probabilities and are difficult to adapt directly to regression settings. Calders et al. proposed using equal means as a fairness metric in linear regression. Fairness was incorporated through constraints in loss functions [18]. Berk et al. developed a series of convex fairness regularizers for linear and logistic regression. They used group fairness and individual fairness analogs in regression settings. Results on six datasets highlight the incompatibility of various fairness metrics and trade-off between accuracy and fairness [11]. Our proposed method was inspired by Berk et al.'s work, but the metrics and the formulation of the loss function are novel, as is the spatial-temporal setting.

3.3 Use Cases

In this section we describe the datasets, pre-processing, and problem formulation for our two mobility use cases.

3.3.1 Datasets

Seattle dockless bikeshare dataset. The city of Seattle requires shared bike operators to submit their data to the Transportation Data Collaborative (TDC) operated by the University of Washington for conducting data ethics related research. The data used in this chapter comes from one of the operators from October 1, 2017 to October 31, 2018, obtained from the TDC. It includes more than 1,600,000 trips and more than 10,000 bikes. The data

contains information about each bike including pickup and drop-off locations, trip start, trip end, and timestamps, as well as information about trips, including trip duration, trip start and end time, trip start location, and trip end location. We mainly use bike pick-up locations and timestamps. We use the number of pickup (trip start) as a proxy for demand as there is no ground truth value for "true demand."

RideAustin dataset. RideAustin [2] is a non-profit ride-hailing service operating in Austin, Texas. Rides data is openly available online [3]. The data used in this chapter spans from August 1, 2016 to April 13, 2017, including over 1,400,000 completed trips. It contains information about each ride including trip duration, trip start time, trip end time, trip start location, and trip end location, and distance travelled, etc. We use the number of rides as a proxy for demand.

Socioeconomic data. Socioeconomic data including population, race, age (under or over 65), and education level for Seattle and Austin at the block group level were obtained from the SimplyAnalytics database [130].

Weather features. Previous studies show that weather conditions are associated with bike demand and ride requests, and can be helpful for prediction [86, 129, 146]. We obtained hourly weather data for Seattle and Austin from the Integrated Surface Dataset from the National Centers for Environmental Information (NCEI) [4]. We included city-level air temperature, sea level pressure, and precipitation as features for prediction. They are all 1D time series as they do not have spatial variations.

Urban features. Urban forms are associated with the access and usage of new mobility systems [147]. We collected 2D features such as bike lanes and steep slopes for Seattle bikeshare demand prediction as they may be associated with bikeshare demand according to existing literature [109, 44, 82]. Likewise, we collected features such as road network and Point of Interest that were suggested by the literature for RideAustin demand prediction

[2] http://www.rideaustin.com/

[3] https://data.world/ride-austin/ride-austin-june-6-april-13

[4] https://www.ncei.noaa.gov/access/search/index

[146, 129]. These urban datasets are all openly available [5].

3.3.2 Data Preparation

Figure 3.2 illustrates the method that we used to process the Seattle bikeshare dataset. The RideAustin dataset was processed in the same way. We place a bounding box around the geographic region of a city and partition the bounding box into equal-sized squares (Figure 3.2(a)). For Seattle bikeshare, we choose a grid size of 1km by 1km. For RideAustin, we choose a grid size of 2km by 2km. The purpose of partitioning the city into square grids rather than using neighbourhoods or block groups as the prediction unit is to prepare the data as a tensor that CNN based models can take. We counted the number of pickup in each hour and in each square region based on pickup locations and timestamps. For each grid cell, resource demand forms a time series as shown in Figure 3.2(b). For each hour, the study area can be likened to a frame in a video and each region can be seen as a pixel with demand as its value (Figure 3.2(c)).

We transformed 2D urban datasets to grid cell representation using the count of features (e.g. Point of Interest, road segments) and the total length of the features (e.g. road segments) within each grid. We calculated socioeconomic attributes for each grid. Mismatches between block group boundaries and grid boundaries were accounted for using proportional allocation based on area.

3.3.3 Prediction Problem Definition

We aim to build fair models to forecast next time step demand for mobility resource for a city based on the demand of previous time steps. For both Seattle bikeshare and RideAustin, we aim to predict hourly demand based on the demand of the last 7 days (168 hours). The prediction problem is similar to predicting next frame based on the previous 168 frames in a video. We generated slices of 169 hours for training and prediction (168 hours for training

[5] `https://data.seattle.gov/` and `https://data.austintexas.gov/`

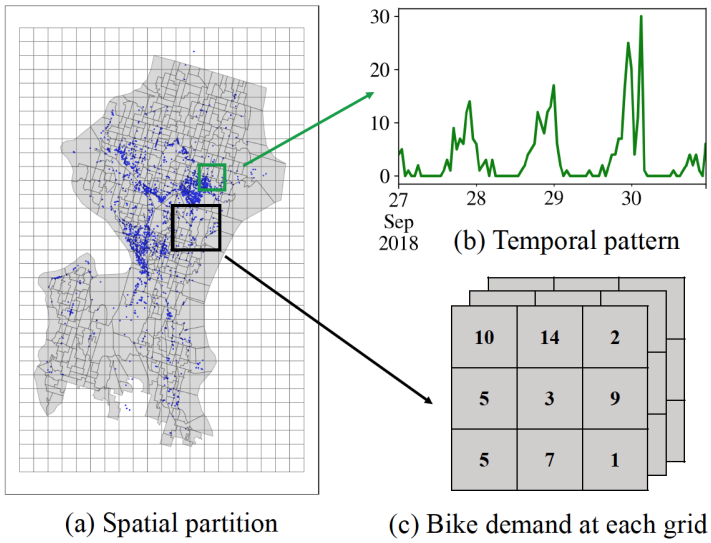

Figure 3.2: Data preprocessing. (a) We partition a city into square grids. (b) For each grid, resource demand forms a time series. (c) Each hour is akin to a frame in a video, with each grid cell as a pixel whose value is the demand.

and to predict the next 1 hour). For Seattle bikeshare, we use the data from October 2017 to August 2018 for training and the data from September to October, 2018 for testing. The training data contains 8040 temporal slices and the test data contains 1464 temporal slices. For RideAustin, we use the data from August 2016 to February 2017 for training and the data from March to April 2017 for testing. The training data contains 5088 temporal slices and the test data contains 1056 slices. The prediction should balance two objectives: minimizing prediction accuracy loss and minimizing fairness loss.

3.4 Model and Fairness Metrics

In this section, we detail our spatial-temporal model architecture and describe our proposed fairness metrics and corresponding fairness regularizers.

3.4.1 Model Architecture

We first introduce 3D convolutions for learning spatial-temporal features, then present the architecture of FairST, followed by the design of the objective function that guides the learning process.

The core building block of FairST is 3D convolution, which models spatial-temporal information [70]. We design a three-stream prediction framework based on 1D, 2D, and 3D CNN to 1) automatically capture the spatio-temporal context, and 2) include external features to help with accuracy. We use a submodel consists of 3D convolution layers to learn from 3D historical demand, a submodel with 1D convolution layers to learn information from 1D time series features, and a submodel with 2D convolution layers to extract information from 2D urban features. The outputs of all submodels were fused together, on top of which additional convolutional layers were applied to achieve the final prediction (See Figure 3.3). Compared to fusing all features before being fed to a single network, this strategy has two main advantages: 1) Integrating semantically related features into one submodel can potentially reinforce the effectiveness of one another [185]. For example, in our setting, 1D features often represent mutually correlated meteorological information, and 2D features reflect the geographic characteristics of the city. 2) Fusing all features early at the dataset level requires 1D and 2D features to be replicated to create 3D tensors. This redundancy brings unnecessary computation overhead and wasted model capacity.

The first submodel is based on 3D convolutions. It takes a time series of resource demand history as input. The submodel consists of three 3D convolutional layers, followed by a 2D convolutional layer, as shown in Figure 3.3. The number of filters of 3D convolutional layers are 16, 32, and 1, respectively. We use $3 \times 3 \times 3$ filters because this is the size that worked best in previous studies [138]. We use padding to ensure the layer outputs are of the same size as inputs. The third 3D convolutional layer adopts 1 filter to achieve dimension reduction [137] and temporal pooling. Finally, a 2D convolution layer is used to integrate temporal information from previous layers and output the feature map for submodel fusion. The

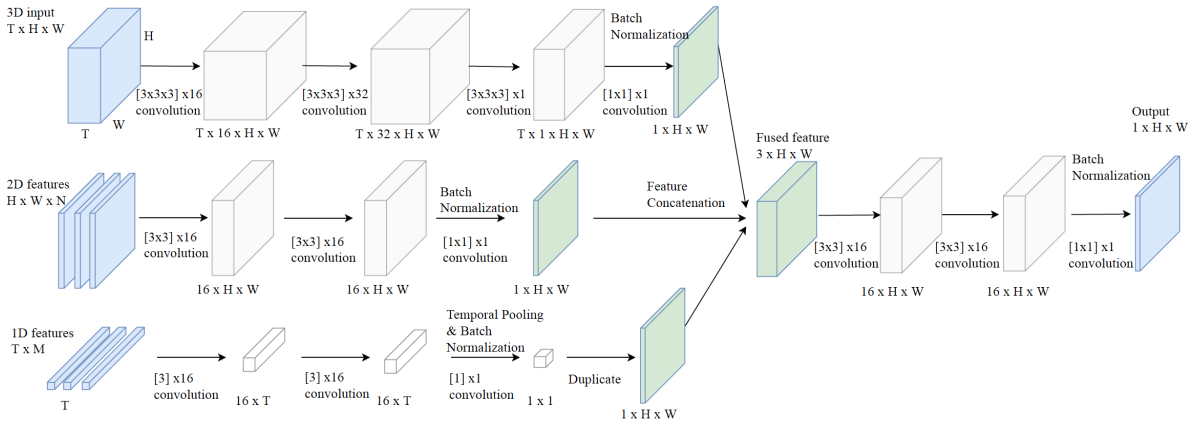

Figure 3.3: A three-stream network architecture. The network input contains three streams, including 1D time series features, 2D urban features, and 3D spatial-temporal input. The network is trained to predict mobility resource demand in an end-to-end fashion. T, H, W are the number of time steps, height of input, and width of input, respectively. N and M are the number of 2D and 1D features, respectively.

second and third submodels are based on 2D and 1D convolutions, respectively. They aim to extract meaningful information from the input features and improve prediction accuracy. We use leaky Relu as activation function for all layers [103]. We keep the model light-weight and skip spatial pooling to avoid deconvolution operations (for maintaining the output size) afterwards, which is typically more prone to overfitting in small training sets.

Training objectives. Our loss function is a weighted sum of an accuracy loss and a fairness loss. The fairness loss acts as a regularizer for the model. We use Mean Absolute Error (MAE) as accuracy loss. The overall loss function is defined as

$$L = L_{accuracy} + \lambda L_{fairness} \tag{3.1}$$

where $L_{accuracy}$ is MAE, $L_{fairness}$ is the fairness loss, and λ is the weight for the fairness loss. In the experiments (Section 3.5), we show the accuracy given different λ values. In the next section, we describe details of the proposed fairness loss.

3.4.2 Fairness Metrics and Regularizers

We consider fairness as individuals of different groups receiving equal resources. In the mobility setting, fair prediction implies adjusting the demand to reduce the difference in per capita resource demand among groups defined by, say, race. Our definition adapts *group fairness* in the machine learning literature that requires the disadvantaged group to receive similar treatment to the advantaged group by experiencing similar predicted outcomes [41], and is informed by *vertical equity* in the transportation literature requiring transportation policies to favor socially disadvantaged groups to compensate for overall inequities [31].

Given this approach to fairness, we propose two fairness metrics: a Region-based Fairness Gap (RFG) and an Individual-based Fairness Gap (IFG). Both RFG and IFG measure the gap between mean per capita demand across two groups over a certain period of time. However, for RFG, each geographic region is assigned a single group label according to some criteria (e.g., Caucasian or non-Caucasian). For IFG, groups are determined based on the demographic distribution in the region, such that the sensitive attribute is numeric (e.g., the percentage of the subpopulation in the region that is Caucasian). In this chapter we focus on a square grid partitioning, these two metrics can be used for any customized partitioning (e.g., census tracts, zip codes, etc.)

Intuition. RFG draws upon the idea that people live in the same region share similar public facilities and economic status, so they may have similar commute patterns and demand for transportation resources. For example, a white person may live in a predominately black community, but she frequents the same bus stops and grocery stores as her neighbors. Therefore, when assessing mobility resource demand equity, policies to distribute resources may primarily consider the majority group. In practice, we can assign each region the group label (e.g., race) with the highest population, or some criteria defined by local governments. However, we caution that a simple discretization of the sensitive attributes by a threshold for each region itself is biased, since the minority population in a region may be underrepresented.

Notation. We start by introducing notation.

- Let s_i be the ith square region of the study area $\mathcal{S}$.
- Let p_i denote the population of square region s_i divided by the total population of the city.
- Let $\hat{y}_{i,t}$ and $y_{i,t}$ be the estimated demand and ground truth demand for region s_i at time t, respectively.
- Let $E_T[\hat{y}_{i,t}]$ be the average predicted value for the ith square region in $\mathcal{S}$ over time period T.

Region-based Fairness Gap (RFG). We now formally define RFG. Let every region s_i be assigned to one of two groups (e.g., Caucasian and non-Caucasian) with regard to one sensitive attribute A (e.g., race), denoted by G^+ (the advantaged group) and G^- (the disadvantaged group). We define RFG between two groups with regard to sensitive attribute A over a period of time T as follow:

$$RFG = \frac{\sum_{i \in G^+} E_T[\hat{y}_{i,t}]}{\sum_{i \in G^+} p_i} - \frac{\sum_{j \in G^-} E_T[\hat{y}_{j,t}]}{\sum_{j \in G^-} p_j} \tag{3.2}$$

The first term can be interpreted as the per capita demand for group G^+ averaged over T. The denominator is the total population (normalized) of G^+. Likewise, the second term is the mean per capita demand in group G^- over T.

Individual-based Fairness Gap (IFG). Let w_i^+ denote the percentage of people in the advantaged group of the sensitive attribute A (e.g., race) in region s_i and let w_i^- denote the percentage of people in the disadvantaged group. For example, if a region s_i is 65% white, then $w_i^+ = 65\%$ and $w_i^- = 35\%$. IFG assumes that given the predicted demand, the number of resources a group will get is proportional to the population percentage of that group. For example, if the predicted demand for bikeshare is 100 bikes for a region and the percentage of white people is 65%, then the demand that allocated to the Caucasian group in that region

is 65 bikes. Formally, we define IFG between two groups with regard to sensitive attribute A over a period of time T as follow:

$$IFG = \frac{\sum_{i \in \mathcal{S}} E_T[\hat{y}_{i,t}] w_i^+}{\sum_{i \in \mathcal{S}} p_i w_i^+} - \frac{\sum_{j \in \mathcal{S}} E_T[\hat{y}_{j,t}] w_j^-}{\sum_{j \in \mathcal{S}} p_j w_j^-} \tag{3.3}$$

The numerator of the first term denotes the predicted total demand of all people in the advantaged group averaged over T. The denominator is the total population (normalized). Then the first term is the predicted per capita demand allocated to the advantaged group averaged over T. The second term can be interpreted similarly.

In summary, for RFG, everyone that lives in the same region is assigned the same group label, whereas IFG assigns group labels proportionally based on the region's demographics.

Fairness loss. Based on the RFG and IFG, we define two fairness loss terms, Region-based Fairness loss (RF loss) and Individual-based Fairness loss (IF loss) to incorporate fairness into training.

The *Region-based Fairness loss (RF loss)* at time t is defined as

$$L_{RF}(t) = \frac{1}{\sum_{i \in \mathcal{S}} y_{i,t}} \left| \frac{\sum_{i \in G^+} \hat{y}_{i,t}}{\sum_{i \in G^+} p_i} - \frac{\sum_{j \in G^-} \hat{y}_{j,t}}{\sum_{j \in G^-} p_j} \right| \tag{3.4}$$

The first term is the estimated per capita demand in group G^+ at time t. Likewise, the second term is for group G^-. $\sum_{i \in \mathcal{S}} y_{i,t}$ is a normalizing factor.

The Individual-based Fairness loss (IF loss) at time t is defined as

$$L_{IF}(t) = \frac{1}{\sum_{i \in \mathcal{S}} y_{i,t}} \left| \frac{\sum_{i \in \mathcal{S}} \hat{y}_{i,t} w_i^+}{\sum_{i \in \mathcal{S}} p_i w_i^+} - \frac{\sum_{j \in \mathcal{S}} \hat{y}_{j,t} w_j^-}{\sum_{j \in \mathcal{S}} p_j w_j^-} \right| \tag{3.5}$$

The first term is the estimated per capita demand for advantaged group at time t. Likewise, the second term is for disadvantaged group.

Multiple sensitive attributes can be represented together in one loss function as the weighed sum of fairness loss of each attribute. Assuming there are $a = \{1, 2, ..., A\}$ sen-

sitive attributes, then the composite loss function is defined as

$$L_{fairness}(t) = \sum_{a=1}^{A} \lambda_a L_{fairness_{(a,t)}} \tag{3.6}$$

where λ_a is the weight term for the ath attribute and $L_{fairness_(a,t)}$ is the fairness loss.

3.5 Experiments

We evaluate our method on the Seattle dockless bikeshare dataset and the RideAustin dataset. First, we compare FairST without fairness loss ($\lambda = 0$) with state-of-the-art spatial-temporal models in terms of prediction accuracy. We then incorporate Region-based Fairness loss (RF loss) and Individual-based Fairness loss (IF loss) into our model. To understand the effectiveness of the two proposed fairness regularizers, we compare against other existing fairness regularizers on a single sensitive attribute (i.e. race). Finally, we integrate the fairness losses for race, age, and education level into FairST to evaluate its capability of reducing unfairness for multiple sensitive attributes in one shot.

3.5.1 Implementation

We implement FairST and the deep-learning based baseline models with TensorFlow Framework [1], and perform training and inference with NVIDIA K80 GPU machines. We use a batch size of 32 and train FairST for 200 epochs for Seattle bikeshare and 350 epochs for RideAustin using Adam optimizer. We use a exponential learning rate decay scheme: the learning rate starts at 0.005 and decays every 5,000 steps with a rate of 0.96.

To implement Region-based Fairness loss, we assign each square region a label for each attribute. We use the overall city statistics as thresholds to discretize the continuous sensitive attributes. For example, the percentage of white population of Seattle in 2018 is 65.74%, we then set the regions with more than 65.74% white population as Caucasian group, otherwise as non-Caucasian group. The same method is used for discretizing age and education level.

3.5.2 Baseline Models

To evaluate the prediction accuracy of our method, we compare FairST with several other models: 1) **Historical Average (HA)**. We compute $\hat{y}_{i,t}$ using the mean values of all previous observations at location s_i at the same time of the day and the same day of the week. 2) **Autoregressive Integrated Moving Average Model (ARIMA)**. ARIMA is a commonly used statistic model for forecasting time series. We develop an independent ARIMA model for each individual grid cell. 3) **Long short-term memory Network (LSTM)** [42]. LSTM is a variant of Recurrent Neural Network that can learn long-term temporal dependencies. We train the LSTM model individually for each square grid. 4) **Convolutional LSTM (ConvLSTM)** [155]. The ConvLSTM network adopts LSTM structure, but replaces fully connected layers with convolutional operations in each cell. As a result, it can capture both spatial and temporal dependencies in one network. We also compare FairST with various 3D CNN models: a **3D CNN** model that is equivalent to FairST without any external features; a **3D CNN + 1D** model that consists of a 3D CNN based submodel and a 1D CNN based submodel; and a **3D CNN + 2D** model that consists of a 3D CNN based submodel and a 2D CNN based submodel.

3.5.3 Baseline Fairness Regularizers

We compare the proposed loss functions (RF loss and IF loss) with two other existing fairness losses [18, 11] in our experiments.

Equal Means Loss (EM Loss). Calders et al. defined Equal Means as a fairness metric for regression [18]. Equal Means enforces the mean prediction to be the same for different groups. This metric is not directly comparable with IFG or RFG as we focus on predicted demand per capita, therefore, we substitute prediction with per capita prediction in Equal Means loss. The modified Equal Means loss is defined as:

$$L_{EM}(t) = \frac{1}{\sum_{i \in \mathcal{S}} y_{i,t}} \left| \frac{\sum_{i \in G^+} \hat{z}_{i,t}}{n^+} - \frac{\sum_{j \in G^-} \hat{z}_{j,t}}{n^-} \right| \tag{3.7}$$

where p_i is the population of region s_i divided by the total population of the city. $\hat{z}_{i,t} = \frac{\hat{y}_{i,t}}{p_i}$, denoting the predicted per capita demand. n^+ and n^- denote the number of advantaged square regions and the number of disadvantaged square regions, respectively.

Pairwise Fairness Loss (Pairwise Loss). Berk et al. defined a family of fairness regularizers that corresponds to individual fairness, group fairness, and hybrid of the two [11]. In all three loss term formations, comparisons across groups are based on cross pairs $i \in G^+$ and $j \in G^-$. Since our metrics are analogs of group fairness, we compare our metrics with Berk's group fairness penalty.

$$L_{PF}(t) = \frac{1}{\sum_{i \in \mathcal{S}} y_{i,t}} \left(\frac{1}{n^+ n^-} \sum_{\substack{i \in G^+ \\ j \in G^-}} d(z_{i,t}, z_{j,t})(\hat{z}_{i,t} - \hat{z}_{j,t}) \right)^2 \tag{3.8}$$

$$d(z_{i,t}, z_{j,t}) = e^{-(z_{i,t} - z_{j,t})^2} \tag{3.9}$$

Similar to the modified Equal Means loss, we substitute prediction with per capita prediction. The model will increase penalty as the difference between $\hat{z}_{i,t}$ and $\hat{z}_{j,t}$ increases, weighted by a similarity function $d(z_{i,t}, z_{j,t})$.

3.5.4 Evaluation Metrics

We evaluate the prediction accuracy of all models with Mean Absolute Error (**MAE**). We evaluate the fairness of prediction outcomes using **RFG** and **IFG**, but we also consider the correlation between the ranked demand and the proportion of the advantaged group. That is, we are considering that city planners are interested in assessing whether the regions with the highest demand also happen to be the wealthy, advantaged neighborhoods. We use Spearman's rank correlation coefficient (**Spearman's rho**) [58], which measures the strength of monotonic correlation between two variables. We calculate Spearman's rho between mean per capita demand over the test period of a grid region and the percent of advantaged population (i.e., percentage of Caucasian, percentage of population under 65 years old, and

percentage of population with a college degree) of that region. A highly positive Spearman's rho with a p-value less than 0.05 suggests disparities in demand.

3.6 Results and Discussion

The primary goal of predicting demand is to guide resource allocation, so it is desirable to make accurate predictions while closing the equity gaps. In this section, we show that proposed fairness regularizers give better performance than baseline regularizers in our problem setting. We also show that FairST is able to achieve better accuracy and less inequity than baseline models.

3.6.1 Demand Prediction Accuracy

We compare prediction accuracy of our model with baselines. Table 3.1 and Table 3.3 show Mean Absolute Error of all models on the Seattle bikeshare dataset and the RideAustin dataset, respectively. It is observed that the 3D CNN based methods (i.e., 3D CNN, 3D CNN + 1D, 3D CNN + 2D, and FairST without fairness penalty) proposed by this chapter achieve higher prediction accuracy than the other methods. HA is a simple and reasonable method to predict demand, however, it overgeneralizes temporal dynamics and does not account for spatial structure. ARIMA assumes input time series is stationary which is often not the case with fluctuating demand. It is also not good at predicting with sparse data where there are many zeros in the series. LSTM achieves better accuracy than ARIMA and HA, but still suffers from inability to learn information from spatial context. ConvLSTM outperforms LSTM due to its capability of learning both spatial and temporal information. The 3D CNN models perform better than ConvLSTM since the 3D CNN is more powerful in terms of capturing strong local spatial-temporal correlations in our problem as compared to the recurrent architectures. Furthermore, the incorporation of external features improves accuracy in both Seattle bikeshare and RideAustin cases.

Table 3.1: FairST compared to baselines for predicting Seattle bikeshare demand (multiple attributes)

	λ	MAE	RFG (race)	RFG (age)	RFG (edu)	IFG (race)	IFG (age)	IFG (edu)	Spearman's rho (race)	Spearman's rho (age)	Spearman's rho (edu)
Ground Truth	/	/	112.568	160.089	37.471	38.969	51.338	30.053	0.016	0.174**	0.338**
HA	/	0.484	194.454	49.494	193.477	79.906	17.641	54.692	0.565**	0.477**	0.500**
ARIMA	/	0.538	319.032	62.793	319.648	129.447	28.170	90.505	0.569**	0.463**	0.489**
LSTM[42]	/	0.468	280.685	61.437	277.938	116.023	23.778	79.162	0.522**	0.441**	0.425**
ConvLSTM [155]	0.000	0.432	74.485	139.666	19.934	22.907	44.459	19.101	0.210**	0.355**	0.324**
3D CNN	0.000	0.408	100.878	169.240	38.873	31.915	53.133	26.851	0.091	0.256**	0.394**
3D CNN + 1D	0.000	0.387	88.587	153.625	19.802	26.791	49.058	20.691	0.291**	0.376**	0.077
3D CNN + 2D	0.000	0.378	93.299	157.025	33.946	28.661	49.792	24.457	0.158**	0.246**	0.370**
FairST	0.000	0.382	83.127	147.437	23.400	25.073	47.403	20.885	0.168**	0.191**	0.328**
FairST + RF	0.005	**0.377**	80.565	146.665	20.855	24.168	46.732	20.184	0.111*	0.262**	0.348**
FairST + RF	0.150	0.437	16.140	35.562	-5.712	4.199	22.543	7.112	-0.019	0.107*	0.321**
FairST + RF	0.250	0.460	**8.650**	**14.242**	**-3.364**	2.226	19.178	6.299	**0.011**	**0.090**	0.231**
FairST + IF	0.100	0.385	67.695	128.010	4.905	17.927	40.811	14.874	0.099	0.231**	0.347**
FairST + IF	0.150	0.394	49.075	110.725	-9.322	11.738	35.410	9.529	0.030	0.181**	0.385**
FairST + IF	0.500	0.439	30.668	53.896	-20.291	3.823	16.536	2.200	0.117*	0.222**	0.085
FairST + IF	0.600	0.460	24.753	34.011	-22.700	**0.898**	**8.855**	**-0.185**	0.060	0.158**	**-0.055**

**. Correlation is significant at the 0.01 level.

*. Correlation is significant at the 0.05 level.

3.6.2 Fair Prediction: Single Attribute

Table 3.2: FairST compared to baselines for Seattle bikeshare demand prediction (single attribute)

	λ	MAE	RFG (race)	IFG (race)	Spearman's rho (race)
ConvLSTM[155]	0.00	0.432	74.485	22.907	0.210**
3D CNN	0.00	0.408	100.878	31.915	0.091
FairST	0.00	0.382	83.127	25.073	0.168**
FairST + RF	0.02	**0.379**	79.570	24.694	0.144**
FairST + RF	0.50	0.404	10.627	3.363	-0.076
FairST + RF	0.90	0.440	0.017	0.473	**-0.005**
FairST + IF	0.20	**0.379**	63.130	15.281	0.085
FairST + IF	1.50	0.406	38.473	4.902	-0.070
FairST + IF	5.00	0.442	**0.004**	**0.266**	-0.046

**. Correlation is significant at the 0.01 level.

*. Correlation is significant at the 0.05 level.

We trained FairST with Region-based Fairness loss (RF), Individual-based Fairness loss (IF), Equal Means loss (Equal Means), and Pairwise loss (Pairwise) respectively, on a single attribute, i.e. race on two datasets. Figure 3.4 illustrates the relationships between MAE and fairness metrics, each point on a curve corresponds to a λ value, which increases from left to right of the curve.

Figure 3.4 (a), (b), (d), and (e) show that RF and IF regularizer are very effective in controlling both RF and IF gaps. Overall, we observe a trade-off between MAE and IFG (or RFG). That is, accuracy decrease as fairness regularizer strength (λ) increases. In Seattle bikeshare case, IF regularizer ($\lambda = 0.2$) brings IFG down from 25.073 to 15.281 while keeping better accuracy than FairST with $\lambda = 0$ (see Table 3.2). This also suggests that the use of fairness loss terms ($\lambda > 0$) may *improve* the MAE over the baseline model ($\lambda = 0$) for small values of λ. The reason is that the addition of fairness terms provides a regularizing effect

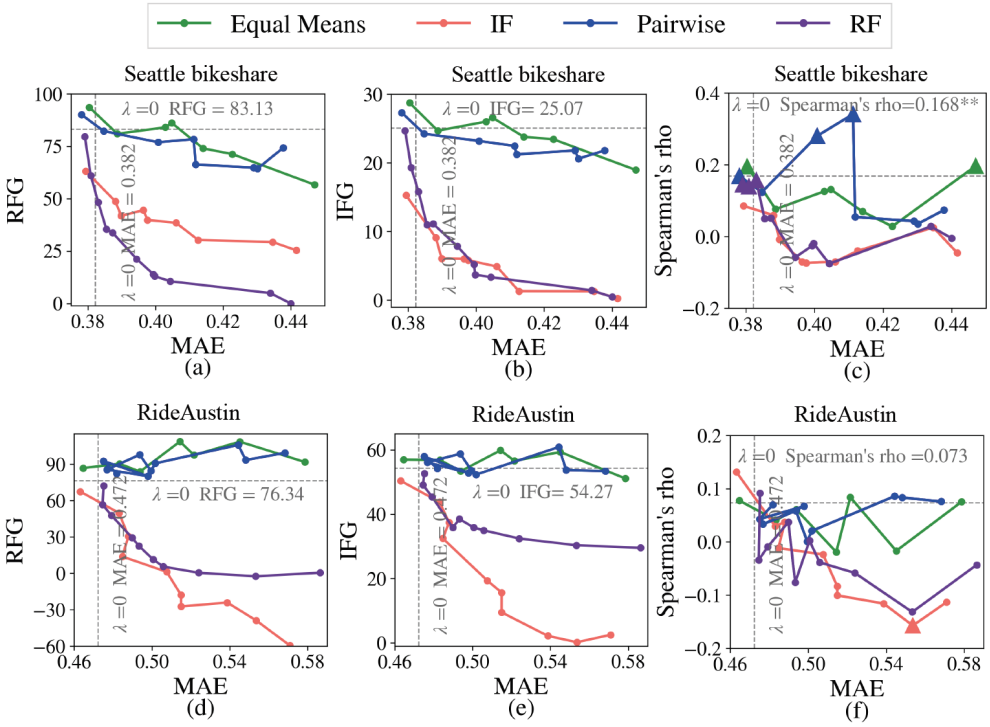

Figure 3.4: Accuracy vs. fairness metrics (single attibute). (a), (b), and (c) show the relationship between MAE vs. RFG, IFG, and Spearman's rho, respectively for Seattle bikeshare. (d), (e), and (f) show the results of RideAustin. Triangles in (c) and (f) represent statistical significance (p-value < 0.01).

on accuracy. In contrast, with the Equal Means regularizer or the pairwise regularizer, the models show no clear patterns in terms of RFG or IFG.

Figure 3.4 (c) and (f) show the fairness of models evaluated by Spearman's rho. Triangles represent statistical significance (p-value < 0.01). In Seattle bikeshare case, FairST ($\lambda = 0$) without fairness would result in an unfair prediction (see Table 3.2). That is, there is a positive monotonic correlation (Spearman's rho $= 0.168$, p-value < 0.01) between the predicted demand and the percent of Caucasian population. Models with an IF or a RF regularizer effectively bring down the Spearman's rho to around zero, and the predictions are no longer significantly correlated with race as λ increases. In contrast, Spearman's coefficients of models with an Equal Means regularizer or a pairwise regularizer stay positive throughout and sometimes show significantly positive correlation between the predicted outcome and race. In the RideAustin case, the predicted outcome of FairST ($\lambda = 0$) does not show a significant correlation with race. The Spearman's coefficients of models with an IF regularizer or a RF regularizer decrease and stay below zero, while the patterns of models with an Equal Means regularizer or a pairwise regularizer are less clear.

Table 3.2 shows the results of FairST compared to baselines for Seattle bikeshare. Both the RF regularizer and IF regularizer bring down about 85% IFG (from 31.915 to 3.363 and 4.902, respectively) while keeping better MAE than 3D CNN (MAE = 0.408). They also bring down IFG and RFG close to zero at MAE = 0.44. Similarly, Table 3.3 shows the results for RideAustin. Compared to 3D CNN, RF regularizer brings down about 99.5% RFG (from 62.004 to 0.347) and IF regularizer brings down 80.5% IFG (from 48.713 to 9.473) while keeping better accuracy.

In summary, in the single sensitive attribute scenario, FairST is able to achieve an accuracy better than the state-of-the-art baseline models while closing more than 80% of fairness gap. The proposed fairness regularizers are more effective than baseline fairness regularizers in reducing unfairness.

Table 3.3: FairST compared to baselines for RideAustin demand prediction (single attribute)

	λ	MAE	RFG (race)	IFG (race)	Spearman's rho (race)
Ground Truth	/	/	80.120	59.742	0.120*
HA	/	0.662	48.457	33.550	0.118*
ARIMA	/	0.597	82.587	61.457	0.117*
LSTM[42]	/	0.570	61.329	42.101	-0.049
ConvLSTM[155]	0.00	0.567	66.428	46.534	0.121
3D CNN	0.00	0.532	62.004	48.713	0.051
3D CNN + 1D	0.00	0.484	69.130	51.048	0.095
3D CNN + 2D	0.00	0.482	71.309	50.630	0.089
FairST	0.00	0.472	76.340	54.274	0.073
FairST + RF	0.05	0.475	56.703	49.092	**-0.034**
FairST + RF	0.80	0.524	**0.347**	32.436	-0.059
FairST + RF	1.00	0.553	-2.499	30.327	-0.132*
FairST + IF	0.06	**0.463**	67.358	50.357	0.131*
FairST + IF	1.20	0.515	-27.397	9.473	-0.100
FairST + IF	2.00	0.554	-38.990	**0.166**	-0.157**

**. Correlation is significant at the 0.01 level.

*. Correlation is significant at the 0.05 level.

3.6.3 Fair Prediction: Multiple Attributes

Having demonstrated the effectiveness of closing fairness gaps with IF and RF regularizers on a single sensitive attribute, we now turn to multiple sensitive attributes. We conduct two experiments on Seattle bikeshare dataset using RF loss and IF loss, respectively according to Equation 3.6. We set λ_a to be 1.0 for all three attributes, i.e. race, age, and education level.

Figure 3.6 shows the results of FairST with RF ((a) and (c)) and IF regularizer ((b) and (d)) evaluated using RFG and IFG. Overall, as λ increases, accuracy decreases and fairness

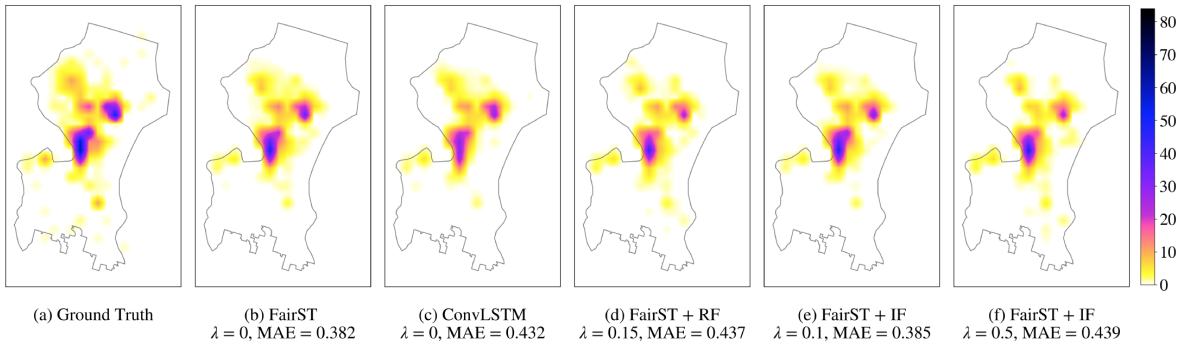

Figure 3.5: Ground truth vs. predictions heat maps for September 27, 2018 16:00 pm - 17:00 pm. (d), (e), (f) are the predictions from FairST using RF or IF regularizier on multiple sensitive attributes.

increases, indicating that both regularizers consistently help the model to approach equity on multiple sensitive attributes without sacrificing too much accuracy.

We now step back to compare FairST and baselines in terms of accuracy and fairness. Table 3.1 shows the results of FairST with RF regularizer and IF regularizer, denoted by FairST + RF and FairST + IF, with different λs. As can be observed, ground truth shows demand gaps between groups, indicating that there were more bikes picked up by whites, young people and college-educated people than the others. There are also significant positive correlations between demand and sensitive attributes (age and education level) as indicated by Spearman's coefficients. Compared to ground truth, all baseline models without fairness consideration amplify inequality in terms of one or more metrics. LSTM achieves good accuracy but drastically enlarges fairness gaps of race and education. ConvLSTM shows better fairness than all baselines in terms of IFG and RFG, but gives higher Spearman's coefficients for race and age than 3D CNN model. FairST with IF or RF regularizer can help reducing inequity in terms of all metrics. For example, compared to ConvLSTM, FairST + RF ($\lambda = 0.15$) and FairST + IF ($\lambda = 0.5$) show comparable accuracy but better fairness in terms of all fairness metrics. FairST + IF ($\lambda = 0.15$) outperforms 3D CNN in both accuracy

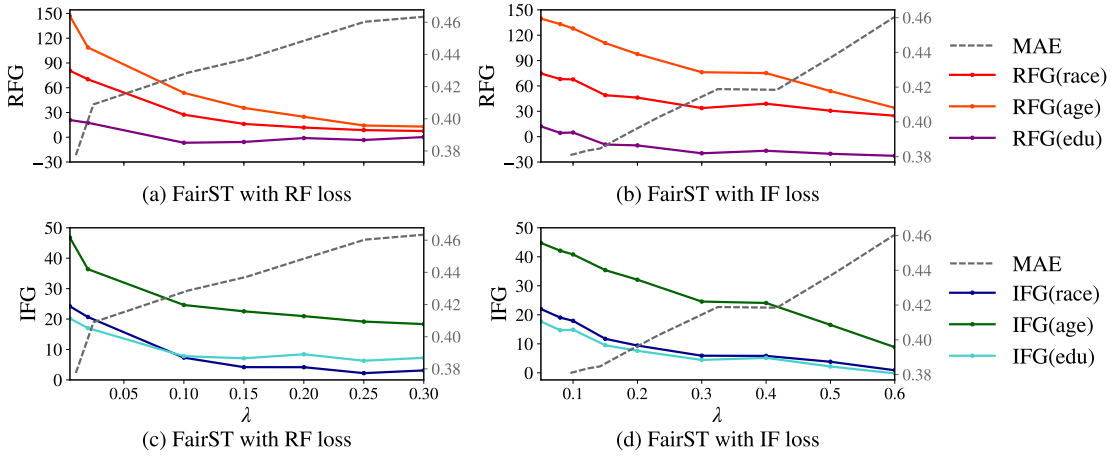

Figure 3.6: λ vs. fairness loss. (a) and (c) show the results of FairST with RF regularizer. (b) and (d) show the results of FairST with IF regularizer.

and fairness.

To understand better how FairST achieves fairness, we visualize the predictions from five different settings as illustrated in Figure 3.5. All five models are capable of learning spatial-temporal dependencies. FairST ($\lambda = 0$) accurately highlights the hot spots. Compared to ConvLSTM, FairSTs are better at capturing fragmented details around major hot spots. Adding fairness regularizers to FairST preserved the major hot spots but "re-weighted" some values in place and "redistributed" demand from some neighborhoods to others. For example, compared to Figure 3.5 (b) which does not consider fairness, Figure 3.5 (d) and Figure 3.5 (f) tend to capture more demand from the south part of the city where the disadvantaged population concentrates, and less demand from the northwest part of city which is dominated by the wealthy and well-educated population.

To summarize, in multiple sensitive attributes scenario, FairST is able to reduce unfairness for all three attributes consistently. With selected regularizer weight, FairST outperforms baseline models in both accuracy and fairness.

3.7 Summary

In this chapter, we proposed FairST, a fairness-aware spatio-temporal prediction model based on 3D convolutional neural network. A key feature of FairST is the integration of fairness regularizers to the model to encourage equitable prediction. We also proposed two fairness metrics that measure equity gaps between social groups for urban mobility systems. Experiments on two real-world new mobility datasets demonstrate that FairST is able to close more than 80% of fairness gap while achieving *better* accuracy than state-of-the-art but fairness-oblivious baseline methods. Further experiments show that FairST is able to reduce unfairness for multiple attributes without sacrificing much accuracy.

Chapter 4

LEARNING FAIR INTEGRATIONS OF OPEN URBAN DATA

In Chapter 3, we introduced a spatio-temporal prediction algorithm, called FairST, to correct biases during model training. FairST achieves good trade-offs between utility and fairness, but requires access to the raw data and the entire training pipeline of a prediction task. Therefore, FairST is suitable for decision makers and application algorithm/software developers who have full control over the ingredients of their prediction tasks. For data owners who hope to release their datasets for public use, they may concern that the potential biases embedded in their data will propagate into downstream applications, which they have little control over. In this chapter, we introduce a method to remove discriminatory signals from heterogeneous spatio-temporal urban datasets. Specifically, we develop an unsupervised algorithm to learn an integrated and fair data representation, called EquiTensor, from an array of urban datasets. EquiTensor aims to retain maximum information of the original datasets but reduce the discriminatory effects. In this way, data publishers such as government agencies can release EquiTensors without worrying about introducing much additional biases to the downstream applications [1].

4.1 Introduction

Predicting urban dynamics using deep learning based spatio-temporal methods is increasingly recognized as a critical capability in the public and private sector. These architectures have been applied to prediction problems for rideshare demand [160, 129], citywide crowd flow [181], traffic conditions [102, 173], accident patterns [174], public safety [74], and more [23]. All of these prediction problems are potentially influenced by a common set of

[1] The content of this chapter will appear in the following publication [162].

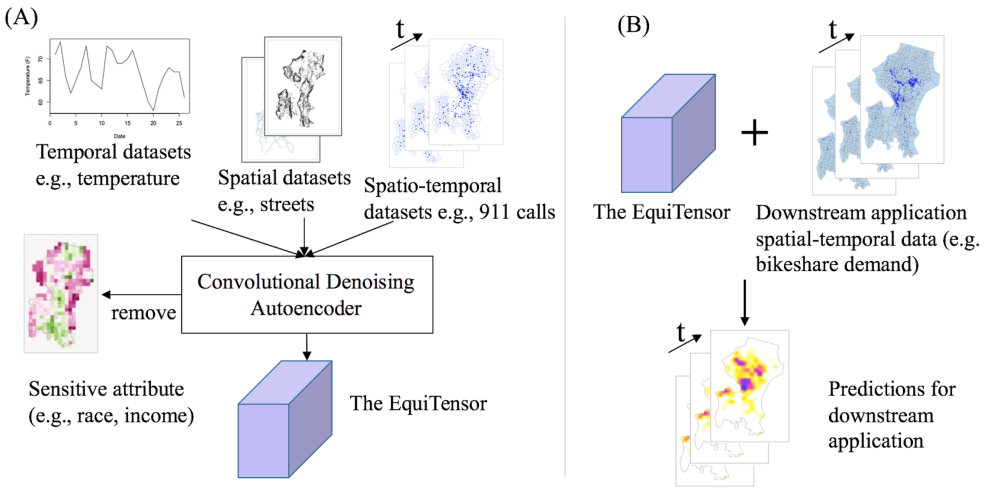

Figure 4.1: (A) An EquiTensor is a learned representation of heterogeneous spatio-temporal features that are free of sensitive demographic information and can be (B) shared across multiple prediction tasks to improve performance.

spatio-temporal factors (e.g., weather, housing prices, traffic, road networks). For example, predicting bikeshare demand depends on weather, topography, and traffic [160, 115], but the same data sources are also helpful for predicting citywide crowd flow and accident patterns [181, 174].

The use of exogenous features can significantly improve model accuracy [146, 129], but selecting and properly integrating potentially a large number of exogenous datasets requires both domain knowledge and substantial redundant engineering effort across applications; it is notoriously difficult to make effective use of open data [116]. More insidiously, the use of exogenous data will almost always reinforce systemic discrimination. For example, housing prices reflect historical discriminatory urban development policies [8] and transportation data reflects biased policies toward wealthy neighborhoods [123]. These sources of bias are propagated into prediction tasks, resulting in unfair predictions [186, 6] and exacerbating structural inequity.

In this chapter, we present an unsupervised learning architecture (Figure 4.1) to integrate heterogeneous spatio-temporal data and counteract bias, producing fair, accurate, and reusable (FAR) representations called EquiTensors. EquiTensors can be incorporated directly in a variety of urban applications to improve accuracy while limiting their exposure to additional discriminatory signals. The proposed architecture addresses three main challenges in learning representations for cities: reusability, accuracy, and fairness.

Reusability. We posit that a learned data representation is considered reusable when it can benefit multiple downstream prediction tasks. In urban settings, reusable representations should be generic across applications and contain rich and meaningful information about a city. Our solution is to design an unsupervised model that integrates as many potentially beneficial spatio-temporal datasets as possible without requiring application-specific feature engineering. Applications should be able to use these pre-trained features (representations) without sacrificing much performance relative to custom supervised "oracle" models that use specialized objective functions and hand-selected relevant datasets. One major challenge lies in the heterogeneity of multi-source open datasets. Urban datasets have varying dimensionality (e.g., topography does not vary with time, while regional-scale temperature does not vary with space), varying resolution (e.g., point events, city blocks), and varying coverage. To address this challenge, we align all datasets to a common spatio-temporal grid, then use a convolutional denoising autoencoder (CDAE) trained by backpropagating reconstruction error across all datasets. The CDAE uses separate encoder for each dataset at input and fuses the intermediate representations to generate a single integrated representation, from which individual decoders are applied to reconstruct each dataset. The unsupervised CDAE, along with task-agnostic pre-processing, is naturally robust to heterogeneous data sources. We refer this approach as the *core integrative model*.

Accuracy. The core integrative model is designed to take dozens and even hundreds of heterogeneous urban datasets. However, encoding a large number of potentially irrelevant datasets into one single representation can overwhelm a model with noise, resulting in sub-optimal reconstruction accuracy. To overcome this challenge, we consider an *adaptive*

weighting approach, which aims to balance the learning of multiple datasets. The core integrative model assigns equal weight to all datasets during training, but the learning process can be dominated by "easy" datasets with strong signals. The adaptive weighting dynamically adjusts the influence of each dataset on the total reconstruction loss based on its learning progress, focusing on slower-learning datasets and finding more general solutions. This approach is informed by similar approaches [94, 22] in multi-task learning, adapted for our reconstruction tasks.

Fairness. Most urban datasets are polluted by systemic socioeconomic and racial discrimination. For example, police incident reports are used to predict crimes, but their location and frequency reflect *only* policing practices rather than criminal activity [126]. To address bias, we incorporate an adversarial model that learns to detect a sensitive attribute (e.g., race, income, etc.) from the learned features; the core integrative model is rewarded for high adversarial error. We also pass the sensitive attribute to the decoder during reconstruction, forcing the decoder to further disentangle the sensitive attribute from other information in the latent space [62, 80, 106]. This approach combines adversarial learning for fairness [159, 157, 128] and learning disentangled representations [96, 29], but adapt them for an unsupervised data integration setting with continuous and distributed sensitive attributes (e.g., percent of high income residents in a region) as opposed to categorical attributes.

4.2 Related Work

Recent work in data management has recognized the challenges in organizing large open data repositories [24, 116]; our focus is making open data directly usable in prediction and learning tasks in urban computing [129, 181, 102, 173, 174, 23, 74, 185, 54, 12, 30]. The machine learning community has made remarkable progress in generalized representation learning [10], multi-task learning [125], and managing bias and discrimination [38, 178]. While our work adapts relevant techniques from these areas where appropriate, our specific context of spatio-temporal prediction, multi-dimensional heterogeneous input, and fairness-sensitive applications motivated the design of an end-to-end architecture specialized for this

setting. In this section, we position our approach in the broader context of related work across urban computing, machine learning, and data management.

Integration of Urban Data. Research on integrating open data [116, 113, 24] focuses on finding structural join and union relationships; we assume the only relationship between datasets is a common spatio-temporal domain and emphasize improving performance for prediction tasks. Representation learning has been effective for specific urban applications [149, 68, 105, 47, 69]; for example, Wang et al.[149] used an autoencoder architecture on GPS trajectories to study driver behavior. Our focus is on understanding the limitations of representation learning if we relax assumptions about the features, architecture, and objective of the target application.

Multi-task learning. Multi-task learning handles multiple related tasks simultaneously, aiming to achieve better performance than learning each task independently [125]. Some models use relationships among tasks to optimize feature sharing [100, 168, 114, 131]. For example, Vandenhende et al. [142] proposed a tree-like network based on task affinity scores, which were derived from the usefulness of the features of one task for the other. However, these approaches usually result in complex models that grow with the number of tasks [94]. Another approach is to balance the loss terms across tasks [94, 77, 22]. For example, Liu et al. [94] proposed a Dynamic Weight Average that adjusts task weights based on learning progress, showing that their method outperforms competitive methods including Uncertainty Weighting [77]. Our adaptive weighting approach is related to that of Liu et al. [94], but our setting of multiple inputs as well as multiple tasks admits new techniques, as we will describe in Section 4.5.1.

Fairness in Machine Learning. There exists extensive literature on fair machine learning [38, 178, 35, 57, 26], but few of them consider spatio-temporal applications. Yan and Howe [161] presented a fairness-aware prediction framework for urban mobility by incorporating fairness as a regularizer, but their approach relied on supervised learning. Unsupervised learning [178, 127, 99] and adversarial learning [159, 157, 128, 145] have been used to learn fair representations. For example, Madras et al. [104] proposed an encoder-decoder struc-

ture to learn a representation Z that predicts a supervised target and reconstructs the input while an adversary attempts to predict the sensitive information from Z. We adapt methods in adversarial learning for fairness [104] and image transformation [80] to predict continuous and spatially distributed sensitive attributes (e.g., a map of income) as opposed to only a categorical value (e.g., gender).

Overall, no existing methods attempt to integrate many heterogeneous datasets for broad reuse in many downstream urban applications. Further, learning fair representations for spatio-temporal settings have not been explored by current literature. We consider a primary contribution to be the scoping and definition of the problem of fair, unsupervised integration of heterogeneous urban data to make uncurated open data repositories safe and usable.

In this chapter, we introduce methods on learning reusable, accurate, and fair representations. We describe the core integrative model for combining heterogeneous urban datasets in Section 4.4. We then describe the adaptive weighting for improving the reconstruction accuracy over the core integrative model in Section 4.5. Finally, we describe the fairness module that seeks to remove discriminatory effects from the data representations generated by the integrative models.

4.3 Preliminaries

We use the City of Seattle as a case study to evaluate the proposed EquiTensors. This section details the datasets used to construct EquiTensors, four downstream prediction tasks used to evaluate the utility of EquiTensors, and the data processing methods.

4.3.1 Datasets

We collected 23 datasets from various online data portals, most of which are open data portals (Table 4.1). We included them because they are commonly used in urban studies [147, 109, 146, 129]. Meteorological data such as air quality is recorded city-wide, and are considered temporal (1D) datasets. Datasets that do not vary significantly over time, such as road networks, are considered spatial (2D) datasets. We included three spatio-temporal (3D)

datasets that vary in both space and time. We restrict these datasets according to Seattle city boundary. We chose the study period to be February 2014 to May 2019 as this period was covered by all temporal and spatio-temporal datasets. Socioeconomic data (percent of White residents and percent of Seattle households with income $\geq$ 100k in 2018) are defined at the block group level and were obtained from the SimplyAnalytics database [130]. We produced a race map and an income map based on 1km by 1km grids.

4.3.2 Downstream tasks

We consider four downstream tasks: three spatio-temporal predictions and one time series prediction (Table 4.2). They are:

- **Dockless bikeshare demand prediction (3D).** We collected Seattle dockless bikeshare data from the Transportation Data Collaborative. We use the number of bike pickup as a proxy for demand. The task is to predict next-hour bike demand for the city given the demand of last 7 days.

- **Reported crime incidents prediction (3D).** We obtained crime reports in Seattle from the City of Seattle Open Data [2]. The task is to predict the accumulated number of crime reports within three days in the next 3 hours based on the data of last 7 days.

- **Fire prediction (3D).** We obtained Seattle Fire Department 911 dispatches from the City of Seattle Open Data [3]. The task setup is the same as that of the crime reports prediction.

- **Bike count prediction (1D).** We obtained the number of bikes that cross the Fremont bridge from the City of Seattle Open Data Portal [4]. The task is to predict

[2] `https://data.seattle.gov/Public-Safety/Crime-Data/4fs7-3vj5`

[3] `https://data.seattle.gov/Public-Safety/Seattle-Real-Time-Fire-911-Calls/kzjm-xkqj`

[4] `https://data.seattle.gov/Transportation/Fremont-Bridge-Bicycle-Counter/65db-xm6k`

Table 4.1: Datasets for generating the Seattle EquiTensor

Name	Type	Source
Temperature	Temporal	NCEI
Precipitation	Temporal	NCEI
Pressure	Temporal	NCEI
Air quality	Temporal	Puget Sound Clear Air Agency
House price	Spatial	Zillow Home Value Index
POI (business)	Spatial	King County GIS data portal
POI (food)	Spatial	King County GIS data portal
POI (government)	Spatial	King County GIS data portal
POI (hospitals)	Spatial	King County GIS data portal
POI (public services)	Spatial	King County GIS data portal
POI (recreation areas)	Spatial	King County GIS data portal
POI (schools)	Spatial	King County GIS data portal
POI (transportation)	Spatial	King County GIS data portal
Transit routes	Spatial	King County GIS data portal
Transit signals	Spatial	King County GIS data portal
Transit stops	Spatial	King County GIS data portal
Seattle streets	Spatial	City of Seattle Open Data portal
Total flow count	Spatial	City of Seattle Open Data portal
Steep slopes	Spatial	City of Seattle Open Data portal
Bikelanes	Spatial	UW library GIS Data
Building permits	Spatio-temporal	City of Seattle Open Data portal
Traffic collisions	Spatio-temporal	City of Seattle Open Data portal
Seattle call data	Spatio-temporal	City of Seattle Open Data portal

Table 4.2: Downstream tasks for evaluation

	Task type	Time range	Known predictive "oracle" features
Bikeshare	Spatio-temporal	10/2017 - 10/2018	precipitation, pressure, temperature, slope, bikelane
Reported crime	Spatio-temporal	02/2014 - 05/2019	precipitation, pressure, temperature, house price, POI business, POI food, Seattle street, Seattle 911 calls,
Fire 911 calls	Spatio-temporal	02/2014 - 05/2019	precipitation, pressure, temperature, house price, POI business, POI food, Seattle street, total flow count, slope
Bike count	Temporal	02/2014 - 05/2019	precipitation, pressure, temperature

the hourly bike count for the next 6 hours based on the data of last 7 days. This tasks is a simple time series prediction, as the bridge is only one point in space.

4.3.3 Data pre-processing

To integrate heterogeneous spatio-temporal urban datasets, we reformat all datasets into a common rectilinear grid consisting of $W(width) \times H(height) \times T(time)$ non-overlapping cells, impute missing values with local average, and normalize values using max absolute scaling [5]. More sophisticated methods of imputation, feature engineering, and normalization exist, but we do not consider them in this dissertation.

Our input is N 1D datasets $D1_1, D1_2, ..., D1_N$ (time-varying, but not space-varying, such as weather), M 2D datasets $D2_1, D2_2, ..., D2_M$ (space-varying, but not time-varying, such as road networks), and L 3D datasets $D3_1, D3_2, ..., D3_L$ (varying in both space and time). A 1D dataset $D1_i$ with C_i attributes is aggregated into 1-hour intervals to produce a tensor of shape $T \times C_i$. Each 2D dataset may be either a set of point values (e.g., restaurant locations)

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html`

or regional values (e.g., housing prices aggregated by census tract). We rasterize point data by counting the events within each target cell. We rasterize regional data using proportional allocation based on area. A dataset $D2_j$ with C_j attributes therefore produces a tensor of shape $W \times H \times C_j$. A 3D dataset $D3_k$ with C_k attributes is aggregated into 1-hour intervals like a 1D dataset and rasterized into a spatial grid like a 2D dataset to produce a tensor of shape $W \times H \times T \times C_k$. The output of pre-processing is a set of training samples, where each training sample represents a 24-hour period. The training samples overlap: hours 0 to 23, 1 to 24, and so on are separate samples. Each training sample includes of all M 2D tensors, a 24-hour slice of each of N 1D tensors, and a 24-hour slice of each of L 3D tensors.

4.4 Reusability: Core Integrative Model

This section presents an algorithm, called *the core integrative model*, to learn reusable representation from heterogeneous multi-dimensional urban data. We begin with algorithm design, followed by experiments, results, and discussion.

4.4.1 Method

The core integrative model uses a convolutional denoising autoencoder (CDAE) that maps input datasets into a compact representation Z, then attempts to reconstruct all input datasets from Z. Compared to an autoencoder, a denoising autoencoder [80] reduces overfitting by dropping values at random in the input, forcing the encoder to learn a more general and robust mapping [10].

The encoder for the proposed model is illustrated in Figure 4.2(A). The input for this step is the set of training samples produced by pre-processing. Each training sample corresponds to a 24-hour period, and consists of (slices of) each of N 1D tensors, all M 2D tensors, and (slices of) each of L 3D tensors. To implement the denoising autoencoder, we corrupt each input tensor by setting 15% of the cell values to -1, at random. For each training sample, we then pass each corrupted tensor through three convolutional layers to learn intra-dataset patterns and collapse multiple attributes to a single feature. That is, each 1D input tensor

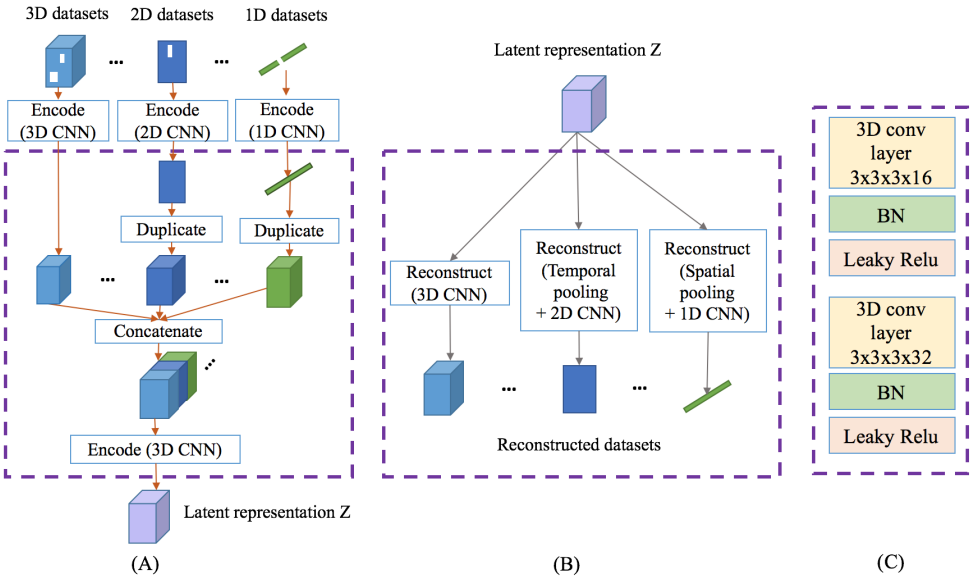

Figure 4.2: The core integrative model. (A): The encoder integrates 1D, 2D, and 3D datasets. (B): The decoder backpropagates the reconstruction error across all input datasets. (C): 3D CNN layers for encoding / decoding from the latent representation.

of shape $24 \times C_i$ (24 hours of a dataset with C_i attributes) is mapped to a representation of shape 24×1. The 2D and 3D cases are handled similarly, producing tensors of shape $W \times H \times 1$ and $W \times H \times 24 \times 1$ respectively. This design choice is consistent with the "late fusion" principle of learning individual representations before concatenating different datasets [184].

We then make the shapes of all datasets consistent by expanding 1D and 2D tensors to the 3D shape $W \times H \times 24$: 1D tensors are duplicated in space, and 2D tensors are duplicated in time. Then all of the $N + M + L$ tensors are concatenated into one large $W \times H \times 24 \times (N + M + L)$ tensor representing all features across all datasets. This concatenated tensor is then passed through three additional convolutional layers to produce a shared representation Z of shape $W \times H \times 24 \times K$ for $K \leq N + M + L$. Although we could use any shape for the representation, retaining the spatial and temporal dimensions allows direct visualization of the learned features, and also simplifies integration in downstream prediction tasks by affording straightforward restriction of the features to a particular sub-region or time period of interest.

The decoder is illustrated in Figure 4.2(B). We use three layers of 3D convolutional layers to reconstruct 3D datasets. For 1D data, we perform average pooling to reduce the spatial information and then apply three layers of 1D CNN. Similarly, we perform temporal pooling before three 2D CNN layers to reconstruct 2D datasets. For training, we use Mean Absolute Error (MAE) as accuracy loss. The reconstruction loss is a sum of MAE of each dataset.

Formally, let $\mathcal{X}$ be the input domain and $\mathcal{X}'$ be the corrupted input. Let n be the number of datasets and m be the number of training samples. The ith input for the CDAE is defined as $\mathcal{X}'^i = \{x'^i_1, x'^i_2, ..., x'^i_n\}$. The encoder Enc 'encodes' these corrupted tensors $\mathcal{X}'^i$ into a latent representation Z^i, from which each input tensor can be reconstructed as $\hat{\mathcal{X}}^i$ by a decoder Dec.

For training, we use Mean Absolute Error (MAE) as accuracy loss. The reconstruction loss is a sum of MAE of each dataset.

Let n be the number of datasets and m be the number of training samples.

$$L_{rec} = \frac{1}{m} \sum_{i=0}^{m} \sum_{j=0}^{n} |Dec(Enc(x_j'^i)) - x_j^i| \tag{4.1}$$

4.4.2 Experiments

We evaluate our core integrative model against several baseline methods, comparing the prediction accuracy for four downstream applications.

Baselines. We evaluate the integrative model by prediction accuracy. We use four baselines for comparison.

- **No exogenous data**: a 3D CNN based prediction model that only trains on historical data without any exogenous data (see Section 3.4.1 of Chapter 3).
- **Oracle features**: a network that makes use of hand-selected exogenous features, known to be predictive from the domain literature (Table 4.2). The "oracle" network for the three 3D tasks adopts the structure described in Section 3.4.1 of Chapter 3 [161], which is based on 1D, 2D, and 3D CNNs. For the 1D temporal prediction task, we use the seq-to-seq LSTM model as described in [136].
- **Principal component analysis (PCA)**: We generate a latent representation that summarizes the 23 datasets using PCA [153], which is then used in downstream tasks.
- **Early fusion**: We produce a latent representation with a CDAE. Instead of encoding each dataset separately, the early fusion CDAE concatenates all datasets as a single tensor, then applies 3D CNN layers. The decoder then reconstructs the concatenated tensor from the latent representation, which is then used in downstream tasks.

Implementation details. We implement all deep learning based models with TensorFlow [1], and perform training and inference with NVIDIA V100 Tensor Core GPUs. For the core integrative model, we adopt Adam optimizers using an exponential learning rate

decay strategy: the learning rate starts at a predefined value and decays every 5,000 steps with a rate of 0.96. We use a batch size of 32 and the initial learning rate is set to 0.01.

We generated training data by slicing the 1D and 3D datasets into 24-hour blocks with a sliding window of one hour, generating $n = 45960$ training samples covering 2014-02-01 to 2019-05-01. Each sample consists of m datasets: a slice of each 1D and 3D dataset and all 2D datasets. To construct the final representation, we chunk the 1D and 3D datasets into 24-hour sequences. Along with the 2D datasets, we pass the samples to the trained model. The non-overlapping part of the output from each sample were concatenated along temporal dimension to form the final output representation. In our experiments we compress the 23 datasets into 5 channels, so the shape of EquiTensors is $32(H) \times 20(W) \times 45960(T) \times 5$. The same shape is used for the two baselines (PCA and early fusion CDAE).

4.4.3 Results and Discussion

We show that the proposed core integrative model benefits downstream applications significantly and outperform the latent representations generated by the data integration baselines in terms of downstream prediction accuracy.

Integrated representations improve downstream task performance. Table 4.3 shows that the oracle networks with hand-selected features outperform the models without exogenous data in four cases, indicating that adding exogenous features is worthwhile. All representations learned by unsupervised methods including PCA, early fusion CDAE, and our method benefit the downstream tasks. Although some of the 23 datasets may not be relevant to the downstream tasks, predictions using the integrated representations still noticeably outperform the "No exogenous data" baselines. It suggests that the integration of multiple urban datasets can capture generic information that is reusable to an array of tasks. Downstream application developers can therefore choose to use the integrated representations instead of hand-picked features to avoid data collection across repositories, multiple pre-processing steps such as slicing and imputation, and feature selection. The integrated representations could also serve as baselines for evaluating feature selection strategies if the

Table 4.3: Prediction accuracy of downstream tasks. Accuracy is assessed with MAE and improvement over 'No exogenous data' model is provided in the brackets.

Model	Bikeshare	Crime	Fire	Bike count
No exogenous data [139, 160]	0.408 (/)	0.137 (/)	0.133 (/)	12.057 (/)
Oracle features [161]	0.382 (6.24%)	0.111 (18.72%)	0.110 (17.08%)	10.983 (8.91%)
PCA [153]	0.402 (1.39%)	0.121 (11.54%)	0.124 (6.62%)	11.099 (7.95%)
Early fusion	0.390 (4.36%)	0.119 (13.07%)	0.123 (7.53%)	11.266 (6.56%)
Core model	0.385 (5.51%)	0.113 (17.80%)	0.112 (15.66%)	11.050 (8.35%)

development cost is not a bottleneck.

Core integrative model outperforms baselines. Table 4.3 shows that the proposed model (Core model) outperform PCA and early fusion CDAE on all four tasks, and are competitive with the "best possible" oracle networks. PCA is simple and fast compared to deep-learning based methods, but it lacks the ability to model complex non-linear relationships. Early fusion CDAE takes the advantages of 3D CNN and shows superior performance to PCA. However, early fusion may not be effective in modeling intra-dataset dynamics, since all datasets are concatenated before being passed to the network [175]. Our method encodes each dataset separately at the input, allowing better modeling of individual datasets. Then the intermediate outputs are concatenated and fed to additional encoding layers, where the interactions among datasets are captured. As such, the proposed models outperform the early fusion CDAE in all four applications.

4.5 Accuracy: Adaptive Weighting

The core integrative model is designed to take dozens and even hundreds of heterogeneous urban datasets. However, encoding a large number of potentially irrelevant datasets into one single representation can overwhelm a model with noise, resulting in sub-optimal reconstruction accuracy. To overcome this challenge, we consider an *adaptive weighting* approach,

which aims to balance the learning of multiple datasets. The adaptive weighting approach does not change the network structure, but only adjusts the weight of the loss of each dataset during training according to its learning progress.

4.5.1 Method

The core integrative model assigns equal weight to all datasets during training (Equation 4.1), but the learning process can be dominated by "easy" datasets with strong signals. In particular, 1D and 2D datasets have repetition in their 3D representations, making them easier to learn. To alleviate this problem, we use an adaptive weighting scheme that adjusts the weight of the loss of each individual dataset dynamically during training according to its learning progress by assigning *larger* weights to datasets that "still have a long way to go" before they converge. This idea is inspired by recent work in multi-task learning [22, 94] in which the learning of a number of supervised sub-tasks needs to be balanced.

Chen et al. [22] calculate the weight of each task loss on every iteration, but this approach requires an additional backpropagation pass that slows down training. Our approach is informed by the Dynamic Weight Average method of Liu et al. [94], which adjusts the weights directly without manipulating the gradients. The main difference is that Liu et al. determine the weight of a task i based on the ratio of the loss of current step ($L(t)^i$) to the loss of the previous step ($L(t-1)^i$). When this ratio is low, the learning progress is (locally) high. However, this definition of progress over-emphasizes local variability in learning progress as opposed to global differences in the data sources.

Instead, we determine the weight based on the ratio of $L(t)^i$ to an "optimal" loss for that dataset, $L(opt)^i$, approximated by the reconstruction error of a CDAE trained separately for that specific dataset alone. When the loss for timestep t ($L(t)^i$) is high relative to the optimal loss, that dataset receives a higher weight. As the loss gets closer to the optimal loss, the weight is lower. With this approach, we accommodate the differences in loss scales across datasets, encouraging the model to minimize reconstruction error across different datasets in a balanced and coordinated way. We consider this adaptive weighting scheme a variation

of the Dynamic Weight Average [94].

Specifically, we define the weight $w^i(t)$ at training epoch t as:

$$w^i(t) = n\frac{exp(r^i(t)/\alpha)}{\sum_{j=0}^{n} exp(r^j(t)/\alpha)} \tag{4.2}$$

where n is the number of datasets. α is a parameter controlling the degree to which learning progress influences the weights [94]. Larger α leads to more equal weights among datasets. $r^i(t)$ is relative learning progress for dataset i at epoch t [22]. The learning progress $LP^i(t)$ is normalized by the average learning progress of datasets and is written as:

$$r^i(t) = LP^i(t)/E_n[LP^i(t)], \quad LP^i(t) = L(t)^i/L(opt)^i \tag{4.3}$$

where $L(t)^i$ is the loss for dataset i at epoch t, which we calculate as the mean loss of the first 50 steps of each epoch in our implementation. $E_n[LP^i(t)]$ is the average learning progress of all datasets. The weights are initialized to 1.0 at the first epoch and updated once every epoch.

4.5.2 Experiments

We evaluate the effectiveness of the adaptive weighting scheme on total reconstruction error of the integrative model. We show that the adaptive weighting scheme improves the reconstruction accuracy over the core integrative model.

4.5.3 Results and Discussion

Adaptive weighting dynamically adjusts the weight for the reconstruction loss of each dataset based on its learning progress. The influence of learning progress on the weights is controlled by a strength factor α (Equation 4.2). Figure 4.3 shows how the total reconstruction error varies with α. Compared to the core model (dashed grey line), our adaptive weighting (blue line) helps reduce the total reconstruction error. Larger α values result in more equal weights, approximating the core model performance (MAE = 0.217). Compared to the Dynamic

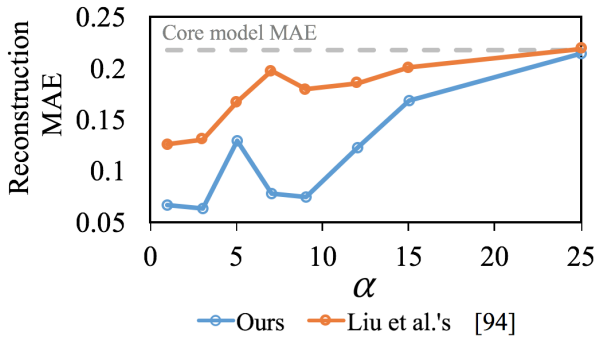

Figure 4.3: Total reconstruction error vs. α. Our adaptive weighting versus Dynamic Weight Average [94].

Weight Average by Liu et al. [94] (orange line) [6], our adaptive weighting consistently achieves higher total reconstruction accuracy for a range of α values. We use $\alpha = 3$ (MAE = 0.063) for the rest of our experiments, as it produced the best reconstruction accuracy.

Figure 4.4 illustrates how adaptive weighting influences the reconstruction accuracy for three datasets. During the training of the core model, the reconstruction loss for Collisions and Building permits quickly plateaued (blue lines). The adaptive weighting scheme increased their weights (grey lines) in the first few epochs to encourage them to learn faster. Once their errors dropped, the weights went down to about 1.0. For Slope, both models were making steady progress, so the weights remained at about 1.0. We observe that the datasets that benefit the most from adaptive weighting are 3D datasets, as they embody more complex spatial or temporal correlations than 1D and 2D datasets.

Table 4.4 compares the downstream accuracy of four tasks using the representations generated by the core integrative model with or without adaptive weighting ($\alpha = 3$). We see that the performance of the four downstream applications was not obviously affected by the use of adaptive weighting, although the total reconstruction accuracy was significantly im-

[6]We implemented a slightly different version from the original Dynamic Weight Average. We computed $LP^i(t)$ according to Equation 4.3, but replaced $LP^i(t) = L(t)^i/L(opt)^i$ by $LP^i(t) = L(t)^i/L(t-1)^i$

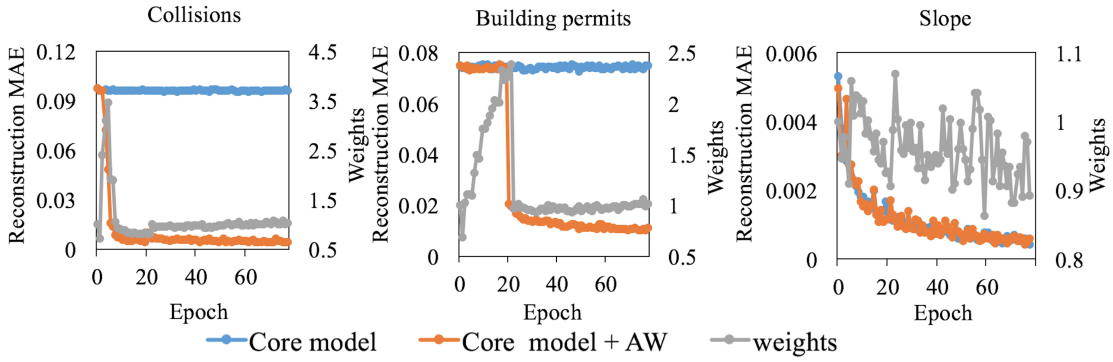

Figure 4.4: Reconstruction loss curves and adaptive weight curves on three datasets ($\alpha = 3$). Under adaptive weighting scheme (Core model + AW), weights for individual datasets change with their reconstruction accuracy.

Table 4.4: Prediction accuracy (MAE) of downstream tasks using the core integrative model and the core integrative model with adaptive weighting ($\alpha = 3$).

Model	Bikeshare	Crime	Fire	Bike count
Core model	**0.385**	0.113	**0.112**	11.050
Core model + AW	0.387	**0.106**	0.114	**11.049**

proved. The reason could be that the improvement of total reconstruction accuracy primarily comes from datasets that are not predictive for the four applications. It is also possible that even though the datasets relevant for downstream tasks get better reconstruction accuracy, the improvement upstream does not necessarily transfer to downstream performance. The relationship between the accuracy of pre-trained representations and downstream accuracy remains open and needs further research.

4.6 *Fairness: Learning Fair Representations*

This section describes the fairness module that seeks to remove discriminatory effects from the data representations generated by the integrative models. We refer the integrative models with fairness treatment as EquiTensor models and the resulting representations as *EquiTensors*. In this chapter, we use the core integrative model (and the core integrative model with adaptive weighting) as the basis of the EquiTensor model.

4.6.1 Method

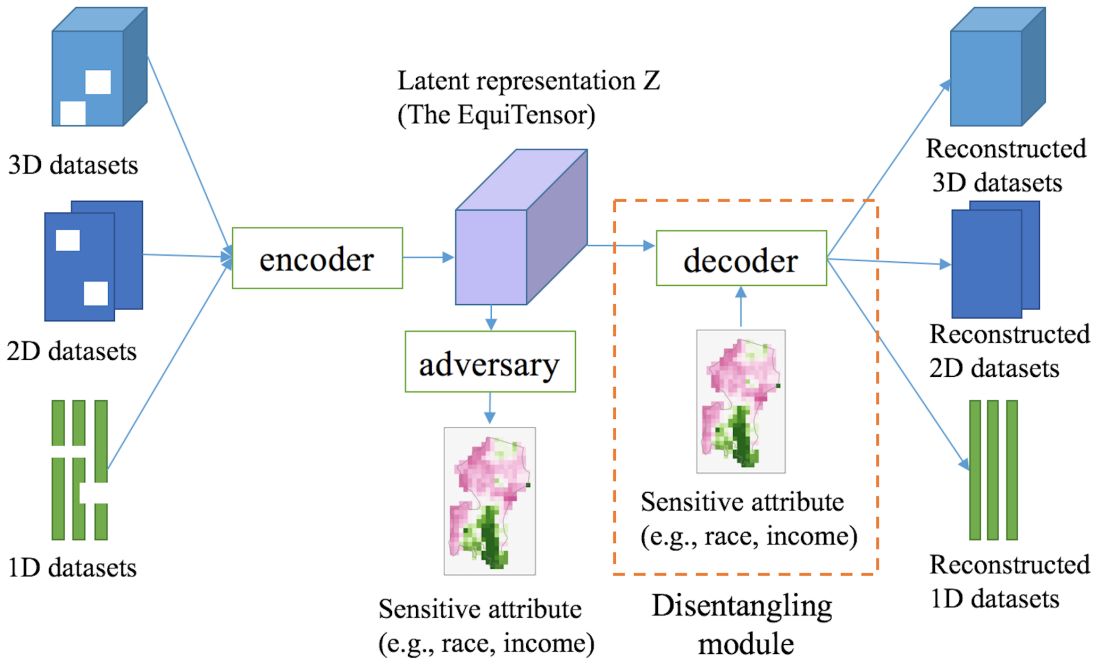

Figure 4.5: The EquiTensor model architecture. The encoder and decoder learn a latent representation Z (the EquiTensor) by minimizing reconstruction error. The sensitive attribute S (e.g., race) is passed to the decoder to disentangle S from other information in Z. The adversary learns to predict S given Z, penalizing the encoder.

Figure 4.5 shows the architecture of the EquiTensor model. It is based on a core integra-

tive model (i.e., CDAE) and a fairness module. The fairness module consists of two parts: a *disentangling module* and an *adversarial model*. First, we use a disentangling module to separate the sensitive attribute S (e.g., race) from other information in the latent space during the reconstruction phase. The decoder uses both the latent representation Z and S to reconstruct the input, learning to disentangle S from Z [106, 80, 104]. Second, an adversarial model A learns to predict S from Z, and the core integrative model is penalized accordingly. The idea of using adversarial model for fairness is inspired by the success of Generative adversarial networks (GAN) [53] in many applications such as generating fake faces [75] and image style transfer [27]. The GAN network typically consists of a generator and a discriminator, which are trained together to work against each other [53]. For example, in generating fake faces, a generator tries to generate fake images as real as possible to fool the discriminator, where the discriminator learns to tell the fake from the real ones. In learning fair representation, a typical network architecture is that a generator tries to learn a representation that can reconstruct the input data or predict the target (depending on the application), whereas a discriminator tries to predict the categorical sensitive attribute (e.g., high income or low income individual, under or over 70 years old) from the representation. Further details of adversarial learning are not discussed here. Interested readers are referred to the seminal work by Goodfellow et al. [53]. The main difference between our approach and other similar methods in various application domains [80, 104] is that we use a regression model as the "discriminator" to predict continuous and spatially distributed sensitive attributes (e.g., percentage of high income households in a region), as opposed to a binary classifier which is used in most of the adversarial models.

The adversarial approach is particularly desirable in our setting of integrating multiple datasets, since a single model can simultaneously remove the effects of a sensitive attribute that are encoded in many different input datasets. The adversary loss is defined as:

$$L_A = \frac{1}{m} \sum_{i=0}^{m} |A(Z^i) - S| \tag{4.4}$$

where m is the number of training samples. Z^i is the representation learned from the ith training sample.

Final objective function for CDAE. The CDAE has two objectives: minimizing the reconstruction error while being penalized by the adversary (Figure 4.5). The loss for CDAE is written as:

$$L_{AE} = \frac{1}{m}\sum_{i=0}^{m}\sum_{j=0}^{n}|Dec(Enc(x_j'^i), S) - x_j^i| + \lambda(1 - \frac{1}{m}\sum_{i=0}^{m}|A(Z^i) - S|) \tag{4.5}$$

where the first term is the reconstruction error and the second term is the negative adversarial loss $1 - L_A$. A parameter λ controls the trade-off between the two terms.

The adversary consists of three 3D CNN layers. The CDAE is trained jointly with the adversary in alternating periods. For each mini-batch of the training data, we 1) update the encoder and decoder while fixing the adversary to minimize L_{AE}, and 2) update the adversary while fixing the encoder and decoder to minimize L_A.

4.6.2 Measuring Fairness

We evaluate the fairness of EquiTensors by 1) measuring the adversarial model's ability to discern the sensitive attribute and 2) measuring the fairness of downstream prediction applications that use the EquiTensor.

The independent adversary. We train a separate adversarial model F instead of directly using the adversarial model A of Equation 4.4 because training a separate model achieves higher accuracy (and is therefore a more stringent evaluation). The higher the MAE of F, the better the protection against unfairness in the EquiTensor.

Fairness metrics. We measure the fairness of downstream predictions using fairness metrics. We measure the disparities by the gap in prediction errors across an advantaged group G^+ and a disadvantaged group G^-. Our unsupervised setting makes no assumptions about downstream applications, so overestimation may be beneficial (e.g., overestimation of bikeshare demand leads to more availability of bikes) or harmful (e.g., overestimation

of law enforcement incidents could lead to increased police presence.) We adapt residual difference (RD) and its positive (PRD) and negative (NRD) variants [18, 60, 167] to our spatio-temporal setting since most fairness metrics apply only to classification settings and are not applicable here.

Let s_i be the ith rectilinear cell of the study area $\mathcal{S}$. Let $\hat{y}_{i,t}$ and $y_{i,t}$ be the prediction and ground truth for cell s_i at time t, respectively. We denote G^+ as the advantaged group and G^- as the disadvantaged group, with regard to one sensitive attribute S (e.g., race). $|G^+|$ and $|G^-|$ are the number of cells in G^+ and G^- group, respectively. We denote H as a hinge function where $H(x) = max\{0, x\}$. We define *positive residual (PR)* for cell s_i at time t as $PR_{i,t} = H(\hat{y}_{i,t} - y_{i,t})$, *negative residual (NR)* as $NR_{i,t} = H(y_{i,t} - \hat{y}_{i,t})$, and *residual* as $R_{i,t} = \hat{y}_{i,t} - y_{i,t}$ [167].

Positive residual difference (PRD) is written as:

$$PRD = \frac{1}{|G^+|} \sum_{t=0}^{T} \sum_{i \in G^+} PR_{i,t} - \frac{1}{|G^-|} \sum_{t=0}^{T} \sum_{j \in G^-} PR_{j,t} \tag{4.6}$$

The first term is the overestimation for each square region in G^+ over a time period T and the second term is the overestimation for G^- over T.

Negative residual difference (NRD) and the symmetric *Residual difference (RD)* can be defined similarly by replacing $PR_{x,t}$ with $NR_{x,t}$ and $R_{x,t}$ respectively. RD measures the difference between the overall overestimation (or underestimation) across two groups.

4.6.3 Experiments

We generate EquiTensors using the framework in Figure 4.5 to remove the influence of sensitive information. We evaluate fairness and accuracy on two downstream tasks: reported crime incidence prediction and bikeshare demand prediction, and compare with two competing baselines.

Fair Representation Baselines. We compare the EquiTensor with a state-of-the-art method for producing fair spatio-temporal representations in supervised and non-integrative

settings that we adapted for our purposes, and a simpler version of our own method.

- **Fair CDAE**: Based on the core integrative model (i.e., CDAE), Fair CDAE uses an additional prediction head H that is trained to learn the sensitive information from the latent representation. Instead of using adversarial training, H is trained together with the CDAE. Fair CDAE minimizes the reconstruction error and simultaneously maximizes the MAE of H. The idea is inspired by Wadsworth et al.[145], but we adapted their method to our unsupervised learning scenario, and we applied a *gradient reversal* layer proposed by [49] on H, so that the minimax optimization can be achieved using standard back-propagation.

- **EquiTensor without the disentangling model (Core + Fair w/o disent.)**: This method is equivalent to the EquiTensor architecture without the disentangling module.

Implementation details. To calculate the three fairness metrics PRD, NRD, and RD (Section 4.6.2), we need to define the advantaged group G^+ and the disadvantaged group G^- with respect to a sensitive attribute. We use the mean city statistics as thresholds to label a square region as either G^+ or G^-. For example, since 65.74% of the overall population of Seattle is white, we label the regions with $\geq$65.74% white as G^+ and the others as G^-. We discretized income level using the same method.

The independent adversary is used for evaluating the fairness of a latent representation. The idea is to predict a sensitive attribute S from the latent representation Z. Higher accuracy (smaller MAE) suggests possible unfairness. Figure 4.6 shows the architecture of the adversary. It takes a 24-hour slice of the latent representation as input and tries to predict S, which is formatted as a map duplicated 24 times along the temporal axis. In our implementation, we generated training data for the adversary by slicing the latent representation ($32 \times 20 \times 45960 \times 5$) along temporal dimension by a step of 24 hours with a sliding window of one hour. We use a batch size of 32 and train for 30 epochs. The initial

learning rate is set to 0.01. The architecture of the independent adversary is the same as the adversary embedded in the EquiTensor model (Figure 4.5).

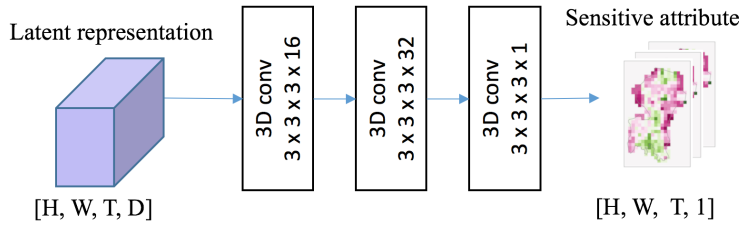

Figure 4.6: The architecture for the independent adversary. H, W, T, D denotes height, width, time steps, and dimension.

4.6.4 Results and Discussion

We evaluate the EquiTensors using two case studies. For reported crimes we remove the effects of race and for bikeshare we remove the effects of income. We show that EquiTensors are more fair than baseline representations and can help the downstream tasks to achieve accuracy that is competitive with the oracle networks.

EquiTensors are fairer than baseline representations. Table 4.5 shows the accuracy of predicting sensitive information S (i.e., race and income) from latent representations generated by different models: Our core model with fairness (Core model + Fair and Core model + Fair + AW), the integrative model baselines (PCA, Early fusion, Core model, and Core model + AW), and two competing fairness approaches: Fair CDAE and EquiTensors without the disentangling module (Core model + Fair w/o disent.).

The sensitive information S (i.e., race and income) was detectable with much lower error in the non-fairness-treated representations (i.e., PCA, Core model, etc.) than with EquiTensors (Core model + Fair and its variants). This result suggests that EquiTensors can improve fairness over fairness-oblivious baselines. The low error of Fair CDAEs suggests that Fair CDAEs are not effective in removing the influence of S. The reason is that Fair

Table 4.5: Accuracy (MAE) of predicting sensitive attribute (i.e., race and income) from various integrated representations.

	λ	Race MAE	Income MAE
PCA [153]	/	0.005	0.005
Early fusion	/	0.001	0.001
Core	/	0.001	0.001
Core + AW	/	0.001	0.001
Fair CDAE [145, 49]	1.0	0.002	0.002
Fair CDAE [145, 49]	10.0	0.001	0.001
Core + Fair w/o disent.	0.6	0.002	0.001
Core + Fair w/o disent.	1.0	0.029	0.053
Core + Fair w/o disent.	2.0	0.076	0.112
Core + Fair	0.6	0.052	0.021
Core + Fair	1.0	0.067	0.073
Core + Fair	2.0	**0.129**	0.112
Core + Fair + AW	0.6	0.038	0.052
Core + Fair + AW	1.0	0.037	0.079
Core + Fair + AW	2.0	0.094	**0.113**

Table 4.6: Accuracy and fairness of crime predictions with different integrated representations in the form of mean(std).

	λ	Accuracy	Fairness	
		Prediction MAE	RD	PRD
No exo. data	/	0.135 (0.002)	-23.138 (3.624)	-27.678 (1.931)
Oracle	/	0.110 (0.007)	-11.994 (5.868)	-20.267 (4.282)
PCA	/	0.117 (0.004)	-12.962 (3.639)	-20.159 (3.517)
Early fusion	/	0.115 (0.004)	-12.499 (3.207)	-20.297 (2.442)
Core	/	0.111 (0.002)	-8.908 (5.900)	-17.850 (4.318)
Core + AW	/	**0.106 (0.004)**	-5.642 (7.194)	-15.197 (5.122)
Core + Fair	0.6	0.114 (0.004)	-4.248 (4.751)	-14.255 (3.016)
Core + Fair	1.0	0.112 (0.006)	6.919 (9.902)	**-7.771 (6.957)**
Core + Fair	2.0	0.112 (0.004)	4.669 (7.492)	-9.303 (4.392)
Core + Fair + AW	0.6	0.111 (0.002)	-3.910 (3.081)	-14.576 (2.118)
Core + Fair + AW	1.0	0.110 (0.005)	5.098 (6.243)	-9.160 (4.650)
Core + Fair + AW	2.0	0.109 (0.004)	**-3.865 (5.843)**	-14.882 (3.453)

CDAE uses a prediction head H embedded within the network, such that a single set of parameters must be found that simultaneously maximizes the accuracy of H and minimizes the reconstruction error [13]. In contrast, we train our adversary separately, so that it only focuses on recovering S from without accommodating reconstruction errors like Fair CDAE. Finally, we see that our models better separate the sensitive information than the models without the disentanglement (Core model + Fair w/o disent.).

Table 4.6 shows the results of crime report predictions using latent representations generated by different models. We reported mean values and standard deviation (parentheses in Table 4.6) of five repeated runs for each model. In a perfectly fair scenario, the three metrics NRD, PRD, and RD should be zero. In the crime report prediction case, The PRD for "No exogenous data" model is -27.68, indicating that over the testing period the model overestimates the reported crime incidents for each cell in non-white neighborhoods by 28 cases more than the overestimation for white neighborhoods. The RD for 'No exogenous data'

Table 4.7: Accuracy and fairness of bikeshare predictions with various integrated representations in the form of mean(std).

	λ	Accuracy	Fairness	
		Prediction MAE	RD	NRD
No exo. data	/	0.408 (0.002)	8.661(20.542)	-151.975 (10.178)
Oracle	/	**0.382(0.002)**	73.313 (29.693)	-180.119 (17.752)
PCA	/	0.400 (0.003)	55.263 (36.637)	-181.624 (21.494)
Early fusion	/	0.390 (0.006)	75.140 (35.765)	-183.178 (21.721)
Core	/	0.385 (0.001)	58.392 (19.104)	-172.065 (9.345)
Core + AW	/	0.388 (0.002)	32.206 (13.569)	-160.273 (7.937)
Core + Fair	0.6	0.392 (0.002)	23.052 (25.576)	-158.482 (11.258)
Core + Fair	1.0	0.395 (0.006)	15.189 (32.480)	-154.733 (16.416)
Core + Fair	2.0	0.394 (0.005)	6.251 (28.036)	-151.316 (12.040)
Core + Fair + AW	0.6	0.390 (0.003)	**-5.320 (18.273)**	**-142.137 (7.890)**
Core + Fair + AW	1.0	0.394 (0.005)	11.203 (43.638)	-153.330 (20.472)
Core + Fair + AW	2.0	0.398 (0.003)	28.371 (32.938)	-161.627 (16.367)

model is -23.14, indicating that the residual (prediction minus ground truth) is 23 more for the non-white regions than white regions. In other words, the "No exogenous data" model is discriminatory by the definition of residual difference, and the reported crime data itself contains correlations with race that can be amplified by the prediction model. In this case, incorporating exogenous features mitigate these biases, as shown by the RD and PRD being closer to zero relative to the baselines. Crime report predictions with EquiTensors further improves RD and PRD, leading to overall fairer predictions.

Table 4.7 shows the results of bikeshare predictions with various representations. In the bikeshare case, we focus on RD and NRD, because underestimating the bike demand of underrepresented communities is considered more harmful than overestimating that of the advantaged groups. The negative NRD of the "No exogenous data" model indicates that the model underestimates the bikeshare demand in low-income regions more than the high-income regions. Nevertheless, its mean RD value (8.66) shows that the prediction errors

are roughly balanced. In contrast, all baseline models without fairness treatment (Oracle models and the core models) show larger disparities than the "No exogenous data" model, as suggested by their RD values. This implies that external features may introduce additional biases into the prediction models. With EquiTensors, fairness for bikeshare predictions are improved compared to the fairness-oblivious baselines, as the mean values of RD and NRD are overall closer to zero relative to those of the baselines. We observe that RD and NRD of bikeshare predictions show large standard deviations. The reason could be that the fairness metrics (i.e., RD and NRD) are not very precise and influenced by the threshold that defines the advantaged and disadvantaged groups. It is also possible that the transferability of EquiTensors' fairness is not strong or stable for some applications as we do not explicitly optimize over RD or NRD in the downstream predictions. The downstream models sometimes may still be able to pick up and exaggerate the remaining information about the sensitive attribute if the predictions heavily rely on it, although the EquiTensors make them increasingly harder to do so as λ increases. The uncertainties of transferability showed by the unsupervised representations are expected and need further investigation. However, despite the large deviation, we see that EquiTensors can offset the bias introduced by the exogenous data and show overall fairer results than the fairness-oblivious baselines. In summary, these results suggest that the improved fairness of EquiTensors may help prevent amplifying socioeconomic disparities in downstream predictions.

EquiTensors help improve downstream accuracy. Table 4.6 shows that crime predictions with EquiTensors (Core + Fair and Core + Fair + AW) show accuracy that is comparable to the oracle networks. Nevertheless, we did observe that predictions tend to overfit as λ increases. The reason could be that the EquiTensor becomes increasingly noisy to crime prediction as more sensitive information is removed. We overcame this issue by early stopping. Table 4.7 shows that even with the fairness treatment, EquiTensors can help the bikeshare predictions achieve higher accuracy than the "No Exogenous features" baseline. Therefore, downstream tasks with EquiTensors tend to achieve overall fairer predictions than using fairness-oblivious features without sacrificing much accuracy.

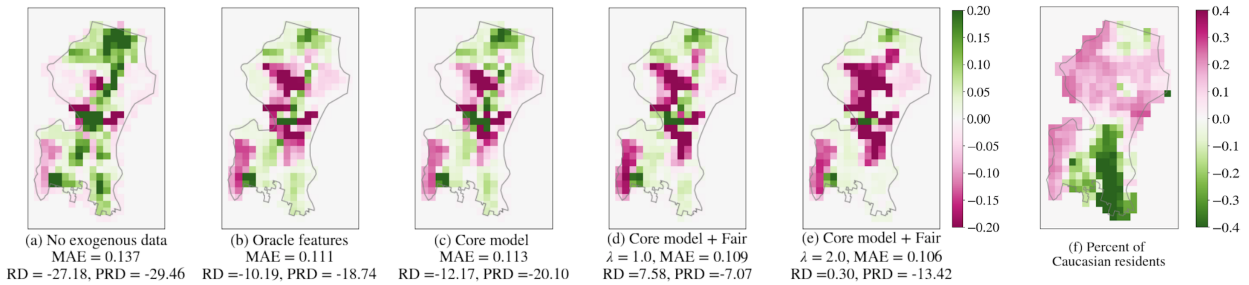

Figure 4.7: Residual (prediction - ground truth) map of crime prediction on November 28, 2018 18:00 pm - 21:00 pm. Green grids are overestimated and pink ones are underestimated. (a), (b), and (c) show the predictions from models without fairness treatment. (d) and (e) are the predictions from models with EquiTensors. (f) shows the normalized map (subtracted by city mean) of percentage of Caucasian residents. Green grids have higher percentage of non-Caucasian residents than pink ones.

Qualitative results. Figure 4.7 provides a qualitative understanding of how EquiTensors influence the spatial distribution of crime prediction results. Figure 4.7 (a), (b), and (c) are predictions from models without fairness treatment. They show clear overestimation for the southeast and the north part of the city, where non-Caucasian residents concentrate (green grids in Figure 4.7 (f)); and underestimation in some Caucasian neighborhoods (pink grids in Figure 4.7 (f)). Noticeably, predictions with EquiTensors (Figure 4.7 (d) and (e)) show much less overestimation for non-Caucasian neighborhoods in the north and south.

Parameter sensitivity analysis. The parameter λ in the loss function for EquiTensor (Equation 4.5) controls the weight of the bias mitigation. Figure 4.8 (A) and (B) show the MAEs of adversary with changing λs for race prediction and income prediction, respectively. We use a tensor filled with Gaussian noise as a baseline (dashed grey lines); the model should exhibit high error when attempting to discern S from noise. We observe in both the race and income prediction cases that the best value of λ is around 2.0. Increasing λ from 0 to 2.0 leads to increasingly fairer EquiTensors. When λ is around 2.0, adversary MAEs approach

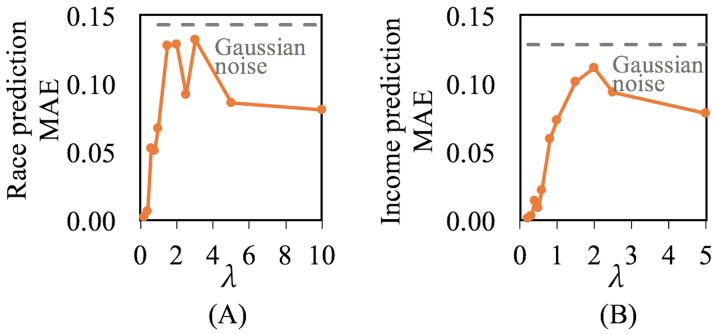

Figure 4.8: Adversary MAE VS. λ. At $\lambda \approx 2$, EquiTensors prevent discerning the sensitive attribute nearly as well as Gaussian noise.

the MAEs of Gaussian noise, suggesting that the influence of the sensitive attribute can be largely removed. As λ increases, fairness levels off. Larger values of lambda cause the adversary loss to dominate reconstruction error in Equation 4.5, leading to noisy EquiTensors with lower utility.

Limitations and future work. There are several limitations in this work. First, we included as many as possible spatio-temporal datasets we could find in learning EquiTensors. There may be datasets we included but are not useful to some urban tasks. Future work could develop protocols to select certain data to include and investigate what kinds of downstream tasks we could support with the selected data. The selection criteria could also explore aspects such as data quality, spatio-temporal coverage, context of use, uniqueness, and popularity. It is particularly interesting to investigate how data quality problems induced by data repurposing, cleaning, and integration influence the reusability of the integrated representation. Second, partitioning all datasets into a common square grid may risk losing important information of the original data. For example, representing road network data using grids cannot capture highway directions and intersections, therefore losing its physical meaning [84]. In addition, grid representation is not ideal for sparse datasets such as train stations in a city. Future work involves handling sparse datasets and managing

datasets with varying coverage, resolution, and formats (e.g., networks, points, etc.). Finally, fairness metrics (i.e., RD and NRD) of bikeshare predictions show large standard deviations, suggesting that the metrics are not very precise or the transferability of EquiTensors' fairness is not strong or stable. Future work could develop finer-grained fairness measurements that consider spatial sub-regions and temporal dynamics, explore the transferability of fair representations, and understand which applications are most likely to benefit from the improved fairness of EquiTensors.

4.7 Summary

The use of exogenous features can significantly improve the accuracy of prediction models [146, 129]. Many urban prediction tasks rely on a common set of external features such as weather and road networks. We propose to learn an integrated representation from a variety of commonly used open datasets, aiming to benefit multiple downstream tasks without assuming specific application domains. We first developed a core integrative model based on denoising convolutional autoencoder to learn reusable representation from heterogeneous and multi-dimensional urban datasets. Our results show that the core integrative model can help an array of downstream tasks to achieve prediction accuracy comparable to "oracle" networks that are trained with hand-selected relevant features.

Drawing on recent literature on multi-task learning, we developed an adaptive weighting scheme for balancing the training of the core integrative model. Our results show that adaptive weighting can achieve superior reconstruction accuracy to that of the core integrative model. We found that the improved reconstruction accuracy of the integrated representations does not necessarily lead to better downstream performance. The relationship between reconstruction accuracy of the integrated representations and the downstream performance remains open for future research.

To combat discriminatory signals embedded in the data, we used a fairness module based on *adversarial learning* on top the integrative models to produce fair and integrated representations (i.e., the EquiTensors). We showed that EquiTensors contain weaker correlations

with the sensitive attributes than the baseline representations. Our results on real-world datasets and applications suggest that EquiTensors could help prevent the discrimination embodied in the data from propagating to the downstream prediction models and at the same time help them to achieve prediction accuracy comparable to 'oracle' networks that trained with hand-selected features.

In summary, we introduced an unsupervised approach to learn fair, accurate, and reusable (FAR) data representations, the EquiTensors, for heterogeneous and multi-dimensional urban datasets. EquiTensors can be trained and released by government agencies or trusted data brokers over both public open data and unreleased data. It presents a novel way to allow downstream applications a means of improving accuracy, avoiding data discovery and pre-processing, and limiting their exposure to new sources of discriminatory bias.

Chapter 5
CONCLUSIONS AND DISCUSSION

This final chapter summaries the contributions of the thesis and discusses the limitations and future directions.

5.1 Summary of Contributions

The main focus of this dissertation is on fairness-aware spatio-temporal prediction for cities. Although we focus on urban applications, most of the methods developed in this context can be generalized to other domains. The results in this manuscript suggest that prediction algorithms without fairness treatment can be discriminatory against underrepresented minorities in applications such as mobility resource allocation and public safety. Moreover, the use of exogenous features in the algorithms may introduce additional bias to the prediction results. To achieve accurate and bias-free prediction, we addressed four questions in this dissertation: How to accurately model the spatio-temporal dynamics of urban activities? How to measure the fairness of spatio-temporal predictions in various urban applications? How to discover and remove discriminatory signals from urban data? How to design fairness-aware spatio-temporal machine learning algorithms?

Generic framework for accurate spatio-temporal prediction. There are two key challenges in predicting urban activities: capturing rich spatio-temporal context and extracting information from heterogeneous urban data. Traditional methods such as time series and conventional machine learning algorithms fall short of modeling complex spatio-temporal interactions and rely on hand-crafted features. While various deep learning based prediction frameworks have been proposed for urban settings, the generalizability of them outside their intended application domains is unclear.

We proposed a three-stream spatio-temporal prediction algorithm based on 3D CNN to model spatio-temporal dynamics and integrate exogenous features. We demonstrated with two mobility demand prediction cases that our method outperforms traditional methods and several state-of-art deep learning based methods (Chapter 3). With two more prediction tasks in public safety, each with different set of exogenous features, we demonstrated the flexibility and generizability of the proposed framework (Chapter 4).

Fairness metrics for spatio-temporal settings. Although a plethora of fairness metrics have been developed [52, 64], most of them are inapplicable in spatio-temporal settings. First, the prediction targets in spatio-temporal settings (e.g., ride-hailing demand) are typically continuous whereas many fairness metrics are designed for discrete classification settings [57]. Second, among the few studies that propose fairness metrics for regression, a categorical sensitive attribute is typically required [18, 11]. Finally, fairness assessment can be application specific, but fairness metrics thus far are generic across applications.

In Chapter 3, we proposed two fairness metrics: *region-based fairness gap (RFG)* and *individual-based fairness gap (IFG)* for urban mobility. Both metrics measure the gap between mean per capita prediction values (e.g., taxi demand) across demographic groups over a certain period of time. The difference lies in that RFG focuses on discrete sensitive attributes while IFG deals with continuous attributes. We showed that prediction algorithms without fairness treatment can be discriminatory against underrepresented minorities. Chapter 4 adapted three fairness metrics to generic spatio-temporal settings: residual difference (RD), positive residual difference (PRD), and negative residual difference (NRD). IFG and RFG assume that the prediction values should be independent of the sensitive attributes, whereas RD, PRD, and NRD assume the prediction errors should be independent of the sensitive attributes. We showed that the use of exogenous features may introduce additional bias to the prediction results.

Fairness-aware spatio-temporal prediction algorithms. Most of the existing fairness-aware algorithms were designed for classification and only tested in conventional machine learning methods. Based on the aforementioned prediction framework and fairness metrics,

Chapter 3 proposed *FairST*, a fairness-aware spatio-temporal prediction framework. It enforces independence between the predictions and the sensitive attributes. FairST integrates RFG or IFG as regularizers into the loss minimization pipelines to encourage fair prediction. Experiments on two real-world new mobility datasets demonstrated that FairST is able to close more than 80% of fairness gap for a single sensitive attribute and at the same time achieve *better* accuracy than state-of-the-art but fairness-oblivious baseline methods. Further experiments showed that FairST is able to reduce unfairness for multiple attributes, outperforming baselines in both accuracy and fairness.

Removing biases from heterogeneous open urban datasets. Representation learning is promising in removing discriminatory signals in data, because 1) it is naturally compatible with neural networks [10], 2) it does not hand-tune the datasets according to some ad-hoc bias-removing rules, and 3) it can handle multiple datasets simultaneously. One major challenge to learning fair representation for cities is how to remove discriminatory effects in spatio-temporal settings where the sensitive attributes are continuous and distributed across the space. We also observed that urban prediction tasks usually rely on a common set of open urban data such as weather from National Centers for Environmental Information and road networks from local open government data portals. It is therefore cost-effective to learn integrated and fair representations from these datasets, so that such representations can potentially benefit a wide range of downstream tasks, including those fairness-sensitive tasks. Nevertheless, there is no generalized approach for integrating open urban data in support of fair spatio-temporal prediction.

In Chapter 4, we proposed an algorithm framework to learn fair data representations, called *EquiTensors*, from a wide range of commonly used urban datasets. We introduced FAR (Fairness, Accuracy, and Reusability) principles for evaluating such pre-trained data representations. The proposed EquiTensor model consists of two main components: an integrative model that summarizes a large number of multi-dimensional open urban datasets as a single data representation and a fairness module based on adversarial learning and disentangled representation to remove discriminatory effects from the learned representation.

Experimental results in Chapter 4 demonstrated that the integrated representation produced by the integrative model can capture generic urban information that is reusable across an array of prediction tasks. The results also showed that the proposed adaptive weighting scheme can significantly improve the total reconstruction accuracy of the integrative model. Experimental results on two fairness-sensitive downstream tasks showed that EquiTensors could help mitigate the effect of discriminatory bias embodied in the external data and at the same time help the downstream tasks to achieve prediction accuracy comparable to an 'oracle' networks that trained with hand-selected features.

To summarize, this dissertation is a pioneering study to explore a suite of fairness-aware spatio-temporal prediction methods for cities, focusing on five aspects: utility, metrics, algorithms, data, and applications. This work makes methodological contributions to urban data science and machine learning research. The proposed methods could provide fairness assessment measures and bias-removal strategies for stakeholders such as public resource distributors and government agencies, allowing for intelligent and responsible decision-making that benefits all citizens.

5.2 Limitations and Future Work

There are a number of opportunities to expand on this dissertation. This section lists the limitations of this study and recommends directions for future research.

5.2.1 Representing Different Types of Urban Data

In this dissertation, we partition all urban datasets into a fixed spatio-temporal grid. With the grid partition, we are able to use convolutional neural networks, and it is easier to define fairness metrics based on a uniform partition. However, grid partition is not ideal for discrete, sparse, or network-like data, and fixed grids can not make use of multi-resolution information.

Future work could explore fairness-aware methods and data integration approaches in spatio-temporal settings using other types of data representations. One alternative rep-

resentation method is the discrete representation of individual regions (e.g., block groups, neighborhoods) or locations (e.g., points of interest, weather stations). For example, Liao et al. [87] proposed an LSTM based model for predicting future traffic speed given previous traffic speed for a road segment in a city. Discrete representation does not require a uniform square grid and is suitable for sparse data, but it needs extra treatment to model structure (e.g., graph RNN [66], embedding method [165]) or features to capture spatial dependencies. For example, Liao et al. [87] incorporate the spatial dependencies by embedding the traffic speed of neighboring road segments using graph convolutional neural network (graph CNN) [117] and concatenating the learned embeddings to the encoder network of the LSTM model. Another popular way for modeling spatio-temporal data is to use graph representation [51, 30]. For example, Sun et al. [134] uses graph convolutional neural network to forecast crowd flows for arbitrary-shaped regions in cities. How to integrate data with various representation methods and enforce fairness over them remain open.

5.2.2 Interpretability of Deep Learning Models

The fairness-aware methods proposed by this dissertation are based on deep neural networks. Compared to process-based models informed by domain knowledge or conventional machine learning models, deep learning models are less interpretable. For example, the importance of hand-designed features in regression models can be explained through their coefficients. Yet deep learning models extract latent features automatically, and it is often difficult to attach any semantic meaning to the learned features. Although feature importance can be qualitatively estimated through ablation studies, it is still far from establishing any causal relationship between features and prediction targets. Moreover, this dissertation proposes to use integrated data representations as external signals to boost prediction model performance. However, compared to hand-crafted features, the learned data representations are less interpretable. For applications that emphasize interpretability, we envision that integrated representations can still be valuable as context features used along with hand-crafted features or as baselines for feature engineering.

Interpretability is related to transparency, fairness, and accountability when machine learning algorithms are deployed in real-world systems to support decision-making. Instead of relying on a "black-box", decision-makers may want to understand the mechanism behind the algorithms and what factors contribute to an estimated outcome [88]. At present, the interpretability of deep learning is an ongoing research field, a breakthrough of which may vastly increase the applicability of deep learning to a wider range of scenarios.

For fairness-adjusted deep learning algorithms, decision-makers may hope to understand the fairness of the predictions beyond the fairness metric(s) the algorithm optimizes. This dissertation visually compares the prediction outcomes between fairness-oblivious methods and fairness-aware models, providing a qualitative understanding of how the spatial distribution of data (e.g., bike demand) has changed under the influence of fairness treatment. Nevertheless, more research is needed to establish a clearer picture of how optimizing over one or a few fairness metrics could influence the fairness of subgroups of the population, for different spatial regions, and for different temporal periods.

5.2.3 Reusability of Integrated Representations

We propose to pre-train multiple urban datasets into an integrated representation to support a range of downstream prediction tasks without assuming application domains. Our results show that the pre-trained representations can improve accuracy for downstream tasks of different domains, suggesting that the integrated representations can capture general urban information that is reusable across tasks. However, we integrated all datasets for Seattle we could find in open data portals into EquiTensor without considering the data quality or the use context of the source datasets. To better understand the reusability of such data representations, future work could explore two areas: data quality and transferability.

Data quality. Strong et al. defined high-quality data as "data that are fit for use by data consumers" [133]. This definition suggests that data should be contextually appropriate for the task. In the case of integrating multi-source open data for unconstrained downstream reuse, it is particularly interesting to investigate how data quality problems inherent to the

datasets such as lack of documentation, and any additional issues induced by data repurposing, cleaning, and integration influence the reusability of the integrated representation in machine learning applications. Data profiling could greatly facilitate our understanding of the context of production and intended use of a dataset, informing us about the fitness for use of this dataset for a particular task. For a simple example, the metadata could help us identify outdated datasets. An old restaurant map produced in 1980 is not ideal for predicting real-time crowd flows in 2020. Moreover, data cleaning, processing, and integration may cause additional quality issues such as bias and loss of information [79]. For example, datasets are required to be transformed into image-like form to be used in CNN models. However, rasterization of road networks using the total length of road segments in a grid removes important information on highway intersections. As a consequence, it is not possible to accurately estimate the traffic speed of two directions of the same location using this rasterized data [84]. Future work could develop protocols to 1) select certain data to include, considering various aspects of data quality such as completeness, consistency, and intended use; and 2) evaluate the potential risks brought by subsequent data processing techniques.

Transferability of utility. Transferability is closely related to reusability in the context of reusing integrated representations for downstream tasks. We found that the improvement of total reconstruction accuracy of pre-trained representation (upstream) does not necessarily transfer to downstream performance. Nevertheless, we did not delve deeper into this topic in this dissertation.

Future work could explore two questions. First, what is the relationship between the upstream accuracy and the downstream performance in the context of learning representation from a single dataset? Most recently, several studies have examined related topics [91, 131]. For example, one study found out that the transferability of network parameters pre-trained using ImageNet first increases then declines during training, suggesting that the best upstream accuracy does not necessarily benefit the downstream the most [92]. In our case of pre-training datasets as external features for downstream applications, we expect the pre-training accuracy is only weakly associated with the downstream performance. More research

is needed to verify this hypothesis. Second, given a pool of datasets and a set of downstream tasks, what would be a reliable and fast way to determine which datasets are predictive to which tasks? The transferability of integrated representations is likely dependent on the correlations (if any) between the datasets included and the downstream tasks. Some datasets are strongly associated with some tasks but not predictive for others. For example, bike lanes and bike facilities may have stronger correlations with bike demand than with fire alarm incidence. A possible solution is to estimate the magnitude of association between a dataset and a task using simpler models such as linear regression. Nevertheless, these conventional models may fail to capture the implicit non-linear associations that deep learning models instead could discover. Future work could explore this direction, possibly drawing on recent work on feature selection and interpretability of deep learning models. The results could shed light on the transferability of integrated representations, guide the selection of datasets to include, and reveal the kinds of downstream tasks such representations can support.

5.2.4 Transferability of Fairness

The above section discussed transferability of *integrated* representations to downstream applications. We now turn to the transferability of fairness-aware models and fair representations. In this dissertation, we mainly use Seattle as a case study and focus on spatio–temporal prediction tasks. We have not explored how we can transfer FairST and EquiTensors to other cities and to other applications such as land use classification. For example, FairST was trained on the Seattle dockless bikeshare data. Can we utilize this model to help the prediction of ride-hailing demand of Seattle or the prediction of bikeshare demand in Los Angeles? A few nascent studies in urban computing have examined transferring knowledge between cities. For instance, Liu et al. [95] presented an inspiring example of detecting shared bikes parking hot spots in a new city. However, it is not clear yet whether fairness could be transferred using the existing techniques.

The transferability of fair representation is still an ongoing research topic [26]. In this dissertation, we proposed a fair representation, called EquiTensor, for spatio-temporal data.

We showed that EquiTensors contain weaker correlations with the sensitive information and that downstream predictions with EquiTensors are overall fairer than those with fairness-oblivious features. This suggests that fair representations may help restrict the propagation of sensitive information to downstream applications. However, we also observed large variations of RD and PRD/NRD values of repeated runs of the same model setting for bikeshare predictions. The reason could be that the fairness metrics (i.e., RD and NRD/PRD) are not precise and influenced by the threshold that defines the advantaged and disadvantaged groups. It is also possible that the transferability of EquiTensors' fairness is not strong or stable for some applications, as we do not explicitly optimize over RD or NRD in the downstream predictions. The downstream models sometimes may still be able to pick up and exaggerate the remaining information about the sensitive attribute if the predictions heavily rely on it. Therefore, our results on the transferability of EquiTensor are only suggestive. More research is needed to verify whether the transferability of fair representations holds for other applications and for other non-spatio-temporal domains.

5.2.5 Theoretical Analysis of Fairness-Aware Prediction Methods

This dissertation relies primarily on empirical experiments and observational data. We did not provide a theoretical analysis on the effectiveness of the proposed methods — this limits the generalizability of our conclusions. Future work involves developing theoretical guarantees for fairness-aware deep learning models. At the same time, more ablation studies could also help evaluate the robustness of the proposed methods and understand in what scenarios they are most likely to succeed. Some immediate follow-up experiments include 1) applying FairST and EquiTensors to other cities and other urban applications; 2) tuning the hyper-parameters (e.g., size of the network) of FairST and EquiTensors; and 3) producing EquiTensors using different subsets of the available datasets to see whether adding more data to the EquiTensors will lead to higher utility.

5.2.6 Informing Model Design with Dataset Relationship

The EquiTensor integrates a lot of urban data, but it does not make use of the relationship between the datasets. How to optimize the current integration model to use the dataset relationship remains open. We have experimented with a few approaches, including grouping datasets by dimension, by semantics, and by similarity metrics. But it is not clear yet which one is the best to represent the relationships between data for the purpose of optimizing model architecture. Below, we report the design, experiments, results, and discussion of a hierarchical grouping strategy that aims to produce a better data representation than the core integrative model described in Section 4.4.

Method. The core integrative model can produce a data representation that contains information from multiple datasets, but it ignores the correlations among datasets. To obtain a representation that encodes the relationships among datasets, we use a hierarchical encoder, the structure of which is informed by the correlations among the datasets or their encoded representations [182]. Meanwhile, we use a hierarchical decoder that mirrors the encoder structure. The idea is similar to using structured decoders in multi-task learning. Grouping similar datasets together helps the reconstruction of each dataset focus only on relevant information [100]. We evaluate different grouping strategies:

- **Grouping by semantic meaning.** Based on domain knowledge or metadata, we can group urban datasets according to their semantic properties. For example, meteorological knowledge suggests that pressure, precipitation, and air temperature correlate with each other.

- **Grouping by dimension.** Since expert advice or metadata is not always available, another natural way to group the datasets is to organize them by dimension. Dimension itself sometimes suggests important datasets relationship. For example, 1D datasets such as citywide air quality do not associate with time-invariant data such as steep slopes.

- **Automatic grouping.** Although dataset dimension can provide useful prior knowledge of dataset relationship, we hypothesize that it may result in coarse-grained groupings when the number of datasets of a particular dimension is large. We can further computationally determine a grouping strategy based on quantitative similarity within datasets of a certain dimension (e.g., all 2D datasets). Specifically,

  - Based on a trained core integrative model, we obtain the intermediate representations (feature maps) of each dataset from the last layer of the encoding part. We then calculate pairwise dataset relationships using some distance metric among the encoded representations of the same dimension. We adopt two commonly used distance metrics: cosine similarity and earth mover's distance (EMD). In this way, we can obtain an *Affinity Matrix*, in which each dataset is represented by its distance with all datasets including itself. We then average out the affinity matrix for training samples to obtain a final affinity matrix for the entire study period for this particular dimension. The idea of representing dataset relationship using affinity matrix is informed by recent work in supervised learning [131, 142, 182] and task taxonomy where the relationships among multiple predictions tasks are studied for transfer learning [177, 34]. However, our setting is unsupervised, and we use symmetric grouping structure in both the encoder and decoder.
  - We perform clustering based on the affinity matrix, resulting in a grouping strategy of all datasets of a particular dimension. We repeat the process to obtain three sets of clustering results for 1D, 2D, and 3D data, respectively. We use the Affinity Propagation algorithm [45] as it automatically determines the number of clusters.

- **Multi-level automatic grouping.** The three methods described above provide two-level grouping strategies. We explore a dynamic multi-level hierarchical grouping strategy based on the relationship among encoded representations generated during

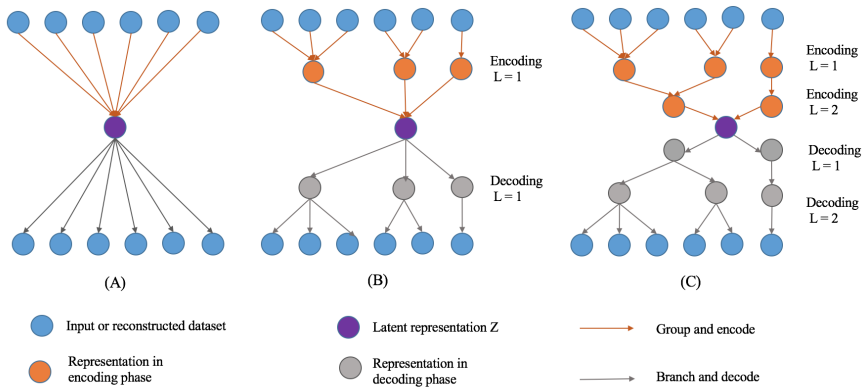

Figure 5.1: (A): The core integrative model to start with. (B): Add one level of encoder and decoder after the first iteration. (C): Add one more level of encoder and decoder after the second iteration.

autoencoder training, aiming to 1) encode hierarchical relationships into the output latent representation; and 2) allow the decoder of each dataset to gradually focus on relevant information. As Figure 5.1 illustrates, we start from the core integrative model and develop the autoencoder greedily based on the clustering results of encoded representations from the last encoding layer of the current architecture. This idea is related to the Adaptive Model Widening by Lu et al. [100] and automatic branched multi-task network by Vandenhende1 et al. [142] for multi-task learning.

Experiments Through experiments, we seek to answer the following questions: 1) Do the grouping algorithms lower reconstruction error relative to ungrouped architectures? And 2) Does the multi-level grouping algorithm outperform the simpler two-level grouping strategies? Specifically, we compare six grouping plans in terms of their reconstruction accuracy and training time.

1. Core model: The core integrative model without grouping.
2. Grouping by dimension: A two-level hierarchical plan that groups the inputs by dimension.

3. Grouping by semantic meaning: A two-level hierarchical plan that groups the inputs by semantic meaning.
4. Random grouping: A two-level hierarchical plan that groups the inputs into clusters randomly.
5. Automatic grouping: A two-level hierarchical plan that groups the datasets by dimension first and then by feature map similarity. We tested two distance metrics: cosine similarity and earth mover's distance (EMD). Feature maps were flattened to 1D vectors when calculating cosine similarity and EMD.
6. Multi-level automatic grouping: A multi-level hierarchical autoencoder that constructs the network dynamically based on feature map similarity. We experimented with a three-level architecture with cosine similarity and earth mover's distance, respectively.

Results and Discussion. Table 5.1 shows the total reconstruction accuracy and per sample training time of different models. We observe that all models with grouping strategies (except random grouping) outperform the ungrouped model (core model) in terms of reconstruction accuracy. This suggests that supplying the model with some information on dataset relationship may help. But we caution that the performance gain of hierarchical strategies over the core model may come from the additional model capacity brought by the model structure change. We observe that the simple strategy, grouping by dimension, outperforms all other models including the multi-level strategies in both accuracy and training time. This implies that dimension information is very informative in our case. Compared to two-level grouping strategies, multi-level grouping shows inferior model performance despite larger network capacity (more parameters to learn). This implies that the grouping results produced by the automatic plans are not ideal. The reason could be that simple distance metrics such as cosines similarity are not sufficient to capture spatial correlations among datasets or that averaging over all training samples ignores the temporal dynamics of our data. It is also unclear whether the multi-level grouping architecture will improve upon the two-level automatic grouping ones if a more appropriate grouping strategy is applied.

Table 5.1: Total reconstruction accuracy (MAE) and per sample training time (ms) of models with different grouping strategies.

Model	Reconstruction accuracy	Per sample training time
Core model	0.217	**12.88**
By dim	**0.062**	14.18
By semantic meaning	0.202	16.44
Random grouping	0.226	16.50
Auto group (cosine similarity)	0.117	24.72
Auto group (EMD)	0.067	23.01
Multi-level auto group (cosine similarity)	0.125	28.93
Multi-level auto group (EMD)	0.074	27.44

One limitation of the hierarchical design is that the complexity of the networks usually grows with the number of groups, so as the training time. Besides, we need to make additional design choices such as the output dimensions of the intermediate representations for each level of grouping. It is not clear yet how these choices may affect model performance. To summarize, the results suggest that a hierarchical grouping design based on dataset relationship could potentially help improve the reconstruction accuracy of the integrative model. Nevertheless, what is the optimal grouping strategy remains an open question.

5.2.7 Addressing Multiple Sensitive Attributes

Our results in Chapter 3 demonstrated that the proposed fairness-aware algorithm (FairST) can reduce unfairness for multiple attributes, outperforming some state-of-the-art fairness-agnostic models in both accuracy and fairness. However, we have not addressed multiple sensitive attributes in EquiTensor. The current version of EquiTensor can only remove the association with one sensitive attribute. This means that we need to train separate EquiTensors for different sensitive attributes. The reusability of the EquiTensor could be improved if the influence of multiple sensitive attributes could be removed all at once. One way to achieve this goal is to use multiple adversaries, each corresponding to a sensitive

attribute. A weighted sum of the adversarial loss can then be incorporated into the training process to optimize for fairness of all sensitive attributes. Nevertheless, this approach may remove too much information from the resulting EquiTensor, leading to a significant loss of utility. Empirical experiments are needed to evaluate this method. Future work could also develop other approaches to address multiple sensitive attributes in fair representations without sacrificing much utility.

5.2.8 Combining Fair algorithms and Bias-free data

This dissertation explored two fairness methods: enforcing equity through modifying the objective function of the prediction model (FairST) and removing biased signals from training data (EquiTensor). The former focuses on optimizing the model for fairness, regardless of whether the input data is bias-free or not. The latter removes from training data the association with a sensitive attribute without modifying the prediction models. Nevertheless, we did not explore combining the two approaches. Our fairness-aware algorithm, FairST, requires full access to the entire training pipeline, which is not always feasible. The EquiTensor does not require access to the training model, but the fairness of the prediction models that use the EquiTensors is not guaranteed as fairness is not explicitly optimized in the models and that other training data used by the models may still contain bias. Therefore, it is desirable to apply other fairness techniques in addition to de-biasing the input data. One promising solution that this dissertation did not cover is the post-processing method: adjusting the prediction outcomes for fairness directly without changing the model training processing [57]. This approach can be used together with the bias-free data to achieve fairness without the need for retraining. The limitation of post-processing approach is that it often leads to a significant loss of utility [154, 4]. Future work could explore combinations of fairness techniques to achieve good fairness utility trade-offs for use cases under different constraints (e.g., limited access to training data or training models).

5.2.9 Defining Fairness for Urban Applications

Current research in fair machine learning focuses on the technical aspects of algorithmic fairness. This dissertation takes one step further to bridge fair machine learning methods and application domains (i.e., predicting urban activities). However, many issues deserve further investigation in order to make fair-aware machine learning systems more useful in the real world. For example, this dissertation relies on fairness metrics to evaluate the equity of urban applications. Nevertheless, there are several limitations to fairness metrics. To begin with, a plethora of fairness metrics has been proposed and many of them are intrinsically incompatible with each other [5]. There is a lack of consensus on how to define and measure the fairness of an algorithm [186]. At the same time, measuring fairness by metrics may be an oversimplification of the problem. Studies showed that metrics derived from observational data ignore causality, thus are not sufficient in identifying discrimination [57, 97]. In reality, fairness is not a technical nor statistical concept; Satisfying fairness metrics does not necessarily meet equity goals in practice [25, 5]. This dissertation inherited these limitations of fairness metrics.

Although we improved upon the generic fairness metrics for spatio-temporal settings and mobility cases, but the metrics we developed are still an over-simplification of equity in the real world. There are still open questions that deserve further exploration. Specifically, we developed two fairness metrics in Chapter 3: *region-based fairness gap (RFG)* and *individual-based fairness gap (IFG)* for new mobility drawing on the notion of statistical parity [35] and vertical equity [31]. We assume that a perfectly fair scenario has zero demand gap for new mobility resources between two demographic groups. Our goal is to encourage fair resource allocation (e.g., bike re-balancing) by making predictions that minimize the demand gap. Here, we implicitly assume that the operators rely on demand predictions to allocate resources and that bridging the predicted demand gap aligns with the equity goal in practice. However, we caution that the demand gap reflected by our fairness metrics is not necessarily a result of discriminatory practices of the operators. Although bikeshare or rideshare operators

need to intervene supply/demand (by re-balancing bikes or directing drivers to high-demand areas), the spatio-temporal distributions of demand are also determined by user needs. This is different from public transportation planning cases where permanent facilities are set by agencies. Therefore, it is likely that the fairness metrics reflect a combination of multiple factors: the user needs, the practices of operators, and the setting of existing transportation facilities, etc. But this dissertation did not examine which factors contributed to the observed fairness gaps, which is an interesting topic for future investigation. In the long run, more research is needed to examine the implications of various fairness definitions, and develop measures that are not only feasible with technical systems, but can satisfy the specific equity goals defined by stakeholders of particular applications as well.

5.2.10 Understanding the Feedback Loops of Fair-Aware Algorithms

The vast majority of research in fair machine learning, including this dissertation, focuses on static problems using historical data for both training and testing. In reality, data-driven systems dynamically affect their environment and stakeholders, which in turn influences future data collection [26, 5]. For example, suppose a data-driven system is used to estimate dockless bikeshare demand and inform bike redistribution. In that case, an underestimation of demand of an area may result in insufficient supply to that area. The system then collects the actual usage, which is upper bounded by the supply for future prediction. Moreover, if the residents of an area often experience unmet demand, they may abandon the service, reinforcing the low-demand and low-supply feedback loop. It is not yet well understood how to mitigate the dynamic feedback effects.

BIBLIOGRAPHY

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

[3] Abolfazl Asudeh, Tanya Berger-Wolf, Bhaskar DasGupta, and Anastasios Sidiropoulos. Maximizing coverage while ensuring fairness: a tale of conflicting objective. *arXiv preprint arXiv:2007.08069*, 2020.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[6] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[7] Anna Maria Barry-Jester, Ben Casselman, and Dana Goldstein. The new science of sentencing. *The Marshall Project*, 4:2015, 2015.

[8] Patrick Bayer, Marcus Casey, Fernando Ferreira, and Robert McMillan. Racial and ethnic price differentials in the housing market. *Journal of Urban Economics*, 102:91–105, 2017.

[9] Franziska Bell and Slawek Smyl. Forecasting at uber: An introduction. *Uber Engineering*, 2018.

[10] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[11] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *CoRR*, abs/1706.02409, 2017.

[12] Toon Bogaerts, Antonio D Masegosa, Juan S Angarita-Zapata, Enrique Onieva, and Peter Hellinckx. A graph cnn-lstm neural network for short and long-term traffic forecasting based on trajectory data. *Transportation Research Part C: Emerging Technologies*, 112:62–77, 2020.

[13] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[14] Jonathan Bright, Helen Zerlina Margetts, Ning Wang, and Scott A Hale. Explaining usage patterns in open government data: the case of data. gov. uk. *Gov. UK (June 3, 2015)*, 2015.

[15] Anne Elizabeth Brown. *Ridehail Revolution: Ridehail Travel and Equity in Los Angeles*. PhD thesis, UCLA, 2018.

[16] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018.

[17] Leonardo Caggiani, Rosalia Camporeale, and Michele Ottomanelli. Facing equity in transportation network design problem: A flexible constraints based model. *Transport Policy*, 55:9–17, 2017.

[18] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 71–80. IEEE, 2013.

[19] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[20] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[21] Biz Carson. Lyft doubled rides in 2017 as its rival uber stumbled. forbes, 2018.

[22] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803, 2018.

[23] Xingyi Cheng, Ruiqing Zhang, Jie Zhou, and Wei Xu. Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[24] Fernando Chirigati, Harish Doraiswamy, Theodoros Damoulas, and Juliana Freire. Data polygamy: the many-many relationships among urban spatio-temporal data sets. In *Proceedings of the 2016 International Conference on Management of Data*, pages 1011–1025, 2016.

[25] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[26] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[27] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017.

[28] Nancy S Cole and Michael J Zieky. The new faces of fairness. *Journal of Educational Measurement*, 38(4):369–382, 2001.

[29] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

[30] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[31] Alexa Delbosc and Graham Currie. Using lorenz curves to assess public transport equity. *Journal of Transport Geography*, 19(6):1252–1259, 2011.

[32] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

[33] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[34] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12387–12396, 2019.

[35] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.

[36] The U.S. EEOC. Uniform guidelines on employee selection procedures. *March 2, 1979*, 43:111–122, 1979.

[37] Hillel J Einhorn and Alan R Bass. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, 75(4):261, 1971.

[38] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 35–47, 2018.

[39] Wafic El-Assi, Mohamed Salah Mahmoud, and Khandker Nurul Habib. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in toronto. *Transportation*, 44(3):589–613, 2017.

[40] Sheena Erete, Emily Ryou, Geoff Smith, Khristina Marie Fassett, and Sarah Duda. Storytelling with data: Examining the use of data by non-profit organizations. In *Proceedings of the 19th ACM conference on Computer-Supported cooperative work & social computing*, pages 1273–1283, 2016.

[41] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.

[42] A Gers Felix, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[43] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[44] Inês Frade and Anabela Ribeiro. Bicycle sharing systems demand. *Procedia-Social and Behavioral Sciences*, 111:518–527, 2014.

[45] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[46] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[47] Yanjie Fu, Guannan Liu, Yong Ge, Pengyang Wang, Hengshu Zhu, Chunxiao Li, and Hui Xiong. Representing urban forms: A collective learning model with heterogeneous human mobility data. *IEEE transactions on knowledge and data engineering*, 31(3):535–548, 2018.

[48] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Xiaolin Li. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 906–913, 2019.

[49] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189, 2015.

[50] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[51] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3656–3663, 2019.

[52] Bruce Glymour and Jonathan Herington. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 269–278, New York, NY, USA, 2019. ACM.

[53] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[54] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019.

[55] Michael B Gurstein. Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 2011.

[56] Jessica Guynn. Google photos labeled black people'gorillas'. *USA Today*, 1, 2015.

[57] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3323–3331, USA, 2016. Curran Associates Inc.

[58] Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011.

[59] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[60] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.

[61] Kate Hosford and Meghan Winters. Who are public bicycle share programs serving? an evaluation of the equity of spatial access to bicycle share service areas in canadian cities. *Transportation research record*, page 0361198118783107, 2018.

[62] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pages 1878–1889, 2017.

[63] Ryan Hughes and Don MacKenzie. Transportation network company wait times in greater seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography*, 56:36–44, 2016.

[64] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 49–58, New York, NY, USA, 2019. ACM.

[65] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it. *Bloomberg*, 2016.

[66] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016.

[67] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Information systems management*, 29(4):258–268, 2012.

[68] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. Unsupervised representation learning of spatial data via multimodal embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1993–2002, 2019.

[69] Shenggong Ji, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Spatio-temporal feature fusion for dynamic taxi route recommendation via deep reinforcement learning. *Knowledge-Based Systems*, 205:106302, 2020.

[70] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[71] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

[72] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE, 2010.

[73] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644, 2013.

[74] Hyeon-Woo Kang and Hang-Bong Kang. Prediction of crime occurrence from multimodal data using deep learning. *PloS one*, 12(4), 2017.

[75] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[76] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.

[77] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[78] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2737–2746, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[79] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. Activeclean: Interactive data cleaning while learning convex loss models. *arXiv preprint arXiv:1601.03797*, 2016.

[80] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in neural information processing systems*, pages 5967–5976, 2017.

[81] Bruno Lepri, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver. The tyranny of data? the bright and dark sides of data-driven decision-making for social good. In *Transparent data mining for big and small data*, pages 3–24. Springer, 2017.

[82] Xuefeng Li, Yong Zhang, Li Sun, and Qiyang Liu. Free-floating bike sharing in jiangsu: Users' behaviors and influencing factors. *Energies*, 11(7):1664, 2018.

[83] Yaguang Li, Kun Fu, Zheng Wang, Cyrus Shahabi, Jieping Ye, and Yan Liu. Multi-task representation learning for travel time estimation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1695–1704, 2018.

[84] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[85] Yexin Li, Yu Zheng, and Qiang Yang. Dynamic bike reposition: A spatio-temporal reinforcement learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pages 1724–1733, New York, NY, USA, 2018. ACM.

[86] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, pages 33:1–33:10, New York, NY, USA, 2015. ACM.

[87] Binbing Liao, Jingqing Zhang, Chao Wu, Douglas McIlwraith, Tong Chen, Shengwen Yang, Yike Guo, and Fei Wu. Deep sequence learning with auxiliary information for

traffic prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 537–546. ACM, 2018.

[88] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[89] Todd Litman. *Evaluating transportation equity*. Victoria Transport Policy Institute, 2018.

[90] Hao Liu, Ting Li, Renjun Hu, Yanjie Fu, Jingjing Gu, and Hui Xiong. Joint representation learning for multi-modal transportation recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1036–1043, 2019.

[91] Hong Liu, Jeff Z HaoChen, Colin Wei, and Tengyu Ma. Meta-learning transferable representations with a single target domain. *arXiv preprint arXiv:2011.01418*, 2020.

[92] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Towards understanding the transferability of deep representations. *arXiv preprint arXiv:1909.12031*, 2019.

[93] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI*, pages 194–200, 2016.

[94] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019.

[95] Zhaoyang Liu, Yanyan Shen, and Yanmin Zhu. Where will dockless shared bikes be stacked?:—parking hotspots detection in a new city. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 566–575. ACM, 2018.

[96] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pages 14611–14624, 2019.

[97] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[98] Steve Lohr. For big-data scientists,'janitor work'is key hurdle to insights. *New York Times*, 17:B4, 2014.

[99] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[100] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5334–5343, 2017.

[101] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[102] Xiaolei Ma, Zhuang Dai, Zhengbing He, Jihui Ma, Yong Wang, and Yunpeng Wang. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, 17(4):818, 2017.

[103] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[104] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[105] Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. *arXiv preprint arXiv:2003.00824*, 2020.

[106] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[107] Koray Mancuhan and Chris Clifton. Combating discrimination using bayesian networks. *Artificial intelligence and law*, 22(2):211–238, 2014.

[108] Sébastien Martin, Muriel Foulonneau, Slim Turki, and Madjid Ihadjadene. Risk analysis to overcome barriers to open data. *Electronic Journal of e-Government*, 11(1):348, 2013.

[109] N McNeil, J Dill, J MacArthur, J Broach, and S Howland. Breaking barriers to bike share: Insights from residents of traditionally underserved neighborhoods. nitc-rr-884b. *National Institute for Transportation and Communities: Portland, ME, USA*, 2017.

[110] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[111] Albert Meijer and Martijn Wessels. Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, 42(12):1031–1039, 2019.

[112] Claire Cain Miller. When algorithms discriminate. *The New York Times*, 9, 2015.

[113] Renée J Miller. Open data integration. *Proceedings of the VLDB Endowment*, 11(12):2130–2139, 2018.

[114] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

[115] Stephen J Mooney, Kate Hosford, Bill Howe, An Yan, Meghan Winters, Alon Bassok, and Jana A Hirsch. Freedom from the station: Spatial equity in access to dockless bike share. *Journal of Transport Geography*, 74:91–96, 2019.

[116] Fatemeh Nargesian, Erkang Zhu, Renée J Miller, Ken Q Pu, and Patricia C Arocena. Data lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment*, 12(12):1986–1989, 2019.

[117] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023, 2016.

[118] Seattle Department of Transportation. The new mobility playbook. *The city of Seattle*, 2017.

[119] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[120] Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund, and Katherine Lamberth. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

[121] Stephen M Powers and Stephanie E Hampton. Open science, reproducibility, and transparency in ecology. *Ecological applications*, 29(1):e01822, 2019.

[122] Vignesh Ramanathan, Kevin Tang, Greg Mori, and Li Fei-Fei. Learning temporal embeddings for complex video analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4479, 2015.

[123] Anthony Michael Ricciardi, Jianhong Cecilia Xia, and Graham Currie. Exploring public transport equity between separate disadvantaged cohorts: a case study in perth, australia. *Journal of transport geography*, 43:111–122, 2015.

[124] Rashida Richardson, Jason M Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, 94:15, 2019.

[125] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[126] Cynthia Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine, August*, 2013.

[127] Anian Ruoss, Mislav Balunović, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. *arXiv preprint arXiv:2002.10312*, 2020.

[128] Bashir Sadeghi and Vishnu Naresh Boddeti. Imparting fairness to pre-trained biased representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2020.

[129] Bilong Shen, Xiaodan Liang, Yufeng Ouyang, Miaofeng Liu, Weimin Zheng, and Kathleen M Carley. Stepdeep: A novel spatial-temporal mobility event prediction framework based on deep neural network. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 724–733. ACM, 2018.

[130] SimplyAnalytics. Easi/mri census us. 2018.

[131] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020.

[132] Jennifer Stark and Nicholas Diakopoulos. Uber seems to offer better service in areas with more white people. that raises some tough questions. *The Washington Post*, 2016.

[133] Diane M Strong, Yang W Lee, and Richard Y Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[134] Junkai Sun, Junbo Zhang, Qiaofei Li, Xiuwen Yi, Yuxuan Liang, and Yu Zheng. Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[135] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[136] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[137] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[138] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society.

[139] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[140] Julia Ursaki and Lisa Aultman-Hall. Quantifying the equity of bikeshare access in us cities. In *95th Annual Meeting of the Transportation Research Board, Washington, DC*, 2016.

[141] WMP van der Aalst, M Bichler, and A Heinzl. Responsible data science. *Business & Information Systems Engineering*, 59(5):311–313, 2017.

[142] Simon Vandenhende, Bert De Brabandere, and Luc Van Gool. Branched multi-task networks: Deciding what layers to share. *arXiv preprint arXiv:1904.02920*, 2019.

[143] Tyler J VanderWeele and Miguel A Hernán. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *American journal of epidemiology*, 175(12):1303–1310, 2012.

[144] Patrick Vogel, Torsten Greiser, and Dirk Christian Mattfeld. Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20:514–523, 2011.

[145] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

[146] Dong Wang, Wei Cao, Jian Li, and Jieping Ye. Deepsd: supply-demand prediction for online car-hailing services using deep neural networks. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 243–254. IEEE, 2017.

[147] Mingshu Wang and Lan Mu. Spatial disparities of uber accessibility: An exploratory analysis in atlanta, usa. *Computers, Environment and Urban Systems*, 67:169–175, 2018.

[148] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Xiaolin Li, and Dan Lin. Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(6):63, 2018.

[149] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Pengfei Wang, Yu Zheng, and Charu Aggarwal. You are how you drive: Peer and temporal-aware representation learning for driving behavior analysis. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2457–2466. ACM, 2018.

[150] Victoria Wang and David Shepherd. Exploring the extent of openness of open government data–a critique of open government datasets in the uk. *Government Information Quarterly*, 37(1):101405, 2020.

[151] Jan Whittington, Ryan Calo, Mike Simon, Jesse Woo, Meg Young, and Peter Schmiedeskamp. Push, pull, and spill: A transdisciplinary case study in municipal open government. *Berkeley Technology Law Journal*, 30(3):1899–1966, 2015.

[152] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[153] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[154] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

[155] Shi Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[156] Chengcheng Xu, Junyi Ji, and Pan Liu. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation Research Part C: Emerging Technologies*, 95:47–60, 2018.

[157] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *IJCAI*, pages 1452–1458, 2019.

[158] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.

[159] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1401–1406. IEEE, 2019.

[160] An Yan and Bill Howe. Fairst: Equitable spatial and temporal demand prediction for new mobility systems. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 552–555, 2019.

[161] An Yan and Bill Howe. Fairness-aware demand prediction for new mobility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1079–1087, 2020.

[162] An Yan and Bill Howe. Equitensors: Learning fair integrations of heterogeneous urban data. In *Proceedings of the 2021 International Conference on Management of Data*, 2021.

[163] An Yan, Caihong Huang, Jian-Sin Lee, and Carole L Palmer. Cross-disciplinary data practices in earth system science: Aligning services with reuse and reproducibility priorities. *Proceedings of the Association for Information Science and Technology*, 57(1):e218, 2020.

[164] An Yan and Nicholas Weber. Mining open government data used in scientific research. In *International Conference on Information*, pages 303–313. Springer, 2018.

[165] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2017.

[166] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *AAAI*, 2018.

[167] Sirui Yao and Bert Huang. New fairness metrics for recommendation that embrace differences. *arXiv preprint arXiv:1706.09838*, 2017.

[168] Yaqiang Yao, Jie Cao, and Huanhuan Chen. Robust task grouping with representative tasks for clustered multi-task learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1408–1417. ACM, 2019.

[169] John Yinger. Measuring racial discrimination with fair housing audits: Caught in the act. *The American Economic Review*, pages 881–893, 1986.

[170] Ji Won Yoon, Fabio Pinelli, and Francesco Calabrese. Cityride: a predictive bike sharing journey advisor. In *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, pages 306–311. IEEE, 2012.

[171] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 191–200. ACM, 2019.

[172] Meg Young and An Yan. Civic hackers' user experiences and expectations of seattle's open municipal data program. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[173] Haiyang Yu, Zhihai Wu, Shuqin Wang, Yunpeng Wang, and Xiaolei Ma. Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors*, 17(7):1501, 2017.

[174] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 984–992. ACM, 2018.

[175] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[176] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[177] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

[178] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–325–III–333. JMLR.org, 2013.

[179] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

[180] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*, pages 361–370. International World Wide Web Conferences Steering Committee, 2017.

[181] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, and Tianrui Li. Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artificial Intelligence*, 259:147–166, 2018.

[182] Yawen Zhang, Qin Lv, Duanfeng Gao, Si Shen, Robert P Dick, Michael Hannigan, and Qi Liu. Multi-group encoder-decoder networks to fuse heterogeneous data for next-day air quality prediction. In *IJCAI*, pages 4341–4347, 2019.

[183] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[184] Yu Zheng. Methodologies for cross-domain data fusion: An overview. *IEEE transactions on big data*, 1(1):16–34, 2015.

[185] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

[186] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.