EPJ manuscript No.
(will be inserted by the editor)

Integrative Urban AI to Expand Coverage, Access, and Equity of Urban Data

Bill Howe[1,a], Jackson Maxfield Brown[1,b], Bin Han[1,c], Bernease Herman[1,d], Nic Weber[1,e], An Yan[2,f], Sean Yang[1,g], and Yiwei Yang[1,h]

[1] University of Washington
[2] Meta

Abstract. We consider the use of AI techniques to expand the coverage, access, and equity of urban data. We aim to enable holistic research on city dynamics, steering AI research attention away from profit-oriented, societally harmful applications (e.g., facial recognition) and towards foundational questions in mobility, participatory governance, and justice. By making available high-quality, multi-variate, cross-scale data for research, we aim to link the macro study of cities as complex systems with the reductionist view of cities as an assembly of independent prediction tasks. We identify four research areas in AI for cities as key enablers: interpolation and extrapolation of spatio-temporal data, using NLP techniques to model speech- and text-intensive governance activities, exploiting ontology modeling in learning tasks, and understanding the interaction of fairness and interpretability in sensitive contexts.

1 Introduction

Cities are complex systems: collections of interacting agents that exhibit non-trivial collective behavior [2,19]. This observation has guided research in general principles of city planning that can govern the behavior of the complex adaptive system the city manifests. Early work by Jacobs proposed ideal sizes and specific guidelines for city neighborhoods [25], and more recently researchers have begun to empirically validate these ideas using mobile phone data [46]. West and Kempes model scaling behavior for cities as balancing sublinear growth in resource consumption (as a function of population) against the superlinear growth of socioeconomic effects, both positive (per capita wages) and negative (inequity, disease) as a power law with an exponent of about 0.15 (as opposed to, for example, the exponent of 0.25 identified in many

[a] e-mail: billhowe@uw.edu
[b] e-mail: jmxbrown@uw.edu
[c] e-mail: bh193@uw.edu
[d] e-mail: bernease@uw.edu
[e] e-mail: nmweber@uw.edu
[f] e-mail: annieyan9033@fb.com
[g] e-mail: seanyang38@uw.edu
[h] e-mail: yanyiwei@uw.edu

biological processes) [28,67]. More recently, the rate of COVID-19 spread has been shown to approximate the same power law [57]. As West and Kempes argue "Cities are machines we evolved to facilitate, accelerate, amplify, and densify social interactions."

The holistic study of cities as complex systems complements the rapid (yet ultimately opportunistic) proliferation of artificial intelligence technology in the public sector. Although conventional machine learning techniques are common in urban applications [65,71,44], neural architectures are opening new opportunities by adapting convolutional, recurrent, and transformer architectures to spatiotemporal data [35, 38,56,73,70,76,75]; see Grekousis 2020 for a recent survey [17].

These two lines of inquiry — top-down modeling of cities as complex systems and bottom-up modeling of specific urban systems using deep learning — are difficult to reconcile. Complex systems are not amenable to reductionist statistical experiments: comparing the results of an agent-based model with observed data (e.g., for autonomous vehicle research [11]) is often the best we can do, despite the challenges of addressing the inverse problems implied [8,59]. The central issue is that observational micro-data for cities are inconsistent in availability and quality, limiting the opportunity for validation of sophisticated models.

This inconsistency persists despite significant investments in open data. Over the last two decades, cities have increasingly released datasets publicly on the web, proactively, in response to transparency regulation. For example, in the US, all 50 states and the District of Columbia have passed some version of the federal Freedom of Information (FOI) Act. While this first wave of open data was driven by FOI laws and made national government data available primarily to journalists, lawyers, and activists, a second wave of open data, enabled by the advent of open source and web 2.0 technologies, was characterized by an attempt to make data "open by default" to civic technologists, government agencies, and corporations [64]. While open data has indeed made significant data assets available online, their uptake and use has been weaker than anticipated [64], an effect many attribute to inconsistent availability of high-value data across cities [32]. Ultimately, open data exhibit convenience sampling effects.

In this paper, we consider four research thrusts all aimed at using AI techniques to improve the coverage, access, and equity of urban data, and thereby reduce barriers and attract attention to the study of critical questions in city dynamics and socioeconomic interactions. Machine learning research is broadly recognized to be too narrow in applications and datasets, focusing on opportunistic, discriminatory, and profit-oriented applications [53,68,50]. By making high-quality urban data available across cities, across variables, across time-scales, and at multiple resolutions, we aim to make AI research on societally important problems the path of least resistance. But to accomplish this long-term goal, we need to address specific challenges in working with urban data.

Expanding existing sources By simultaneously modeling multiple heterogeneous datasets [69], we aim to identify the underlying relationships and interactions between urban systems as a middle path between reductionist, application-specific prediction tasks and holistic, simulation-oriented inference. But where our earlier work assumed uniform data coverage, we now need to apply advanced learning techniques to interpolate and extrapolate dense spatiotemporal datasets to account for inconsistent coverage (Section 2). These techniques help expand the utility and reach of data-hungry predictive models to counteract the sparse and inconsistent availability of public and private data. As an example, we show how the interpolation of urban transportation data is remarkably amenable to deep learning architectures developed for image inpainting on the web.

Developing new sources Open data in urban contexts is typically either spatiotemporal (vector or raster) or administrative (structured). But by investing in infrastructure, we can develop and make available data sources around governance, economics, decision-making, public participation. As an example, we show how transcripts from public meetings are amenable to computational processing to increase oversight and participation, if we can first establish an infrastructure and appropriate standards to collect and manage this data (Section 5).

Exploiting rich ontologies The use of large, noisy, and heterogeneous data motivates investment in data curation: associating contextual information with the data to mediate its collection and use. But manual curation activities (e.g., human labeling of data) scale poorly. In complex domains, human expertise is better invested developing richer labeling schemes than actually labeling data. For example, ontologies have been developed for electric mobility [55], humanitarian [3], and smart city applications [13,1,61,18,9]. But categorizing public data (e.g., social media posts) using these ontologies requires new techniques in hierarchical multi-label classification. As an example, we show how graph encoding techniques can be used to significantly improve performance in these contexts (Section 4).

Incorporating fairness and interpretability In every application of urban machine learning, prediction and modeling carries enormous risk of exacerbating inequity and opacity [21,49,42,43]. Building on recent advances in fair and explainable AI, we consider the interactions between accuracy, fairness, and explainability in urban applications. We then propose new methods for controlling these tradeoffs in response to emerging regulation (Section 3).

2 Interpolation of Spatiotemporal Data Using Deep Learning

Image inpainting is a task of synthesizing missing pixels in images. In computer vision, there are two board branches to inpaint images. The first branch contains diffusion-based or patch-based methods that utilize low-level image features to reconstruct the missing regions. The second branch contains learning-based methods that involve the training of deep learning models. Traditional diffusion-based methods transfer information from the valid regions to the missing regions, which are convenient to apply but limited to small missing regions only. Learning-based approaches aim to recover the images based on the patterns learned from large amount of training data. Such methods include context encoders by Pathak et al. [47], global and local image inpainting by Lizuka et al.[22], partial convolution method by Liu et al. [36] etc.

Image inpainting techniques have wide application potentials, including the geospatial domain that works frequently with satellite images. Zhang et al. [77] proposed a unified spatial-temporal-spectral deep convolutional neural network (CNN) image inpainting architecture to recover information obscured by poor atmospheric conditions in satellite images. Kang et al. [27] modified the architecture from [72] to restore the missing patterns of sea surface temperature from satellite images. Tasnim and Mondal [60] also applied the inpainting architecture from [72] to remove redundancies in satellite images and restore the imagery.

We build on prior work from our group in learning fair integrations of heterogeneous urban data [69]. We originally assumed uniform spatial and temporal coverage data, but in practice urban datasets are spatially imbalanced: one neighborhood may

be missing a variable of interest defined everywhere else, undermining trust in the results. Conventional statistical approaches to impute missing data, such as global/local mean imputing, interpolations, spatial regression models are limited in their ability to capture non-linear interactions, where deep learning methods, including image inpainting techniques in geo-spatial imputation, excel.

Given the similar nature of images and gridded urban data, we conjecture that image inpainting techniques can be adapted to impute missing urban data, improving coverage and quality, and therefore usability. As far as we know, no prior work that has exploited image inpainting techniques to reconstruct missing values in raster urban data. In this section, we present our preliminary experiments and results of utilizing an image inpainting technique to compute missing values in gridded urban data.

2.1 Example: Interpolating Urban Mobility Data

We use taxi trip data as a representative example of urban data, though the coverage of urban data is much broader. We used NYC taxi trip data from 2011 to 2016 from NYC Open Data Portal [10]. The years 2011 — 2014 cover the trips throughout the entire year, while 2015 and 2016 only cover half of the year. The raw data are collected tabular format, where each record/row contains the information of each taxi trip, including the longitude and latitude of the starting location. We considered the demand prediction problem, interpreting each record as an indicator of demand following Mooney et al. [43]. We processed the tabular data into raster format given the following steps:

- We defined a rectangular subset of the greater metropolitan area of New York City representing lower Manhattan. We only consider the taxi trips that began within this rectangular region.
- We imposed a 32x32 grid over our selected region. This choice of dimensions is somewhat arbitrary, balancing fidelity (reducing the need to upsample or downsample datasets too much), computational efficiency, and interpretability (1 grid cell is approximately 1 km^2.) For each year, each unique date, and each unique hour, we count how many taxi trips are within each grid and interpret these values as a an estimate of taxi demand in that cell, at that time.
- In total, we have 32,616 samples to model with, each having 32x32 dimension. 70% of samples (23,482) are used as training data, 10% (2,610) as validation set and the rest 20% (6,524) as test data.

2.2 Modeling & Results

We implemented the architecture from Liu et al.[36]. Many prior works in image inpainting only considered rectangular-shaped missing regions, but rarely are the patterns of missing data so regular. In urban data, the missing regions could be scattered or in irregular shapes corresponding to irregular political boundaries or inconsistent data collection. Therefore, Liu et al.'s work fits well into the urban scheme. Liu et al. used the summation of four different losses as the objective function, to account for different factors related to the perception of the resulting image, which was appropriate for web images but less appropriate for quantitative urban data. We only used the ℓ_1 regularization loss between the original data and inpainted data. The model hyper-parameters are set to be consistent with Liu's work. The learning rate is set to $1e^{-4}$ flat. The batch size is set to 32. The maximum iteration is 10,000 and we evaluated the model on validation set every 100 iterations.

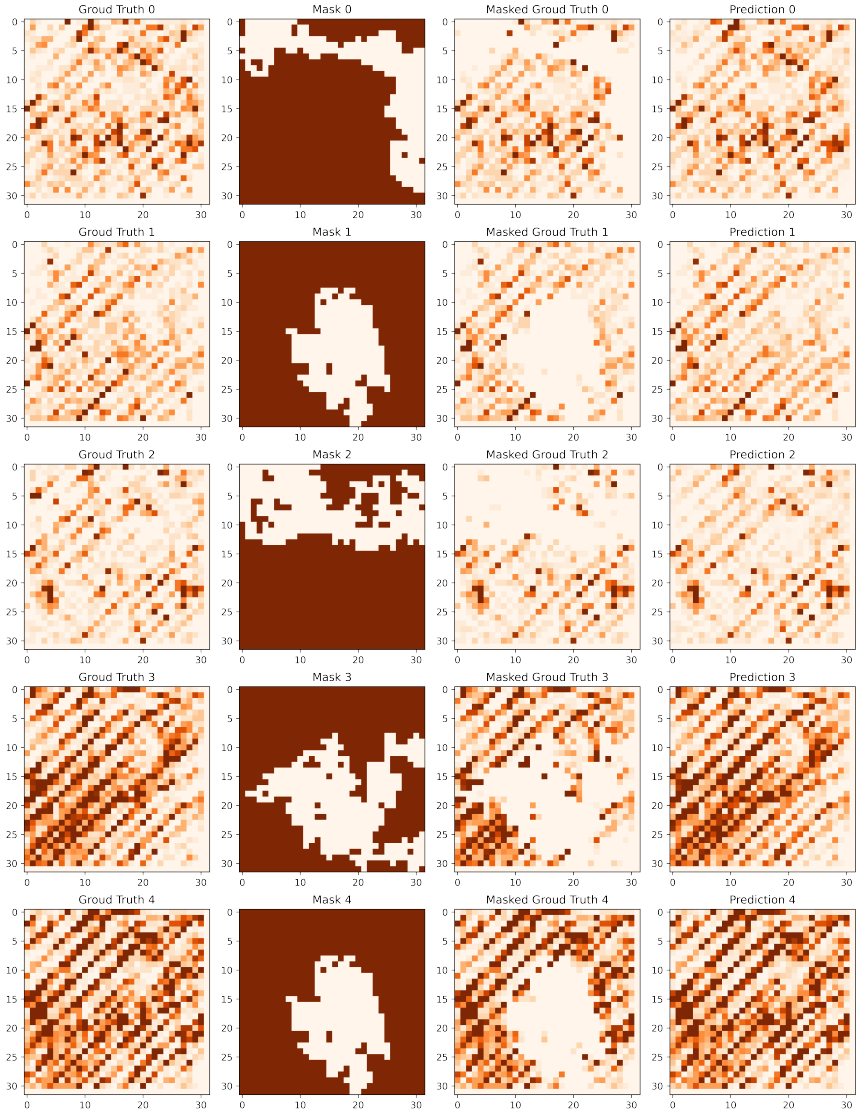

Fig. 1. Inpainting results of taxi trip data. From left to right, the columns are: ground truth images; the irregular masks; the masked ground truth; the final inpainting results.

Five inpainted examples are visually presented in Figure 1. We can see that the inpainting technique can be naturally applied to gridded urban data and yield promising results. Imputing the missing values in urban setting could also be viewed an a type of synthesis. The synthesis of partial urban data could improve the applicability and usability of urban data, but will require future work in multiple areas:

– Though deep learning methods are powerful, we need rigorous evaluation against traditional imputation techniques to see if these complex methods are warranted. Additionally, visual similarities are subjective, which is appropriate for web images but not if we intend to use these datasets for quantitative analysis. Additional quantitative measurements should be incorporated.

- The region of the experiment (NYC) and the dimensions of the urban grid (32x32) are both limited. Expanding the region to cover more area would capture more urban dynamics, while evaluating the effect of different grid sizes, will be necessary to test generalizability.
- We treat each date and each hour as a unique sample. But in reality, the current hour timestamp is closely related to the demand from the previous hour. Modeling each sample separately ignores such dependency. Therefore, we hypothesize that modifying the architecture to work with temporal blocks would help improve performances.

3 Trade-off among Distributive and Procedural Fairness

Real world datasets often contain societal biases, which are perpetuated in the machine learning models, leading to discriminatory decisions in high-stake domains. In response, many methods were developed to mitigate fairness by achieving some statistical measure of equity between majority and minority groups (e.g. equalized odds, equality of opportunity)[4,74,7,30,39]. This line of work is guided primarily by the notion of distributive fairness, which emphasizes on a fair allocation of resources.

However, prior work has shown that *procedural* fairness, the perceived fairness of the process that leads to the outcome, is equally as important as distributive fairness[5,63]. For example, in court systems, studies have shown that "most people care more about procedural fairness ... than they do about winning or losing the particular case." [63] Recent studies have also shown that procedural fairness is critical to automated decision systems [33,40]. For instance, through a cross-sectional survey study at a large German university, Marsinkowski et al. found that both distributive and procedural fairness have significant implications on higher education admission that uses an automated decision system[40].

The interaction between distributive fairness, interpretability, and procedural fairness are rapidly becoming a compliance issue. In April 2021, the EU released a proposal for sweeping regulation of algorithmic bias [14]. The same week, the Federal Trade Commission released a blog post [26] that described a legal framework for evaluating AI bias, foreshadowing enforcement. In the California, a bill regulating automated decision systems is in committee [23].

Drawing from procedural fairness theory, we propose Explanation Loss (See Equation 1), a novel fairness metric that measures procedural fairness and a method to optimize for it[34]. In particular, this metric measures the neutrality of the decision process to different demographic groups. Since complex black-box models (E.g. deep neural networks, tree-based ensemble models) are often used due to their high predictive power, we use interpretability methods to generate explanations that reveal the model decision process for each datum. The metric then computes the average absolute differences of the explanations between all possible pairs of input samples, one from the minority group and one from the majority group. The intuition is that the difference in the model's explanation for two groups can be approximated by the average differences of all pairs of individual explanations. Therefore, Explanation Loss measures how far away the decision process is from being perfectly neutral. In the following sections, we describe the data we used, the method that optimizes for the metric, and the preliminary results we obtained.

3.1 Data

We used the COMPAS dataset [31] for the preliminary study. COMPAS dataset, which contains attributes of criminal defendants, is often being used to study (deeply

flawed) recidivism models: whether a person will reoffend within 2 years). It is known to exhibit severe biases against minority groups. Specifically, studies have shown that models trained on COMPAS tend to overpredict recidivism for black defendants and underpredict recidivism for white defendants [31].

We preprocessed the dataset following Rieger et al. [52]. The dataset contains a total of 7214 samples. We filtered 1042 due to missing information about the recidivism. We categorized age into under 25, between 25 and 45, inclusively, and above 45. We categorized sex into Male and Female. We also categorized the crime description based on matching words, resulting in categories Possession (of drugs), Driving, Violence, Theft, and No Charge. For example, descriptions that are matched with "theft" or "burglary" are categorized as Theft. We then one-hot encoded all categorical variables, and used the numeric variables as is. We focused on equalizing explanations of the Black and Caucasian records, since these two are the predominant groups of the data.

We split the data into train, validation, and test sets, with a ratio of 80/10/10.

3.2 Method

Interpretability techniques aim to generate explanations for a model's individual predictions. A popular class of such techniques is known as feature attribution, which, given an input, the model, and prediction, assigns a number to each feature of the input to represent how much it contributes to the prediction [37,45,51,58]. There are two reasons that feature attribution methods are appropriate for our study: 1) they allow us to compare model's explanation for each prediction at the feature level, which is especially important for fairness since certain features are more sensitive than others 2) feature attribution vectors can be interpreted as attribution priors to incorporate the notion of procedural fairness in the model.

We propose the following regularization to achieve procedural fairness:

$$\hat{\theta} = \arg\min_{\theta} \sum_{x_i, y_i \in D} \mathcal{L}(f_\theta(x_i), y_i) + \lambda \frac{1}{|D_{s1}||D_{s2}|} \sum_{x_j \in D_{s1}, x_k \in D_{s2}} |expl(x_j) - expl(x_k)| \quad (1)$$

The regularization computes the average L1 norm of difference between explanations of every pair of instances (one from each group), $x \in D_{s1}$, $x' \in D_{s2}$. Each explanation, $expl(x)$, is a vector of feature importance scores with dimension equal to the number of features of the input. This vector is generated using any feature attribution method. In this study, we used Contextual Decomposition (CD) as the feature attribution method [45]. This regularization term takes the exact form of our proposed metric for procedural fairness.

We trained a simple multi-layer neural network model with 1 hidden fully connected layer of 100 neurons and ReLU activation, and varied a weight for the regularization term of 0 (no explanation loss), 0.2, 0.4, 0.6, 0.8, and 1.0. The model was trained with a batch size of 256 and a learning rate of 0.001 for 5 seeds, and the average results were reported. While the regularization equation refers to explanations of the entire dataset, in practice, this term is computed per batch for faster convergence. Specifically, for each batch, we partition the instances into two groups, then for every possible pair (one from each group), we compute the L1 norm of the difference of the feature attributions, and lastly we average the differences. An ablation study of the effect of batch size on results remains future work.

Reg Rate	Explanation Loss	Accuracy	Equality of Opportunity	Equalized Odds
0	1.68	70.40	0.19	0.50
0.2	0.05	70.11	0.22	0.56
0.4	0.03	70.16	0.22	0.58
0.6	0.02	70.90	0.21	0.56
0.8	0.01	70.65	0.23	0.56
1	0.02	69.70	0.24	0.60

Table 1. The effect of equal explanations on accuracy and fairness distances of the model.

3.3 Metrics

In addition to our proposed metric of procedural fairness (Explanation Loss), the metrics we used to evaluate the model include 1) accuracy and 2) fairness. We considered two popular fairness metrics: equality of opportunity [20] and equalized odds [54]. A model is said to satisfy equality of opportunity if the false positive rates are equivalent across two demographic groups. Similarly, equalized odds requires false negative rates to be equivalent across two groups in addition to false positive rates.

The loss term represents the notion of fairness *distance*: how far away the model is from perfectly fair. The fairness distance we consider is based on equality of opportunity [], and measures the absolute difference between the false positive rates of one demographic group (FPR1) and another (FPR2): $|FPR1 - FPR2|$. On the other hand, fairness distance based on equalized odds measures [] $|FPR1 - FPR2|$ + $|FNR1 - FNR2|$, which adds an additional absolute difference between the false negative rates.

3.4 Results

The results are summarized in Table 1. From the first two columns of the table, we can see that the regularization effectively encourages the model to predict with similar explanations across two demographic groups, which we interpret as improved procedural fairness by penalizing the tendency for a model to essentially learn two separate submodels, one for each group. Second, adding the regularization term does not reduce the accuracy of the model. Third, equalizing the explanations has a minor effect on fairness of the outcomes, causing a slight increase on fairness distances.

4 Hierarchical Multi-label classification

We demonstrate hierarchical multi-label classification (HMC) in the urban domain. HMC tasks involve a large set of labels organized into parent-child relationships, typically representing increasing specificity or isa relationships. Each input record is associated with multiple labels in the hierarchy, representing the uncertainty associated with a large label space in a complex domain. HMC has received increasing attention with the adoption of neural networks [16,78,15,66], often in contexts requiring significant human expertise, making large-scale labeling exercise infeasibly expensive. In other words, human expertise is invested in modeling the world through a complex ontology rather than labeling data using that ontology. As a result, the machine learning tasks represent a different set of requirements: the number of labels can be large relative to the number of labeled items, but there is structure among the labels that algorithms can exploit for distance supervision.

Ontology development is common in urban planning, where the complexity of the domain and multiplicity of perspectives require building consensus around a universe of discourse. For example, ontologies have been developed by teams of experts to describe electric mobility [55], humanitarian efforts [3], and smart city applications [13,1,61,18,9][1]. The HMC literature rarely considers these urban applications, instead favoring biological and scientific domains where public data is more readily available.

We are exploring new approaches for HMC that involve learning reusable representations of the ontology itself (using graph encoding techniques) to tame the complexity, then using these learned representations as the labels when training a classifier. We show that using these ontologies as a source of supervision can significantly improve the classification performance over other HMC techniques, motivating greater investment in developing comprehensive ontologies to represent the complex urban domain as a whole rather than expending resources on creating expensive labeled datasets for myriad specific applications.

4.1 Case Study: Community Listener

We worked with a local non-profit organization to identify the community needs from several sources of the data, such as social media (Twitter, Reddit, and Facebook conversations) and long-form survey responses. We classify these discourses into the Sustainable Development Goals Ontology (SDG)[3][2] and the Social Progress Index (SPI)[3]. The data and the predicted labels are then aggregated and visualized on an online dashboard serving policymakers and entrepreneurs. The Sustainable Development Goals Interface Ontology (SDG) was developed by United National Environment Programme to support the achievement of the 17 United National Sustainable Development Goals to promote human rights and equity. The ontology includes 169 nodes with 3 levels. **Social Progress Index (SPI)** was introduced by Social Progress Imperative to promote improvement and actions for social progress. They define Social Progress as "the capacity of a society to meet the basic human needs of its citizens, establish the building blocks that allow citizens and communities to enhance and sustain the quality of their lives, and create the conditions for all individuals to reach their full potential." SPI includes 3 levels with 124 nodes.

4.2 Experimental Settings

Table 2. Dataset Statistics

Programs			Organization		
Train	Val	Test	Train	Val	Test
6412	801	802	4558	570	570

There are two datasets used in this experiment, Programs and Organizations. Programs is a list of descriptions of humanitarian programs; the task is to determine

[1] https://techcommunity.microsoft.com/t5/internet-of-things/smart-cities-ontology-for-digital-twins/ba-p/2166585

[2] https://www.unep.org/explore-topics/sustainable-development-goals/what-we-do/monitoring-progress/sdg-interface-ontology

[3] https://www.socialprogress.org/2020-Social-Progress-Index-Methodology.pdf

which areas of humanitarian need are intended by the Program. The description typically mentions the mission and areas of focus for the program, which we anticipated would make Programs relatively easy to classify. The Organizations dataset is a list of companies and non-profits that may work in areas of interest for humanitarian causes. In this case, the descriptions are less likely to explicitly mention areas of humanitarian need. For both datasets, we associate each record with zero or more labels from the SDG and SPI ontologies. Statistics of the two datasets are shown in Table 2. We split each dataset into training, validation, and test set with 8:1:1 ratio. The models are optimized with validation set and the experimental results are reported from the test set.

We experimented with different text embedders and classification models to find the best combinations. Because the organization did not have abundant computation resources, we limited our choices within computation efficient models. We chose TF-IDF and Glove [48] as our text embedders. For the classification model, we adopted two frameworks for classification, one considered the hierarchical structure with graph encoding (named **Ontology**) within the labels and the other did not (named **naive**). The **naive** model considered the labels as a flat list. The model consisted of two fully connected layers and iwas optimized with Binary cross entropy, which is often used for multi-label classification. The diagram of the **ontology** framework is shown in Figure 2. The framework learned a representation for the label ontology using a graph autoencoder [29]. Then, the model considered the node embeddings and mapped the input instances onto the node embedding space with cosine similarity. Finally, the model was optimized with binary cross entropy and produce probability confidence as output. The threshold for classification is set to be 0.5. Finally, we evaluated all models with Precision (P), Recall (R), and F1 score which are commonly used in multi-label classification community. Following the literature, a data record is considered correctly classified when the predicted leaves match the ground truth exactly: there is no partial credit for siblings, for example.

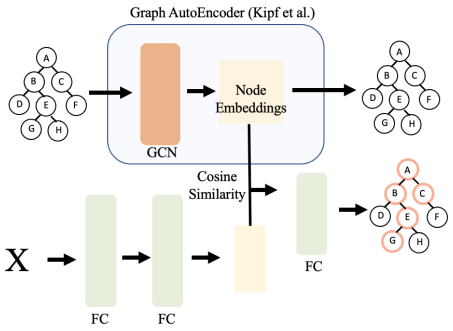

Fig. 2. Illustration of our framework.

Experimental Results We demonstrate our experimental results in Table 3. Because these two datasets are custom and not publicly available, we provide results from two baseline methods — majority vote and random selection. We can observe that considering the label ontology significantly improve the results. The trained model then allows us to tag the discourses on social media based on the humanitarian ontologies from SPI and SDG and to visualize within an online dashboard serving policymakers and entrepreneurs. As a result, we can organize public discourse and participation to capture levels of interest in various topics.

This approach is potentially critical for addressing data scarcity in practice. As we have argued, in complex domains, obtaining labeled data is expensive and requires significant human expertise. For example, determining whether a potential project is related to a goal to *enhance inclusive and sustainable urbanization* (SDG 11.3), *achieve sustainable management of resources* (SDG 12.2), *encourage adoption of sustainable practices* (SDG 12.6), or all three, requires significant expertise with the SDG ontol-

Table 3. Experimental results on the Program and Organization datasets. **Acc** indicates accuracy, **P** means precision, **R** is recall, and **F1** suggest F1 score. Because these are custom datasets that are not publicly available, we also provide results from two baseline methods, majority vote and random selection. We can observe that considering ontology improves the results significantly.

	Programs				Organization			
	Acc	P	R	F1	Acc	P	R	F1
Majority	0.075	0.006	0.075	0.01	0.056	0.003	0.056	0.006
Random	0.003	0.016	0.003	0.003	0.006	0.018	0.006	0.008
TFIDF + naive	0.102	0.085	0.132	0.087	0.090	0.032	0.090	0.030
Glove + naive	0.202	0.143	0.201	0.158	0.249	0.141	0.249	0.171
TFIDF + Ontology	0.145	0.134	0.167	0.159	0.143	0.095	0.123	0.117
Glove + Ontology	**0.245**	**0.175**	**0.254**	**0.219**	**0.286**	**0.176**	**0.287**	**0.256**

ogy, municipal government practices, and the data being labeled. Moreover, labeled datasets can be rendered obsolete with only minor changes to the ontology, requiring an expensive re-labeling exercise. To enable comprehensive cross-sector models that can be deployed in a variety of contexts, we need to make efficient use of the human attention invested in creating the ontology.

5 Modeling Governance Behaviors

The source of municipal democracy in the United States is found in city halls across the country. Even as our collective work in the analysis of urban space is used to create, debate, and ultimately enact urban policy, there is a lack of large-scale quantitative studies on municipal government. Comparative research into municipal governance in the USA is often prohibitively difficult due to a broad federal system where states, counties, and cities divide legislative powers differently. This power distribution has contributed to the lack of necessary research into the procedural elements of administrative and legislative processes because it affords each municipality to each have their own standards for archival and publishing of municipal data [62].

To better study the complexities of municipal councils across the county, multiple tools are needed to standardize and aggregate data into large research databases and access portals. The data from municipal government meetings (videos, transcripts, voting records, etc.) must be made more accessible to both the general public and to researchers, and, such tools must be deployed in multiple municipalities across the nation so that data can be used in aggregate to study the spread of policy, topic coverage, public sentiment, and more.

Once this infrastructure is available, it becomes possible to conduct large-scale quantitative studies on the dynamics of discourse in policy deliberation and enactment, quantifying how much of policy is decided upon using community sentiment as the policy basis, how such policy is supported or not from the public, and how similar policy proposals in different municipalities (or levels of government) are discussed and either enacted or rejected.

5.1 Council Data Project

To enable such large-scale studies, we have begun work on "Council Data Project," [6] a suite of tools for deploying and managing infrastructure for rapidly generating,

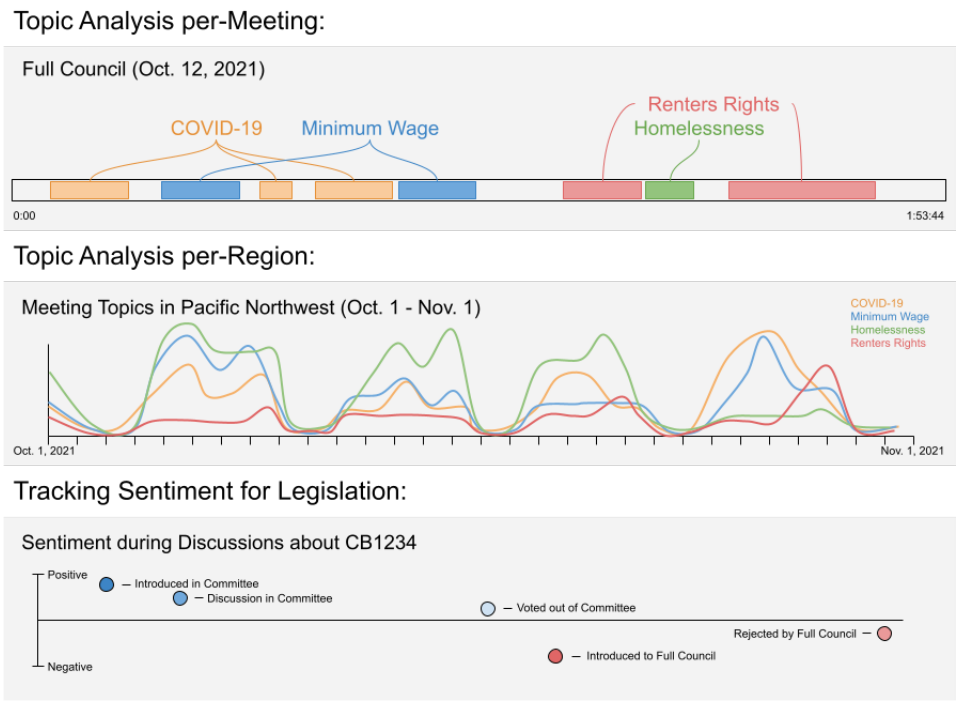

Fig. 3. Examples of analysis made possible through CDP infrastructure. Using the produced transcripts we can build topic models to tag topics both in a single meeting's transcript and track topic trends over time. With multiple CDP instances we can show how these trends hold (and spread — investigating the topical latency between municipalities) over entire regions. Additionally, building models for tracking the sentiment of discussions regarding specific pieces of legislation as they move through council.

archiving, and analyzing transcript datasets of municipal council meeting content. Council Data Project (CDP) is easily deployable and generalizes to many different meeting venues but is specifically built with municipal council meetings in mind.

For each meeting a CDP deployment processes, our tools generate a transcript of timestamped sentences, and archives the produced transcript and all attached metadata (minutes items, presentations and attachments, voting records, etc.). CDP deployments additionally create a keyword based index multiple times a week to enable plain-text search of events.

To further the utility of the CDP produced corpus we are creating audio classification models for labeling each sentence with the classified speaker, aligning sentences in the transcript to the provided list of minutes item, re-using the generated keyword based index for a municipality level n-gram viewer [41], and much more. Such work will enable the creation of datasets such as a dataset of discussions where only a set of specific councilmembers are present, or a dataset of discussions regarding specific pieces of legislation (minutes items).

Council Data Project enables large-scale quantitative studies by generating standardized municipal governance corpora — including legislative voting records, timestamped transcripts, and full legislative matter attachments (related reports, presentations, amendments, etc.). CDP enables the reproduction of political science research such as studying the effects of gender, ideology, and seniority in council deliberation

[24], and, studying the effects that adopting information communication technologies have on the civic participation process [12].

In constructing CDP to be as easily deployable as possible, we enable studies to understand how these behaviors generalize (or act as outliers) in different municipalities and settings.

Effective use of this new source of data motivates research in adapting deep learning techniques to multi-speaker, structured settings. Tasks include identifying speakers, topic and sentiment labeling by speaker to understand political positions, labeling speech by agenda topic, summarizing public sentiment to guide outreach and communication investments. These problems appear to be within the capabilities of emerging deep learning techniques, but require research attention in formulating the problem and evaluating competing techniques, which in turn require access to high-quality labeled datasets. Moreover, linking public discourse over social media with formal discourse in public hearings, administrative data collected through municipal service delivery, and geospatial data collected through sensing technologies will be required to meet our goal of a holistic study of the science of cities.

6 Conclusions

We aim to improve the coverage, access, and equity of urban data to advance understanding of city dynamics, unifying a top-down, holistic view of cities as a complex system and bottom-up, application-oriented view of cities as an assembly of independent subsystems. We aim to combat the disproportionate attention received by online advertising, face recognition, image labeling, and NLP tasks that dominate the machine learning literature by making high-quality, comprehensive urban datasets available for research. We identify four areas of research, with promising preliminary results, that involve the application of AI in urban contexts: spatiotemporal interpolation of data, unifying fairness and interpretability in the context of emerging regulation of algorithms, accommodating the complex domain models that are necessary to describe cities holistically, and engaging with new sources of data at the intersection of public discourse and policymaking.

References

1. Tarek Abid, Hafed Zarzour, Mohamed Ridda Laouar, and Mohamed Tarek Khadir. Towards a smart city ontology. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE, 2016.
2. Peter M Allen. *Cities and regions as self-organizing systems: models of complexity*. Routledge, 2012.
3. Ki-moon Ban. Sustainable development goals, 2016.
4. Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression, 2017.
5. Steven L. Blader and Tom R. Tyler. A four-component model of procedural justice: Defining the meaning of a "fair" process. *Personality and Social Psychology Bulletin*, 29(6):747–758, 2003. PMID: 15189630.
6. Jackson Maxfield Brown, To Huynh, Isaac Na, Brian Ledbetter, Hawk Ticehurst, Sarah Liu, Emily Gilles, Katlyn M. f. Greene, Sung Cho, Shak Ragoler, and Nicholas Weber. Council Data Project: Software for Municipal Data Collection, Analysis, and Publication. *Journal of Open Source Software*, 6(68):3904, 2021.
7. Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized data pre-processing for discrimination prevention, 2017.

8. Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.
9. Antonio De Nicola and Maria Luisa Villani. Smart city ontologies and their applications: A systematic literature review. *Sustainability*, 13(10):5578, 2021.
10. Brian Donovan and DB Work. New york city taxi trip data (2010-2013). *Univ. Illinois Urbana-Champaign, Champaign, IL, USA, Tech. Rep*, 2014.
11. Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator, 2017.
12. Katherine Levine Einstein, David Glick, Luisa Godinez Puig, and Maxwell Palmer. Zoom does not reduce unequal participation: Evidence from public meeting minutes. 2021.
13. Paola Espinoza-Arias, María Poveda-Villalón, Raúl García-Castro, and Oscar Corcho. Ontological representation of smart city data: From devices to cities. *Applied Sciences*, 9(1):32, 2019.
14. EUR-Lex. Proposal for a regulation of the european parliament and of the council. laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. `https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206`, April 2021.
15. Shou Feng, Ping Fu, and Wenbin Zheng. A hierarchical multi-label classification method based on neural networks for gene function prediction. *Biotechnology & Biotechnological Equipment*, 32(6):1613–1621, 2018.
16. Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. *NeurIPS 2020*, 33, 2020.
17. George Grekousis. Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. *Computers, Environment and Urban Systems*, 74:244–256, 2019.
18. Amelie Gyrard, Antoine Zimmermann, and Amit Sheth. Building iot-based applications for smart cities: How can ontology catalogs help? *IEEE Internet of Things Journal*, 5(5):3978–3990, 2018.
19. Seungwoong Ha and Hawoong Jeong. Unraveling hidden interactions in complex systems with deep learning. *Scientific reports*, 11(1):1–13, 2021.
20. Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.
21. 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.
22. Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4), July 2017.
23. California Legislative Information. Ab-13 public contracts: automated decision systems. (2021-2022). `https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB13`, July 2021.
24. Tonja Jacobi and Dylan Schweers. Justice, interrupted: The effect of gender, ideology, and seniority at supreme court oral arguments. *Va. L. Rev.*, 103:1379, 2017.
25. Jane Jacobs. *The death and life of great American cities*. Vintage, 2016.
26. Elisa Jillson. Aiming for truth, fairness, and equity in your company's use of ai. 2021.
27. Song-Hee Kang, Youngjin Choi, and Jae Young Choi. Restoration of missing patterns on satellite infrared sea surface temperature images due to cloud coverage using deep generative inpainting network. *Journal of Marine Science and Engineering*, 9(3), 2021.
28. Christopher P Kempes and Geoffrey B West. The simplicity and complexity of cities. *The Bridge*, 50, 2020.
29. Thomas N Kipf and Max Welling. Variational graph auto-encoders. In *Bayesian Deep Learning Workshop (NeurIPS 2016)*, 2016.
30. Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning, 2020.
31. Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm, 2016.

32. Melissa Lee, Esteve Almirall, and Jonathan Wareham. Open data and civic apps: first-generation failures, second-generation improvements. *Communications of the ACM*, 59(1):82–89, 2015.
33. Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.
34. Gerald S. Leventhal. *What Should Be Done with Equity Theory?*, pages 27–55. Springer US, Boston, MA, 1980.
35. 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.
36. Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions, 2018.
37. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
38. 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.
39. David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations, 2018.
40. Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. Implications of ai (un-)fairness in higher education admissions: The effects of perceived ai (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 122–130, New York, NY, USA, 2020. Association for Computing Machinery.
41. Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.
42. Claire Cain Miller. When algorithms discriminate. *The New York Times*, 9, 2015.
43. 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.
44. Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas. Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1393–1402, 2013.
45. W. James Murdoch, Peter J. Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms, 2018.
46. Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: A mobile phone data perspective. *CoRR*, abs/1603.04012, 2016.
47. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016.
48. Jeffrey Pennington, Richard Socher, and Christopher D 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.
49. Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund, and Katherine Lamberth. Algorithms and bias: What lenders need to know. *White & Case*, 2017.
50. Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

51. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.
52. Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge, 2020.
53. Cynthia Rudin and Kiri L Wagstaff. Machine learning for science and society, 2014.
54. Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. Transfer of machine learning fairness across domains, 2019.
55. Mario Scrocca, Ilaria Baroni, and Irene Celino. Urban iot ontologies for sharing and electric mobility. *Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal*, 2021.
56. 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.
57. Andrew J. Stier, Marc G. Berman, and Luis M. A. Bettencourt. Covid-19 attack rate increases with city size, 2020.
58. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.
59. Matt Taddy, Herbert K. H. Lee, and Bruno Sansó. Fast bayesian inference for computer simulation inverse problems. Technical Report AMS2008-3, 2008.
60. Jarin Tasnim and Debajyoti Mondal. Data reduction and deep-learning based recovery for geospatial visualization and satellite imagery. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5276–5285, 2020.
61. Jacques Teller, Abdel Kader Keita, Catherine Roussey, and Robert Laurini. Urban ontologies for an improved communication in urban civil engineering projects. presentation of the cost urban civil engineering action c21 "towntology". *Cybergeo: European Journal of Geography*, 2007.
62. Jessica Trounstine. All politics is local: The reemergence of the study of city politics. *Perspectives on Politics*, 7(3):611–618, 2009.
63. Tom R. Tyler. Procedural justice and the courts. *Court Review 26*, 2007.
64. Stefaan Verhulst, Andrew Young, Andrew Zahuranec, Susan Ariel Aaronson, Ania Calderon, and Matt Gee. The emergence of a third wave of open data. 2020.
65. 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.
66. Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *ICML 2018*, pages 5075–5084, 2018.
67. Geoffrey B West. *Scale: the universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies*. Penguin, 2017.
68. Meredith Whittaker. The steep cost of capture. *Interactions*, 28(6):50–55, 2021.
69. An Yan and Bill Howe. Equitensors: Learning fair integrations of heterogeneous urban data. In *ACM SIGMOD*, pages 2338–2347, 2021.
70. Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, Yanwei Yu, and Zhenhui Li. Modeling spatial-temporal dynamics for traffic prediction. *arXiv preprint arXiv:1803.01254*, 2018.
71. Ji Won Yoon, Fabio Pinelli, and Francesco Calabrese. Cityride: a predictive bike sharing journey advisor. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 306–311. IEEE, 2012.
72. Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.
73. 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.

74. Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
75. Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. Dnn-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 92. ACM, 2016.
76. 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.
77. Qiang Zhang, Qiangqiang Yuan, Chao Zeng, Xinghua Li, and Yancong Wei. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4274–4288, 2018.
78. Zhenzhen Zou, Shuye Tian, Xin Gao, and Yu Li. mldeepre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. *Frontiers in genetics*, 9:714, 2019.