NYU TANDON SCHOOL OF ENGINEERING

Visual Analytics Tools for the Interactive Exploration and Management of Multimodal Data

Joao Rulff

Background

Joao Rulff

- PhD student at NYU, VIDA Lab
  - Advised by Claudio Silva
- NYU Urban Doctoral Fellow
- Bachelor’s degree from UFF-Brasil
- Worked at SLAC-Stanford and IBM

- Research Interests
  - Large-Scale Visual Analytics
  - Data Management
  - Human-Centered Machine Learning

Outline

- **Urban**
  - Urban Rhapsody: Human-centered exploration of soundscapes
- **Art History**
  - ARIES: Art Image Exploration Space
- **ESports**
  - GGViz: Accelerating Large-Scale Esports Analysis

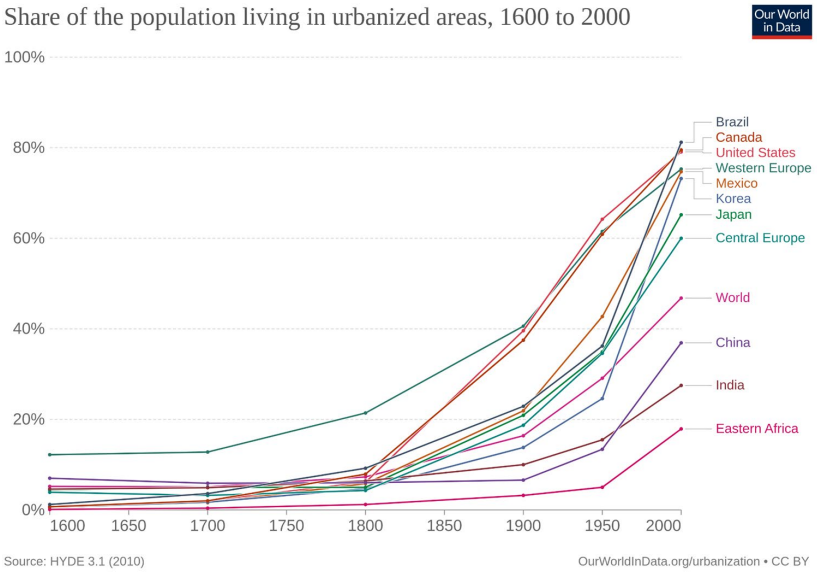
Urban Rhapsody

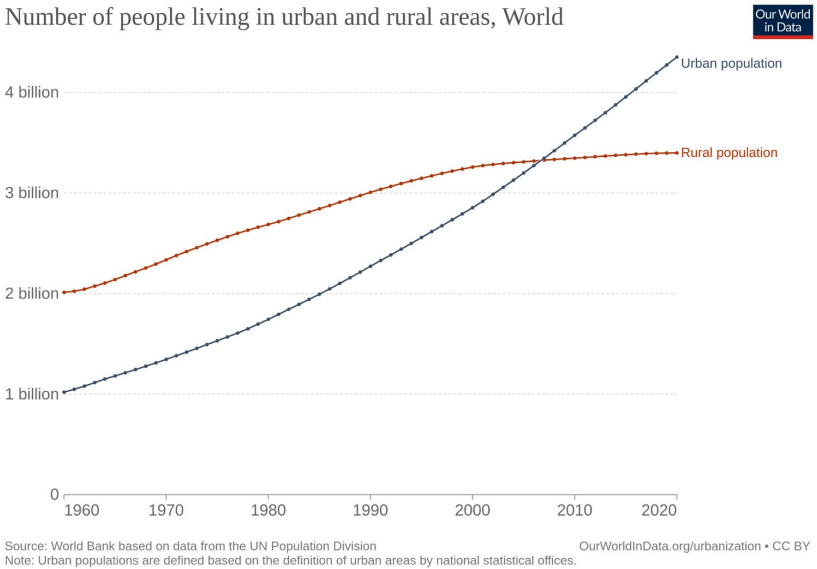
Urban Rhapsody

- Motivation
- Related Work
- SONYC
- Challenges
- Requirements
- Urban Rhapsody
- Case Studies
- Ongoing Work

Motivation

Urbanization





- Some countries, like Brazil and Canada, have more than 80% of its population living in cities;
- The urbanization process is happening across all continents

Urban Sensing

Audio

Sounds of New York City

SONYC Home

Urban Sensing

Audio

Images

Google Street View

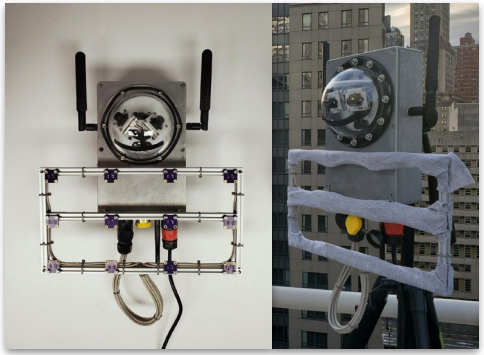
Carmera

Urban Sensing

Audio

Images

Videos



REIP

VisDrone

Sounds of Cities

Sounds of Cities

Soundscape

- The soundscape of a city is composed of **countless layers of sound events** emitted by different sources (traffic, construction, industrial, social activities).
- Importantly, the soundscape also includes the **perception of sounds by those living within it**.
- However, big cities suffer from having a lot of **noise pollution** in the soundscapes

The New York Times https://nyti.ms/2vBlsjh

N.Y. / REGION

New York Becomes the Cit
That Never Shuts Up

By WINNIE HU JULY 19, 2017

Richard T. McIntosh has never heard such a racket outsi

Traffic roars through his neighborhood on the Upp
all hours. The whine of refrigerated grocery trucks by the
And construction of a new apartment tower across the st
own home. There is the deafening rat-a-tat of jack hamm
banging and high-pitched wail of construction equipmen

"I've had two years of absolute violation of my right
Mr. McIntosh, a television producer who has lived on the
than five decades. "I think it's against the Geneva Conver
noise."

New York City has never been kind to human ears, from its screeching subways and honking taxis to wailing police sirens. But even at its loudest, there were always relatively tranquil pockets like the Upper East Side that offered some relief from the day-to-day cacophony of the big city. Those pockets are vanishing. As the city grows more crowded, with a record 8.5 million residents and a forest of new buildings, finding respite from loud cellphone chatter, rooftop parties, backhoes digging foundations, or any other aural assault has become harder and harder.

In other words, New York is really living up to its reputation as the city that never sleeps.

Noise Pollution

- NYC alone, 9 out of 10 adults are exposed to excessive noise levels (i.e., beyond the limit of what the EPA considers safe)

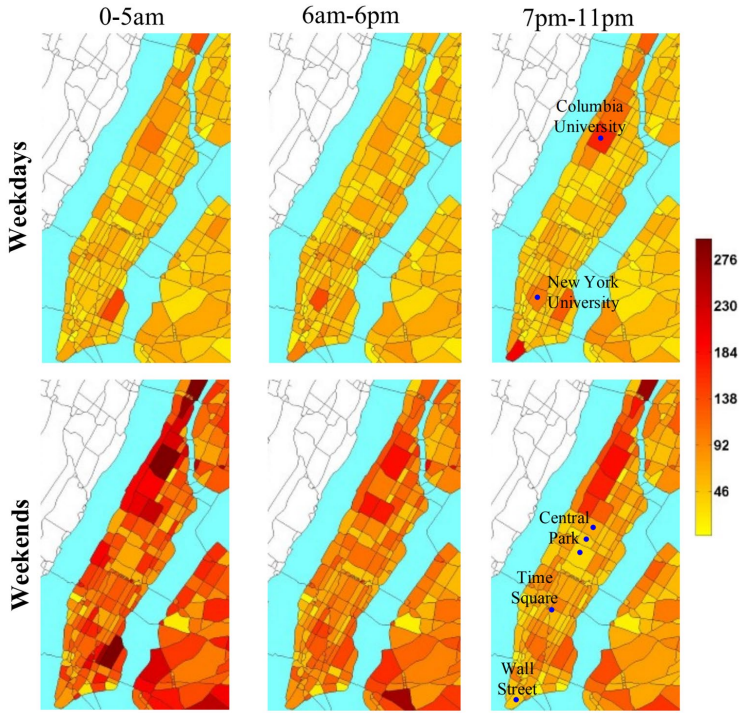
- Noise Pollution can result in several problems such as **sleep loss, stress, hearing loss, reduced productivity, learning impairment.**

- Noise pollution also has an incredibly negative economic impact; according to the World Health Organization, in Western Europe alone more than **1 million healthy life-years are lost annually** to environmental noise pollution.

Related Work

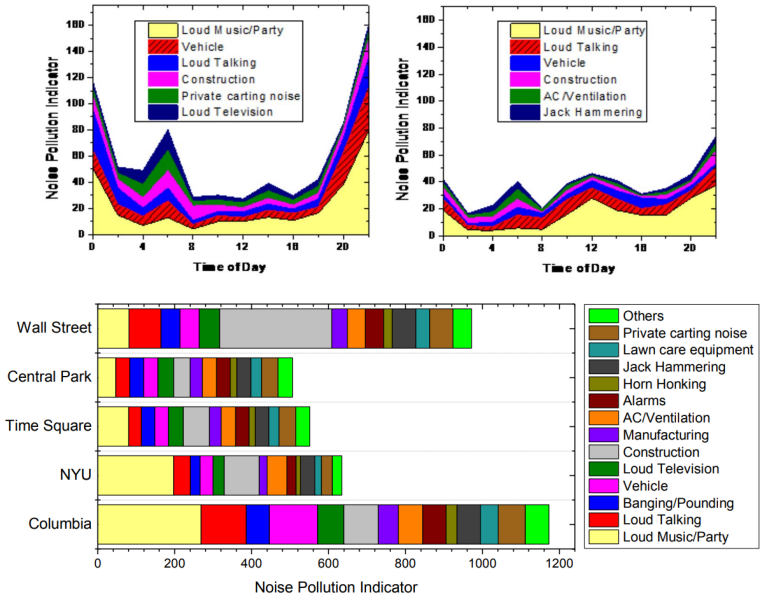
Related Work

- Quantification of noise pollution in New York City.
- Uses **311 data** as proxy for the noise situation in NYC
- Example of classes are: **loud music, construction, human talking**

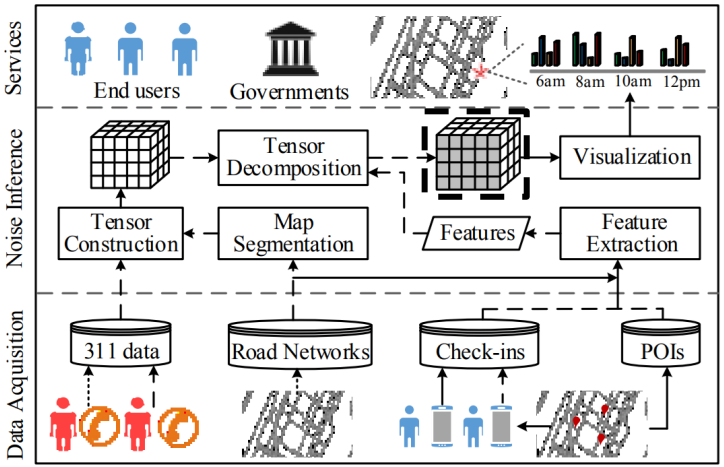


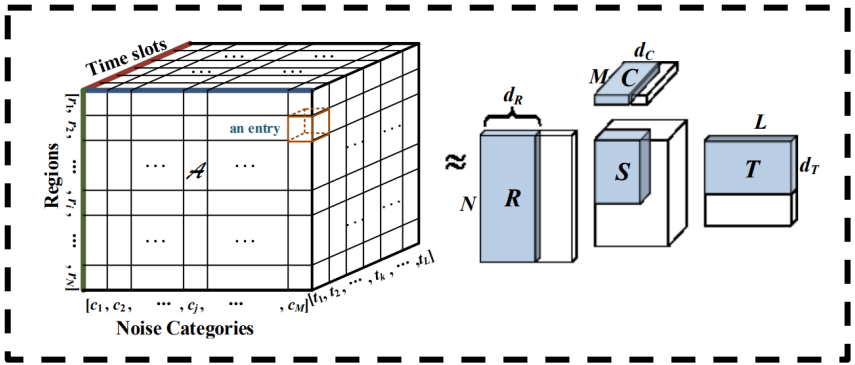
Related Work

- Quantification of noise pollution in New York City.
- Uses **311 data** as proxy for the noise situation in NYC
- Example of classes are: **loud music, construction, human talking**



Related Work

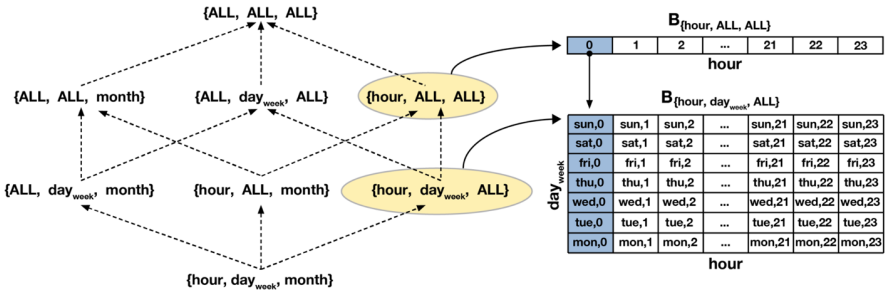




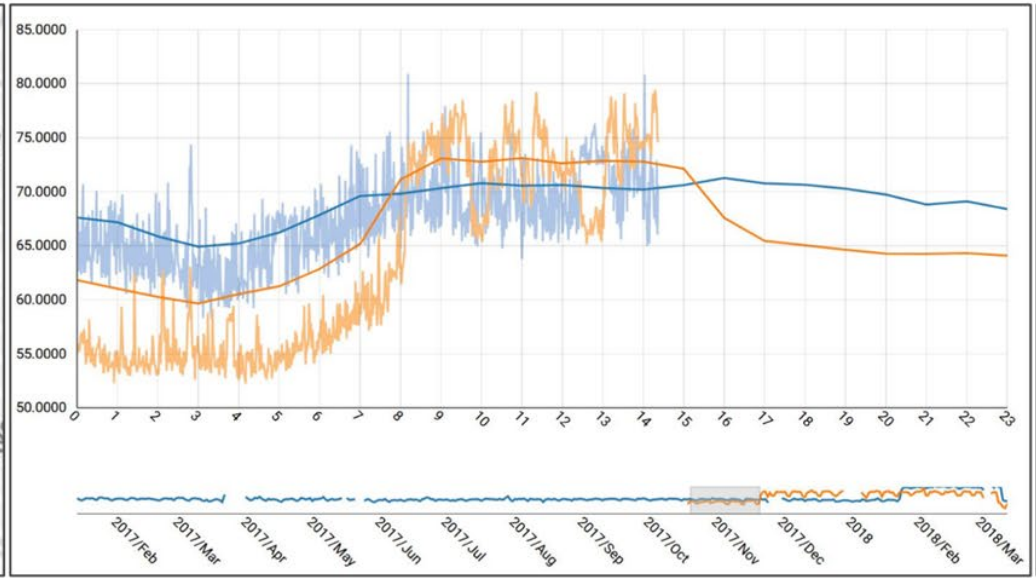
- This work proposes the creation of a spatiotemporal noise map of New York City based on the classes present on the 311 dataset

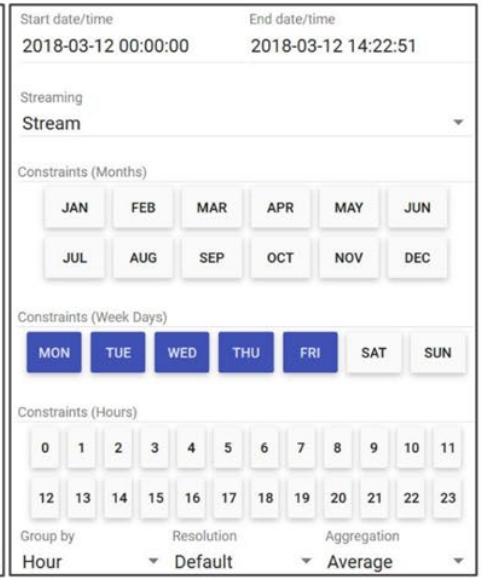
Related Work

- Proposes a novel data structure that allows interactive aggregation and OLAP Queries
- Shortcomings:
  - Only considers SPL data (loudness)
  - Lacks feedback to inform the users



Related Work



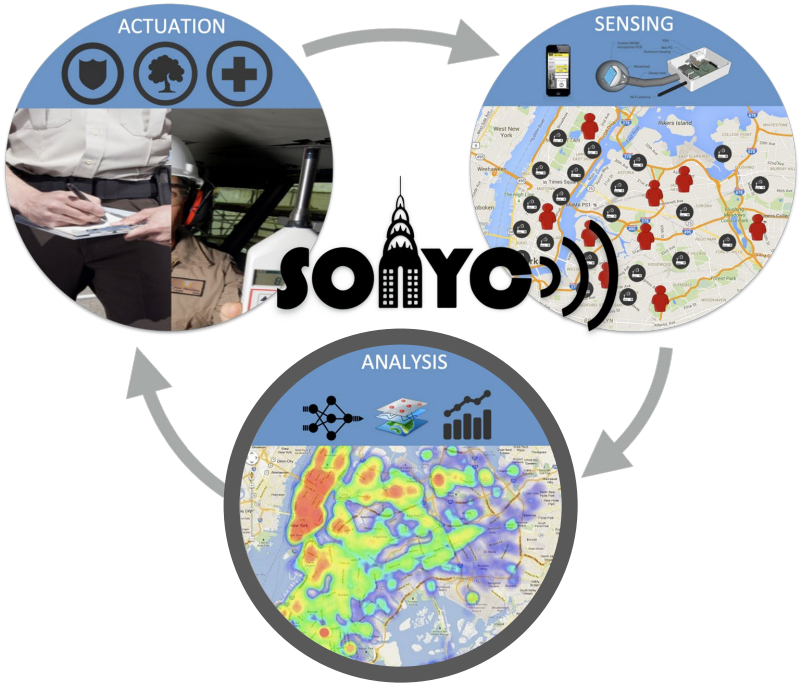


- Lacks user feedback to better understand the semantic aspect of the soundscape

SONYC

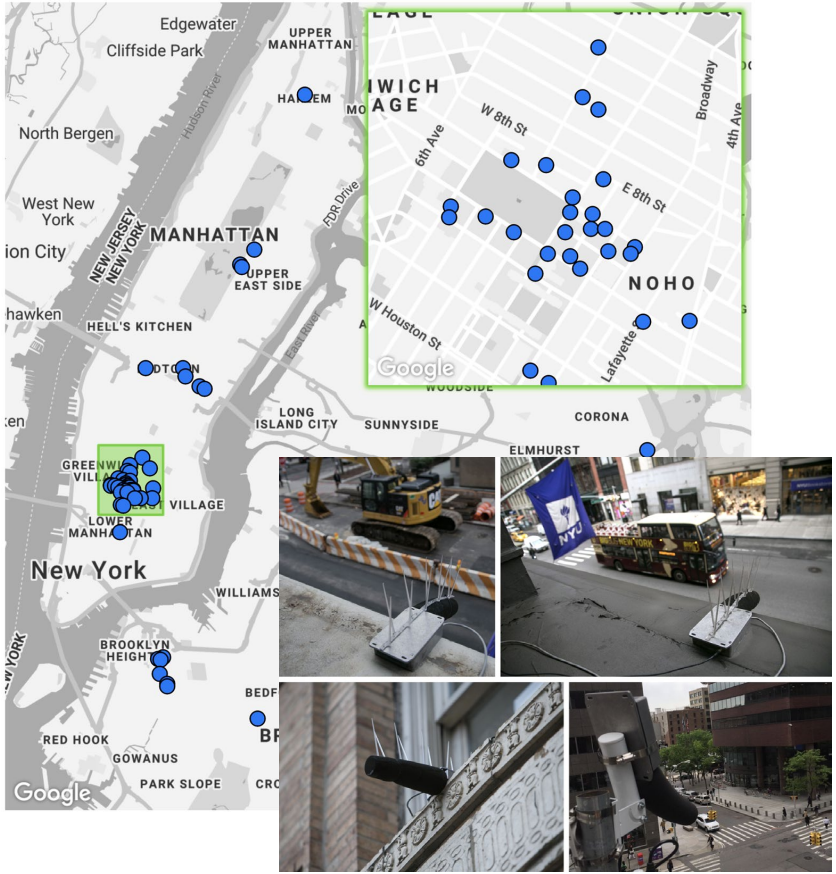
SONYC

- Given high noise pollution levels in New York City, a solution that supports **data-driven analysis** of the soundscape is needed.
- SONYC (Sounds of New York City), aims to address limitations of previous work through an integrated **cyber-physical system** to approach noise pollution
- Through a sensor network SONYC is able to **monitor, analyze and mitigate** noise pollution in NYC.



Deployment

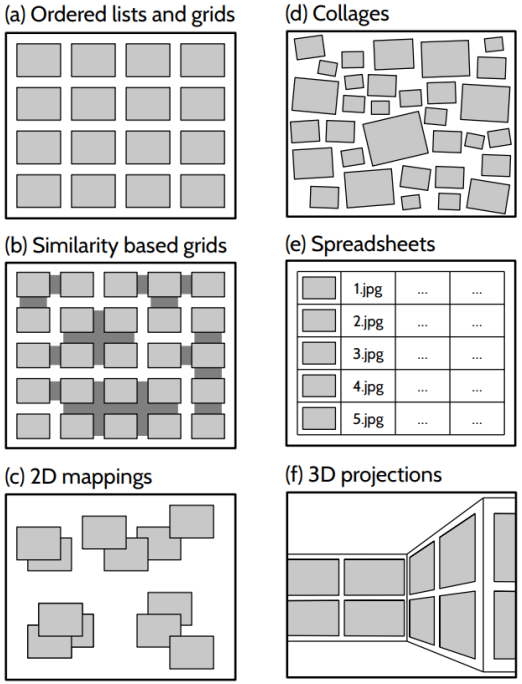
- **55 sensors** deployed over **5 years**;
- **150 sensor years** of decibel data;
- **75 years** of audio data;
- **200M audio recordings** which accounts for **75 Terabytes of data**;
- **75B** decibel rows



Challenges

Sound Exploration

- Unlike images, there is no clear pictorial representation of audio data.
- Humans are able to visualize and understand sets of images in a rather ***parallel approach***.
- Organization approaches for images were proposed in the past. These approaches try to optimize the observation of specific **patterns present in collections of images.**

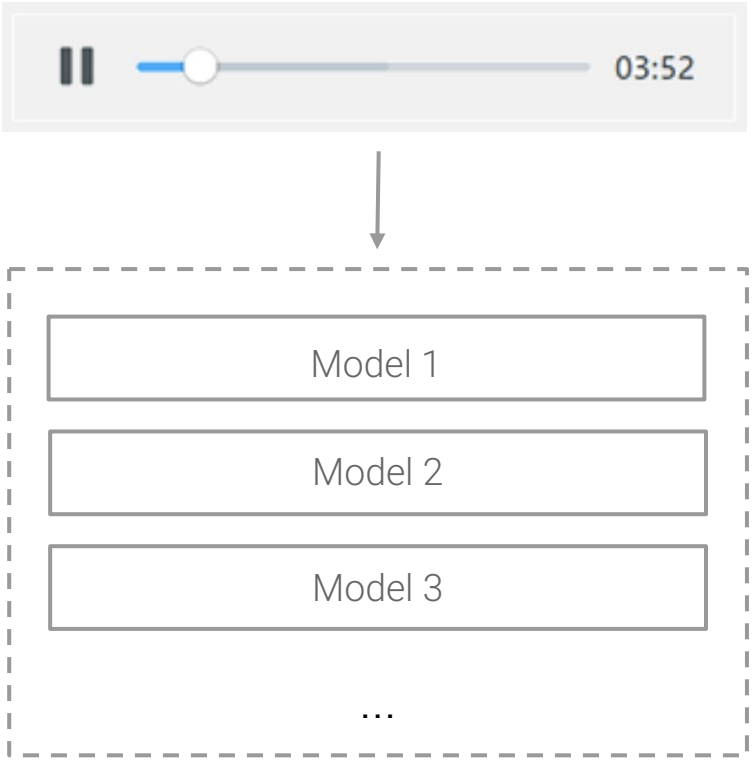


Sound Exploration

- Audio recordings are consumed in a **serial** way by humans.
- To understand events happening in a 10-second audio snippet, users **must listen through the entire recording**
- Although the **visualization of specific frequencies or loudness** can help identify interest periods of the recording, it is still difficult to build a **semantic understanding** of the recording.

Sound Representation

- The **scarcity of labeled urban audio data** makes it hard to generate models capable of transforming audio into representations that can represent different audio classes.
- Also, the **complexity of the urban soundscape** makes it even harder for models to be representative of such a dynamic environment.

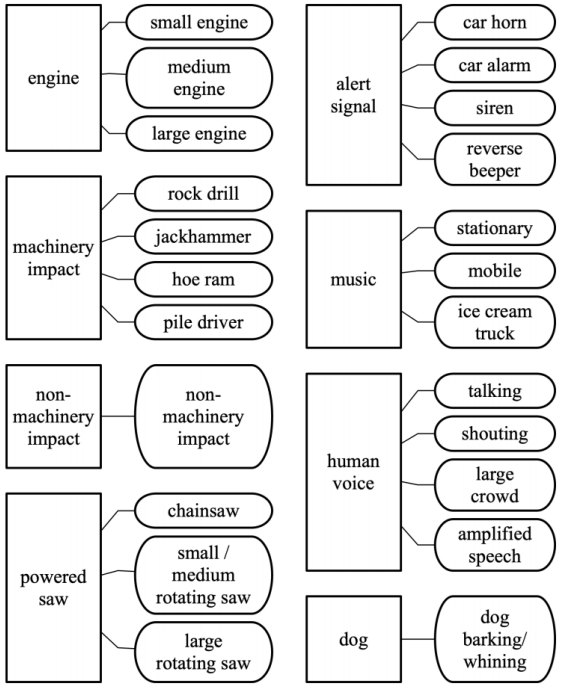


Mixture of Sounds

- 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.
- at any given instant in time, a sound recording might have a mixture of background (birds, dog barks) and foreground sounds (party, sirens).

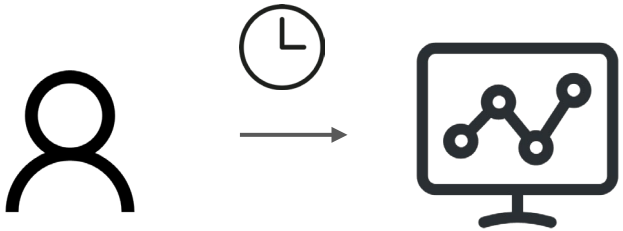
Sound Labeling and Classification

- Previously proposed classifiers provide a reasonable link between embeddings and human-understandable vocabulary;
- However, their **class vocabularies are limited**, providing a **narrow view of the rich and varied soundscape** of the city.



Data Size

- SONYC has captured more than **70TB** of audio recordings throughout **5 years**.
- As past work studied, interactive systems must provide reasonably **low response times** for querying datasets
- One limiting factor for domain experts, usually professionals with no computer science background, is to generate useful insights from a large and complex dataset such as the one produced by SONYC.



Requirements

Requirements

[R1] **Interactive Identification and Labeling of Similar Sounds**: Given the highly complex acoustic environment we observe in cities, audio representations cannot clearly encode specific audio events that users might be interested in. finding similar audio frames based on user perception is one of the system's requirements

[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.

[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).

[R4] **Local and Global Sound Perspectives**: Audio embeddings might possess certain characteristics that only become clear when they are analyzed locally or globally

[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.

[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.

Requirements

[R1] **Interactive Identification and Labeling of Similar Sounds**: Given the highly complex acoustic environment we observe in cities, audio representations cannot clearly encode specific audio events that users might be interested in. finding similar audio frames based on user perception is one of the system's requirements

[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.

[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).

[R4] **Local and Global Sound Perspectives**: Audio embeddings might possess certain characteristics that only become clear when they are analyzed locally or globally

[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.

[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.

Requirements

[R1] **Interactive Identification and Labeling of Similar Sounds**: Given the highly complex acoustic environment we observe in cities, audio representations cannot clearly encode specific audio events that users might be interested in. finding similar audio frames based on user perception is one of the system's requirements

[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.

[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).

[R4] **Local and Global Sound Perspectives**: Audio embeddings might possess certain characteristics that only become clear when they are analyzed locally or globally

[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.

[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.

Requirements

[R1] **Interactive Identification and Labeling of Similar Sounds**: Given the highly complex acoustic environment we observe in cities, audio representations cannot clearly encode specific audio events that users might be interested in. finding similar audio frames based on user perception is one of the system's requirements

[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.

[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).

[R4] **Local and Global Sound Perspectives**: Audio embeddings might possess certain characteristics that only become clear when they are analyzed locally or globally

[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.

[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.

Requirements

[R1] **Interactive Identification and Labeling of Similar Sounds**: Given the highly complex acoustic environment we observe in cities, audio representations cannot clearly encode specific audio events that users might be interested in. finding similar audio frames based on user perception is one of the system's requirements

[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.

[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).

[R4] **Local and Global Sound Perspectives**: Audio embeddings might possess certain characteristics that only become clear when they are analyzed locally or globally

[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.

[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.

Requirements

[R1] **Interactive Identification and Labeling of Similar Sounds**: Given the highly complex acoustic environment we observe in cities, audio representations cannot clearly encode specific audio events that users might be interested in. finding similar audio frames based on user perception is one of the system's requirements

[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.

[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).

[R4] **Local and Global Sound Perspectives**: Audio embeddings might possess certain characteristics that only become clear when they are analyzed locally or globally

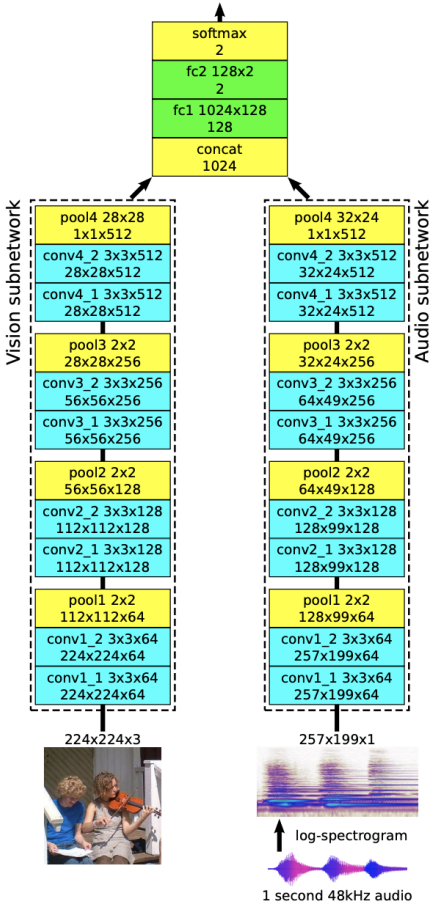
[R5] **Match between Audio and Visual Representations**: Visualizing audio files in the frequency domain is important for the user when assessing the accuracy of both the embeddings and classifications.

[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.

Urban Rhapsody

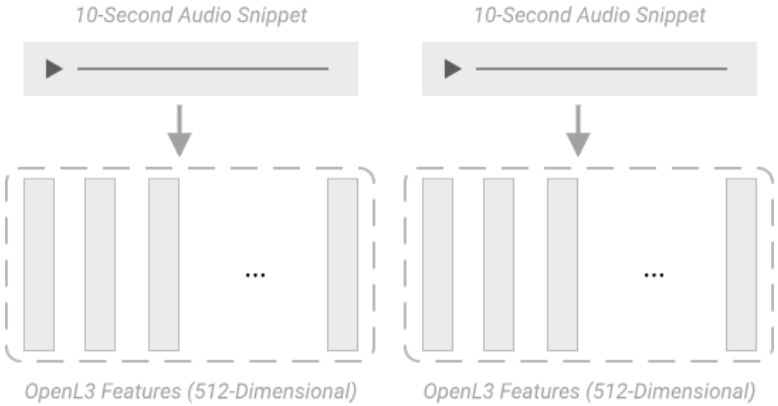
Audio Representation

- In Urban Rhapsody, uses an unsupervised feature extractor model called OpenL3 as the audio representation.
- OpenL3 is an open source implementation of the L3-Net trained with AudioSet.
- The network takes as input the image frames of a given YouTube video and it's corresponding audio in the format of a spectrogram.



Audio Representation

- We want to make sure we capture both short-period audio events and long-period audio events.
- For this reason, we decided to segment our audio snippets into 1-second chunks to use as
- For each 1-second period of audio extracted from the 10-second audio snippets, we generate one 512-Dimensional feature vector representation of the audio clip



Similarity Search

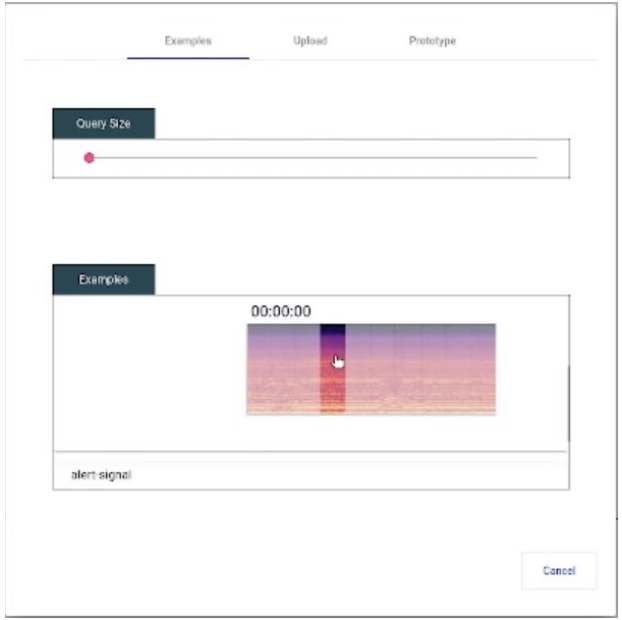
- Urban Rhapsody allows users to search for similar audio events captured by a sensor in the space of one year.
- Our definition of similarity here is based on the angular distance between two vectors. Where each vector represents a 1-second period of audio.
- Not feasible to calculate this for 1 year of recordings for a given sensor

$$\alpha_{1,2} = \cos^{-1}\left(\frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1||\vec{v}_2|}\right)$$

[R1] Interactive identification and labeling **of similar concepts**

Similarity Search

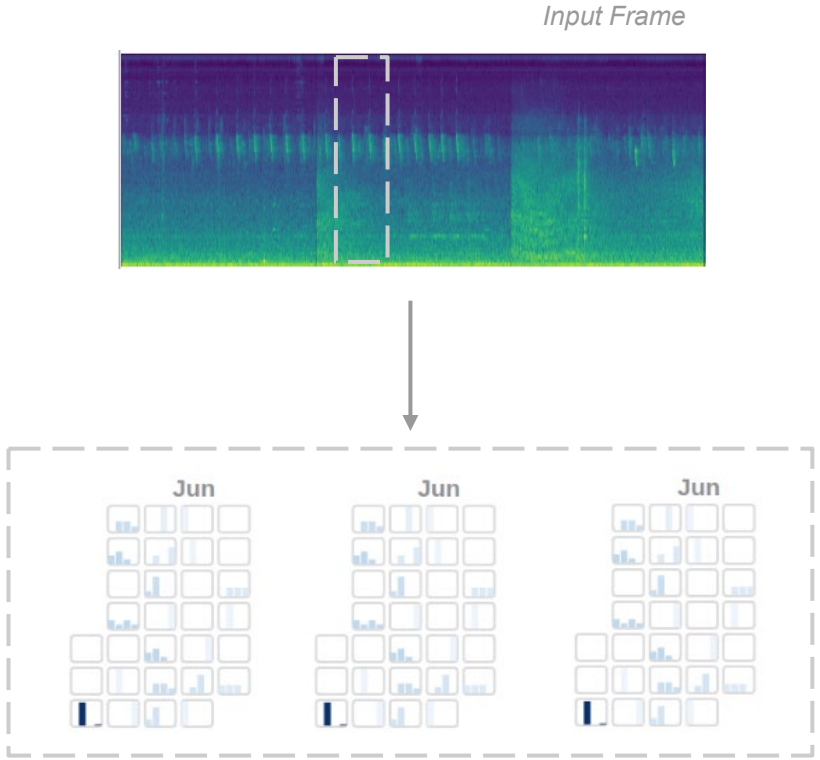
- To allow for interactivity when querying for similar concepts, we calculate LSH-based indices for each sensor-year set of recordings.
- This way, Urban Rhapsody allows for interactive times when the user is searching for similar audio excerpts.
- Urban Rhapsody allows for two different query approaches: **by example** and **by concept**



[R1] Interactive identification and labeling **of similar concepts**

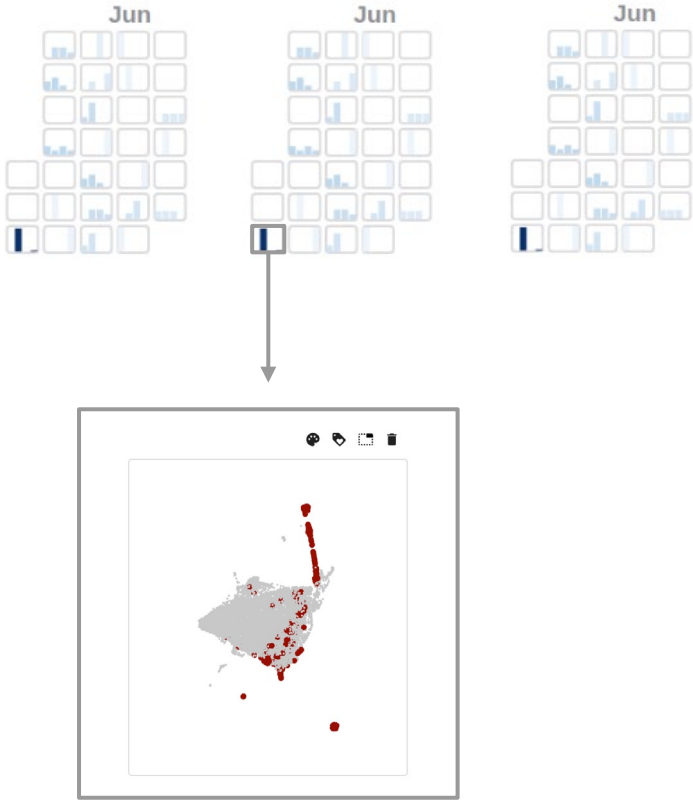
Events Distribution

- Once the user runs a similarity query on Urban Rhapsody database using our query system, Urban Rhapsody provides feedback regarding the distribution of that specific event throughout a given year and within the days of the year.
- The color scale encodes the density of the events across the year.
- The distribution plot shows the distribution of the event within days.



Day View

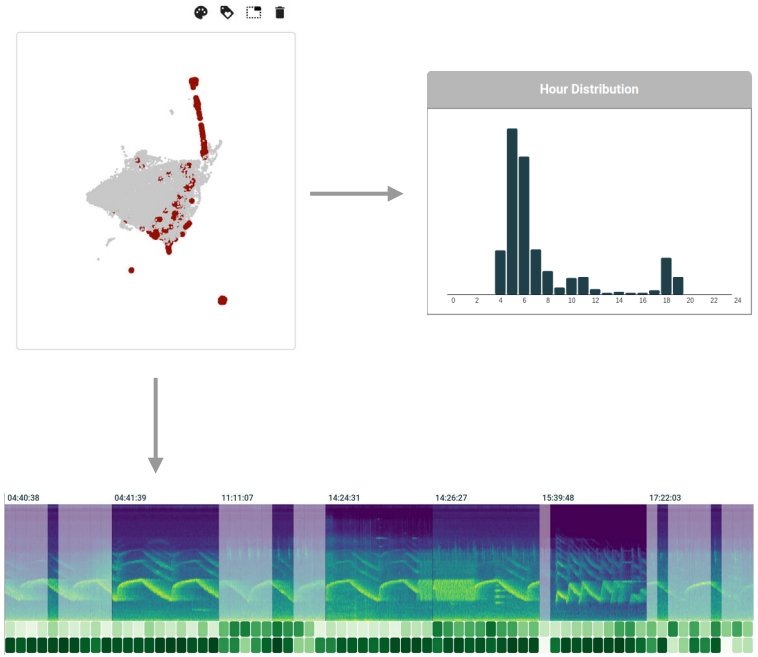
- Following Schneiderman's visualization-seek mantra: **Overview first. Zoom and Filter. Details on Demand.** Urban Rhapsody allows users to analyze specific days by loading the whole data available for a given day.
- To allow users visualize and explore all the feature vectors for a given day, we use well known projection techniques such as UMAP.



[R4] Local and Global **Audio Perspectives**

Focused View

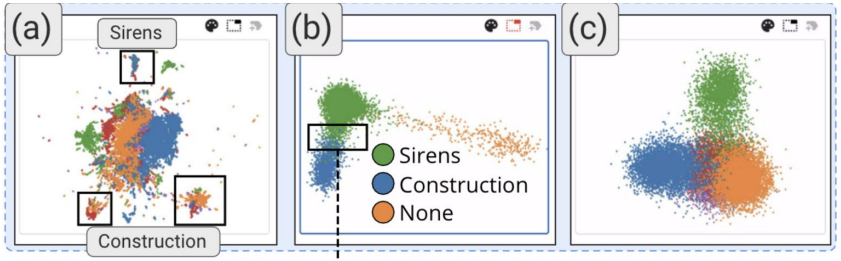
- When a specific day is loaded, the user can interact with the projected scatterplots by selecting either individual points or subsets of the day through bounding boxes.
- Once a selection is made, the user is able to see the spectrograms relative to the selection and the output of classification models built on the fly using Urban Rhapsody.
- Also, when selecting points the user is able to see how those audio events are distributed over a the day



[R4] Local and Global **Audio Perspectives**
[R2] Match between Audio and Visual Representations

Projection Steering

- Although the the day view helps the users to zoom in specific days by filtering points from the rest of the year. It is still hard to explore a whole day of audio recordings.
- While users are selecting and listening to specific audio frames, they can generate new projections that takes the acquired knowledge as input.
- They users can perform three operations: **steer, focus and remove**.



[R2] Projection Steering based on user perspective

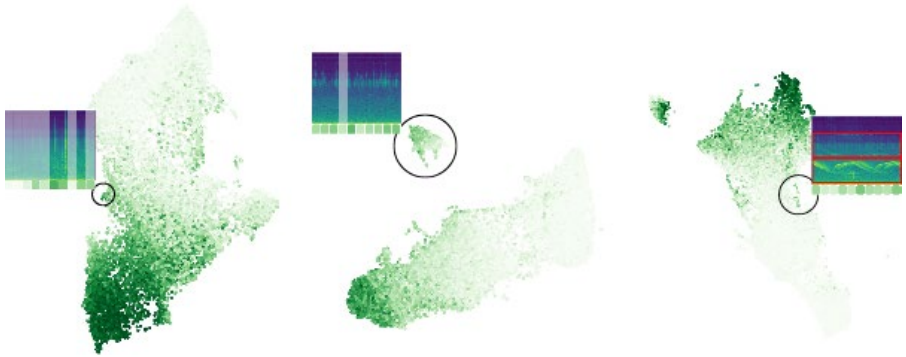
Labeling

- When interacting with the scatterplots, Urban Rhapsody allows users to label either sets of points or individual points with labels.
- Labels can be of two kinds: positive labels and negative labels.
- A free text field is provided for label definition.

[R1] Interactive identification and **labeling of similar concepts**

Model Building

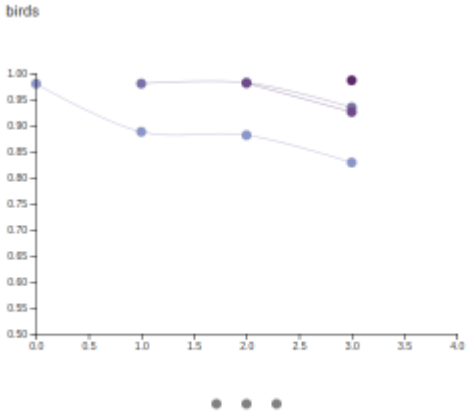
- Once the user labels points as part of an specific concept, Urban Rhapsody uses these samples to build a binary classification model that will help in the identification and exploration of audio excerpts belonging to that specific concept
- Urban Rhapsody is agnostic of learning algorithm. However, for the current version of the system we train the classification models using Random Forest.
- The classification model is trained using a random sample of the whole dataset as negative labels, explicit negative labels and positive labels

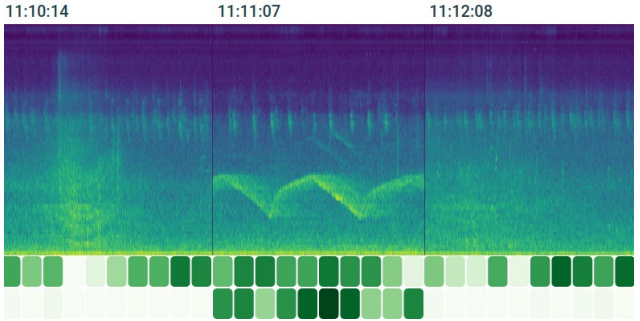


[R3] Iterative Creation of Classification Models

Model Assessment

- We can assess the model performance historically by comparing with older models
- Also, a local assessment of how the model behaves for specific audio frames is provided in the Urban Rhapsody interface.

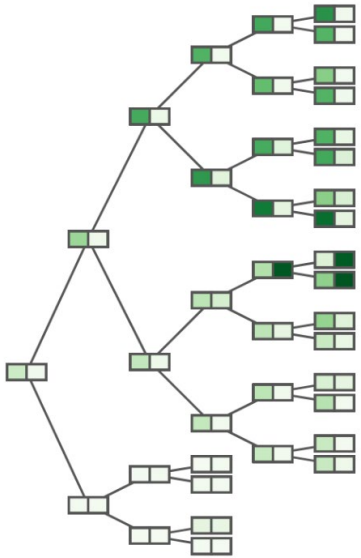




[R3] Iterative Creation of Classification Models

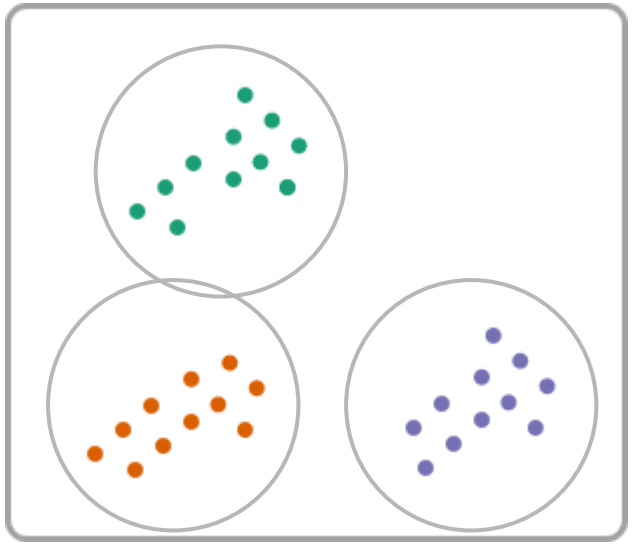
Mixture Explorer

- Although the day view of Urban Rhapsody provides an easier way to inspect instances from a given day, go through hours of recordings is still laborious.
- Moreover, finding regions of the embeddings space that contains mixtures of sounds is a difficult task
- We took a visual approach to facilitate the selection and identification of regions that potentially contain mixture of sounds



Concept Query

- Concepts like construction may be complex. Construction noise is usually a combination of many different audio sources.
- Powered Saw, Large Engines, Drilling Machines are examples of sounds that you can find in a construction site.
- Urban Rhapsody allows users to run similarity queries based on these complex concepts.
- We extract the centroid of each cluster and use as input for the similarity query.

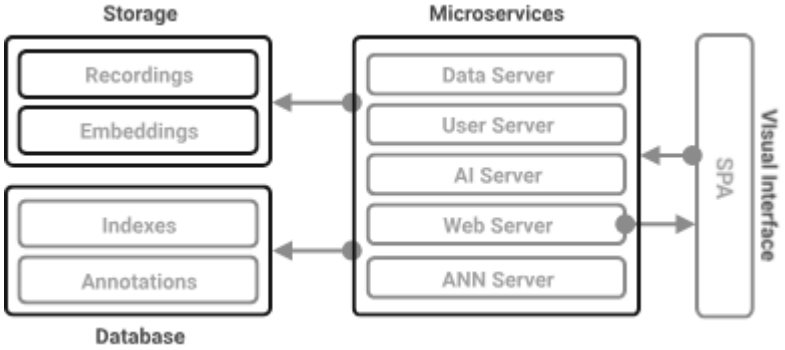


[R1] Interactive identification and labeling **of similar concepts**

Interface Overview

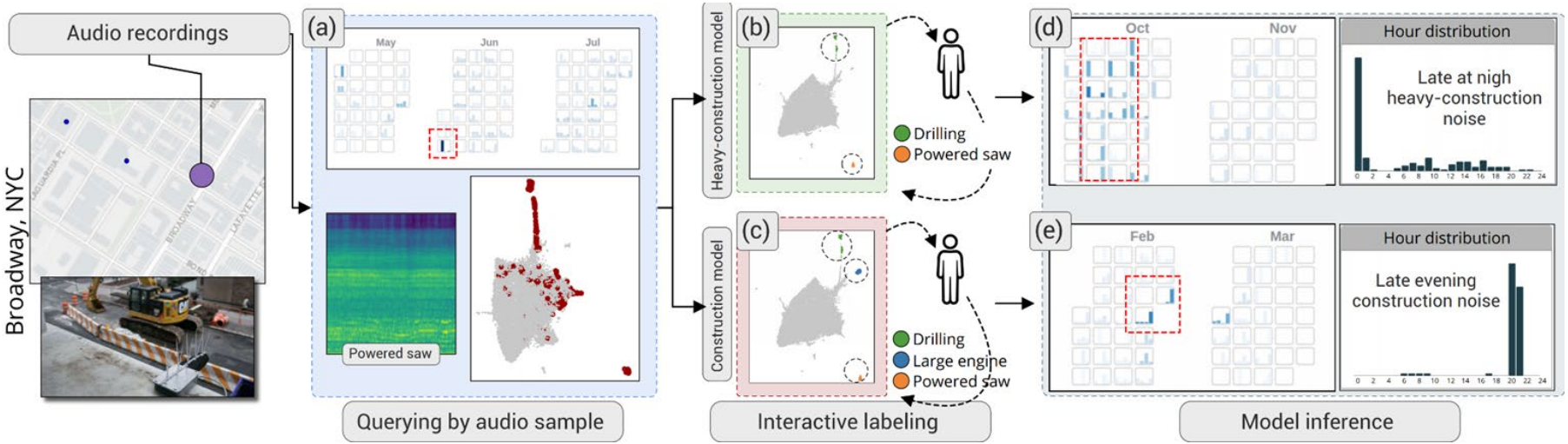
Architecture

- We designed Urban Rhapsody following the microservices architecture, where each module should be responsible for very specific tasks.
- This approach also helps us to extend the system and easily replace modules to test different implementations that can lead to performance improvements.
- The interface was built using Angular framework to facilitate the reuse of modules, WebGL for data-intensive visualizations, and D3 for lightweight visualizations.

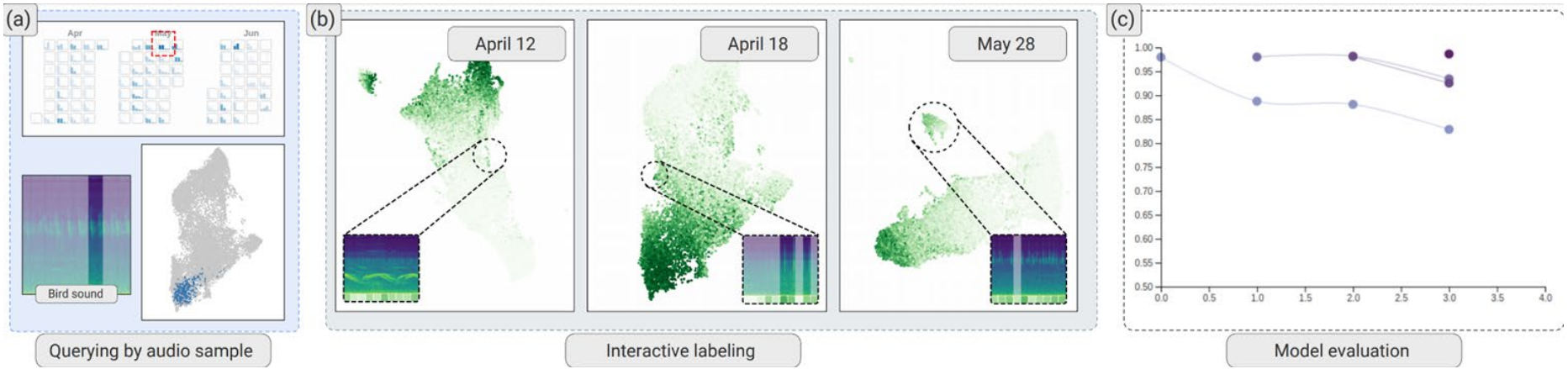


Case Studies

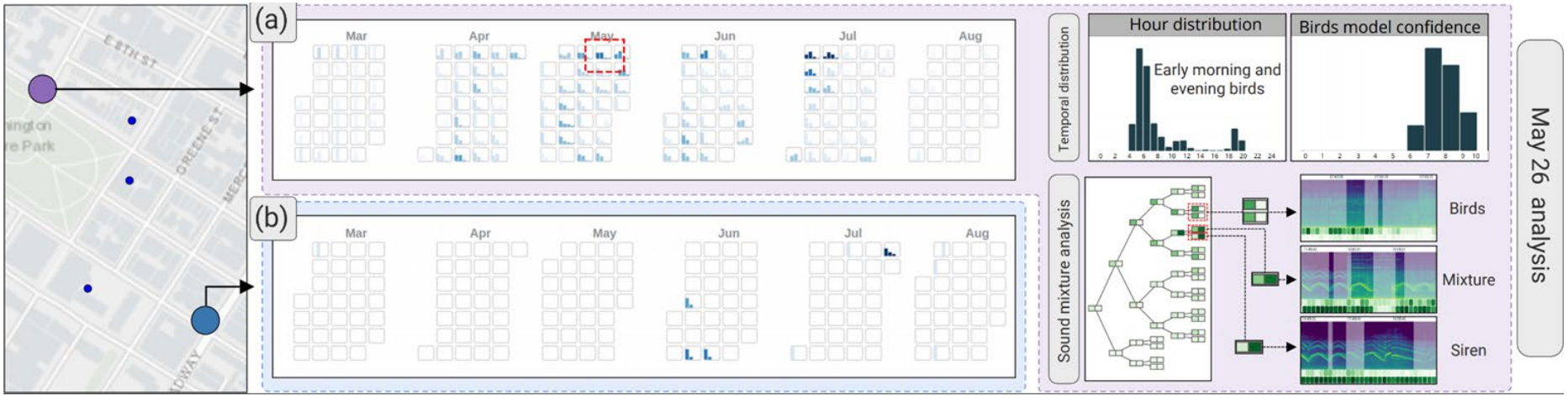
Construction in New York City



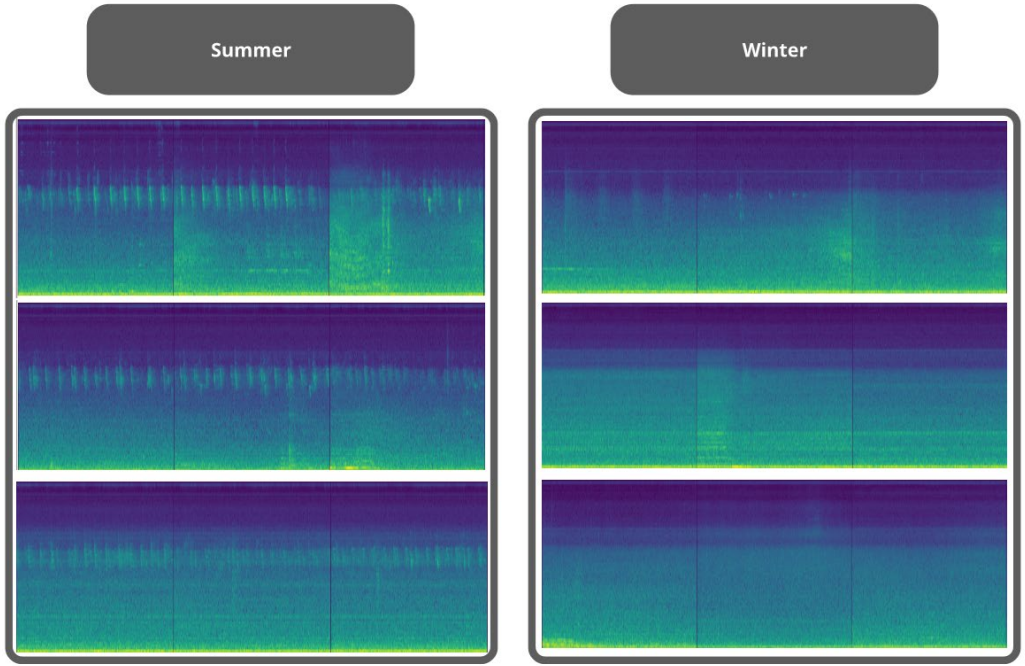
Birds Model


(a)
Apr
Jun
Bird sound
Querying by audio sample
(b)
April 12
April 18
May 28
Interactive labeling
(c)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Model evaluation

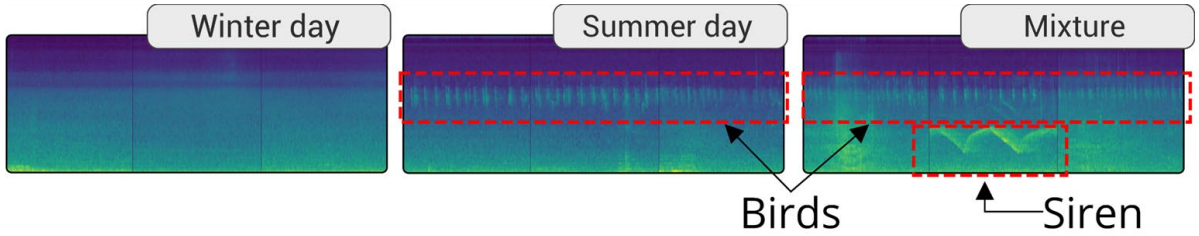
Birds Distribution



Birds Distribution



Birds x Sirens



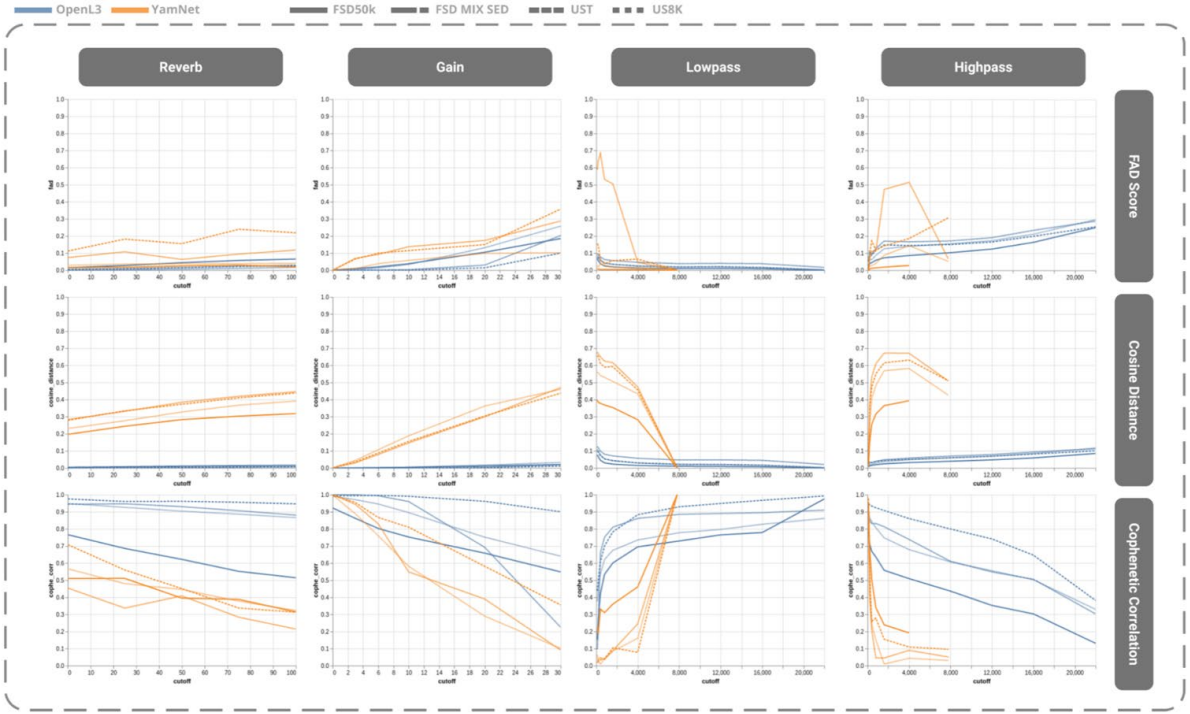
Video

SONYC: Sounds of New York City

SONYC
SOUNDS OF NEW YORK CITY

Ongoing Work

Robustness


OpenL3
YamNet
FSD50k
FSD MIX SED
UST
US8K
Reverb
Gain
Lowpass
Highpass
FAD Score
Cosine Distance
Cophenetic Correlation
cutoff

Model Interpretation

Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples

Angie Boggust[†1], Brandon Carter[†1], and Arvind Satyanarayan[1]
[1]CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA

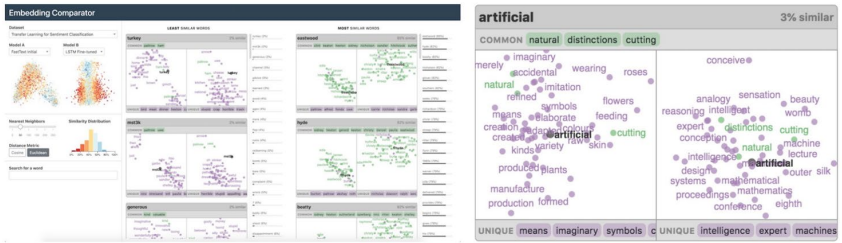


Figure 1: The Embedding Comparator (left) facilitates comparisons of embedding spaces via *local neighborhood dominoes*: small multiple visualizations depicting local substructures (right).

embComp: Visual Interactive Comparison of Vector Embeddings

Florian Heimerl, Christoph Kralj, Torsten Möller and Michael Gleicher