EquiTensors: Learning Fair Integrations of Heterogeneous Urban Data

An Yan
University of Washington
yanan15@uw.edu

Bill Howe
University of Washington
billhowe@uw.edu

ABSTRACT

Neural methods are state-of-the-art for urban prediction problems such as transportation resource demand, accident risk, crowd mobility, and public safety. Model performance can be improved by integrating exogenous features from open data repositories (e.g., weather, housing prices, traffic, etc.), but these uncurated sources are often too noisy, incomplete, and biased to use directly. We propose to learn integrated representations, called EquiTensors, from heterogeneous datasets that can be reused across a variety of tasks. We align datasets to a consistent spatio-temporal domain, then describe an unsupervised model based on convolutional denoising autoencoders to learn shared representations. We extend this core integrative model with adaptive weighting to prevent certain datasets from dominating the signal. To combat discriminatory bias, we use adversarial learning to remove correlations with a sensitive attribute (e.g., race or income). Experiments with 23 input datasets and 4 real applications show 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.

CCS CONCEPTS

• **Information systems** → **Information integration**; *Spatial-temporal systems*; *Data analytics.*

KEYWORDS

neural networks, spatial-temporal predictions, data integration, open data, fairness in machine learning

ACM Reference Format:
An Yan and Bill Howe. 2021. EquiTensors: Learning Fair Integrations of Heterogeneous Urban Data. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), June 20–25, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3448016.3452777

1 INTRODUCTION

Predicting urban dynamics using spatio-temporal neural 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 [46, 57], citywide crowd flow [65], traffic conditions [31, 61], accident patterns [62], public safety [24], and more [9]. All of these 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 [38, 57], but these same data sources are also helpful for predicting citywide crowd flow and accident patterns [62, 65].

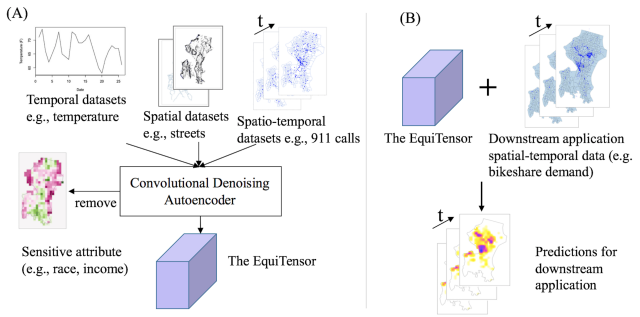

Figure 1: (A) An EquiTensor is a learned representation of heterogeneous spatio-temporal datasets with sensitive demographic information removed and can be (B) shared across multiple prediction tasks to reduce development costs and improve performance.

The use of exogenous datasets can significantly improve model accuracy [46, 51], but selecting and properly integrating a potentially 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 [39]. More insidiously, the use of exogenous data can reinforce systemic discrimination. For example, housing prices reflect historical discriminatory urban development policies [3], public safety data reflects racist policing practices [43], and transportation data reflects biased policies toward wealthy neighborhoods [40]. These sources of bias are propagated into prediction tasks, resulting in unfair predictions [2, 68] and exacerbating structural inequity.

In this paper, we present an unsupervised learning architecture (Figure 1) to integrate heterogeneous spatio-temporal data and counteract bias, producing fair and reusable representations called EquiTensors that can be incorporated directly in a variety of urban applications to improve accuracy while limiting exposure to additional bias. The proposed architecture addresses three main challenges: heterogeneity, selection, and fairness.

Heterogeneity. 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.,

SIGMOD '21, June 20–25, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8343-1/21/06…$15.00
https://doi.org/10.1145/3448016.3452777

point events, city blocks), and varying coverage. The goal is to design an unsupervised model that accepts heterogeneous and multi-dimensional datasets as input without requiring application-specific feature engineering. Applications should be able to use these pre-trained features without sacrificing much performance relative to "oracle" models that use hand-selected relevant datasets. We propose to align all datasets to a common spatio-temporal grid, then use a convolutional denoising autoencoder (CDAE) as our *core integrative model* to learn a shared representation from all datasets. The unsupervised CDAE, along with task-agnostic pre-processing, is naturally robust to heterogeneous uncurated data sources.

Selection. Determining which datasets will be predictive for which applications is non-trivial. To address this problem, we propose to incorporate all available open urban datasets into our core integrative model. However, it is challenging to coordinate the learning of a large number of datasets. To address the issue, we use an adaptive weighting scheme that 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 recent work [8, 27] in multi-task learning, adapted for our unsupervised setting.

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 [43]. To address bias, we incorporate an adversarial model that learns to detect a sensitive attribute (race, income, etc.) from the learned representations; 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 learn representations that are "disentangled" from the sensitive attribute [21, 26, 32, 34]. This approach combines adversarial learning for fairness [45, 55, 56] and learning disentangled representations [12, 28], but adapt them for an unsupervised data integration setting with continuous and spatially distributed sensitive attributes (e.g., a map of income) as opposed to only a categorical value (e.g., gender).

Example. Consider dockless bike share, where bikes can be left and re-rented anywhere. The operating company must redistribute bikes to align supply with demand; the corresponding prediction problem is central to the company's business model. Future bike demand can be predicted from past demand to inform redistribution. However, heterogeneous data sources including weather, traffic, demographics, and more can influence demand (heterogeneity), but often in surprising ways (selection): demand is only weakly associated with precipitation, yet highly associated with socioeconomic factors [41]; models that reinforce these biases can violate city, state, and federal policies [38] (fairness).

Our main contributions are summarized as follows:

- We propose an unsupervised, task-agnostic model for learning integrated representations of many heterogeneous spatio-temporal datasets for reuse across multiple urban prediction problems.
- We describe an adaptive weighting scheme that improves reconstruction error across highly heterogeneous datasets, building on prior approaches for multi-task learning [27] by exploiting our specific setting.
- We present an architecture for learning fair representations for continuous spatio-temporal data, extending recent work in adversarial learning and disentangled representation learning developed for categorical sensitive variables [26, 32].
- We provide experimental results using 23 real-world urban datasets and evaluate on 4 applications, showing that EquiTensors could improve the downstream prediction performance while limiting the exposure to discriminatory bias from the exogenous data.

2 RELATED WORK

Recent work in data management has recognized the challenges in organizing large open data repositories [10, 39]; our focus is making open data directly usable in prediction tasks in urban computing [5, 9, 13, 18, 24, 31, 46, 61, 62, 65, 67]. The machine learning community has made remarkable progress in representation learning [4], multi-task learning [42], and managing bias and discrimination [15, 64]. 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 [10, 36, 39] focuses on finding structural join and union relationships; we assume the only relationship between datasets is a common spatio-temporal domain and instead aim to directly benefit downstream prediction tasks. Representation learning has been effective for specific urban applications [16, 22, 23, 33, 53]; for example, Wang et al.[53] used an autoencoder on GPS trajectories to study driver behavior. Our focus is on understanding the limitations of representation learning when we relax assumptions about the features, architecture, and objective of the target application.

Multi-task learning. Multi-task learning trains multiple related tasks simultaneously from a shared input, aiming to achieve better performance than learning each task independently [42]. Some models use task relationships to optimize feature sharing [30, 37, 60], but model complexity usually grows with the number of tasks [27]. Another approach is to balance the loss terms across tasks [8, 25, 27]. For example, Liu et al. [27] proposed a Dynamic Weight Average that adjusts task weights based on learning progress, showing that their method outperforms competitive methods including Uncertainty Weighting [25]. Our adaptive weighting approach is related to that of Liu et al. [27], but our setting of reconstructing multiple inputs admits new techniques, as we will describe in Section 3.3.

Fairness in machine learning. There exists extensive literature on fair machine learning [11, 14, 15, 19, 64], but few that consider spatio-temporal applications. Yan and Howe [58] presented a fairness-aware prediction framework for urban mobility by incorporating fairness as a regularizer, but their approach relied on supervised learning. Unsupervised learning [29, 44, 64] and adversarial learning [45, 50, 55, 56] have been used to learn fair representations. For example, Madras et al. [32] proposed an encoder-decoder structure 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 [32] and image transformation [26] 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, nor learn fair representations for spatio-temporal settings. 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 safer and more usable.

3 THE EQUITENSOR MODEL

The EquiTensor framework for learning integrated and equitable representations for urban datasets consists of three main components: a core integrative model to address heterogeneity, an adaptive weighting scheme to address selection, and a fairness representation component to address fairness. The input is a set of exogenous datasets D (e.g., weather, road networks, 911 calls, etc.) and a dataset S representing a sensitive attribute (e.g., race or income). The output is a tensor Z (the EquiTensor) that encodes the spatio-temporal correlations within and among the data in D with minimal redundancy, while removing any correlations with S.

3.1 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 map values to [0, 1] using max absolute scaling. More sophisticated methods of imputation, feature engineering, and normalization exist, but we do not consider them in this paper.

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 a set of points, lines, or regional values. We rasterize point data by counting the events within each target cell, lines by counting the number of segments, regional data by 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.

3.2 The Core Integrative Model

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.

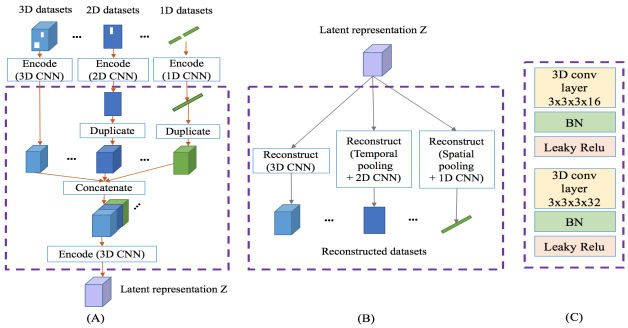

Figure 2: The core integrative model consisting of an encoder (A) that integrates 1D, 2D, and 3D datasets, and a decoder (B) that backpropagates the reconstruction error across all input datasets. The 3D CNN layers for encoding / decoding from the latent representation are shown at (C).

The encoder for the proposed core integrative model is illustrated in Figure 2(A). The input for this step is the set of training samples produced by pre-processing. 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 pass each corrupted tensor through three convolutional layers (with number of filters 16, 32, and 1) to learn intra-dataset patterns and collapse multiple attributes to a single feature. That is, each 1D input tensor 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 [66].

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 2(B). We use three layers of 3D convolutional layers (16, 32, and C_i filters) 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 all layers, we use kernel size of 3 and stride size of 1.

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 $\mathcal{X}'^i = \{x_1'^i, x_2'^i, ..., x_n'^i\}$. The encoder Enc encodes these corrupted tensors

$\mathcal{X}'^i$ into a latent representation Z^i, from which each input tensor can be reconstructed 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:

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

3.3 Adaptive Weighting

The core integrative model assigns equal weight to all datasets during training (Equation 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 related to recent work in multi-task learning [8, 27] in which the learning of a number of supervised sub-tasks needs to be balanced.

Chen et al. [8] 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. [27], 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.

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

$$w^i(t) = n\frac{exp(r^i(t)/\alpha)}{\sum_{j=0}^{n} exp(r^j(t)/\alpha)} \tag{2}$$

where n is the number of datasets. α is a parameter controlling the degree to which learning progress influences the weights [27]. Larger α leads to more equal weights among datasets. $r^i(t)$ is relative learning progress for dataset i at epoch t [8]. The learning progress $LP^i(t)$ is normalized by the average learning progress of datasets and is written as:

$$r^i(t) = LP^i(t)/E_n[LP^i(t)], \quad LP^i(t) = L(t)^i/L(opt)^i \tag{3}$$

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.

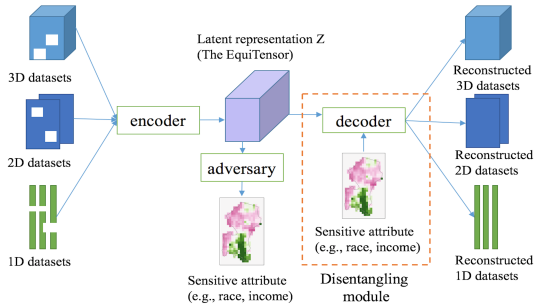

Figure 3: The EquiTensor architecture. The encoder and decoder learn a latent representation Z (the EquiTensor) by minimizing reconstruction error. The sensitive attribute S (e.g., race) is passed to the decoder to disentangle S from other information in Z. The adversary learns to predict S given Z, penalizing the encoder.

3.4 Learning Fair Representations

We now describe two strategies to remove discriminatory effects from our shared representation.

Figure 3 shows the architecture of the EquiTensor model. First, we use a *disentangling module* to separate the sensitive attribute S from other information in the latent space during reconstruction. The decoder uses both the latent representation Z and S to reconstruct the input, learning to disentangle S from Z [26, 32, 34]. Second, we use an adversarial model A to predict S from Z, and the core model is penalized accordingly. 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 (e.g., race) that are encoded in many different input datasets. The adversary loss is defined as:

$$L_A = \frac{1}{m}\sum_{i=0}^{m}|A(Z^i) - S| \tag{4}$$

where m is the number of training samples. Z^i is the representation learned from the ith training sample. Here we duplicated S along the temporal dimension to match the shape of Z^i.

Final objective function for CDAE. The CDAE has two objectives: minimizing the reconstruction error while being penalized by the adversary (Figure 3). The loss for CDAE is written as:

$$L_{AE} = \frac{1}{m}\sum_{i=0}^{m}\sum_{j=0}^{n}|Dec(Enc(x_j'^i), S) - x_j^i| + \lambda(1 - \frac{1}{m}\sum_{i=0}^{m}|A(Z^i) - S|) \tag{5}$$

where the first term is the reconstruction error and the second term is the negative adversarial loss $1-L_A$. Unlike Equation 1, the decoder now has direct access to the sensitive attribute S, encouraging the model to find a "disentangled" solution [26, 32, 34]. A parameter λ controls the tradeoff between the two terms.

The adversary consists of three 3D CNN layers, for which the number of filters is 16, 32, and 1, respectively. 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.

3.5 Measuring Fairness

We evaluate fairness of EquiTensors by measuring the adversarial model's ability to discern the sensitive attribute. We measure the fairness of downstream predictions that use the EquiTensors with three fairness metrics.

We train a separate adversarial model F instead of directly using the adversarial model A of Equation 4 used in training 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 EquiTensors.

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 [7, 20, 59] to our spatio-temporal setting.

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. $|G^+|$ 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 $R_{i,t} = \hat{y}_{i,t} - y_{i,t}$.

Positive residual difference (PRD) is written as:

$$PRD = \frac{1}{|G^+|} \sum_{t=0}^{T} \sum_{i \in G^+} PR_{i,t} - \frac{1}{|G^-|} \sum_{t=0}^{T} \sum_{j \in G^-} PR_{j,t} \tag{6}$$

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 EXPERIMENTS

Using the City of Seattle as a case study, we first evaluate our core integrative model (without considering fairness) against several baseline methods, comparing the prediction accuracy for four downstream applications. We then evaluate the effectiveness of the adaptive weighting scheme on total reconstruction error of the integrative model. Finally, we generate EquiTensors using the framework in Figure 3 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.

Table 1: Downstream tasks for evaluation

	Task type	Time range	Known predictive "oracle" features
Bikeshare	Spatio-temporal	10/2017 - 10/2018	precipitation, pressure, temperature, slope, bikelanes
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

We consider four downstream tasks: three spatio-temporal predictions and one time series prediction (Table 1). They are:

- **Dockless bikeshare demand prediction (3D).** We collected Seattle dockless bikeshare data from the Transportation Data Collaborative. 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. 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. 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. The task is to predict the hourly bike count for the next 6 hours based on the data of last 7 days. This is a time series prediction, as the bridge is only a point in space.

4.1 Datasets

We collected 23 datasets from various online data portals, most of which are open data (Table 2). We included them because they are commonly used in urban studies [35, 46, 51, 52]. 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 the 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 [47]. We produced a race map and an income map based on 1km by 1km grids.

4.2 Integrative Model Baselines

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

- **No exogenous data**: a 3D CNN based prediction model that only trains on historical data without any exogenous data [49]. The model structure is described in [58].
- **Oracle model**: a network that makes use of hand-selected exogenous features, known to be predictive from the domain

Table 2: 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

literature (Table 1). The "oracle" network for the three 3D tasks adopts the structure described in [58], 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 [48].

- **Principal component analysis (PCA)**: We generate a latent representation that summarizes the 23 datasets using PCA [54], which is then used in downstream tasks.
- **Early fusion**: We produce a representation with a CDAE. Instead of encoding each dataset separately, the early fusion CDAE concatenates all datasets as a single tensor at the input, then applies 3D CNN layers. The decoder then reconstructs the concatenated tensor from the learned representation.

4.3 Fair Representation Baselines

We compare the EquiTensor model with a state-of-the-art method for producing fair 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 CDAE framework (Section 3.2), Fair CDAE uses an additional prediction head H 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.[50], but we adapted their method to our unsupervised learning scenario, and we applied a *gradient reversal* layer [17] 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.

4.4 Implementation Details

We implement all deep learning based models with TensorFlow [1], and perform training and inference with NVIDIA V100 Tensor Core GPUs. We adopt Adam optimizers using an exponential learning rate decay strategy.

EquiTensors. We train EquiTensor model with training samples covering 2014-02-01 to 2019-05-01 for 80 epochs using a batch size of 32. We then pass non-overlapping samples to the trained model and concatenate the output representation from each sample along temporal dimension to form the final EquiTensor. In our experiments we compress the 23 datasets into 5 channels, so the shape of EquiTensor is $32(H) \times 20(W) \times 45960(T) \times 5$. The same shape is used for the baselines.

Downstream tasks. For bikeshare demand prediction, we forecast hourly demand. We restricted the EquiTensor along the temporal dimension to match the domain of the bikeshare dataset. We predict criminal reports in the next 3-hour window. We average over 3-hour windows to match the temporal resolution of crime prediction. The model configuration for fire prediction is the same as that of crime prediction. For bike count prediction, we predict hourly bike count for the next 6 hours for a specific location. We query the EquiTensor to extract the time series $(45960(T) \times 5)$ of the corresponding grid cell as features for prediction.

Fairness metrics. To calculate the three fairness metrics PRD, NRD, and RD (Section 3.5), 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.

Evaluation. We evaluate our core integrative model by accuracy of downstream prediction tasks, adaptive weighting by total reconstruction error, and fairness by both the accuracy of the adversary and the three fairness metrics (i.e., RD, PRD or NRD).

5 RESULTS AND DISCUSSION

In this section, we show that the proposed core integrative model benefits downstream predictions significantly and outperform the representations generated by the integrative model baselines in terms of downstream prediction accuracy. The proposed adaptive weighting effectively improves the reconstruction accuracy of the core integrative model. We also show that EquiTensors are fairer than baseline representations and can help the downstream tasks to achieve accuracy that is competitive with the oracle networks.

5.1 Utility of EquiTensors

The results for our core integrative model (Core model) extended with adaptive weighting (Core model + AW) on four downstream prediction tasks are shown in Table 3.

Integrated representations improve performance. The oracle networks with hand-selected datasets outperform the models without exogenous data in four cases, indicating that adding exogenous datasets 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

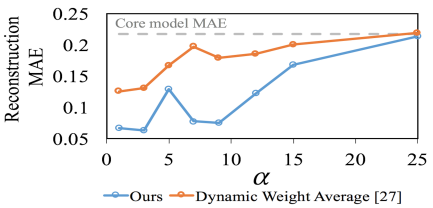

Figure 4: Total reconstruction error vs. α. Our adaptive weighting versus Dynamic Weight Average [27].

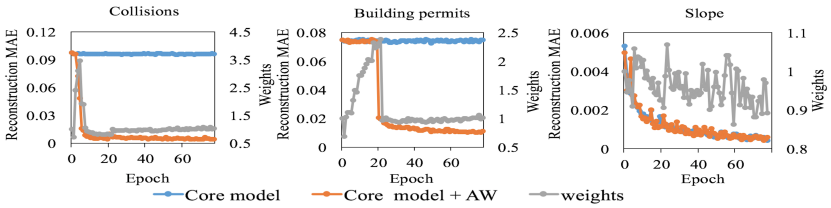

Figure 5: Reconstruction loss curves and adaptive weight curves on three datasets (α = 3). Under adaptive weighting scheme (Core model + AW), weights for individual datasets change with their reconstruction accuracy.

Table 3: Prediction accuracy (MAE) of downstream tasks. Parenthetical numbers are the factor improvement over PCA and early fusion.

Model	Bikeshare	Crime	Fire	Bike count
No exo. data [58]	0.408	0.137	0.133	12.057
Oracle [58]	0.382	0.111	0.110	10.983
PCA [54]	0.402	0.121	0.124	11.099
Early fusion	0.390	0.119	0.123	11.266
Core model	**0.385 (4.0×, 1.3×)**	0.113 (1.5×, 1.4×)	**0.112 (2.4×, 2.1×)**	11.050 (1.1×, 1.3×)
Core model+AW	0.387 (3.6×, 1.1×)	**0.106 (2.0×, 1.7×)**	0.114 (2.1×, 1.8×)	**11.049 (1.1×, 1.3×)**

of multiple urban datasets can capture generic information that is useful to an array of tasks.

Core model outperforms baselines. Table 3 shows that the proposed models (Core model and Core model + AW (α = 3)) outperform PCA and early fusion CDAE on all four tasks, and are competitive with the "best possible" oracle networks. Specifically, the proposed models show factor improvement (parentheses in Table 3) over PCA and early fusion in terms of performance gain of downstream tasks. For example, the core model for bikeshare prediction outperforms the *No exogenous data* baseline by 5.51%, which is 4.0× better than the improvement of PCA (1.39%) and 1.3× better than early fusion (4.36%). Similarly, the core model + AW shows a 2.0× and 1.6× improvement over PCA and early fusion, respectively, for crime prediction.

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 [63]. 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.

Adaptive weighting reduces total reconstruction error. The strength factor α (Equation 2) controls the influence of learning progress on the weight for the reconstruction loss of each dataset. Figure 4 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. The peak at α = 5.0 is not persistent across small changes in α values. Compared to Liu et al. [27] (orange line), our method consistently achieves higher total reconstruction accuracy for a range of α values. We use α = 3 for the rest of our experiments.

Figure 5 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. As such, the learned representation with adaptive weighting is likely to improve accuracy for downstream tasks that depend on these datasets.

5.2 Fairness of EquiTensors

We evaluate the fairness of EquiTensors using two case studies. For reported crimes we remove the effects of race and for bikeshare we remove the effects of income.

EquiTensors counteract bias in input data. Table 4 shows the accuracy of predicting sensitive information S (i.e., race and income) from representations generated by different models: Our core model with fairness (Core + Fair and Core + Fair + AW), the integrative models without fairness (PCA, Early fusion, Core model, and Core + AW), and two competing fairness approaches: Fair CDAE and EquiTensors without the disentangling module (Core model + Fair w/o disent.).

The sensitive information S 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 have weaker correlations with S than fairness-oblivious baselines. The low detection error of Fair CDAEs suggests that Fair CDAEs are not effective in removing the influence of S. The reason is that Fair 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 [6]. Therefore, an external adversary is often still able to recover S. We show that our model can better remove S from the learned representation than those without disentanglement (Core model + Fair w/o disent.).

Table 4: Accuracy of predicting a sensitive attribute (i.e., race or income) from various integrated representations. Higher MAE suggests weaker association with the sensitive attribute.

	λ	Race MAE	Income MAE
PCA [54]	/	0.005	0.005
Early fusion	/	0.001	0.001
Core	/	0.001	0.001
Core + AW	/	0.001	0.001
Fair CDAE [17, 50]	1.0	0.002	0.002
Fair CDAE [17, 50]	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**

Figure 6 (A) and (B) show the MAEs of adversary with changing λs for race prediction and income prediction, respectively. The parameter λ in the loss function for EquiTensor (Equation 5) controls the weight of the bias mitigation. 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 cases that when λ is around 2.0, adversary MAEs approach 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 would cause the adversary loss to dominate reconstruction error in Equation 5, leading to noisy EquiTensors with lower utility.

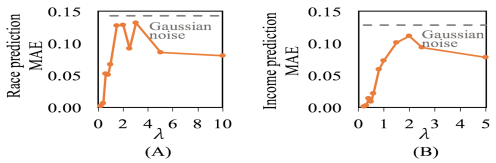

Figure 6: Adversary MAE vs. λ. At $\lambda \approx 2$, EquiTensors prevent discerning the sensitive attribute nearly as well as Gaussian noise.

Table 5 shows the results of crime predictions and bikeshare predictions using representations generated by different models. We reported mean and standard deviation (parentheses in Table 5) 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* is -27.7, indicating that the model overestimates the reported crime incidents for each cell in non-white neighborhoods by 28 cases more than the overestimation for white neighborhoods over the test period. The RD for *No exogenous data* is -23.1, indicating that the residual (prediction minus ground truth) is 23 more for the non-white regions than white regions over the test period. In other words, the reported crime data itself may contain correlations with race that are *amplified* by the *No exogenous data* model. In the bikeshare case,

Table 5: Accuracy and fairness of crime predictions and bikeshare predictions with different integrated representations in the form of mean (std). For RD, PRD, and NRD, lower absolute value suggests fairer predictions.

	λ	Crime prediction			Bikeshare prediction		
		Accuracy	Fairness		Accuracy	Fairness	
		MAE	RD	PRD	MAE	RD	NRD
No exo. data [58]	/	0.135 (0.002)	-23.1 (3.6)	-27.7 (1.9)	0.408 (0.002)	8.7(20.5)	-152.0 (10.2)
Oracle [58]	/	0.110 (0.007)	-12.0 (5.9)	-20.3 (4.3)	**0.382(0.002)**	73.3 (29.7)	-180.1 (17.8)
PCA [54]	/	0.117 (0.004)	-13.0 (3.6)	-20.2 (3.5)	0.400 (0.003)	55.3 (36.6)	-181.6 (21.5)
Early fusion	/	0.115 (0.004)	-12.5 (3.2)	-20.3 (2.4)	0.390 (0.006)	75.1 (35.8)	-183.2 (21.7)
Core	/	0.111 (0.002)	-8.9 (5.9)	-17.9 (4.3)	0.385 (0.001)	58.4 (19.1)	-172.1 (9.3)
Core+AW	/	**0.106 (0.004)**	-5.6 (7.2)	-15.2 (5.1)	0.388 (0.002)	32.2 (13.6)	-160.3 (7.9)
Core+Fair	0.6	0.114 (0.004)	-4.2 (4.8)	-14.3 (3.0)	0.392 (0.002)	23.1 (25.6)	-158.5 (11.3)
Core+Fair	1.0	0.112 (0.006)	6.9 (9.9)	**-7.8 (7.0)**	0.395 (0.006)	15.2 (32.5)	-154.7 (16.4)
Core+Fair	2.0	0.112 (0.004)	4.7 (7.5)	-9.3 (4.4)	0.394 (0.005)	6.3 (28.0)	-151.3 (12.0)
Core+Fair+AW	0.6	0.111 (0.002)	**-3.9 (3.1)**	-14.6 (2.1)	0.390 (0.003)	**-5.3 (18.3)**	**-142.1 (7.9)**
Core+Fair+AW	1.0	0.110 (0.005)	5.1 (6.2)	-9.2 (4.7)	0.394 (0.005)	11.2 (43.6)	-153.3 (20.5)
Core+Fair+AW	2.0	0.109 (0.004)	**-3.9 (5.8)**	-14.9 (3.5)	0.398 (0.003)	28.4 (32.9)	-161.6 (16.4)

we focus on RD and NRD, because underestimating bike demand in underrepresented communities is considered more harmful than overestimating that of the advantaged group. All bikeshare models with fairness-agnostic features show larger disparity (i.e., positive RD) than *No exogenous data*, implying that external features may introduce additional biases into the predictions. Overall, predictions with EquiTensors show improved RD and PRD (or NRD) over the fairness-oblivious baselines in both cases. This suggests that EquiTensors could help prevent introducing *new sources* of systematic bias into downstream predictions, although the downstream predictions will still be affected by biases in their own training data.

Fairness interventions preserve accuracy. Table 5 shows that in the crime prediction case, models with EquiTensors (Core + Fair and Core + Fair + AW) achieve prediction accuracy that is comparable to the oracle network. 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 5 also shows that EquiTensors help the bikeshare predictions achieve higher accuracy than the *No Exogenous data* baseline. In summary, compared to fairness-oblivious features, EquiTensors could help the downstream tasks achieve overall fairer predictions without sacrificing much accuracy.

6 CONCLUSIONS AND FUTURE WORK

We introduced an unsupervised approach to learn integrated and equitable data representations, the EquiTensors, for heterogeneous and multi-dimensional urban datasets. We demonstrated with real-world datasets and applications that EquiTensors could help reduce propagating discrimination from biased data, while still delivering prediction accuracy comparable to oracle networks trained with hand-selected datasets. EquiTensors present a different approach to making open data available, useful, and safe for urban applications: they improve accuracy, avoid the need for data discovery and preprocessing, and can help limit the discriminatory effects of using uncurated data. Future work involves studying the transferability of fair and integrated features to other applications or cities, handling sparse datasets using graph convolutional networks, and studying transparency issues that arise when using learned features.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

[13] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. 2019. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems* 21, 11 (2019), 4883–4894.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[15] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. 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*. 35–47. http://proceedings.mlr.press/v81/ekstrand18a.html

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

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

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

[19] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., USA, 3323–3331. http://dl.acm.org/citation.cfm?id=3157382.3157469

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

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

[22] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised Representation Learning of Spatial Data via Multimodal Embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1993–2002.

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

[24] Hyeon-Woo Kang and Hang-Bong Kang. 2017. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one* 12, 4 (2017), e0176244.

[25] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. 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*. 7482–7491.

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

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

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

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

[30] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. 2017. 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*. 5334–5343.

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

[32] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*. PMLR, 3384–3393.

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

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

[35] N McNeil, J Dill, J MacArthur, J Broach, and S Howland. 2017. 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).

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

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

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

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

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

[41] R Alexander Rixey. 2013. Station-level forecasting of bikesharing ridership: Station Network Effects in Three US Systems. *Transportation research record* 2387, 1 (2013), 46–55.

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

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

[44] Anian Ruoss, Mislav Balunović, Marc Fischer, and Martin Vechev. 2020. Learning Certified Individually Fair Representations. *arXiv preprint arXiv:2002.10312* (2020).

[45] Bashir Sadeghi and Vishnu Naresh Boddeti. 2020. Imparting Fairness to Pre-Trained Biased Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 16–17.

[46] Bilong Shen, Xiaodan Liang, Yufeng Ouyang, Miaofeng Liu, Weimin Zheng, and Kathleen M Carley. 2018. 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*. ACM, 724–733.

[47] SimplyAnalytics. 2018. *EASI/MRI Census US.* Retrieved November 2, 2018 from SimplyAnalytics

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

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

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

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

[52] Mingshu Wang and Lan Mu. 2018. Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. *Computers, Environment and Urban Systems* 67 (2018), 169–175.

[53] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Pengfei Wang, Yu Zheng, and Charu Aggarwal. 2018. 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*. ACM, 2457–2466.

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

[55] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

[56] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. FairGAN+: Achieving Fair Data Generation and Classification through Generative Adversarial Nets. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 1401–1406.

[57] An Yan and Bill Howe. 2019. 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*. 552–555.

[58] An Yan and Bill Howe. 2020. Fairness-Aware Demand Prediction for New Mobility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1079–1087.

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

[60] Yaqiang Yao, Jie Cao, and Huanhuan Chen. 2019. 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*. ACM, 1408–1417.

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

[62] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. 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*. ACM, 984–992.

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

[64] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* (Atlanta, GA, USA) *(ICML'13)*. JMLR.org, III–325–III–333. http://dl.acm.org/citation.cfm?id=3042817.3042973

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

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

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

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