



**JOHANNES KEPLER  
UNIVERSITY LINZ**

# UE MLPC 2025: DATA EXPLORATION TASK



Tara Jadidi, Paul Primus, Florian Schmid

2025-04-07

*Institute of Computational Perception*

# Agenda

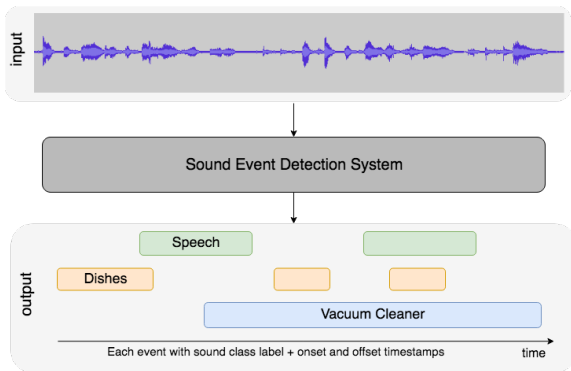
- Project Vision: Where are we?
- Dataset Statistics
- Explaining the Dataset Features
- Data Exploration: What to Look for?

# WHERE ARE WE?



# The project vision

- Goal: Train models on a general-purpose dataset that can detect a set of arbitrary sound events with their respective onsets and offsets

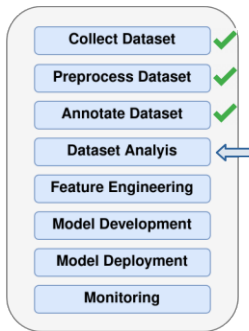


# Where we are



## Events of Interest:

Ship, Train, Helicopter, Truck



**COLLECTED DATA**



# Data Collection Task

## What you had to do:

- Each of you annotated 40 audio snippets, resulting in over 70 hours of annotated audio
- There was an overlap between the annotators (some files are annotated by more than one annotator)

## The final dataset:

- 9026 audio files, annotated by 330 students
- 35826 textual annotations, with at least one annotation per file and four on average




# Overview

```
MLPC2025_Dataset/  
|- audio/  
|  |- 0.mp3  
|  |- 1.mp3  
|  \- ...  
|- audio_features/  
|  |- 0.npz  
|  |- 1.npz  
|  \- ...  
|- metadata.csv  
|- annotations.csv  
|- README.md  
|- annotations_text_embeddings.npz  
|- metadata_keywords_embeddings.npz  
|- metadata_title_embeddings.npz
```

# Overview


```
MLPC2025_Dataset/  
|- audio/  
|  |- 0.mp3  
|  |- 1.mp3  
|  \- ...  
|- audio_features/  
|  |- 0.npz  
|  |- 1.npz  
|  \- ...  
|- metadata.csv  
|- annotations.csv  
|- README.md  
|- annotations_text_embeddings.npz  
|- metadata_keywords_embeddings.npz  
\- metadata_title_embeddings.npz
```



*directory containing the  
audio files*

# Overview

```
MLPC2025_Dataset/  
|- audio/  
|  |- 0.mp3  
|  |- 1.mp3  
|  \- ...  
|- audio_features/  
|  |- 0.npz  
|  |- 1.npz  
|  \- ...  
|- metadata.csv  
|- annotations.csv  
|- README.md  
|- annotations_text_embeddings.npz  
|- metadata_keywords_embeddings.npz  
\- metadata_title_embeddings.npz
```



*The features extracted  
from the audio files  
(more on that in the next  
slides)*

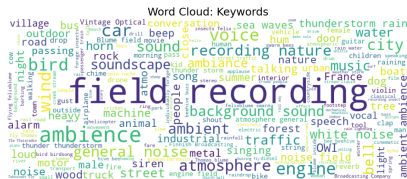
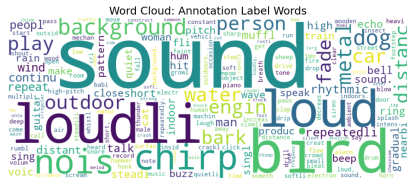
## Textual Data Available: Metadata

- The information included for each file as metadata:
  - keywords and the original title of the uploaded audio file
  - The username of the person providing the file
  - The original description provided for the entire audio file
  - The start and end time of the audio segment
  - A link to the sound and information such as the geotag, number of downloads, and license
- Of these, the keywords and the original titles of the files were embedded using our text embedding model, and can be found in `metadata_keywords_embeddings.npz` and `metadata_title_embeddings.npz` respectively.

## Textual Data Available: Annotations

- The annotations file includes each annotation text as a row with:
  - task ID in the annotation framework
  - The audio file the annotation is for
  - onset and offset in seconds
  - the ID of each annotator
  - the name of the original file
- The annotation embeddings can be found in `annotations_text_embeddings.npz`

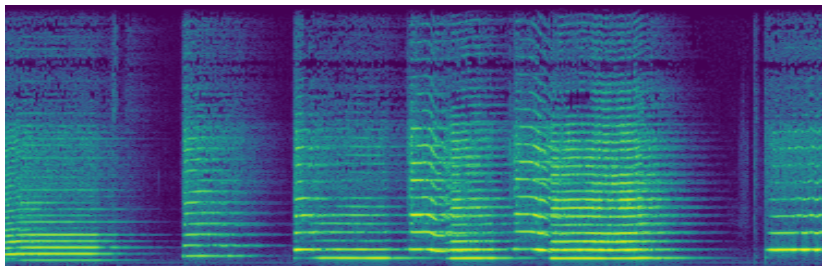
## Comparing Keywords and Annotations



# FEATURES



# Spectrum and Spectrogram



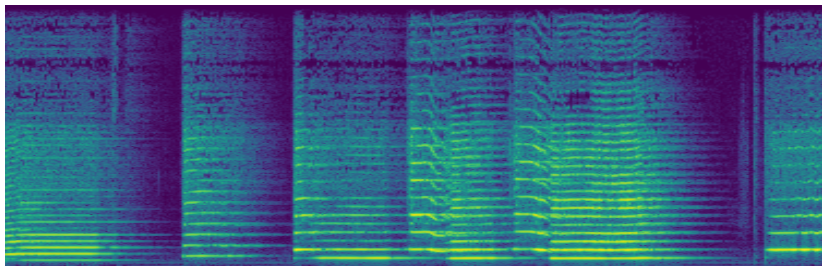
## ■ Computation:

- ☐ take tiny excerpts (e.g., 25 ms) in regular intervals
- ☐ compute **spectrum** of each via Discrete Fourier Transform
- ☐ stack the resulting frames (so that each column in the figure above is a spectrum of one excerpt)

## ■ A column of a spectrogram (or the tiny excerpt of audio it corresponds to) will often be called a **frame**



# Spectrum and Spectrogram

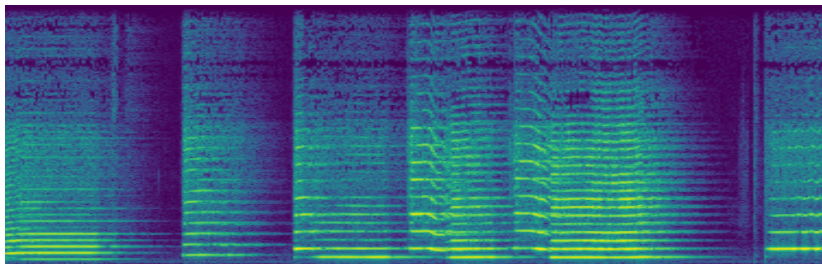


## ■ First step: **spectrogram**

- ☐ time from left to right
- ☐ frequency from bottom to top
- ☐ phase information removed
- ☐ energy per time-frequency bin mapped to color

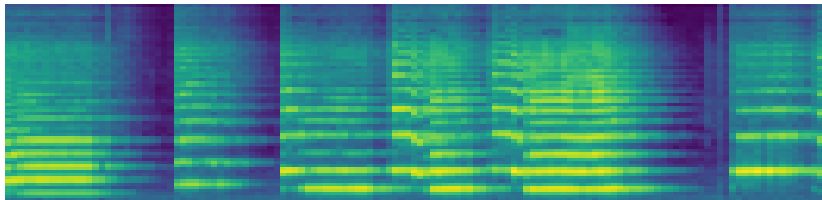
## Feature: Mel spectrogram

- Spectrogram is quite large (201 pixels high), with most information spent on high frequencies
- We do not need so much detail in high frequencies (humans cannot perceive small differences in high frequencies either, but could solve our classification task)

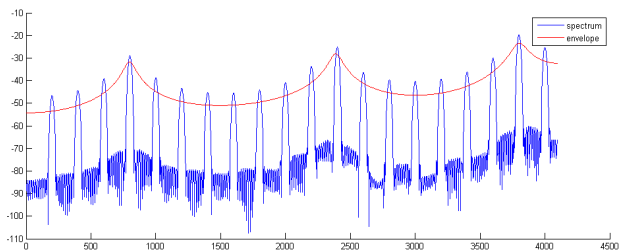
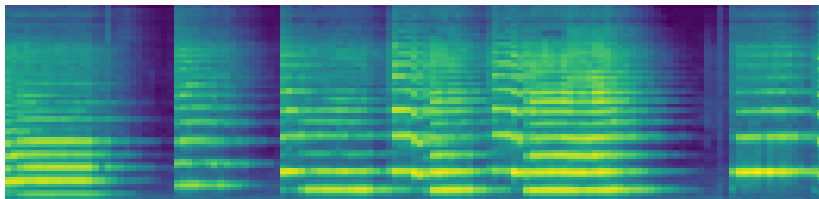


## Feature: Mel spectrogram

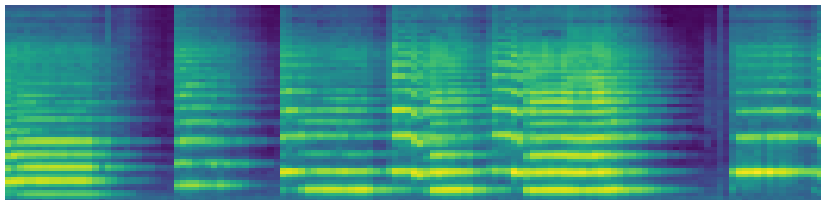
- Spectrogram is quite large (201 pixels high), with most information spent on high frequencies
- We do not need so much detail in high frequencies (humans cannot perceive small differences in high frequencies either, but could solve our classification task)
- ▶ Remap to 64 pixels with high resolution on the bottom, and progressively lower resolution on the top (a **mel scale**)



# Feature: MFCCs



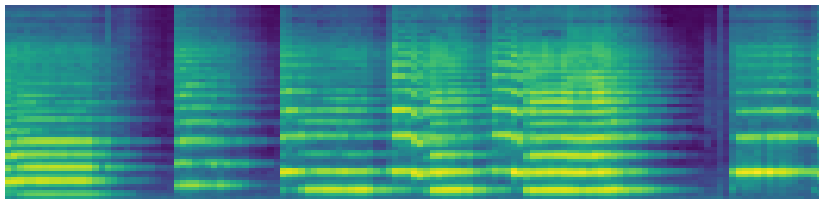
## Feature: MFCCs



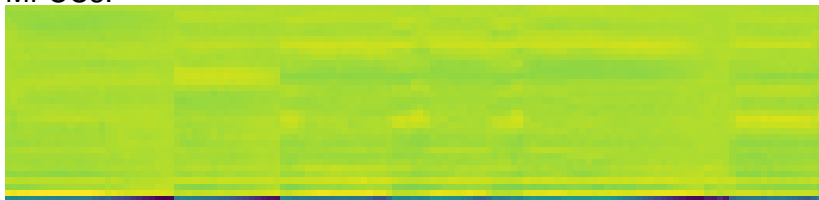
### Mel-Frequency Cepstral Coefficients **MFCCs**

- mel scaling and log magnitudes of spectrogram (both can be perceptually motivated)
- then approximate the 64 Mel-Energies with 32 Discrete Cosine Transform coefficients

## Feature: MFCCs



MFCCs:

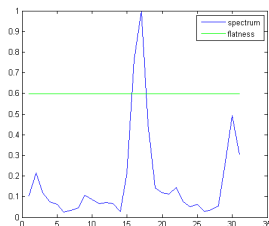
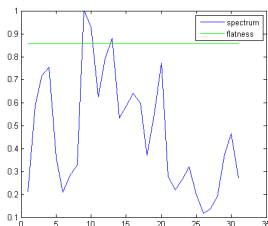


$\Delta$ -MFCCs:

$\Delta\Delta$ -MFCCs ...

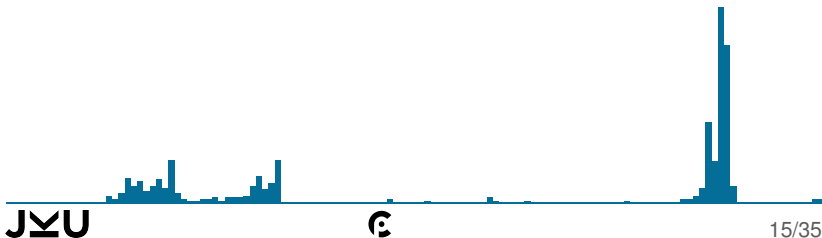
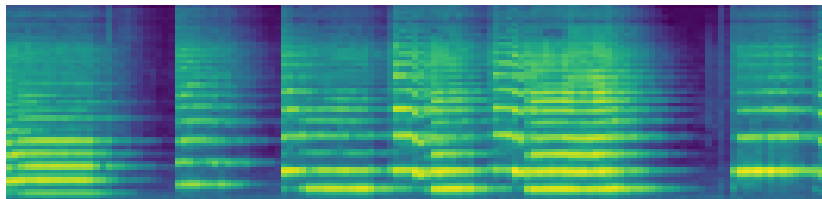
# Feature: Spectral Flatness

- Simple estimator for noisy (1.0, all frequencies have the same energy) vs. tonal (0.0, single peak in spectrum)
- Geometric mean divided by arithmetic mean of spectrum



## Feature: Spectral Flatness

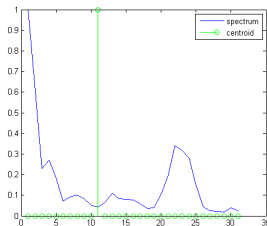
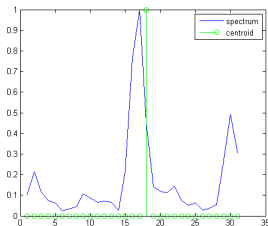
- Simple estimator for noisy (1.0, all frequencies have the same energy) vs. tonal (0.0, single peak in spectrum)
- Geometric mean divided by arithmetic mean of spectrum





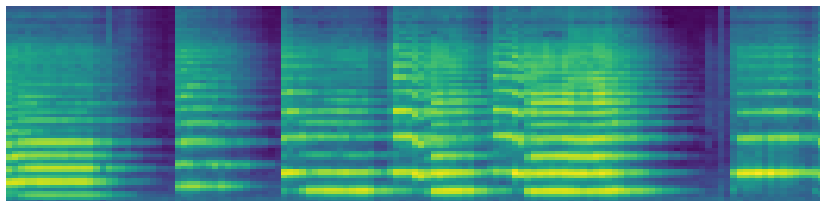
# Feature: Spectral Centroid & Bandwidth

- Where is the “center of mass” in the spectrum? (brightness of sound)
- Mean of the frequency values weighted by their energy
- Bandwidth: spread around the centroid.



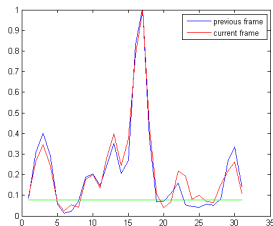
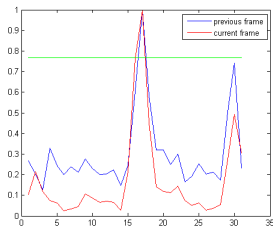
## Feature: Spectral Centroid & Bandwidth

- Where is the “center of mass” in the spectrum? (brightness of sound)
- Mean of the frequency values weighted by their energy
- Bandwidth: spread around the centroid.



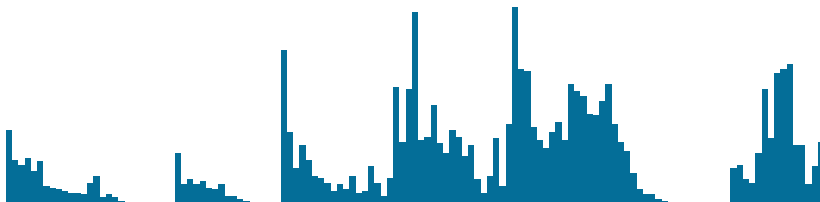
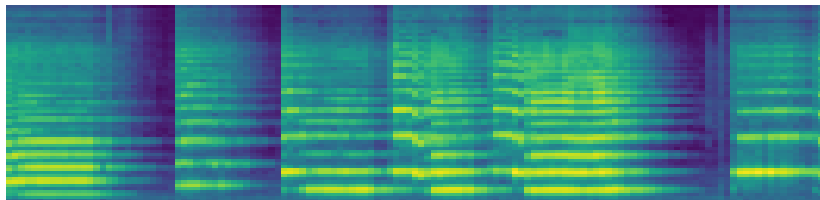
# Feature: Spectral Flux

- How much does the spectrum change over time?
- Euclidean distance of consecutive spectrogram frames



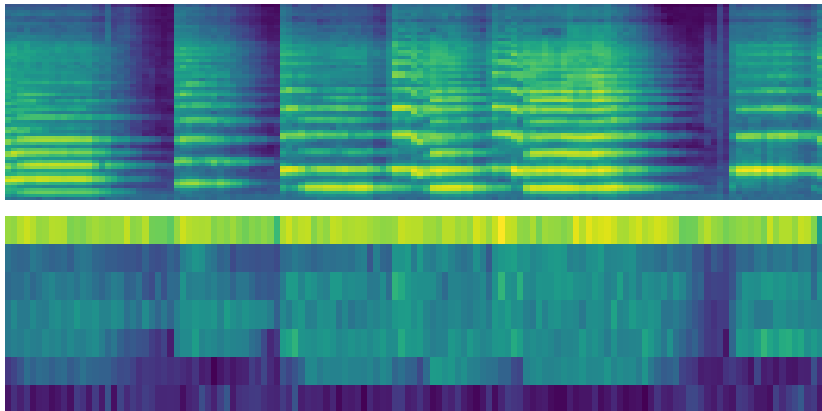
## Feature: Spectral Flux

- How much does the spectrum change over time?
- Euclidean distance of consecutive spectrogram frames

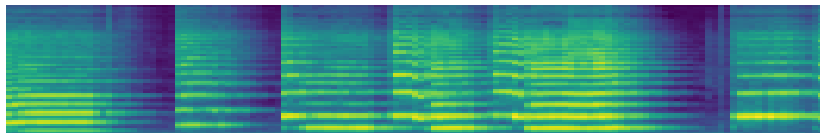


## Feature: Spectral Contrast

- Logarithmic energy difference between peak and valley in a spectral band; high for pitched tones, low for noise
- Computed for 7 intervals (0-200, 200-400, 400-800 ... Hz)



## Features: Energy, Power, ZCR



Energy: sum of bins of spectrum (correlates with loudness)



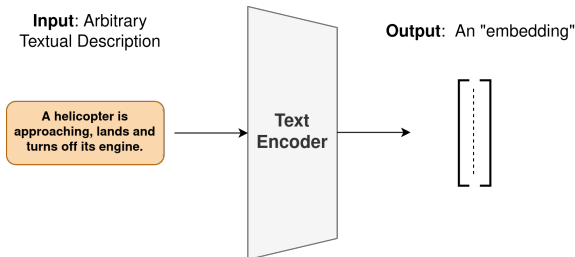
Power: sum of squared bins of spectrum



Zero Crossing Rate: number of zero crossings in waveform



# Text Embeddings



- Reminder: An embedding model can take a text of arbitrary length and map it to a fixed-size vector that represents the semantic meaning of the text. (For a more detailed explanation, refer to our first lecture.)

# Text Embeddings

- There are different ways to create such embeddings. For instance, algorithms like Word2Vec are based on the frequency with which words appear in the context of other words. However, more recent approaches use neural networks trained on large text corpora.
- For the text embeddings in our dataset, we use a the text encoder of a model trained for the task of audio retrieval. For more details on the model, you can refer to the respective paper<sup>1</sup>, but the details are beyond the scope of this course.

---

<sup>1</sup>Primus, P., Schmid, F., & Widmer, G. (2024). Estimated Audio-Caption Correspondences Improve Language-Based Audio Retrieval. arXiv:2408.11641



# Audio Embeddings

- Similar to text embeddings, it is possible to map each audio segment to a fixed sized vector representing its content as well.
- to create the embeddings, we use a model <sup>2</sup> pre-trained on a large corpus of audio files for sound event detection.

---

<sup>2</sup>Schmid et al. (2024). Effective Pre-Training of Audio Transformers for Sound Event Detection.  
arXiv:2409.09546v2

## Feature Summary

- All features are computed for 40ms frames and aggregated over three time steps
  - 768 for the embedding vector dimensions
  - 1 ZCR
  - 64 mel-scaled energies
  - 32 MFCCs, 32  $\Delta$ -MFCCs, 32  $\Delta\Delta$ -MFCCs
  - 1 spectral flatness, 1 centroid, 1 flux, 1 energy, 1 power, 1 bandwidth, 7 contrasts
- This results in 942 features per each 120 ms, and each audio file is between 15 to 30 seconds.
- For classification, we might not need all features in such a high temporal resolution.

# THE EXPLORATION TASK



# Data Explorations: Motivation 1/2

## Ambiguous or incorrect annotations

5	161986736.98833.mp3	1.00089183809168E+077 Commentator speaks. Crowd cheers
6	161986736.98833.mp3	1.00089183809168E+077 Quiet Humming sound of engine
7	161988518.98014.mp3	4.11636573135803E+074 A brief saxophone note rings out, then softly fades away
8	161988518.98014.mp3	7.24707469045091E+076 A saxophone playing, with reverb.
9	161988518.98014.mp3	4.11636573135803E+074 Simple, sharp tones in the upper register being played on the saxophone, lacking melodic structure
10	161988518.98014.mp3	4.11636573135803E+074 A brief saxophone note rings out, then softly fades away
11	161988518.98014.mp3	7.24707469045091E+076 An airy sound rushes.
12	161988518.98014.mp3	7.24707469045091E+076 A saxophone playing, with reverb.
13	161988518.98014.mp3	4.11636573135803E+074 A high-pitched piercing note being played on a saxophone
14	161988518.98014.mp3	7.24707469045091E+076 A saxophone playing, with reverb.
15	161987900.97621.mp3	1.03976418446305E+077 A single chord loudly played on an electric guitar slowly decreasing in volume
16	161987900.97621.mp3	4.54504816665745E+076 Electrical buzz made from electric guitar
49	161979734.95743.mp3	2.0004300421917E+077 Several people talking with occasional laughs and overlapping voices inside a moving vehicle
49	161979734.95743.mp3	1.11470609057027E+077 A dripping sound, repeating rhythmically.
50	161979734.95743.mp3	1.11470609057027E+077 A constant sound of raining.
51	161982094.94763.mp3	9.40579719615464E+076 A wave approaching gradually increasing near the sea
52	161982094.94763.mp3	9.40579719615464E+076 Wind blowing muffled steadily loud near the sea
53	161982094.94763.mp3	7.00003942184145E+076 Ocean waves rolling in and out with a soft wind blowing in the background. The surf is steady and calming, with a natural rhythm.
54	161982094.94763.mp3	9.40579719615464E+076 Waves approaching repeatedly near the sea
55	161986470.94743.mp3	8.14337619464377E+076 Vacuum cleaner powering on, gradually reaching full suction.
56	161986470.94743.mp3	8.14337619464377E+076 Vacuum cleaner powering down, suction sound gradually fades.
57	161986470.94743.mp3	8.14337619464377E+076 Vacuum cleaner running steadily with a consistent suction sound.
58	161983567.94334.mp3	9.49791867083704E+076 footsteps through hard packed snow
59	161976941.94017.mp3	6.89460965255535E+076 A Beeping sound

Figure: The two annotators can agree or disagree on the content of the file

## Data Explorations: Motivation 2/2

### ■ Curse of dimensionality

- Each textual annotation is described with 768 features, and each audio frame with 942
- Many of the features will be unnecessary/ highly correlated
- Maybe removing some features, or combining them could even increase the performance...

# What to Investigate 1/5

- **Case Study** (2 points): Find two interesting recordings with at least two annotators and multiple annotations. Compare the temporal and textual annotations, and try to answer the following questions:
  1. Identify similarities or differences between temporal and textual annotations from different annotators.
  2. To what extent do the annotations rely on or deviate from keywords and textual descriptions in the audio's metadata?
  3. Was the temporal and text annotations done according to the task description?

## What to Investigate 2/5

- **Annotation Quality** (6 points): Use the audio recordings annotated by multiple annotators to answer the following questions:
  1. How precise are the temporal annotations?
  2. How similar are the text annotations that correspond to the same region?

Use the complete data set (or a subset) to address the following points quantitatively.

1. How many annotations did we collect per file? How many distinct sound events per file?
2. Does the quality of annotations vary between different annotators? Are there any obvious inconsistencies, outliers, or poor-quality annotations in the data? Propose a simple method to filter or fix incorrect or poor-quality annotations.

## What to Investigate 3/5

■ **Audio Features** (6 points): Load and analyze the audio features:

1. Which audio features appear useful? Select only the most relevant ones or perform a down projection for the next steps.
2. Extract a fixed-length feature vector for each annotated region as well as for all the silent parts in between. The most straightforward way to do this is to average the audio features of the corresponding region over time, as shown in the tutorial session.
3. Cluster the audio features for the extracted regions. Can you identify meaningful clusters of audio features? Do the feature vectors of the silent regions predominantly fall into one large cluster?



## What to Investigate 4/5

- **Text Features** (6 points): Load and analyze the text features:
  1. Cluster the text features. Can you find meaningful clusters?
  2. Design a labeling function<sup>3</sup> for classes *dog* and *cat*. Do the annotations labeled as dog or cat sounds form tight clusters in the text and audio feature space?
  3. How well do the audio feature clusters align with text clusters?

---

<sup>3</sup>A function that takes a class of interest and a textual annotation (or text features) as input and returns `True` if the annotation refers to the given class and `False` otherwise.

## What to Investigate 5/5

- **Conclusions** (2 points): What conclusions can you draw from your analysis for the next phases of the project?
  1. Is the dataset useful to train general-purpose sound event detectors?
  2. Which biases did we introduce in the data collection and annotation phase?

# Data Exploration Task: Report

Compile your results into a short report.

- Cover **all** of the five previous aspects, the corresponding subquestions, and other interesting things that you've discovered.
- You may use any kind of **statistical computation** or **visualization** that suits your purpose
- use the **L<sup>A</sup>T<sub>E</sub>X template** that is available on Moodle
- max. **5 pages** (including tables, figures)
- max. **60% text** (= 3/5 pages; rest tables, figures, ...)
- include a **statement** about the **contributions** of each team member

# Data Exploration Task: Slides

In addition to the detailed report, compile a short presentation

- Cover **one** of the five previous aspects and subquestions
- Your topic is determined by the first letter of your group name (see table below)
- max. 4 slides + 1 title slide

First Letter of Group Name	Topic
A, C, E, M, Q	Case Study & Conclusion
B, F, I, L, N, P	Annotation Quality
D, G, J, R, T, U, W	Audio Features
H, K, O, S, V, Y, Z	Text Features

# Data Exploration Task: Submission

- Submit your report and slide deck as two separate PDF files via Moodle by April 24th, 23:59.
- Selected groups will be asked to present their results in class on April 28th.
- At least one team member must be available to present in-person or via Zoom.

# Data Exploration Task: Grading

- Completing all tasks **is mandatory** to pass the course!
- The report is worth **22 points** and the slides **3 points**
- **Grading criteria** for the report in the task description on Moodle
- Submitting a day late will cost you  $\frac{1}{3}$  of the total points:
  - ☐ Up to April 24th, 24:00: 100 %
  - ☐ April 25th 00:00–24:00: 66.66%
  - ☐ April 26th 00:00–24:00: 33.33%
  - ☐ Afterwards, we will not accept submissions.

# Data Exploration Task: Summary

- Completion of Task 2 is mandatory
- Answer all **five aspects** and the **corresponding questions** in your **written report**
- Use the **L<sup>A</sup>T<sub>E</sub>X** template, stick to the **page limit** (5 pages, 3 pages text max.) and include a **statement of contributions**
- Create a **slide deck** which tackles the **one aspect assigned to your group**. 4 slides + 1 title slide max.
- Upload both until **April 24th** to get up to **25 points**