Eurographics Conference on Visualization (EuroVis) 2022
R. Borgo, G. E. Marai, and T. Schreck
(Guest Editors)

Volume 41 (2022), Number 3

Urban Rhapsody: Large-scale exploration of urban soundscapes

Joao Rulff[1], Fabio Miranda[2], Maryam Hosseini[1], Marcos Lage[3], Mark Cartwright[4], Graham Dove[1], Juan Bello[1], Claudio T. Silva[1]

[1]New York University, [2]University of Illinois at Chicago, [3]Universidade Federal Fluminense, [4]New Jersey Institute of Technology

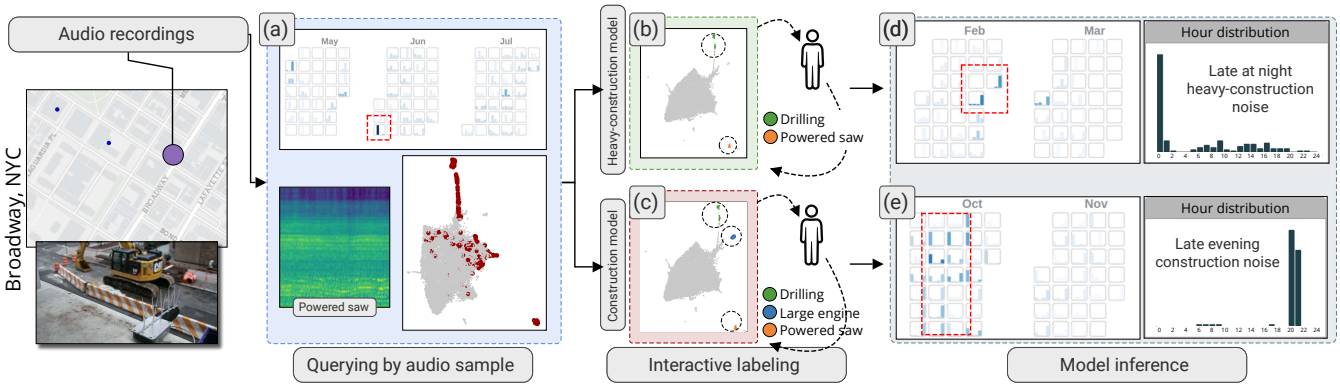

Figure 1: *We use Urban Rhapsody to assess after-hour construction in New York City, first selecting audio recordings captured by sensors deployed around Broadway. Urban Rhapsody allows users to query using an audio sample, and drill down to days containing similar audios (a). Using the interactions provided by the tool, we are able to create classification models according to a user's perception of the soundscape (b,c), and then use these models to classify the entire data set and look for unusual events (d,e).*

Abstract

Noise is one of the primary quality-of-life issues in urban environments. In addition to annoyance, noise negatively impacts public health and educational performance. While low-cost sensors can be deployed to monitor ambient noise levels at high temporal resolutions, the amount of data they produce and the complexity of these data pose significant analytical challenges. One way to address these challenges is through machine listening techniques, which are used to extract features in attempts to classify the source of noise and understand temporal patterns of a city's noise situation. However, the overwhelming number of noise sources in the urban environment and the scarcity of labeled data makes it nearly impossible to create classification models with large enough vocabularies that capture the true dynamism of urban soundscapes. In this paper, we first identify a set of requirements in the yet unexplored domain of urban soundscape exploration. To satisfy the requirements and tackle the identified challenges, we propose Urban Rhapsody, a framework that combines state-of-the-art audio representation, machine learning and visual analytics to allow users to interactively create classification models, understand noise patterns of a city, and quickly retrieve and label audio excerpts in order to create a large high-precision annotated database of urban sound recordings. We demonstrate the tool's utility through case studies performed by domain experts using data generated over the five-year deployment of a one-of-a-kind sensor network in New York City.

CCS Concepts

• ***Human-centered computing*** → ***Visualization systems and tools; Visual analytics;***

1. Introduction

City soundscapes represent a rich source of information about urban systems, such as transportation, civil construction, and social activity. Low-cost sensors can be used to capture aspects of this acoustic environment, and computational methods for large-scale data analysis offer new approaches to characterizing the different contributing sources. Such understanding offers insight into how a city behaves through space and time (e.g., *"what are the typical sounds in a neighborhood during the night?"*), and can help in tackling various urban problems such as noise pollution. The re-

search reported here was undertaken in partnership with researchers from one such sensing initiative, the Sounds of New York City (SONYC) project [BSN*19], who have developed and deployed low-cost sensors to measure and stream real-time sound pressure level (SPL) and audio data. To date, more than fifty sensors have been deployed throughout New York City (NYC), collecting data for over five years (in total, more than 60 TB). To meaningfully understand this data, the project's researchers are developing new machine listening models that 1) extract audio embeddings and 2) classify these sounds based on a set of predefined labels. However, these tasks pose several challenges that impede even state-of-the-art models' effectiveness in capturing the urban soundscape's dynamism. First, audio is complex, a recording typically captures different sound sources (e.g., dogs barking and people talking) simultaneously. Second, sound events are transient (e.g., a honking car horn) but in aggregation can last for hours (e.g., car engines on a busy highway). Third, audio has a temporal aspect, and so unlike images or words, sounds do not have a straightforward pictorial representation, limiting our ability to quickly review a large collection of recordings in parallel. Hundreds of images can be reviewed at the same time, with objects identified in minutes. However, looking for patterns or events in a large collection of audio data often requires listening to hours of individual recordings one after another. Analyzing audio data is time-consuming and hard to scale. This calls for novel techniques and visualization interfaces to facilitate the process, leveraging human expertise.

Motivated by these challenges and the need to gain new insights into the soundscape of the city, we introduce Urban Rhapsody, a framework for the interactive visual analysis of large collections of urban acoustic data. Using recent advances in machine listening to generate audio representations, Urban Rhapsody allows analysts to create a visual representation of the soundscape across different ranges of temporal and geographical granularity. We adopt a human-in-the-loop approach that enables users to interactively label data points, create new classification models based on their expertise of the soundscape, and assess the performance of audio classification tasks. Finally, because noise patterns might happen at different scales (minutes, days, months, etc.) in the urban environment, we employ a multilevel visualization scheme. Using case studies that demonstrate the utility of Urban Rhapsody, we showcase support for fast exploration of similar sounds or concepts, assessment of classification model outputs in different scenarios, geographical and temporal understanding of the embedding space, and summarization of soundscapes by key representative audio frames. Previous approaches to these challenges were either applied in a different context [DBC*17], or constrained to the analysis of sound pressure level (SPL) data [MLD*18], painting an incomplete picture regarding urban noise problems [ZLW*14]. Urban Rhapsody is the first visual analytics framework that enables a comprehensive analysis of urban acoustic environments, going beyond time series to leverage a unique audio data set that enables a more comprehensive analysis. Our contributions can be summarized as follows: (1) A set of requirements, elicited in collaboration with SONYC's audio researchers, for visual exploration of large urban audio sets. (2) A set of visual interactions that enables users to iteratively construct audio machine learning models; (3) An interactive visual analysis framework, Urban Rhapsody, that supports concept-based exploration of large collections of audio recordings (such as the ones generated over the five-year deployment of the SONYC sensor network). We illustrate this with two case studies set in NYC, highlighting how our approach can be useful in tackling issues that have generated intense public debate. Our framework is also available on GitHub (https://github.com/VIDA-NYU/Urban-Rhapsody).

2. Background

According to the World Health Organization, in Western Europe alone, more than 1 million healthy life-years are lost annually to environmental noise pollution [Org11], and in NYC, an estimated 9 out of 10 adults are exposed to excessive noise levels [NGM*12]. This impacts public health [HSN14], social well-being [GCA06] and quality of life [DZD*10], as noise increases stress, sleep disruption, annoyance and distraction [Bro07, Org11, HDVT*08, Muz02]. To mitigate this, governments devise noise codes that typically consider SPL measurements in relation to time of the day/week and location and impose regulations that aim at mitigating the noise at the source (e.g., by erecting sound barriers around major roads or modifying building designs) [tab07, BH10, HSN14]. However, enforcing these codes is time-consuming and costly, requiring trained inspectors to be present at sites to make assessments and capture sound carefully using calibrated equipment [BSN*19].

Beyond this, noise pollution can be highly subjective [dPVCR15], and so quantitative SPL metrics may be insufficient [RLB03, Gua03]. Because of this, there is a shift towards understanding the *source* of the noise, and to consider context in people's perception of sounds [RD05, VKDSVK14]. Such a "soundscape approach" [PDA09, Bro10, DAB*13] views the acoustic environment as composed of both positive and negative sources [Bro12]. Data gathered using SONYC sensors offers a unique opportunity to measure noise pollution quantitatively and additionally gain insights into the acoustic environment's qualitative characteristics. We can therefore conduct structured assessments at scale, accounting for both SPL and sound source. This raises important challenges (outlined in Section 4.2) that we seek to address in this research. Urban Rhapsody is the first step towards allowing domain experts to better understand the soundscape of complex cities such as NYC.

3. Related Work

3.1. Urban visual analytics

Urban areas are a major source of data that have tremendous potentials to improve policy making, enhance the lives of citizens, and pursue sustainable development. Visualization systems have for long been an important tool for the analysis of urban data [ZWC*16, DFL*18]. Several approaches use urban data to study different properties of a city, such as air pollution [ZLH13], public utility service problems [ZYM*14], sunlight access [MDL*19], land use [QS14], human movement patterns [NSL*12, LGTR15, MDL*17], transportation [AA08, WLY*13, FPV*13, IYT*14, ZFA*14], and also the relationship between these data sets [MME*12, CDDF16, DWX*21]. More general tools, such as ArcGIS [JVHKL01], Urbane [FLD*15,

DTZM*18], and Vis-A-Ware [OSS*16] have facilitated the use of multiple urban data sets to help inform urban planning and decision making process.

In our previous work, Time Lattice [MLD*18], we have tackled the problem of noise pollution by proposing a data structure and visual interface that allowed experts to explore a large data set composed of SPL dB measurements from SONYC sensors. We only used SPL measurements without considering that the soundscape of a city is composed of different sources and can be perceived differently by different people. With Urban Rhapsody, our goal is to account for the user's knowledge and perception in the exploratory process of large collections of urban sound recordings in a vocabulary-free approach, meaning that users are free to explore the soundscape according to any concept they create. To the best of our knowledge, Urban Rhapsody is the first visual analytics system specifically designed to allow domain experts to explore a large collection of sound recordings of an urban environment.

3.2. Environmental sound representation

In recent years, several large audio data sets have been released that have moved the field of environmental machine listening forward [GEF*17, FFP*20]. However, many audio classification tasks do not map onto the class vocabulary of these data sets and thus require additional labeling, which is time-consuming and costly. To address this problem, machine listening practitioners have turned to transfer learning [YCBL14] in recent years, which has been shown to be effective for many audio classification tasks [AVT16, AZ17, JPP*18, KKF18, CCSB19, CWSB19, TGdCQR20, GCKT21]. In transfer learning, models are typically pre-trained on large data sets using supervised [AVT16, HCE*17] or self-supervised learning [AZ17, JPP*18, KKF18, CCSB19, TGdCQR20], and the knowledge acquired during pre-training is re-used for tasks where data is limited. A common method of re-using this knowledge is to treat the pre-trained models as feature extractors, utilizing learned latent representations (i.e., embeddings) from within the pre-trained models as the inputs to models with little or no labeled data. Look, Listen, and Learn [AZ17] is one such pre-trained model whose embeddings were shown to be discriminative in several environmental audio classification tasks [AZ17, CWSB19, CCSB19, Wil21]. This model is pre-trained using self-supervision on an auxiliary task of audio-visual correspondence. In this work, we use OpenL3 [CWSB19], an open-source code of Look, Listen, and Learn, as an audio feature extractor to transform each audio recording into a series of embedding representations.

3.3. Machine-learning-aided multimedia exploration

Machine learning has opened a new horizon in data exploration across various fields, with numerous systems making use of the powerful capabilities it provides. For instance, Urban Mosaic [MHL*20] uses deep learning representations to search for patterns in a large collection of street-level images. II-20 [ZWVW20] allows users to generate image classifiers using novel interactions. Previous works tried to explore the semantic meaning of the features extracted by deep learning models, as they do not always map into human-understandable semantic features: Embedding Projector [STN*16] and Latent Space Cartography [LJLH19] enable the analysis of embedding spaces for multimedia data through multidimensional projections [JCC*11, MHM18] to enable users to understand features that might be encoded in the latent representations.

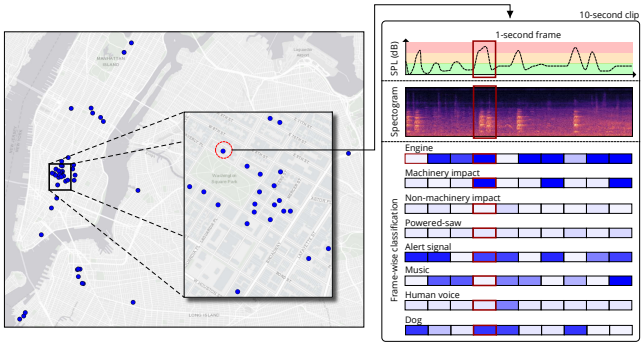

Figure 2: *Spatial distribution of SONYC sensors (left) showing the coverage of the city. Right image illustrates the data from a sensor located near a park in Manhattan. Sensors record both the sound pressure level at each second (SPL dB), as well as the environmental sounds (stored as 10-second clips). For each 1-second frame in the clip (highlighted in red), we compute the classification considering user-crated prototypes. The figure shows classes following standard urban audio taxonomies.*

To create classification models that can recognize human-understandable features in multimedia data sets, previous approaches employ active learning frameworks leveraging the users as oracle annotators to annotate new samples the system identified as the most informative. For example, previous studies investigated the usefulness of active learning for labeling tasks [BHZ*17] and its application in other fields such as anomaly detection [LYC10, LGG*17], commuting flow estimation [YWL*15], and image categorization [ZWVW20]. These approaches often guide the user on choosing the next subsets of the data to label next to improve the performance of the model, which, in our case, can limit the user in applying their previous understanding of the soundscape to label the concepts [NKLM20]. Our proposal leverages a set of techniques presented in previous works to enable users to better understand the spatiotemporal distribution of events in acoustic recordings while accounting for their knowledge of the soundscape to build concept-based classification models to gain insights into the dynamics of the urban environment.

4. Sounds of New York City

The research reported in this paper was undertaken in conjunction with audio and machine listening experts from the SONYC project [BSN*19], and utilizes data generated by the project's sensors. Our collaborators have background in urban science and machine listening [DTZM*18, CCSB19, MHL*20, WBS*21]. In addition, the project communicates their findings to the media [BB20] and works closely with the NYC Dept. of Environmental Protection to understand their needs and investigate new ways of monitoring and mitigating noise pollution.

4.1. A data set of urban sounds

The SONYC acoustic sensor network consists of more than 50 sensors deployed around three boroughs of NYC: Manhattan, Brooklyn, and Queens. Figure 2 shows the spatial distribution of the sensors. These sensors are positioned 15 to 25 feet above the ground. To maintain the privacy of bystanders and prevent the recording of intelligible conversations, the sensors do not record continuously, but rather record 3 10-second recordings at random intervals within each minute of a day (i.e., for a single day, each sensor will record 720 minutes worth of 10-second audio recordings uniformly distributed throughout the 1440 minutes of the day). As of 2021, SONYC has collected approximately 1,700,000 hours of SPL data (stored as second or millisecond resolution timeseries), and 877,000 hours of recorded audio. To extract a discriminative, lower-dimensional representation of each 10-second recording, we employ OpenL3 trained on an environmental sound subset of AudioSet [CWSB19]. OpenL3 is an open-source library for computing deep audio embeddings, developed by researchers from SONYC, and its design choices were informed by the need to classify sounds from urban environments. For each 10-second recording, we use a hop size and window size of 1.0 second (with centered windows) to compute 10 512-dimensional feature vectors. This produces a feature vector that coarsely captures the *general* acoustic aspects of the environment.

4.2. Challenges

The complexity of the urban environment brings several challenges when it comes to analyzing and extracting insights from urban sound data, especially considering such a large data set as the one captured by SONYC.

Sound representation. In complex environments such as cities, many sound classes seem quite similar, such as car alarms and sirens, but are distinct in the noise code and should be treated as such. On top of that, when handling sounds from cities, the acoustic environment changes by location and by time within seasonal cycles. As a self-supervised method, OpenL3 does not need human-generated labels to be trained, while still providing good sensitivity to different urban sounds. However, it falls short of properly accounting for *all* of the complexity of the soundscape of a city.

Mixture of sounds. Unlike images, where visual objects are opaque, sound objects are conceptually *transparent*, meaning that multiple objects (sound sources) can have energy at the same frequency [Wys17]. This is especially true in an environment as complex as cities, where sounds are emitted from multiple sources, creating a soundscape that, albeit quite characteristic, is very difficult to parse and understand. In other words, in a city, at any given instant in time, a sound recording might have a mixture of background (e.g., bird songs, dog barks) and foreground sounds (e.g., engine, party, sirens).

Sound exploration. Again unlike images, there is no clear pictorial representation of audio data. This gap between audio data and visual representation is challenging when building visual analytics systems. Visual objects are *opaque* (a given pixel in a visual image corresponds to only one object), whereas sound objects are *transparent* (multiple objects can have energy at the same frequency). Sounds are therefore serial objects: when assessing an image, we can visually *scan* it to identify each visual object in the scene, creating a visual map of the objects that can help us fully understand the scene. Sounds only exist at one moment in time; once the moment is gone, the sound is also gone. In other words, a user can only observe a sound one moment at a time, unlike images where we can observe multiple objects at a time. In spectrograms representation, similar neighboring pixels cannot be assumed to belong to the same object (i.e., frequencies are non-locally distributed on the spectrogram [Wys17]). As we can notice, creating visual representations of sounds is a challenge, specifically considering a scenario with multiple sound sources, such as urban soundscapes.

Sound labeling and classification. Although previously proposed classifiers provide a reasonable link between embeddings and human-understandable vocabulary, their class vocabularies are limited, providing a narrow view of the rich and varied soundscape of the city, which is comprised of numerous types of sound events. Furthermore, manually labeling sound data to be used as groundtruth for model training is a laborious process. As previously mentioned, sounds are serial objects where the user needs to listen to one at a time, limiting the number of audio files that can be labelled in a short period of time. Purely automated mechanisms, however, are prone to misclassifications given the complexity of soundscapes.

Data size. Over the past five years, SONYC has generated more than 60 TB of data, including high-resolution SPL timeseries and audio recordings. If we consider the embeddings computed with OpenL3, we have 86,400 feature vectors with size 512 (177 MB in total) *per sensor per day*. Any visualization system must properly handle such data size to be interactive [LH14], either by sampling, filtering or aggregating the data.

5. System Requirements

In our collaboration with machine listening researchers, over the course of two years in the context of the SONYC project, we established a set of requirements for a visual analytics tool to facilitate their analysis workflows. We then validated the working system through interactive demo sessions exploring a number of potential use cases. Underlying our work is the necessity to account for user knowledge when exploring the urban soundscape for different concepts. During these meetings, we identified the following main tasks that the experts desire to perform with the tool: 1) Select and listen to sound recordings from a set of sensors, considering different days of the week and time ranges; 2) Considering a query audio, quickly identify a set of possible similar sounds throughout a long period; 3) Create and refine classification models that allow for searching of complex sound scenes; 4) Assess classification performance interactively. To accomplish the listed tasks, we identified the following system requirements:

[R1] Interactive identification and labeling of similar sounds. Given the highly complex acoustic environment we observe in cities, audio representations cannot encode specific audio events that users might be interested in. Moreover, the high-dimensional nature of audio representations makes it hard to visually analyze such data, making multidimensional projection techniques a stan-

dard in this process. However, in many cases, user-perceived similarities between sets of audio frames (i.e., a one-second slice of the ten-second audio snippet) might not be represented in the selected projection technique, e.g., similar frames are far apart in the projected space (low-dimensional space), making it harder for users to find similar audio frames. Hence, finding similar audio frames based on user's perception is one of the requirements of the Urban Rhapsody framework.

[R2] Projection steering based on user perception. When exploring audio embeddings extracted from urban recordings through multidimensional projections, we often recognize clusters that do not represent the user's perception of the soundscape. Based on the user's understanding of the data set expressed through labeled points, the system should provide the capability of producing new projections that better encode the user's perception.

[R3] Iterative creation of classification models. Considering that current machine listening models present certain limitations, the system should provide the capability to iteratively create new classification models based on the data points labeled by the user (and, therefore, the user's perception of the soundscape). The system should also support assessing the evolution of the model's convergence through successive iterations.

[R4] Local and global sound perspectives. Audio embeddings might possess certain characteristics that only become clear when analyzed locally or globally. Then, it is important for the user to assess their local characteristics and to relate one sound to its immediate neighborhood or distant clusters.

[R5] Match between audio and visual representations. Visualizing audio files in the frequency domain is important for the user when assessing the accuracy of both the embeddings and classifications. For instance, two sounds might have very similar spectro-temporal patterns and classifications but completely different embeddings; it is important, therefore, to further assess and create hypotheses on what led to these different outputs.

[R6] Support interactive query times. The system should support interactive queries to enable the easy and quick labeling of data points and the creation of classification models.

6. Urban Rhapsody

To satisfy the previous requirements, we introduce Urban Rhapsody. A visual analytics tool able to provide a human-centered exploration of the urban soundscape using prototypes created on the fly through different interaction mechanisms. Our description of the framework is broadly divided into three parts. First, we describe our approach to generate classification models (or *prototypes*) of different *concepts* denoting complex urban sound scenes. Second, we describe the different components of Urban Rhapsody's visual interface (also see accompanying video), followed by a discussion of its architecture and implementation.

6.1. Prototype-based interaction

In Urban Rhapsody, we would like to support the search for audio events based on concepts and not only based on a single audio event. Here, we use the term *concept* to refer to an abstract idea or a general representation of a category in mind, such as "crowded street", which can be perceived differently by people. In one of our case studies, we describe a case where the user keeps refining their concept of construction while annotating new sounds that together compose the full picture of a construction. To allow for this kind of search, we define *prototypes*, a structure composed of a classification model and a set of representative audio frames that defines a user's understanding of a concept.

The classification model learns how to distinguish between the audio frames that are part of a given concept according to the user's perception represented by annotations made during the interaction process with the system. Once the user starts labeling a specific concept in Urban Rhapsody, they can generate a new classification model that will be trained using annotated frames as input. Since our goal is to find occurrences of specific concepts in our data set, we should train this model with a diverse enough sample of the data so it can generalize well to different scenarios. Given this constraint, we train our model to distinguish between two labels: positive (frame is part of the concept) and negative (not part of the concept). For positive labels, we use all the frames annotated as the concept we are interested in. For negative labels, we use frames explicitly annotated as not being part of a concept and a random sample of all frames in our data set twice as big as our set of positive-labeled frames. The classifier we train in Urban Rhapsody is based on the classic random forest algorithm using a standard parameter setting for audio classification [WMCB19]. However, any classification model capable of outputting a likelihood score of a data point belonging to a class can be used in Urban Rhapsody. In this version, the likelihood function is calculated as the average prediction score across the trees in the forest. This interaction supports requirement **R3**.

Following **R6**, Urban Rhapsody must be capable of providing interactive query times during the exploration process. However, the size of the data set handled by our framework blocks us from filtering interesting audio frames by scanning the entire data set and computing the prediction probability of a given model to generate our visualizations. For this reason, after every model refinement made by the user, we also calculate a set of representative audio frames that will help us sample the data set to a smaller size before filtering interesting points using the aforementioned classification model. We calculate representative points of a concept by selecting all the points annotated as being part of a concept by the user and running a density-based clustering algorithm on the positive-annotated frames for a concept. For each cluster, we calculate the frame closest to its centroid and add it to the set of representative frames of that concept. The representative audio frames also help the users keep track of the concept they are creating through their interaction with the system. We enable the user to use these representative points as query input for a concept search using an approximated nearest neighbors (ANN) query.

6.2. Visual interface

The visual interface was designed to provide the user with the ability to browse through the entire data set, identify and annotate concepts present in audio samples, and, finally, iteratively and interac-

Figure 3: *The Urban Rhapsody system visual interface: (a) Calendar View; (b) Sensor Map and Distribution View; (c) Day View (projections); (d) Focused View (spectrograms); (e) Frame Classification View; (f) Model Summary; (g) Mixture Explorer.*

tively build prototype models that generalize these concepts over the entire data set. Figure 3 shows the different components of the visual interface. Next, we discuss the design of each visualization based on its functionality: provide easy navigation through the audio collection, enable the annotation of audio concepts, allow for the detailed inspection of individual audio samples and facilitate the evaluation of prototype models.

Audio collection navigation. The interface implements several strategies to enable navigating through our data set. The first is the Calendar View (Figure 3(a)). This component presents a calendar of the year with each cell representing a single day. Within each cell, we can visualize a bar chart representing the distribution of frames of a specific concept during the four time slices of a day, allowing for the fast identification of the daily distribution of sounds. The bars of each cell are also colored according to the density of a specific concept in a day (more examples in a day will lead to darker blue bars). If a Calendar View cell is clicked, all the data corresponding to that specific day is loaded and in the day view (Figure 3(c)). Using the Day View, we can visualize the audio frames through the analysis of scatterplots generated by projecting high-dimensional feature vectors (audio embeddings) into a two-dimensional space using UMAP [MHM18]. Although UMAP was the projection technique used for this version of Urban Rhapsody, given its dimensionality reduction capabilities, it is important to notice that Urban Rhapsody is agnostic of projection technique. The adaptation of the system to better accommodate experts' needs in terms of projection techniques is trivial. Here, the users can horizontally stack projections in three ways: reprojecting a subset of the data available for a day (i.e., reproject specific clusters to capture local structures of the data), removing a subset of the data, and reprojecting the remaining points (useful for removing clusters representing sensor failure, for example), and, lastly, steer the projections based on frames annotated by the user using a semi-supervised dimensionality reduction algorithm [SCMD19] that can learn a new low-dimensional space that better encodes the user's perception of the data (i.e., bringing frames with the same labels closer while keeping the different ones distant from each other), therefore supporting **R2**.

The projections in the Day View are linked and allow for selecting points through a bounding box or periods of the day. Selections update the Distribution View as well as the components designed for the individual inspection of audio samples, the Focused and the Frame Classification Views, shown in Figure 3(d, e) are described later in this section. At last, the projected points, each representing an audio frame, can be colored according to a likelihood of belonging to a concept or user annotation. When a day is loaded, Urban Rhapsody automatically calculates a hierarchical clustering of the points and updates the Mixture Explorer, represented in Figure 3(g) by a tree. Each node of the tree represents a cluster found by the algorithm. Each node is subdivided into subnodes, each being one concept that the user previously created. In the example presented in Figure 3(g) each subnode is representing a concept (people talking, birds, and siren from left to right) and is colored based on the average likelihood of the correspondent cluster contain the specific concept (darker green for higher likelihoods). If a node is clicked, the corresponding cluster is selected in the scatterplots and all the components of the interface are updated accordingly. For example, the node where all subnodes are darker green

is where the user is more likely to find frames that contain all created concepts. It is important to notice that hierarchical clustering is a powerful visual strategy that enables the user to explore clusters of different sizes, both locally and globally (**R4**), and gain new insights into sound mixtures by focusing its inspection on cluster where previously created concepts are more likely to be found.

Annotation of audio concepts. One of the requirements elicited with domain experts is regarding the ability to annotate specific audio frames **(R1)**. To satisfy this requirement, Urban Rhapsody provides a mechanism to annotate specific audio frames that works as follows: users can select specific frames by using the selection mechanisms provided by the scatterplots or select a cluster using the hierarchical tree. Once a selection is made, the users can click on the labeling icon on top of the scatterplot to open a dialog that will allow for the annotation of these frames with as many labels as they want (positive labels). Also, users are able to annotate frames with negative labels, to explicitly say that a selection of frames is not part of a specific concept. This will help refine the prototype models when we find false positives during the exploration process.

Inspection of audio samples. To inspect details of an audio recording selected by the user during the exploration of the projections, Urban Rhapsody contains two widgets with visualization metaphors commonly used by audio experts: the Focused View (Figure 3(d)) and the Frame Classification View (Figure 3(e)). The Focused View shows a spectrogram of the audio samples selected in the projection. A spectrogram is a visual representation of the magnitude of the short-time Fourier transform, which describes the signal's energy by frequency as it varies with time. It can be visually encoded in a heatmatrix where each cell represents the intensity of a frequency in a given time. For example, the spectrogram of an audio file containing the sound of a siren contains wave patterns. Previous work investigated the usefulness of spectrograms in representing audio classes for humans and its performance in comparison to others standard audio visualizations [CSS*17]. We use this representation to allow the user to compare different sounds without having to listen to multiple audio files. The Frame Classification View displays the likelihood of observing a concept in the audio sample. In this way, the color of each cell of the matrix represents the probability of observing different sound classes in the associated audio frame. Finally, Urban Rhapsody allows the user to click on the spectrogram to listen to the recording. This interaction is important to bridge the gap between the visual representation and the actual audio (**R5**).

Evaluation of prototypes. As users keep creating and refining prototypes, they can evaluate its performance by making use of several components of our interface. First, for any given selection on the scatterplots, they can check a histogram showing the distribution of a concept's likelihood across the selected points. If the histogram is shifted to the right, it means the selection has a higher chance of belonging to a concept. Besides that, the users can assess the robustness of models in the Model Summary View (Figure 3(f)) where we present the evolution of the prototypes over the course of several refinements. Once we create a new version of a labeled subset for a specific concept, we train a new classification model to be part of the prototype and evaluate old versions of the prototype's classification model to assess the change in prediction over time. At some point, the user can come to a conclusion that labeling more points has no significant impact on the classification model and then stop the process.

6.3. Analysis flow

The exploration process starts with the user querying the data set using any of the three approaches we propose: select a frame from the examples we provide in a Query View as input for the similarity query, upload their own audio snippet and select a frame from this audio snippet, or query using one of the created prototypes. For all three query approaches, the user is able to select the number of frames the query will retrieve. Once the query is processed, the Calendar View is updated, showing the density of a given class, or concept, on each cell throughout the year (color), and its distribution within the day (bar chart). Next, the user can select a specific day and load all the available data for that day to further inspect the day's soundscape using the scatterplots in the Day View. At this point, the user can select specific regions of the scatterplot and listen to the correspondent audio frames, reproject specific regions of the day scatterplot to focus on local structures, remove undesired clusters or steer the scatterplot based on the annotation of frames. Also, color the points by prototype probability or created annotations. These operations will help users in two tasks: assess the performance of the prototypes they are creating and find data points that should be labeled as any concept of interest. Following that, it's possible to create different prototypes and refine existing ones based on new annotations the users are creating, either positive annotations or negative. Meanwhile, when prototypes are created and refined, the Model Summary gets updated, showing the change in prediction probability of the models and the set of representative frames of a given concept. When the user is confident about the prototype they are creating, they can reuse this prototype to query the entire data set and look for specific temporal patterns that a specific concept is happening. This analysis flow denotes the importance of having a user in the loop to evaluate the performance of the prototype models as Urban Rhapsody allows for the creation of concepts that match the user perception of the city's soundscape, which can not be evaluated quantitatively.

6.4. System implementation

We decided to develop Urban Rhapsody following a client-server architecture. We structured our application following microservices guidelines to ensure that we could effortlessly add new features to the tool and scale its deployment to make it available for the general public. The storage component keeps audio recordings and their embeddings located in different folders following the same naming convention for faster localization. Each audio file is also associated with a set of metadata attributes with temporal and spatial information (time of the recording and location of the sensor) that is kept in a separate database. The core of our application is composed of several microservices. The data server is responsible to serve audio files and spectrogram images. The web server provides users with a bundle of our Angular web application. The user server stores annotations on RocksDB [DTZM*18]. The most complex services of our system are the ML server and the ANN

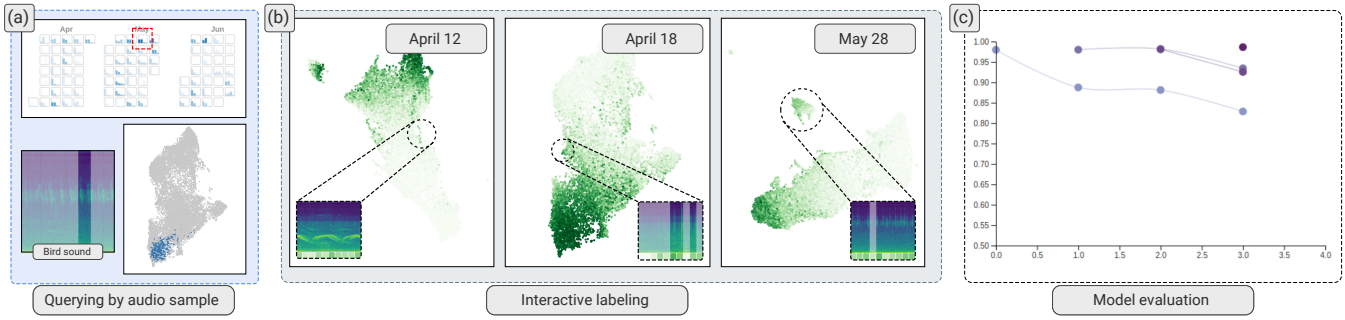

Figure 4: *Interactive monitoring of the training process and refining the model. (a) We run a query using our sample birds' sound and analyze the clusters; (b) Investigating the clusters on different days to detect and re-label false positive and false negative instances, and refine the model; (c) The model evaluation indicates that our prototype models are converging as we do further iterations of refinement.*

server. The first is responsible for all machine-learning-related operations, such as multidimensional projections, hierarchical clustering, and model training. Following **R6**, the operations are processed using GPUs through RAPIDS libraries [Rapa]. CPU-based libraries would not be able to handle such data-intensive operations required by Urban Rhapsody. The ANN server is responsible for computing similarity queries based on the euclidean distance between frames.

7. Case Studies

In this section, we demonstrate the application of Urban Rhapsody through two case studies using data from the SONYC sensors. In doing so, we highlight how the requirements listed in Section 5 are met in different tasks. The first case study explores how Urban Rhapsody can facilitate the interactive labeling and exploration of data for investigating out-of-hours construction noise, a pressing issue facing many large cities. The second one highlights another capability of Urban Rhapsody to facilitate searching for mixture of sounds to explore the impact of anthropogenic noises such as siren on bird songs. These case studies can be of interest to various stakeholders, from the general public and advocacy groups to government agencies, such as the Dept. of Environmental Protection.

7.1. After-hour construction noise

Construction noise is one of the primary sources of noise-related complaints in NYC. As the city grows, new structures are built, old ones get renovated, and economic pressures and deadlines lead developers to request the city for permits allowing them to perform construction outside the regular workday hours (i.e., 8 AM to 5 PM). In the past few years, this has been a major source of dispute between NYC residents and developers [May19], and this problem is increasingly getting worse. In 2018, NYC's Department of Buildings issued around 67,000 after-hour permits, more than double the number of permits issued in 2012. Although developers must follow strict noise guidelines during after-hour constructions, the increase in the number of complaints related to these types of disturbances indicates otherwise. Even though the city constantly issues noise construction fines through manual inspections, the after-hour nature of these noises makes it especially hard to monitor them. This is a significant problem that needs to be addressed by cities and their different departments, with severe political, social, and economic ramifications.

In this study, we use the SONYC network to understand the impact of construction-related noises on the soundscape of NYC. Our first goal is to assess if these noises were captured by our sensors, to facilitate noise code enforcement activities. Secondly, we would like to use examples that we found during our initial exploration to build a prototype capable of pointing us to specific days and times where after-hour construction work might have happened. We start by querying our data set for similar audio snippets using one of the examples provided in the system containing the recording of a powered saw **(R1)**. Using the Calendar View, we can quickly observe a day containing most of the similar audio excerpts according to our ANN model (Figure 1(a, top)). We select that day, and Urban Rhapsody generates a UMAP projection of all the audio frames within that day (Figure 1(a, bottom)). After a quick inspection of the projection scatterplot, we can notice a set of distinctive clusters (highlighted in red). Using the tool's interactions, we start by selecting the one cluster containing most of the points retrieved by the initial similarity query. By listening to a few recordings, we can notice that the points belonging to this cluster are perceptually similar to a powered saw, very common on construction sites **(R2)**. We also notice that most of these audio snippets were recorded around 8 AM, as the hour distribution chart shows us. Figure 1(b,c) highlights the recordings that happened around 8 AM, and it's possible to again see different clusters. After listening to recordings from each cluster, we noticed that each one of them represents different sounds (powered saw, drilling machine, engine). At this point, we can leverage Urban Rhapsody's feature that allows us to create models on the fly and decide whether to include certain sounds in our prototype **(R3)**. Once we label recordings from that specific day, we generate two construction prototypes (with and without large engine noise). We can now use them to guide the exploration through different days of the year. This step allows us to speed up the search for similar sounds, without the need to listen to *hours and hours* of soundscape audio files. Also, during this guided exploration, we can adjust the prototype by labeling more points, either as negative or positive labels, as we assess the model's performance by listening to the recordings. This interactive process is highlighted in Figure 1(b,c), for two different models.

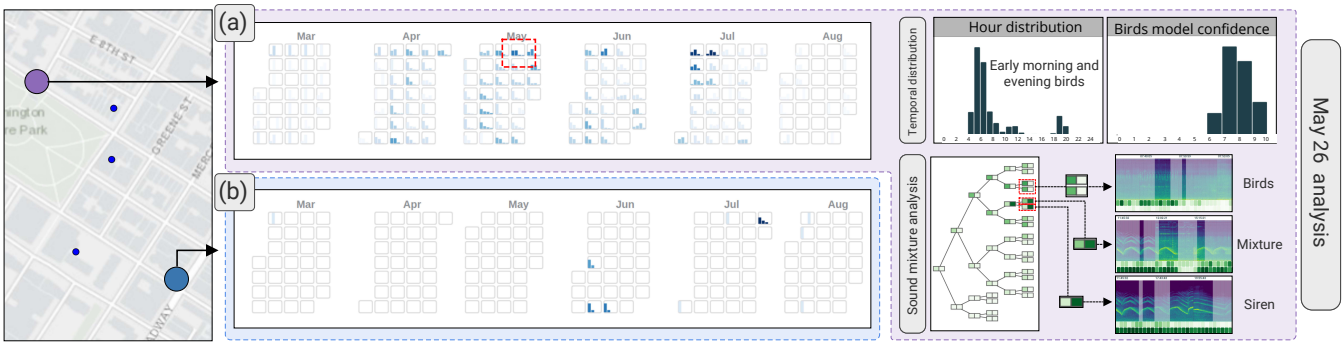

Figure 5: *Looking for bird songs in two different Manhattan locations: (a) Edge of Washington Square Park with high concentration of bird songs and (b) a street corner on Broadway with very few instances of bird songs since we do not have trees for birds to nest.*

After refining our models once, we listened to the representative snippets of our prototypes and used them to look for unusual events. The calendar heatmaps show the results of the prototype queries (Figure 1(d,e)) where we can spot two interesting events. In February, we noticed that during two days, construction work happened during the night (Figure 1(d)). And that, during many days in October, the same engine noise started at 11 PM and lasted for approximately 30 minutes (Figure 1(e)).

To further validate this finding, we used citizen complaints filled through NYC's 311 non-emergency service phone number. Interestingly, there were actually a series of complaints reported on those two specific days of February. The ability to intuitively create prototypes based on audio files listened in the exploratory process sets Urban Rhapsody apart. Findings such as these not only highlight the usefulness of a *passive* network of sensors (as opposed to *active* sensors deployed in inspection visits), but also the usefulness of distinguishing different noises emitted from construction sites. Previous approaches, like Noise Profiler [MLD*18], focus on the SPL characteristics, a useful but crude measurement of noise. By enabling the exploration of specific types of noise, Urban Rhapsody can 1) provide a clearer picture of the soundscape near a construction site, 2) facilitate monitoring tasks carried out by enforcement agencies, and 3) validate the accuracy of 311 complaints.

7.2. Birds in New York City

The impacts of urban noise, air pollution, and the built environment on residents and migrating birds have been extensively studied [SL15]. There is a strand of research that specifically analyze birdsong to discover if exposure to loud urban noise can lead to significant changes in their song traits and the time and frequency of their chorus, specifically since birds use different sounds to communicate, mate, and defend breeding territories and rely on the vocal communication to sustain their lives [MCRP11, Sla13]. One of the main challenges in the majority of bioacoustics and avian behavior studies is the costly and time-consuming nature of working with audio data, which limits the duration and geographical extent of the research. The application of machine learning in bird song classification is not new [MC97], but most of the developed models are trained using specific sets of data, limiting the user to a predefined set of labels, with no control over what the model perceives as bird songs. This is specifically important in bird song studies since the model can classify some sounds, such as whistling, as bird sounds and discard some bird songs which are very different from what it was trained on [XZ19].

In this case study, we demonstrate how Urban Rhapsody can facilitate such studies by providing a robust and easy to use solution where the user can search for specific sounds among hundreds of hours of recordings, refine the results if needed to reach the confidence level of interest, monitor the frequency and changes in the song traits, and investigate the impact of anthropogenic noises on birds. Sitting on the Atlantic Flyway, NYC offers great resting grounds for birds traveling along the north-south migratory route in the Americas [DR15]. We choose Washington Square Park, a popular local park situated in a dense and busy neighborhood of the Manhattan borough, with the natural environment for birds to nest as well as the attributes representing a crowded and noisy urban environment [Was21].

The first step is to build our bird representation model. We start our exploration by using one of the bird song examples provided in the query view. Next, we select a day with high density of similar bird sounds. As shown in Figure 4(a), we generate a UMAP projection of our selected day on the Day View and see that majority of the bird songs are clustered on the bottom region of the projection (blue points). Next, we create our first representation model of bird songs to speed up our search across different days (**R1**). We can find false positives and false negative examples throughout this process, fix those and refine our prototype. For instance, we found out that on April 18th, the model assigned a high likelihood to a small cluster of points (Figure 4(b)). We investigate this

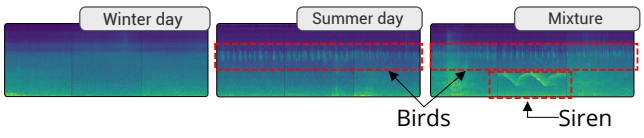

Figure 6: *Spectrogram showing a winter day with no bird song, a summer day with birds' singing and the selected day in summer when birds dawn chorus continued despite loud siren.*

cluster closely and realize they are not bird songs, so we re-label these points, refine our model, make a new prediction with updated weights, and run this process iteratively until the model reaches a robust state **(R2)**. In the Model Summary View (Figure 4(c)), we can see that our new prototypes are converging: Our first model had the worst performance, and as we continued refining, the difference between the prediction probabilities of the labeled birds' data set get smaller after each iteration **(R3)**.

Using our refined model, we run a new query to explore the distribution and patterns of bird songs near Washington Square Park over the course of one year. The retrieved results clearly show two levels of seasonal patterns: a daily pattern with peaks in the mornings and afternoons corresponding to the dawn and dusk chorus times, and another pattern with peaks during spring to early summer, when songbirds usually migrate, as illustrated in Figure 5(a). This signifies the robust performance of the model in classifying birds. We also look at the corner of Broadway and Waverly Pl., where we have no trees on both sides of the street, to see if we can find similar patterns there. As Figure 5(b) shows, we have very few instances of bird songs in that location throughout 2017.

One useful aspect of Urban Rhapsody is the ability to analyze sound mixtures. To investigate how the siren sound can impact or even halt the birds' chorus, we use Urban Rhapsody to query for dawn chorus times (6-11 AM) where siren was also present. This allows us to discover whether loud sirens can halt birds' dawn chorus or whether birds in noisy urban areas like Manhattan local parks are adapted to the level of noise [NB10, NPZ*13]. We can use the Mixture Explorer to differentiate between these two sounds, as illustrated by Figure 5(a, bottom right). Notice that nodes containing bird songs, siren, or mixture of both are clearly distinguishable with our visual encodings **(R4)**. Drilling down to this specific example (Figure 6), we can see that the birds continue singing despite the loud siren **(R5)**. This analysis can create a ground for further research by bioacousticians and researchers in this field to investigate whether this pattern is more prevalent in birds of specific species or whether we can find incidents of ambient noise halting birds singing. Urban Rhapsody helped us to iteratively refine our model, track the sounds of interest and search for a combination of sounds across a large data set, detect the pattern and drill down to the exact moments to listen and investigate more.

8. Discussion and Conclusion

We have presented Urban Rhapsody, a novel interactive system for seamlessly exploring large audio data sets, based on user-generated concepts. Leveraging machine learning techniques, Urban Rhapsody supports labeling and analysis at scale, while our multilevel visualization approach enables the inspection of temporal patterns at varying levels of granularity. By enabling users to interactively label data based on their knowledge, Urban Rhapsody can be used to augment self-supervised methods that might not account for audio complexity. We illustrate its potential through data collected by the SONYC project. However, Urban Rhapsody can be applied to other longitudinal spatiotemporal acoustic data (e.g., bioacustics [MMW, FKL*21]), and to support this we made the tool available on GitHub. We hope this will encourage researchers to use it in many different contexts and further develop the code base.

Limitations. While we define interactivity based on benchmarks for querying large data [BEA*20], we also identify three potential bottlenecks: similarity search, model training, and projection generation. Urban Rhapsody responds to similarity queries by returning up to 10,000 points in less than one second (for the examples provided as initial query seeds [Fai]). However, a one-time preprocessing computation is required to generate indices. This takes on average one hour per sensor/year and needs 9 GB of memory space (for sensors with low rates of missing data). GPU implementations [Rapa, Rapb] achieve response times of under one second when loading Day View selections and for inference of created concept models. Deploying Urban Rhapsody to handle data from alternate sensor networks requires sufficient memory space to handle query indices, GPU capabilities to train models, and connectivity to support client-server architectures.

Expert feedback. Analyzing large collections of audio data is a challenging task, in which views into the data can be limited. The number of classes classifiers detect may be small, not matched to the task at hand, or too coarse-grained. Deep audio embeddings help to distill the semantics of audio to a smaller number of dimensions, but they are still very opaque and not easily interpretable. In addition, translation between modalities (e.g., using visual tools to explore audio data) is also highly challenging, and yet we know that it can be very effective. Our collaborators highlighted that Urban Rhapsody helps overcome these challenges by enabling interactive exploration, labeling, clustering, and reprojection of collections of audio data; and supports insights into models, labeled data, and previously unseen patterns within unlabeled data.

Future work. We plan to investigate whether Urban Rhapsody can accurately and efficiently represent concepts matching the user's mental model of their data. To investigate this we plan to conduct a large-scale user study with machine learning and audio researchers. While previous research [CSS*17] shows that spectrogram visualizations lead to high annotation accuracy at low time and labor costs, further investigation is also needed to explore additional visualization metaphors (e.g., to summarize longer periods of audio recordings). We will also explore how the analyses supported by systems such as Urban Rhapsody can useful to public officials and community representatives.

Conclusion. Urban Rhapsody is an interactive visual analytics tool for gaining insight into large collections of audio data, which we have demonstrated through use cases that characterize the acoustic environment of NYC. We believe that Urban Rhapsody offers an important step in moving beyond simple metrics, such as SPL, and will be of value to researchers in human-centered machine learning, acoustics, and urban science.

Acknowledgements

We would like to thank our colleagues at CUSP (NYU) for their feedback during the development of this work. This research has been supported by NSF awards CNS-1229185, CCF-1533564, CNS-1544753, CNS-1730396, CNS-1828576, CNS-1626098; CNPq grant 305974/2018-1; FAPERJ grants E-26/202.915/2019, E-26/211.134/2019.

References

[AA08] ANDRIENKO G., ANDRIENKO N.: Spatio-temporal aggregation for visual analysis of movements. In *2008 IEEE Symposium on Visual Analytics Science and Technology* (2008), IEEE, pp. 51–58. 2

[AVT16] AYTAR Y., VONDRICK C., TORRALBA A.: Soundnet: Learning sound representations from unlabeled video. *arXiv preprint ID:1610.09001* (2016). 3

[AZ17] ARANDJELOVIC R., ZISSERMAN A.: Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 609–617. 3

[BB20] BUI Q., BADGER E.: The Coronavirus Quieted City Noise. Listen to What's Left. *The New York Times* (May 2020). URL: https://www.nytimes.com/interactive/2020/05/22/upshot/coronavirus-quiet-city-noise.html. 3

[BEA*20] BATTLE L., EICHMANN P., ANGELINI M., CATARCI T., SANTUCCI G., ZHENG Y., BINNIG C., FEKETE J.-D., MORITZ D.: Database benchmarking for supporting real-time interactive querying of large data. In *Proceedings of the 2020 International Conference on Management of Data* (2020), SIGMOD '20, ACM, pp. 1571–1587. 10

[BH10] BRONZAFT A. L., HAGLER L.: Noise: The invisible pollutant that cannot be ignored. In *Emerging Environmental Technologies, Volume II*. Springer, 2010, pp. 75–96. 2

[BHZ*17] BERNARD J., HUTTER M., ZEPPELZAUER M., FELLNER D., SEDLMAIR M.: Comparing visual-interactive labeling with active learning: An experimental study. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 298–308. 3

[Bro07] BRONZAFT A.: Neighborhood noise and its consequences. *Survey Research Unit, School of Public Affairs, Baruch College, New York* (2007). 2

[Bro10] BROWN A. L.: Soundscapes and environmental noise management. *Noise Control Engineering Journal 58*, 5 (2010), 493–500. 2

[Bro12] BROWN A. L.: A review of progress in soundscapes and an approach to soundscape planning. *International Journal of Acoustics and Vibration 17*, 2 (2012), 73–81. 2

[BSN*19] BELLO J. P., SILVA C., NOV O., DUBOIS R. L., ARORA A., SALAMON J., MYDLARZ C., DORAISWAMY H.: Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM 62*, 2 (2019), 68–77. 2, 3

[CCSB19] CARTWRIGHT M., CRAMER J., SALAMON J., BELLO J. P.: TriCycle: Audio representation learning from sensor network data using self-supervision. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2019), IEEE, pp. 278–282. 3

[CDDF16] CHIRIGATI F., DORAISWAMY H., DAMOULAS T., FREIRE J.: Data polygamy: the many-many relationships among urban spatio-temporal data sets. In *Procedings of the 2016 International Conference on Management of Data* (2016), pp. 1011–1025. 2

[CSS*17] CARTWRIGHT M., SEALS A., SALAMON J., WILLIAMS A., MIKLOSKA S., MACCONNELL D., LAW E., BELLO J. P., NOV O.: Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction 1*, CSCW (2017), 1–21. 7, 10

[CWSB19] CRAMER J., WU H.-H., SALAMON J., BELLO J. P.: Look, listen, and learn more: Design choices for deep audio embeddings. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), IEEE, pp. 3852–3856. 3, 4

[DAB*13] DAVIES W. J., ADAMS M. D., BRUCE N. S., CAIN R., CARLYLE A., CUSACK P., HALL D. A., HUME K. I., IRWIN A., JENNINGS P.: Perception of soundscapes: An interdisciplinary approach. *Applied acoustics 74*, 2 (2013), 224–231. 2

[DBC*17] DEMA T., BRERETON M., CAPPADONNA J. L., ROE P., TRUSKINGER A., ZHANG J.: Collaborative exploration and sensemaking of big environmental sound data. *Computer Supported Cooperative Work 26*, 4–6 (2017), 693–731. 2

[DFL*18] DORAISWAMY H., FREIRE J., LAGE M., MIRANDA F., SILVA C.: Spatio-temporal urban data analysis: A visual analytics perspective. *IEEE Computer Graphics and Applications 38*, 5 (2018), 26–35. 2

[dPVCR15] DE PAIVA VIANNA K. M., CARDOSO M. R. A., RODRIGUES R. M. C.: Noise pollution and annoyance: An urban soundscapes study. *Noise & Health 17*, 76 (2015), 125. 2

[DR15] DAY L., RIEPE D.: *Field Guide to the Neighborhood Birds of New York City*. JHU Press, 2015. 9

[DTZM*18] DORAISWAMY H., TZIRITA ZACHARATOU E., MIRANDA F., LAGE M., AILAMAKI A., SILVA C. T., FREIRE J.: Interactive visual exploration of spatio-temporal urban data sets using urbane. In *Proceedings of the 2018 International Conference on Management of Data* (2018), SIGMOD '18, ACM, pp. 1693–1696. 2, 3, 7

[DWX*21] DENG Z., WENG D., XIE X., BAO J., ZHENG Y., XU M., CHEN W., WU Y.: Compass: Towards better causal analysis of urban time series. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2021), 1051–1061. 2

[DZD*10] DRATVA J., ZEMP E., DIETRICH D. F., BRIDEVAUX P.-O., ROCHAT T., SCHINDLER C., GERBASE M. W.: Impact of road traffic noise annoyance on health-related quality of life: Results from a population-based study. *Quality of Life Research 19*, 1 (2010), 37–46. 2

[Fai] Faiss. URL: https://faiss.ai/. 10

[FFP*20] FONSECA E., FAVORY X., PONS J., FONT F., SERRA X.: FSD50k: an open dataset of human-labeled sound events. *arXiv preprint ID:2010.00475* (2020). 3

[FKL*21] FARNSWORTH A., KELLING S., LOSTANLEN V., SALAMON J., CRAMER A., BELLO J. P.: BirdVox-296h: a large-scale dataset for detection and classification of flight calls, Dec. 2021. 10

[FLD*15] FERREIRA N., LAGE M., DORAISWAMY H., VO H., WILSON L., WERNER H., PARK M., SILVA C.: Urbane: A 3D framework to support data driven decision making in urban development. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2015), IEEE, pp. 97–104. 2

[FPV*13] FERREIRA N., POCO J., VO H. T., FREIRE J., SILVA C. T.: Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2149–2158. 2

[GCA06] GUITE H. F., CLARK C., ACKRILL G.: The impact of the physical and urban environment on mental well-being. *Public Health 120*, 12 (2006), 1117–1126. 2

[GCKT21] GROLLMISCH S., CANO E., KEHLING C., TAENZER M.: Analyzing the Potential of Pre-Trained Embeddings for Audio Classification Tasks. In *2020 28th European Signal Processing Conference (EUSIPCO)* (2021), IEEE, pp. 790–794. 3

[GEF*17] GEMMEKE J. F., ELLIS D. P., FREEDMAN D., JANSEN A., LAWRENCE W., MOORE R. C., PLAKAL M., RITTER M.: Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), IEEE, pp. 776–780. 3

[Gua03] GUASTAVINO C.: *Etude sémantique et acoustique de la perception des basses fréquences dans l'environnement sonore urbain*. PhD Thesis, Paris 6, 2003. 2

[HCE*17] HERSHEY S., CHAUDHURI S., ELLIS D. P., GEMMEKE J. F., JANSEN A., MOORE R. C., PLAKAL M., PLATT D., SAUROUS R. A., SEYBOLD B., OTHERS: CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), IEEE, pp. 131–135. 3

[HDVT*08] HARALABIDIS A. S., DIMAKOPOULOU K., VIGNA-TAGLIANTI F., GIAMPAOLO M., BORGINI A., DUDLEY M.-L., PERSHAGEN G., BLUHM G., HOUTHUIJS D., BABISCH W.: Acute effects of night-time noise exposure on blood pressure in populations living near airports. *European Heart Journal 29*, 5 (2008), 658–664. 2

[HSN14] HAMMER M. S., SWINBURN T. K., NEITZEL R. L.: Environmental noise pollution in the United States: developing an effective public health response. *Environmental Health Perspectives 122*, 2 (2014), 115–119. 2

[IYT*14] ITOH M., YOKOYAMA D., TOYODA M., TOMITA Y., KAWAMURA S., KITSUREGAWA M.: Visual fusion of mega-city big data: an application to traffic and tweets data analysis of metro passengers. In *2014 IEEE International Conference on Big Data (Big Data)* (2014), IEEE, pp. 431–440. 2

[JCC*11] JOIA P., COIMBRA D., CUMINATO J. A., PAULOVICH F. V., NONATO L. G.: Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2563–2571. 3

[JPP*18] JANSEN A., PLAKAL M., PANDYA R., ELLIS D. P., HERSHEY S., LIU J., MOORE R. C., SAUROUS R. A.: Unsupervised learning of semantic audio representations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), IEEE, pp. 126–130. 3

[JVHKL01] JOHNSTON K., VER HOEF J. M., KRIVORUCHKO K., LUCAS N.: *Using ArcGIS geostatistical analyst*, vol. 380. Esri Redlands, 2001. 2

[KKF18] KUMAR A., KHADKEVICH M., FÜGEN C.: Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), IEEE, pp. 326–330. 3

[LGG*17] LIN H., GAO S., GOTZ D., DU F., HE J., CAO N.: Rclens: Interactive rare category exploration and identification. *IEEE Transactions on Visualization and Computer Graphics 24*, 7 (2017), 2223–2237. 3

[LGTR15] LENORMAND M., GONÇALVES B., TUGORES A., RAMASCO J. J.: Human diffusion and city influence. *Journal of The Royal Society Interface 12*, 109 (2015), 20150473. 2

[LH14] LIU Z., HEER J.: The effects of interactive latency on exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2122–2131. 4

[LJLH19] LIU Y., JUN E., LI Q., HEER J.: Latent space cartography: Visual analysis of vector space embeddings. *Computer Graphics Forum 38*, 3 (2019), 67–78. 3

[LYC10] LIAO Z., YU Y., CHEN B.: Anomaly detection in gps data based on visual analytics. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), IEEE, pp. 51–58. 3

[May19] MAYS J. C.: Why Construction Noise Is Keeping You Up at 3 A.M. *The New York Times* (Sept. 2019). URL: https://www.nytimes.com/2019/09/27/nyregion/noise-construction-sleep-nyc.html. 8

[MC97] MCILRAITH A., CARD H.: Bird song identification using artificial neural networks and statistical analysis. In *CCECE'97. Canadian Conference on Electrical and Computer Engineering. Engineering Innovation: Voyage of Discovery. Conference Proceedings* (1997), vol. 1, IEEE, pp. 63–66. 9

[MCRP11] MENDES S., COLINO-RABANAL V. J., PERIS S. J.: Bird song variations along an urban gradient: The case of the european blackbird (turdus merula). *Landscape and Urban Planning 99*, 1 (2011), 51–57. 9

[MDL*17] MIRANDA F., DORAISWAMY H., LAGE M., ZHAO K., GONÇALVES B., WILSON L., HSIEH M., SILVA C. T.: Urban Pulse: Capturing the rhythm of cities. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 791–800. 2

[MDL*19] MIRANDA F., DORAISWAMY H., LAGE M., WILSON L., HSIEH M., SILVA C. T.: Shadow Accrual Maps: Efficient accumulation of city-scale shadows over time. *IEEE Transactions on Visualization and Computer Graphics 25*, 3 (2019), 1559–1574. 2

[MHL*20] MIRANDA F., HOSSEINI M., LAGE M., DORAISWAMY H., DOVE G., SILVA C. T.: Urban Mosaic: Visual exploration of streetscapes using large-scale image data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), CHI '20, ACM, p. 1–15. 3

[MHM18] MCINNES L., HEALY J., MELVILLE J.: Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint ID:1802.03426* (2018). 3, 6

[MLD*18] MIRANDA F., LAGE M., DORAISWAMY H., MYDLARZ C., SALAMON J., LOCKERMAN Y., FREIRE J., SILVA C. T.: Time Lattice: A data structure for the interactive visual analysis of large time series. *Computer Graphics Forum 37*, 3 (2018), 23–35. 2, 3, 9

[MME*12] MALIK A., MACIEJEWSKI R., ELMQVIST N., JANG Y., EBERT D. S., HUANG W.: A correlative analysis process in a visual analytics environment. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 33–42. 2

[MMW] MILLER B. S., MILNES M., WHITESIDE S.: Long-term underwater acoustic recordings 2013-2019. URL: https://researchdata.edu.au/long-term-underwater-2013-2019/967510. 10

[Muz02] MUZET A.: The need for a specific noise measurement for population exposed to aircraft noise during night-time. *Noise and Health 4*, 15 (2002), 61. 2

[NB10] NEMETH E., BRUMM H.: Birds and anthropogenic noise: are urban songs adaptive? *The American Naturalist 176*, 4 (2010), 465–475. 10

[NGM*12] NEITZEL R. L., GERSHON R. R., MCALEXANDER T. P., MAGDA L. A., PEARSON J. M.: Exposures to transit and other sources of noise among New York City residents. *Environmental science & technology 46*, 1 (2012), 500–508. 2

[NKLM20] NADJ M., KNAEBLE M., LI M. X., MAEDCHE A.: Power to the oracle? Design principles for interactive labeling systems in machine learning. *KI-Künstliche Intelligenz 34*, 2 (2020), 131–142. 3

[NPZ*13] NEMETH E., PIERETTI N., ZOLLINGER S. A., GEBERZAHN N., PARTECKE J., MIRANDA A. C., BRUMM H.: Bird song and anthropogenic noise: vocal constraints may explain why birds sing higher-frequency songs in cities. *Proceedings of the Royal Society B: Biological Sciences 280*, 1754 (2013), 20122798. 10

[NSL*12] NOULAS A., SCELLATO S., LAMBIOTTE R., PONTIL M., MASCOLO C.: A tale of many cities: universal patterns in human urban mobility. *PloS one 7*, 5 (2012), e37027. 2

[Org11] ORGANIZATION W. H.: *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*. World Health Organization. Regional Office for Europe, 2011. 2

[OSS*16] ORTNER T., SORGER J., STEINLECHNER H., HESINA G., PIRINGER H., GRÖLLER E.: Vis-a-ware: Integrating spatial and non-spatial visualization for visibility-aware urban planning. *IEEE Transactions on Visualization and Computer Graphics 23*, 2 (2016), 1139–1151. 3

[PDA09] PAYNE S. R., DAVIES W. J., ADAMS M. D.: *Research into the practical and policy applications of soundscape concepts and techniques in urban areas*. Tech. rep., University of Salford, 2009. 2

[QS14] QUERCIA D., SAEZ D.: Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing 13*, 2 (2014), 30–36. 2

[Rapa] RAPIDS. URL: https://rapids.ai/start.html. 8, 10

[Rapb] RAPIDS Benchmark. URL: https://www.alcf.anl.gov/sites/default/files/2021-03/NVIDIA_RAPIDS_ANL.pdf. 10

[RD05] RAIMBAULT M., DUBOIS D.: Urban soundscapes: Experiences and knowledge. *Cities 22*, 5 (2005), 339–350. 2

[RLB03] RAIMBAULT M., LAVANDIER C., BÉRENGIER M.: Ambient sound assessment of urban environments: field studies in two French cities. *Applied Acoustics 64*, 12 (2003), 1241–1256. 2

[SCMD19] SZUBERT B., COLE J. E., MONACO C., DROZDOV I.: Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific reports 9*, 1 (2019), 1–10. 6

[SL15] SERESS G., LIKER A.: Habitat urbanization and its effects on birds. *Acta Zoologica Academiae Scientiarum Hungaricae 61*, 4 (2015), 373–408. 9

[Sla13] SLABBEKOORN H.: Songs of the city: noise-dependent spectral plasticity in the acoustic phenotype of urban birds. *Animal Behaviour 85*, 5 (2013), 1089–1099. 9

[STN*16] SMILKOV D., THORAT N., NICHOLSON C., REIF E., VIÉGAS F. B., WATTENBERG M.: Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469* (2016). 3

[tab07] *Technology for a quieter America, National Academy of Engineering*. Tech. rep., Technical report, NAEPR-06-01-A, 2007. 2

[TGdCQR20] TAGLIASACCHI M., GFELLER B., DE CHAUMONT QUITRY F., ROBLEK D.: Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters 27* (2020), 600–604. 3

[VKDSVK14] VAN KEMPEN E., DEVILEE J., SWART W., VAN KAMP I.: Characterizing urban areas with good sound quality: Development of a research protocol. *Noise and Health 16*, 73 (2014), 380. 2

[Was21] WASHINGTON SQUARE PARK ECO PROJECTS: Explore birds, 2021. URL: https://www.wspecoprojects.org/our-projects/explore-birds/. 9

[WBS*21] WANG Y., BRYAN N. J., SALAMON J., CARTWRIGHT M., BELLO J. P.: Who calls the shots? Rethinking few-shot learning for audio. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2021), IEEE, pp. 36–40. 3

[Wil21] WILKINGHOFF K.: On open-set classification with L3-Net embeddings for machine listening applications. In *2020 28th European Signal Processing Conference (EUSIPCO)* (2021), IEEE, pp. 800–804. 3

[WLY*13] WANG Z., LU M., YUAN X., ZHANG J., VAN DE WETERING H.: Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2159–2168. 2

[WMCB19] WANG Y., MENDEZ A. E. M., CARTWRIGHT M., BELLO J. P.: Active learning for efficient audio annotation and classification with a large amount of unlabeled data. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), IEEE, pp. 880–884. 5

[Wys17] WYSE L.: Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint ID:1706.09559* (2017). 4

[XZ19] XIE J., ZHU M.: Handcrafted features and late fusion with deep learning for bird sound classification. *Ecological Informatics 52* (2019), 74–81. 9

[YCBL14] YOSINSKI J., CLUNE J., BENGIO Y., LIPSON H.: How Transferable Are Features in Deep Neural Networks? In *Procedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (2014), NIPS'14, MIT Press, pp. 3320–3328. 3

[YWL*15] YU L., WU W., LI X., LI G., NG W. S., NG S.-K., HUANG Z., ARUNAN A., WATT H. M.: iviztrans: Interactive visual learning for home and work place detection from massive public transportation data. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2015), IEEE, pp. 49–56. 3

[ZFA*14] ZENG W., FU C.-W., ARISONA S. M., ERATH A., QU H.: Visualizing mobility of public transportation system. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1833–1842. 2

[ZLH13] ZHENG Y., LIU F., HSIEH H.-P.: U-air: When urban air quality inference meets big data. In *Procedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), pp. 1436–1444. 2

[ZLW*14] ZHENG Y., LIU T., WANG Y., ZHU Y., LIU Y., CHANG E.: Diagnosing new york city's noises with ubiquitous data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014), pp. 715–725. 2

[ZWC*16] ZHENG Y., WU W., CHEN Y., QU H., NI L. M.: Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data 2*, 3 (2016), 276–296. 2

[ZWVW20] ZAHÁLKA J., WORRING M., VAN WIJK J. J.: Ii-20: Intelligent and pragmatic analytic categorization of image collections. *IEEE Transactions on Visualization and Computer Graphics* (2020). 3

[ZYM*14] ZHANG J., YANLI E., MA J., ZHAO Y., XU B., SUN L., CHEN J., YUAN X.: Visual analysis of public utility service problems in a metropolis. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1843–1852. 2