# MONYC: MUSIC OF NEW YORK CITY DATASET

*Magdalena Fuentes*[1]*, Danielle Zhao*[1]*, Vincent Lostanlen*[2]*,*
*Mark Cartwright*[3]*, Charlie Mydlarz*[1]*, Juan Pablo Bello*[1]*,*

[1] New York University, New York, NY, USA
[2] CNRS, Laboratoire des Sciences du Numérique de Nantes (LS2N), France
[3] New Jersey Institute of Technology, New Jersey, NY, USA

## ABSTRACT

Music plays an important role in human cultures and constitutes an integral part of urban soundscapes. In order to make sense of these soundscapes, machine listening models should be able to detect and classify street music. Yet, the lack of well-curated resources for training and evaluating these models currently hinders their development. We present MONYC, an open dataset of 1.5k music clips as recorded by the sensors of the Sounds of New York City (SONYC) project. MONYC contains audio data and spatiotemporal metadata, i.e., coarse sensor location and timestamps. In addition, we provide multilabel genre tags from four annotators as well as four binary tags: whether the music is live or recorded; loud or quiet; single-instrument or multi-instrument; and whether non-musical sources are also present. The originality of MONYC is that it reveals how music manifests itself in a real-world setting among social interactions in an urban context. We perform a detailed qualitative analysis of MONYC, show its spatiotemporal trends, and discuss the scope of research questions that it can answer in the future.

***Index Terms—*** Audio databases, environmental music, sound event detection, spatiotemporal context, street music, urban sound

## 1. INTRODUCTION

Although music is a fundamental component of urban life, our understanding of the musical soundscape of cities remains limited. Few prior sources address the detection of music in noisy environments [1, 2, 3], and even fewer the retrieval of music information. It is currently impossible to automate the indexation of street music, whether recorded by sensor networks or by smartphones. Such a limitation results from the lack of resources for training and evaluating dedicated machine listening models.

UrbanSound8k [2], now regarded as a standard benchmark for audio classifiers, was first in introducing a *street music* class as part of its taxonomy with the aim of detecting the presence of music. More recently, the SONYC-UST dataset [1] included *music* as one of the categories of its taxonomy. Yet, the musical samples in both of these datasets do not have detailed annotations; did not preserve the spatiotemporal distribution of street music; and were not representative in terms of acoustic content. In recent years, the field of audio event recognition has shifted its benchmarks to larger datasets, notably AudioSet [4] (derivative of YouTube) and FSD50k [5] (derivative of FreeSound). Although these datasets contain millions of audio clips, neither of them accommodates a *street music* category or provides any details about urban musical content. As a consequence, AudioSet-based classifiers such as YAMNet[1] cannot reliably identify urban music samples from YouTube or FreeSound, let alone from an acoustic sensor network. To the best of our knowledge, there is no open-source dataset for environmental music analysis that allows for developing models for understanding noisy music urban recordings.

The SONYC project [6] has recorded more than 700k hours of audio data from the streets of New York City. This data tells us a lot about the city's dynamics. The sounds of the city reflect its rhythm, and the major events that happened in the last years. A big part of the soundscape of the city is the music people listen to. People use music to manifest ideas or promote activities. Nightlife, festivals, social demonstrations, street celebrations, restaurants, bars and shops usually play music. Even music played loudly from the speakers of a car are manifestations of people's behaviour. Understanding music in an urban context gives us a deeper perspective on human behaviour, which is harder to obtain from other environmental sound events. From an acoustic viewpoint, the music recordings present in SONYC's archive are recorded in open spaces, in day-to-day conditions, and differ tremendously from commercial studio-recordings or artistic, close-field street recordings such as non-professional music videos. Recordings have low Signal-to-Noise Ratio (SNR), are picked up by the sensors at far-field distances (ranging from approximately ten to fifty feet), and they present big differences in their acoustic characteristics due to the different locations of the sensors: some are located in parks, some in commercial districts; some sensors have buildings close by, and others face towards open spaces. Additionally, the recordings have variant levels of noise from other sources present in the streets such as cars or people talking.

We present MONYC, a manually-annotated dataset of 10-second music clips recorded from the sensors of the SONYC project in the streets of New York City. MONYC was created using a combination of urban sound tagging; self-supervised learning; point process modeling; and human labeling. It conveys very rich metadata including timestamps and spatial location of clips, along with binary scene descriptors to assess models in different conditions (e.g. high interference of non-musical sources). By framing the detection and classification of musical events in the context of environmental acoustics, our goal is not to advance the state of the art in music technology; but rather, to discover relational, spatiotemporal, and behavioral trends in urban sounds at large.

## 2. DATA COLLECTION AND CURATION

As shown in Table 1, MONYC proceeds from SONYC by several stages of data filtering: from 250M audio clips acquired by SONYC sensors to 1.5k after agreement by multiple annotators. This section summarizes the process of data curation which led to MONYC.

**Audio acquisition with SONYC sensors**: SONYC sensors are

[1] https://www.tensorflow.org/hub/tutorials/yamnet

| *stage* | SONYC | 2017 | subsampled | 15 sensors | music | uniform | DPP | annotation | agreement |
|---|---|---|---|---|---|---|---|---|---|
| *# clips* | 250M | 30M | 10M | 5.8M | 94k | 30k | 3k | 1.7k | 1.5k |

Table 1: Number of clips at different stages of curation of MONYC.

placed on second-storey window ledges, at a typical height of seven meters [7]. They record street sounds under the form of 10-second audio clips at a sample rate of 48 kHz. For privacy reasons, the acquisition schedule of audio clips is randomized, with three clips per minute on average. Since May 2016, SONYC has acquired over 250M audio clips, making it one of the largest datasets of urban sounds worldwide.

**Spatiotemporal filtering**: We restrict our study to the year 2017 since it is the first year of the SONYC archive with complete data for the entire period, yielding 30M clips from 26 sensors. We subsample these 30M clips in time down to one clip per minute, yielding 10M clips. Then, we manually select 15 of the 26 sensors in diverse locations: e.g., on a small road, on a main avenue, in different corners of a park, next to a concert venue. Most of these sensors are located in Lower Manhattan, nearby Washington Square Park (WSP) and 5th Avenue; with a few other near Central Park and in downtown Brooklyn. Note that WSP is a well-known spot for street musicians while downtown Brooklyn hosts weekly outdoor concerts in the summer. Reducing the number of sensors from 26 to 15 yields 5.8M clips.

**Urban sound tagging**: We now proceed to retrieve street music within these 5.8M unlabeled clips. To this end, we run a deep learning model for urban sound tagging (UST) named SONYC-UST. SONYC-UST was trained on an open dataset of 19k SONYC clips from years 2016–2019 [1]. This dataset set was annotated by citizen scientists[2] and part of it was verified by experts. Among the 23 classes of the SONYC-UST taxonomy, three of them form the coarse category *music*. The SONYC-UST model has served as the baseline system for Task 5 of the DCASE 2020 challenge[3]. SONYC-UST does not take raw audio as input but relies on a pretrained feature extractor: Open-$L^3$ [8]. Open-$L^3$ is a deep convolutional network which was trained in a self-supervised way on unlabeled YouTube videos [8]. It is based on a mel-frequency spectrogram representation and produces 128-dimensional embeddings at a rate of 1 Hz, i.e. 10 embeddings per clip. SONYC-UST consists of two convolutional layers with a receptive field of 1 and ReLU nonlinearities, followed by AutoPool [9] to aggregate the 10 frame-level predictions.

We keep all clips where the model's output likelihood of *music* was above a threshold that corresponds to the validation set of [1]. This threshold is relatively low since at this stage we prioritized not discarding music clips that might be interesting but the model might not be confident with. At this point, we are down to 94k clips potentially containing street music.

**Uniform spatiotemporal sampling**: We compute the distribution of number of clips through the whole year for each sensor as a reference of the seasonal patterns spotted in that sensor. We keep this distribution as we downsize the number of clips for annotation in the following stage. Meanwhile, we select 2k samples per year per sensor at random (respecting the monthly distribution of music recordings), hence a total of 30k clips.

**Diverse sampling with determinantal point processes**: For our final sample we want not only to keep this seasonal distribution but to have as much diversity of acoustic conditions (e.g. instrumentation, genre, recording conditions) as we can within each sensor. For this, we use a determinantal point process (DPP).

Formerly known as fermionic processes, determinantal point processes (DPPs) are probabilistic models which initially arose in quantum physics to represent repulsive interactions between particles [10]. DPPs has gained attention in machine learning research [11] over the past decade, with the overarching goal of modeling the relative diversity of all possible subsets of a dataset. Since then, it has found many applications, e.g., text summarization [12], video recommendation [13], and news threading [14].

To curate the MONYC dataset, we consider a modified form of DPP known as $K$-DPP [15]. The key idea behind $K$-DPPs is to select a subset of $K$ items from a larger collection $\Omega$ while striking a tradeoff between relevance and diversity. We use the DPP implementation from the DPPy package [16]. We use OpenL3 embeddings as the representation and the music likelihood output of the SONYC-UST model as relevance. We down-sampled each month of data per sensor following the yearly distribution explained below, for a total of 200 samples per sensor, and 3k samples for all sensors.

**Annotations**: Once we had the 3k music clips from the data-driven sampling, we performed the the manual annotation in four stages: 1) pre-selection of musical recordings by one annotator, 2) confirmation by three other annotators, 3) detailed annotation by the four annotators, and 4) annotation agreement and conflict solving. Annotating urban sound recording is a particularly hard endeavor, and street music clips are no exception. Many clips present low signal-to-noise ratio and sometimes music is faint. Other times it is distant and the interference of other sources makes it difficult to disambiguate if there is music or what instruments are present. Besides, the clips' duration of 10s is another challenging aspect, particularly for music annotations since the music can be captured at the least identifiable moment (e.g., an intro) making it hard to determine what music genre is being played. The four annotators are one student and three experienced machine listening researchers, all with musical training.

In a first stage, the student annotator curated all of the 3k clips. This annotator filtered out clips with no music or music too faint, leaving around 1.7k clips. Then each of the remaining annotators annotated one third of this data, with no overlap with each other. These annotators were asked again to confirm if there was music in the clip. Only clips where both annotators said there was music and no sensor faults were included in MONYC, for a total of 1587 clips.

Each annotator was then asked to provide for each clip multi-label free-form genre tags, and a set of binary indicators: whether the music is live or recorded, whether it is loud or quiet, whether the clip has a single instrument, and whether there is high interference from non-musical sources. To consider the uncertainty of annotating hard clips, the annotators used the label *unclear* for particularly hard examples. This free-form tagging was chosen to allow the discovery of music genres in urban music, and the process lead to a total of 114 tags, where the different annotators provided different level of granularity of sub-genres depending on their music preferences and knowledge. The annotators went then through a stage of agreement on a set of "sibling genres" that collapsed some of the annotated sub-

[2]https://www.zooniverse.org/projects/anaelisa24/sounds-of-new-york-city-sonyc

[3]https://github.com/sonyc-project/dcase2020task5-uststc-baseline

genres in a new set of agreed annotations. For instance, tags such as "ragtime", "bebop", "cool jazz", or "free jazz" were collapsed into "jazz". As many as 82 genres were collapsed to bigger categories, for a total of 41 genres in the agreed annotations taxonomy. After the agreement on this set of genres, recordings with conflicting annotations, either because annotators did not overlapped in any tag or they disagreed in the binary flags, were audited by at least two annotators to provide the final set of annotations for MONYC. One of the machine listening experts participated in all conflict solving to ensure consistency in the final annotations set. When there was no agreement among the two annotators, a third annotator was consulted. A total of 924 recordings were audited for agreement (58% of the total). Two sets of annotations are released: a set of agreed, audited annotations intended for developing machine listening models, and a set of pre-agreement annotations with more variance, to illustrate the difficulty and subjective nature of the data.

## 3. DATASET OVERVIEW

MONYC is an open source dataset of environmental music from the streets of New York City. The dataset taxonomy, all annotations and data are available online[4]. It consists of 1587 clips with manual annotations of multi-label free-form genre tags from four annotators, binary descriptors and non-exhaustive instrument annotations, as well as spatiotemporal metadata.

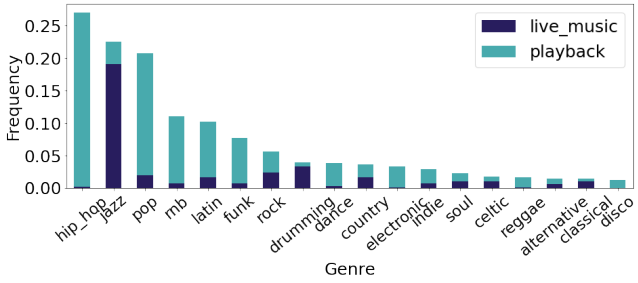


Figure 1: Distribution of 20 top genre labels.

**Music genres in MONYC**: The genre distribution of MONYC, as depicted in Figure 1, has several particularities. Firstly, unlike datasets such as AllMusic, Discorgs, Lastfm or Tagtraum (all part of AcousticBrainz [18]), where the genres with more appearances are rock or pop, the top genre in MONYC is hip hop, pop being the third one and rock just making it to the top ten. Looking at the binary indicator of whether the performance is live or playback, we can look at the relation between genre and live music in the context of street music. We see that most genres are playbacked, sometimes from cars passing, sometimes from shops, or speakers outside homes. The exception are two genres: jazz and drumming, which are both mostly live. Genres such as rock or country are more evenly spread between the two categories, being good candidates to asses the performance of models to identify live vs. playback music.

**Spatiotemporal information**: One of the unique features of MONYC is that each clip contains contextual information of where and when this clip was collected by the sensor network. To maintain privacy and following [1], we quantize the spatial information to the block level and the temporal information to the hour level. For the spatial information we provide borough and block identifiers,

[4]See https://magdalenafuentes.github.io/monyc/ and Soundata [17].

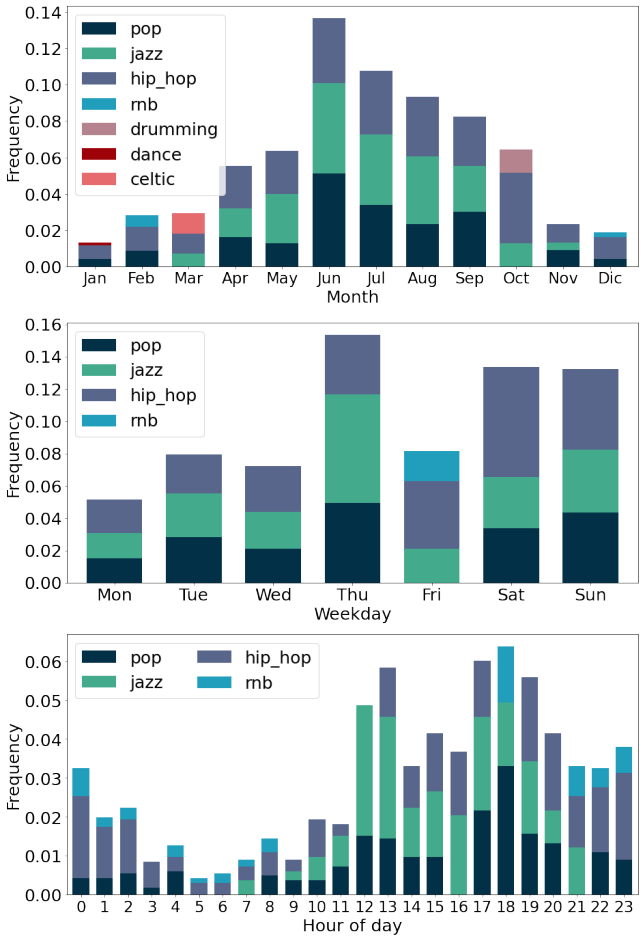


Figure 2: Temporal distribution of music clips at the month (top), weekday (middle) and hour (bottom) level.

as used in NYC's parcel number system known as Borough, Block, Lot (BBL) [19]. This is a common identifier used in NYC datasets, which facilitates the contrast of the sensor data to other open city data [20]. Figure 2 shows the temporal distribution of music clips at the month, day, and hour level broken down by the three top genres in each temporal scale. A first observation is that the amount of music clips increases towards the Summer months (June, July) an decreases considerably in Winter (November, December and January). This makes sense considering that the sensors are capturing environmental music, and in summer there are more concerts, more cars passing playing music with their windows open and more people in the street in general. The percentage of live music oscillates from less than 10% during Winter up to 35-40% in Summer, when it is at its highest. Music genres also change with seasons. As we saw previously, given its live nature, jazz has more presence in the streets in warmer months, with its peak being the summer. Hip hop and pop are season-less, being part of the City's music scene all year round. The appearance of other genres in the monthly top 3 is usually correlated with events that happened in those months. For example, the high presence of celtic music in March is highly explained by St. Patrick's day celebrations and parade on March 17th.

The weekday and hourly distributions also show interesting patterns. The first observation is that there is less street music at

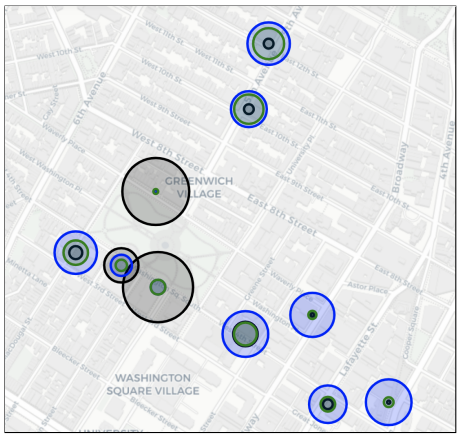


Figure 3: Spatial distribution of hip hop (blue), jazz (black) and pop (green). Circle sizes indicate number of recordings.

the beginning of the week, and an increase towards the weekend, with the exception of Thursdays which show a lot of music activity. When looking at this closer we noticed this is due to two sensors in Downtown Brooklyn which are facing towards a park, which in summer has live concerts every week on Thursdays[5], between noon and 2PM (which also explains the increase of music recordings around that time in the hourly distribution, with concerts having a big component of jazz as can be seen on the genre breakdown). Except for those two sensors, the rest follow a distribution that has more clips during weekends and late-afternoon/evenings. We included these two sensors in the dataset as they are a good example of how the location affect the observations in environmental music.

There are some genres that are constant through the week: pop, jazz and hip hop are present every day, but at different amounts per weekday. For instance, we observe that jazz has more presence on Thursdays, which can be related to the live concerts mentioned before. Hip hop is stable through the week with an increase towards the weekend. Rnb explains the increase of music clips on Fridays. More can be spotted by looking at e.g. the top five genres at different scales, we discussed the top three for better visualization.

An additional aspect of having the spatiotemporal data is that we can explore the spatial distribution of the different genres, i.e. where do we see more instances of one genre or another. An example of that is shown in Figure 3, which shows the spatial distribution of hip hop (blue), jazz (black) and pop (green). The map is zoomed in around Washington Square Park, where the density of sensors is the biggest for the SONYC network. Ten of the 15 sensors of MONYC are around the area of Greenwich Village. In the map we see the music events appearing in the same locations, which correspond to sensor deployments. The first observation is that hip hop and pop are more spread out across roads while jazz is more concentrated around the park. This makes sense considering that jazz is mostly live as in Figure 1, and it is being played in settings suitable for live performances such as a park. Hip hop and pop music are often played by passing cars or in gatherings outside homes, which agrees with the observations in the maps.

**Music tagging in MONYC**: We consider MONYC to be a challenging and interesting scenario to test tagging models, especially when focused on characterizing the music being played. This type of problem is not simply solved by using standard music taggers that where trained in high SNR recordings with no interfering sources, but requires models dedicated to environmental music. We include as a first example an experiment using the off-the-shelf music genre tagger *musicnn* [21], which is a convolutional neural network for out-of-the-box audio music tagging. We use the model trained with the Million Song Dataset (MSD) [22] since it is the one with bigger overlapping with MONYC's taxonomy out of the available models. We evaluate the model by computing the area under receiver operating characteristic curve (ROC-AUC) using the scikit-learn [23] implementation. We selected the subset of MONYC's clips that had genre annotations overlapping with the models' vocabulary, that is the 85% of the data. Twelve genres overlapped in MONYC's and MSD's taxonomies: *rock, pop, alternative, indie, dance, jazz, soul, electronica, folk, 90s, blues, hip hop, country, funk, and rnb*. The overall performance of the model in MONYC is considerably lower than in other datasets [24] (in the range of 90%), with a median ROC-AUC score of 50%. Figure 4 shows that the performance varies widely depending on the genre. Popular street genres such as hip hop, which are usually underrepresented when training music taggers, have very low performance. Looking at recordings from the three most common genres in the data (hip hop, jazz, pop), we noticed that the model performed 8-12% worse in average in those recordings labeled with high interference of sources. We hypothesize that this type of systematic error and the low overall performance could be corrected by re-training or fine tuning such models on MONYC data, which is out of the scope of this paper. This result presents a compelling first look at the type of errors such systems make in environmental music, and the steps we can now take towards making them more robust.

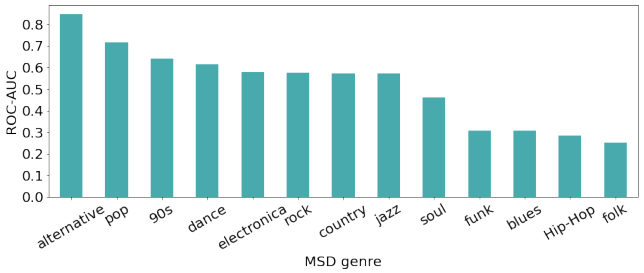


Figure 4: Median ROC tagging results of off-the-shelf music tagger breakdown per genre in MONYC.

## 4. CONCLUSIONS AND FUTURE WORK

We presented MONYC, the first-of-its-kind open dataset of music in urban settings. The dataset was created from the SONYC sensor network archive, delivering data-driven and self-supervised methods for sampling and curating a diverse set of music clips. It consists of a total of four hours of street music audio data along with highly rich annotations: multi-label genre tags from four annotators; spatiotemporal data consisting of location and timestamps of clips; and binary scene descriptors such as whether the music is live or recorded.

We hope this dataset provides the foundations for the development of machine listening models for environmental music, and we plan to expand the dataset with more recordings from the SONYC archive in the future by exploiting the current annotations, as well as similar data-driven methods.

[5] https://www.bam.org/media/9456156/Metrotech-2017_final.pdf

## 5. REFERENCES


[1] M. Cartwright, J. Cramer, A. E. M. Mendez, Y. Wang, H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, O. Nov, and J. P. Bello, "SONYC-UST-V2: An Urban Sound Tagging Dataset with Spatiotemporal Context," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, ser. DCASE, 2020.

[2] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[3] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[5] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an Open Dataset of Human-Labeled Sound Events," 2020.

[6] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[7] C. Mydlarz, M. Sharma, Y. Lockerman, B. Steers, C. Silva, and J. P. Bello, "The Life of a New York City Noise Sensor Network," *Sensors*, vol. 19, no. 6, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/6/1415

[8] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[9] B. McFee, J. Salamon, and J. P. Bello, "Adaptive Pooling Pperators for Weakly Labeled Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[10] O. Macchi, "The Coincidence Approach to Stochastic Point Processes," *Advances in Applied Probability*, vol. 7, no. 1, pp. 83–122, 1975.

[11] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Machine Learning*, vol. 5, no. 2-3, pp. 123–286, 2012.

[12] ——, "Learning Determinantal Point Processes," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2011, pp. 419–427.

[13] M. Wilhelm, A. Ramanathan, A. Bonomo, S. Jain, E. H. Chi, and J. Gillenwater, "Practical Diversified Recommendations on YouTube with Determinantal Point Processes," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2018, pp. 2165–2173.

[14] J. Gillenwater, A. Kulesza, and B. Taskar, "Discovering Diverse and Salient Threads in Document Collections," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*. Association for Computational Linguistics, 2012, pp. 710–720.

[15] A. Kulesza and B. Taskar, "k-DPPs: Fixed-size determinantal point processes," in *Proc. ICML*, 2011.

[16] G. Gautier, G. Polito, R. Bardenet, and M. Valko, "DPPy: DPP Sampling with Python," *Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS)*, 2019, code at http://github.com/guilgautier/DPPy/ Documentation at http://dppy.readthedocs.io/. [Online]. Available: http://jmlr.org/papers/v20/19-179.html

[17] M. Fuentes, J. Salamon, P. Zinemanas, M. Rocamora, G. Plaja, I. R. Román, R. Bittner, M. Miron, X. Serra, and J. P. Bello, "Soundata: A Python library for reproducible use of audio datasets," *http://arxiv.org/abs/2109.12690*, 2021.

[18] D. Bogdanov, A. Porter, H. Schreiber, J. Urbano, and S. Oramas, "The AcousticBrainz Genre Dataset: Multi-source, Multi-level, Multi-label, and Large-scale," in *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019): 2019 Nov 4-8; Delft, The Netherlands.[Canada]: ISMIR; 2019.* International Society for Music Information Retrieval (ISMIR), 2019.

[19] "Borough, block, lot lookup." [Online]. Available: https://portal.311.nyc.gov/article/?kanumber=KA-01247

[20] "NYC Open Data." [Online]. Available: https://opendata.cityofnewyork.us/

[21] J. Pons and X. Serra, "musicnn: Pre-trained Convolutional Neural Networks for Music Audio Tagging," *arXiv preprint arXiv:1909.06654*, 2019.

[22] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," 2011.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-based Automatic Music Tagging Models," *arXiv preprint arXiv:2006.00751*, 2020.