

# Sound events classification with CNN and data augmentation

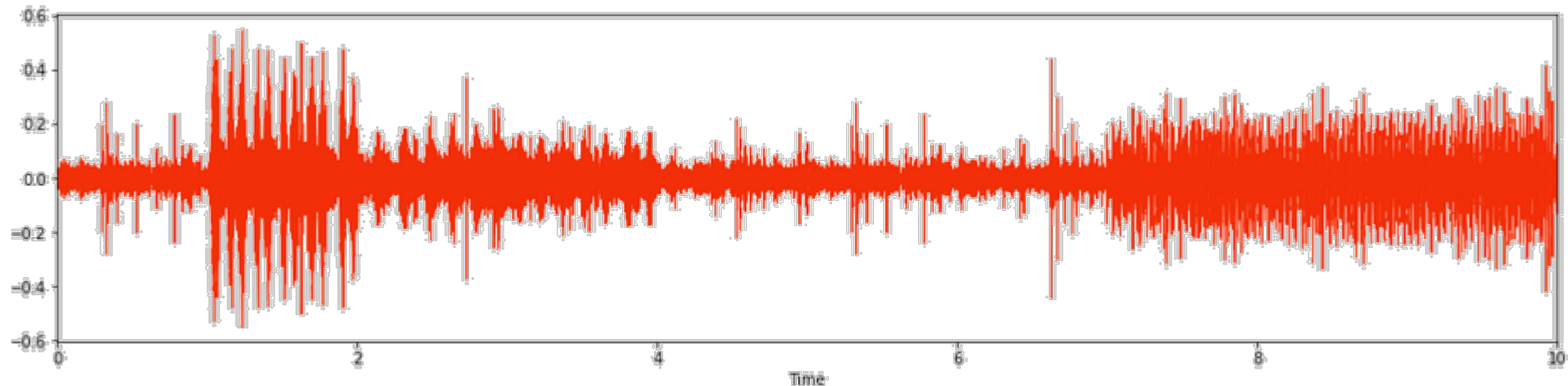
Machine-Learning-Artificial-Intelligence Meetup

Bern, Monday, August 27th 2018

Christophe Lesimple, Bernafon AG, EPFL

# | Sound Event Classification

- Source identification or event retrieval
- Sound event segmentation



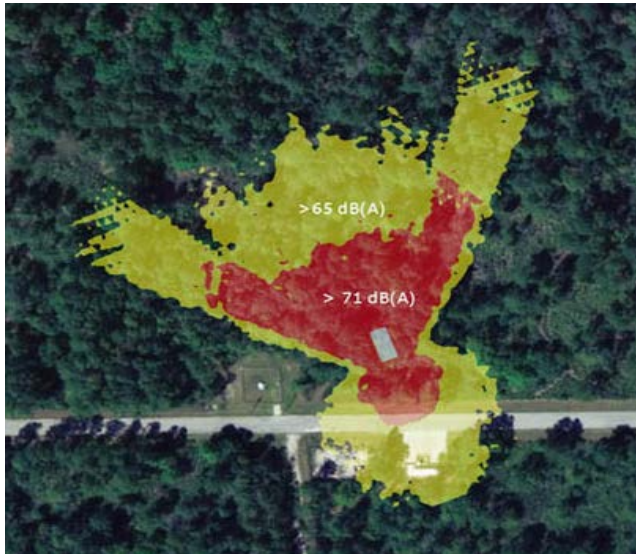
bird

rain

car

# | Sound Classification: Applications

- Environmental sound/noise: qualitative measures <sup>1</sup>
- Medicine / Machine: diagnostic of pathological sounds <sup>2</sup>
- Hearing device: adjustment of amplification <sup>3</sup>



# | Hierarchy of classes

- Within class <sup>4</sup>



- Between classes <sup>5</sup>

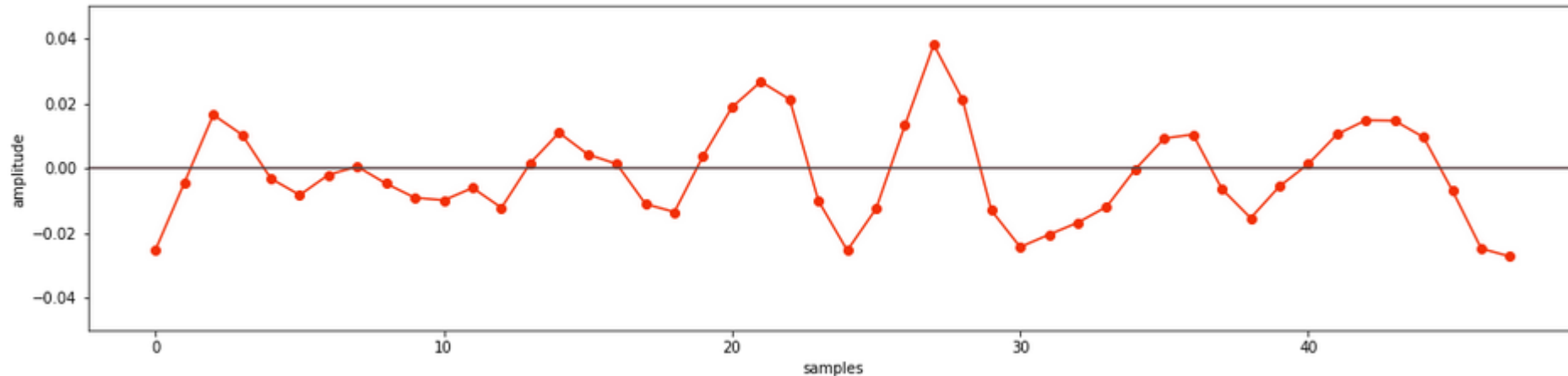


- Combined classes as soundscapes <sup>6</sup>



# | From a sound to 1D data

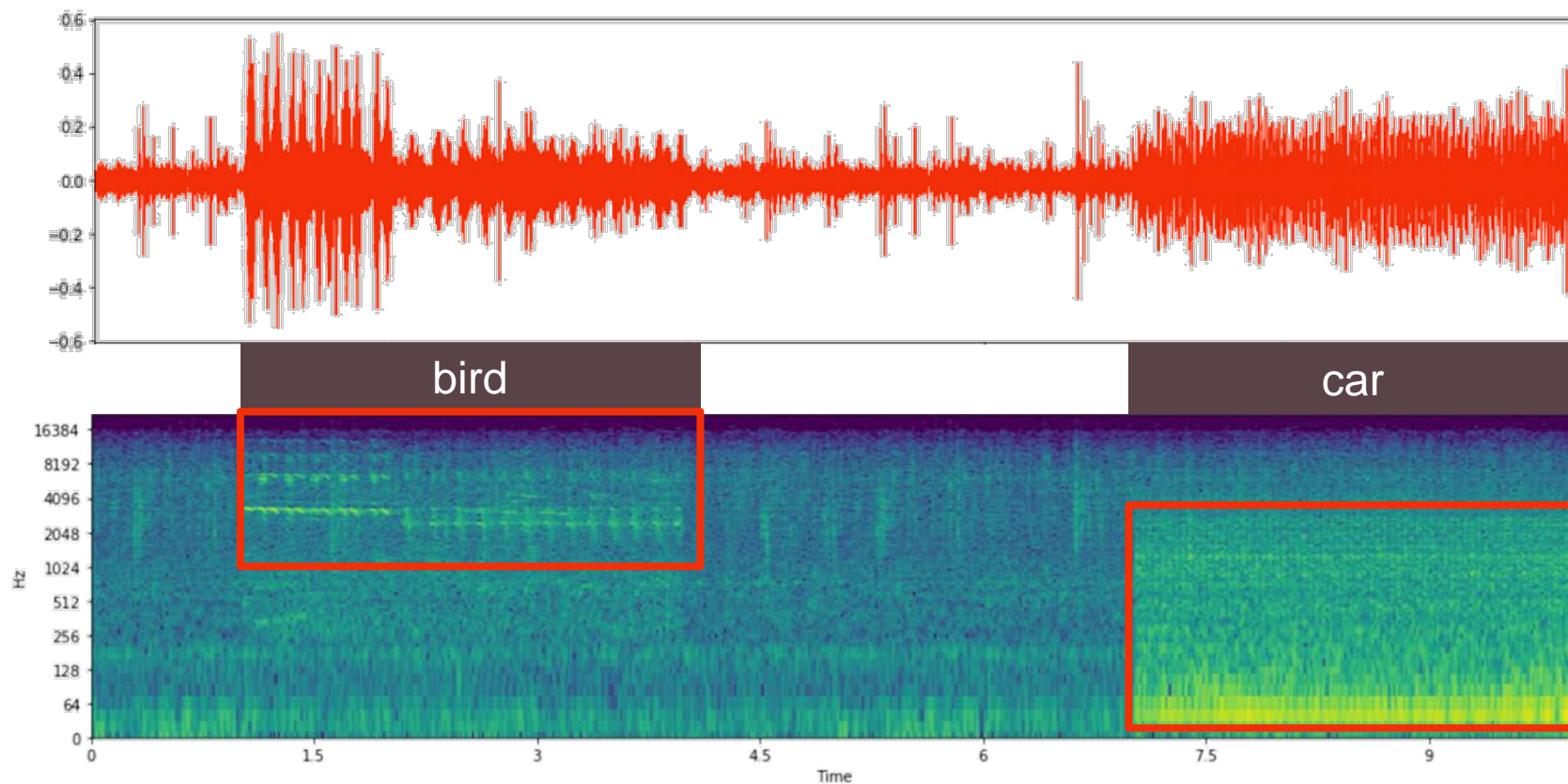
- Wavefile: amplitude variations over time
- Sampling frequency:
  - time resolution @ 44.1 kHz, 50 samples ~ 1.1 ms
  - influence the bandwidth @ 44.1 kHz,  $f_{\max} = 22.05\text{kHz}$
  - large vectors without all the information





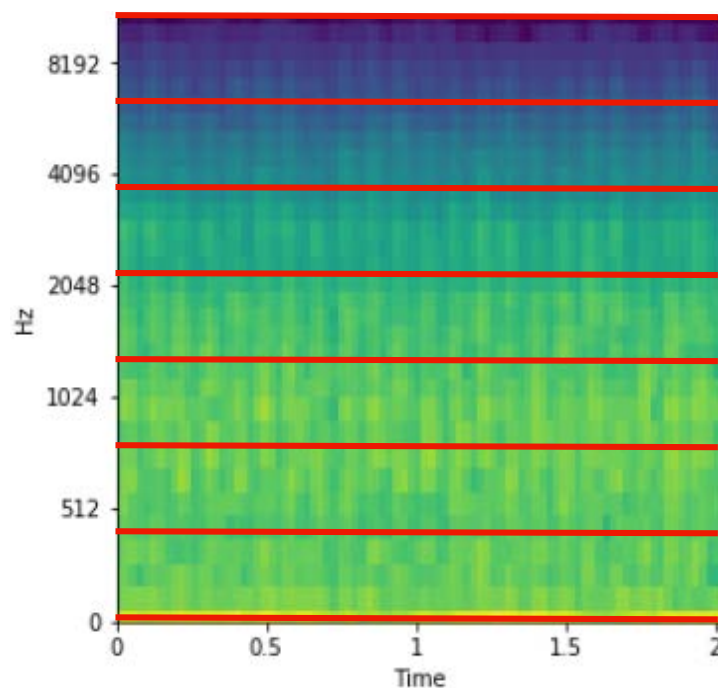
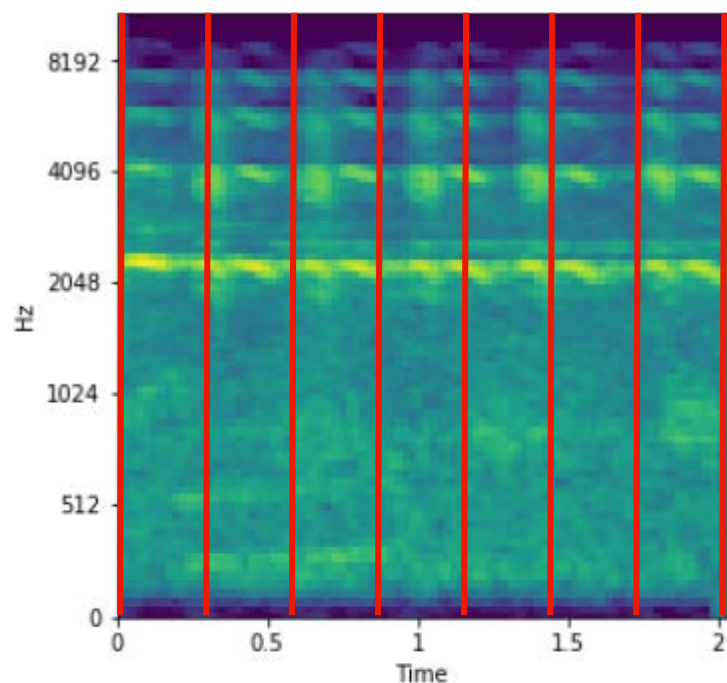
# | From sound to 2D data

- Convert the acoustic signal in time-frequency domain <sup>7</sup>:



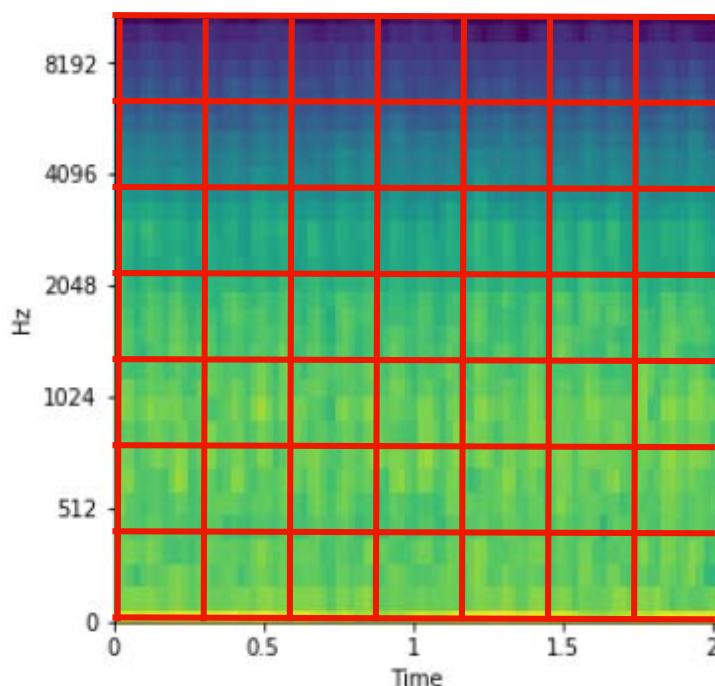
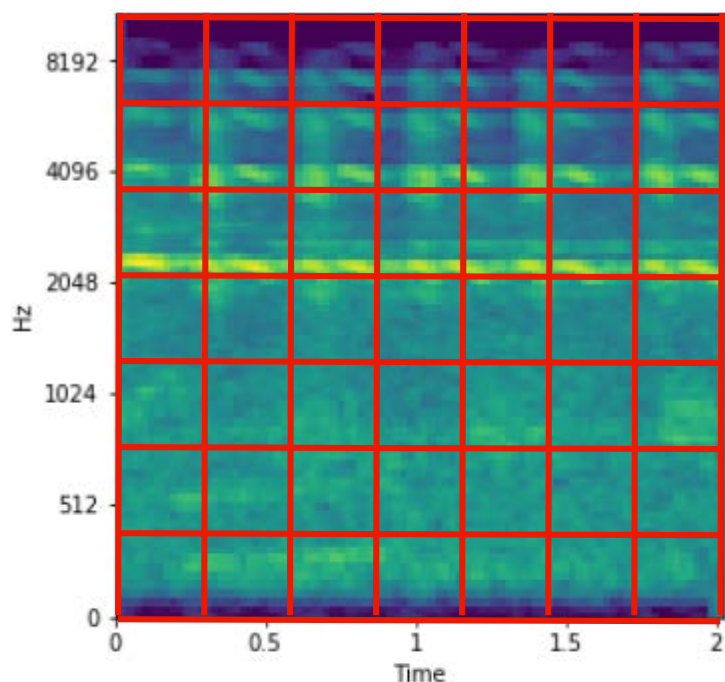
# | Data dimension vs. time/frequency

- Time resolution with ``hop_length`` → frames
- Frequency resolution with ``n_mfcc`` → bins



# | Data dimension vs. time/frequency

- Time resolution with `hop\_length` → frames
- Frequency resolution with `n\_mfcc` → bins



Adjust parameters based on sound source attributes e.g.:

- Modulation rate
- Frequency content

Same process for classification as image <sup>8</sup>

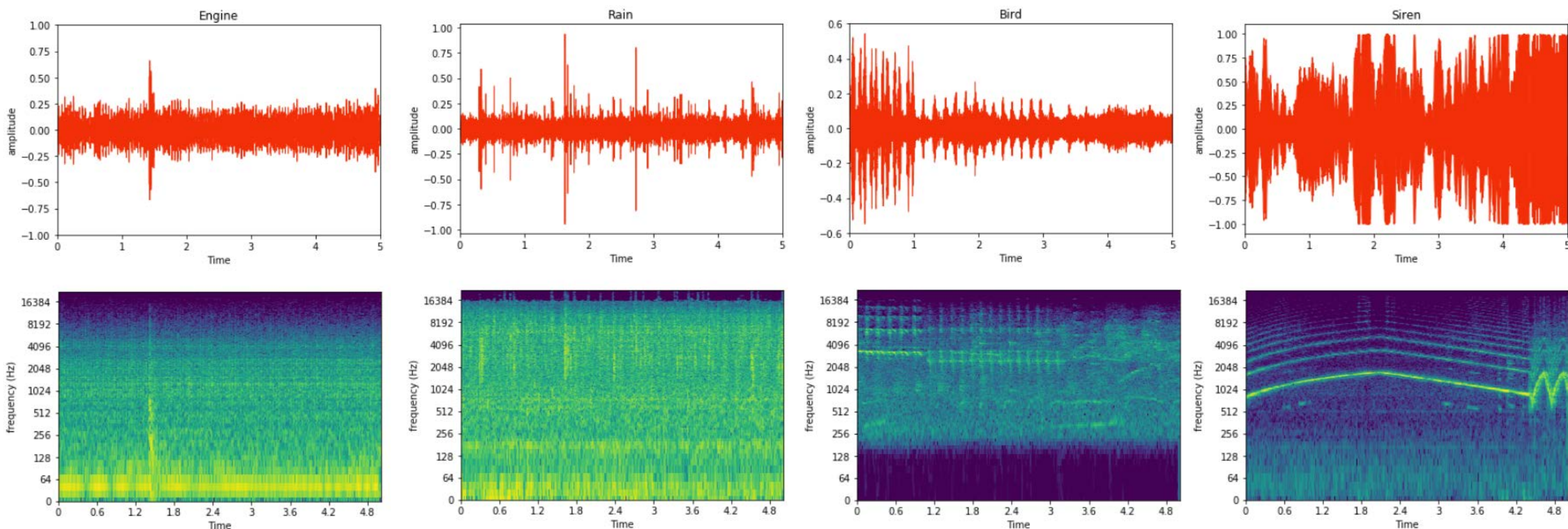


# | Dataset from ESC 50

- Sounds from the [freesound.org](https://freesound.org) project, 5 seconds long
- Selection of 10 classes:
  - rain, sea waves, wind, crickets, birds,
  - car horn, train, siren, engine, church bells.
- Source<sup>5</sup>: [github](#)
- 40 samples per classes split in 80/20:
  - Training / validation set,
  - Test set.

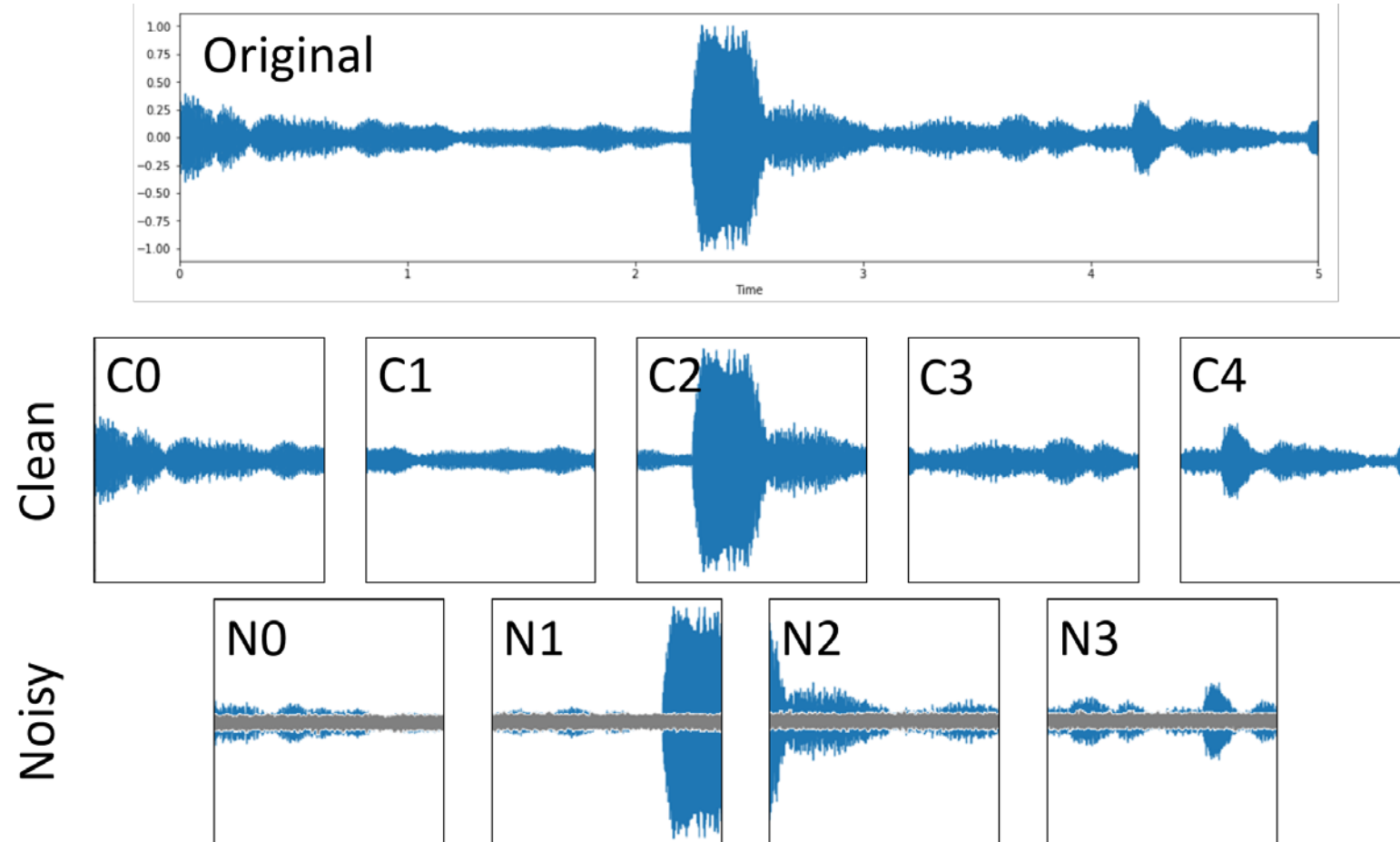
# | Supervised learning approach

- Features extracted from the wavefile
- Mapping between features and labels



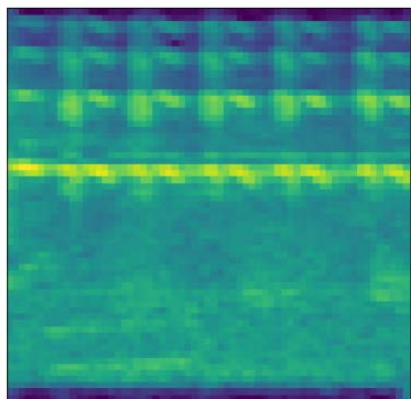
# | Data segmentation / augmentation

- 5s original file segmented in 9 files, 1s each,
- Data augmentation<sup>9</sup> by adding noise to each odd sample,
- Helps the CNN to see more relevant patterns at once and faster convergence.

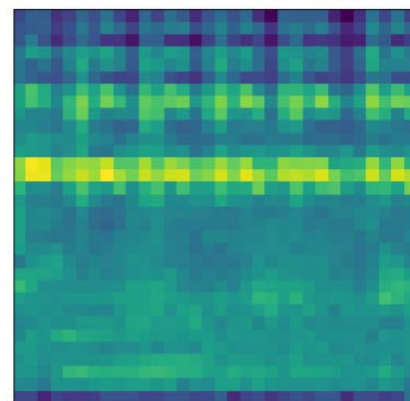


# | Features dimensions

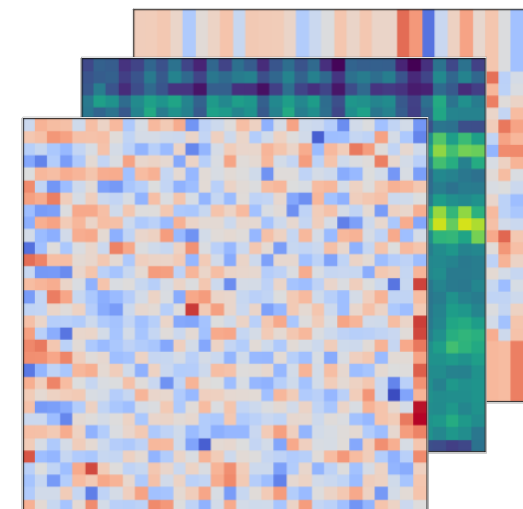
- Model 1: 63 x 63
- Model 2: 32 x 32
- Model 3 <sup>10</sup>: 32 x 32 x 3



- hop\_length 512
- 63 MFCCs



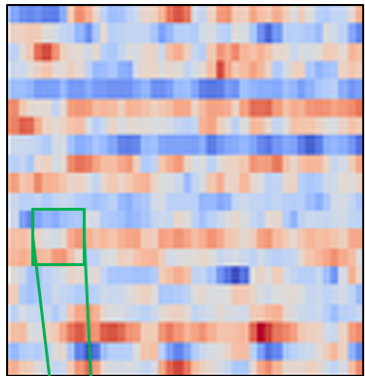
- hop\_length 1024
- 32 MFCCs



- Model 2 +
- Delta MFCCs
- Mel-spectrogram

Mel-Frequency Cepstral Coefficient

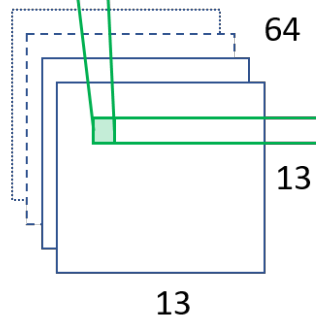
# CNN with 2d convolution



Input: scaled MFCCs from selected 1s short sequence with data augmentation.  
Dimensions 63 x 63 x 1

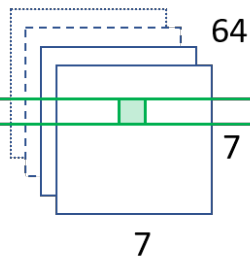
## Convolution 1:

kernel: 7 x 7  
stride: 5 x 5  
Padding: same  
ReLU activation



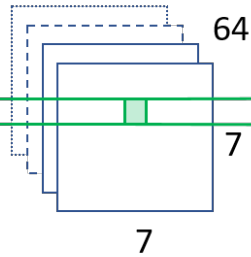
## Max pooling 1:

Pool size: 2 x 2  
stride: 2 x 2  
Padding: same



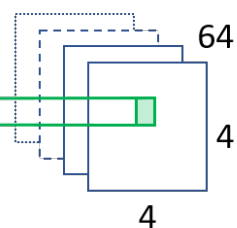
## Convolution 2:

kernel: 3 x 3  
stride: 1 x 1  
Padding: same  
ReLU activation

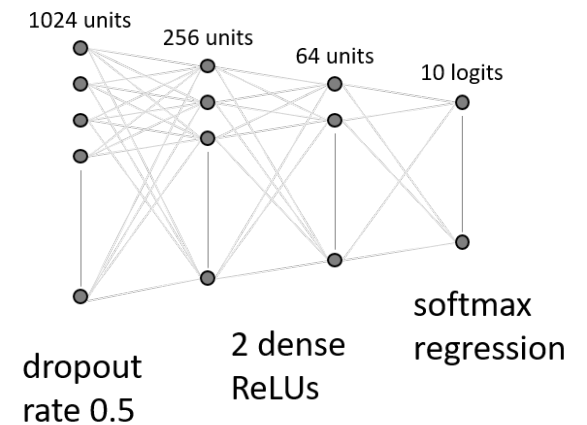
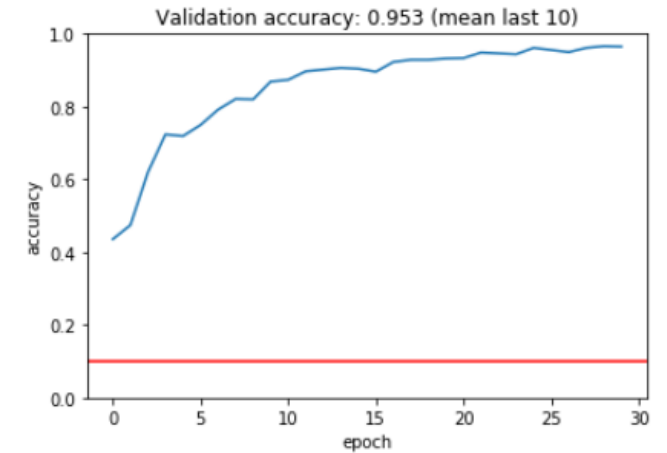


## Max pooling 2:

Pool size: 2 x 2  
stride: 2 x 2  
Padding: same



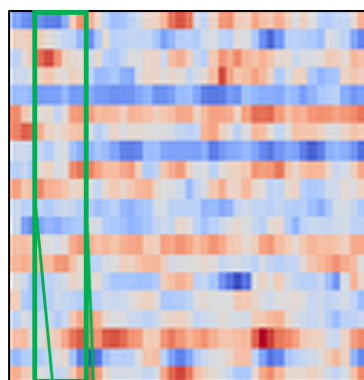
2 convolutional/max pooling layers



dropout - 3 fully connected layers



# CNN with 1d convolution



Input: scaled MFCCs from selected 1s short sequence  
with data augmentation.  
Dimensions 63 x 63

Convolution 1:

kernel: 5  
stride: 3  
Padding: same  
ReLU activation

64

21

Max pooling 1:

Pool size: 2  
stride: 2  
Padding: same

64

11

Convolution 2:

kernel: 3  
stride: 1  
Padding: same  
ReLU activation

64

11

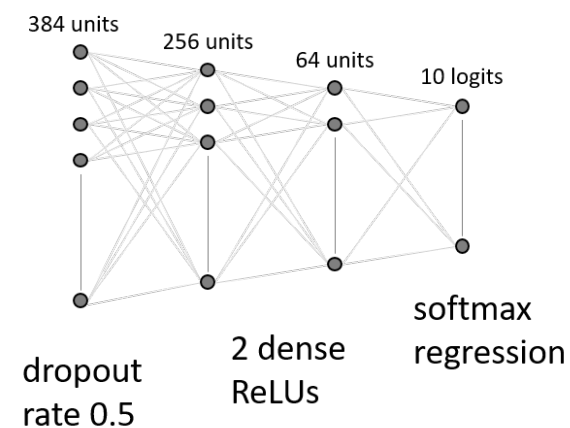
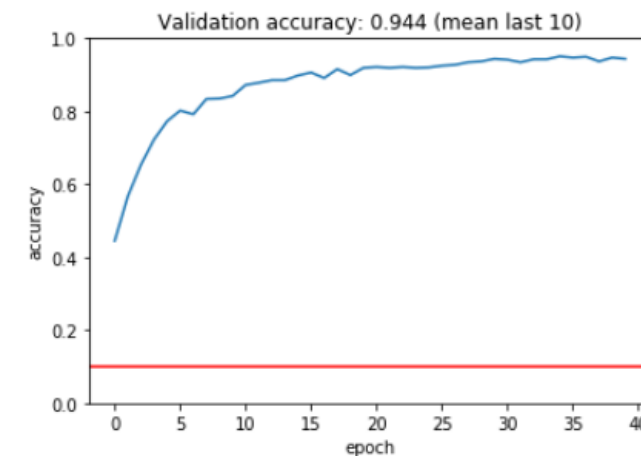
Max pooling 2:

Pool size: 2  
stride: 2  
Padding: same

64

6

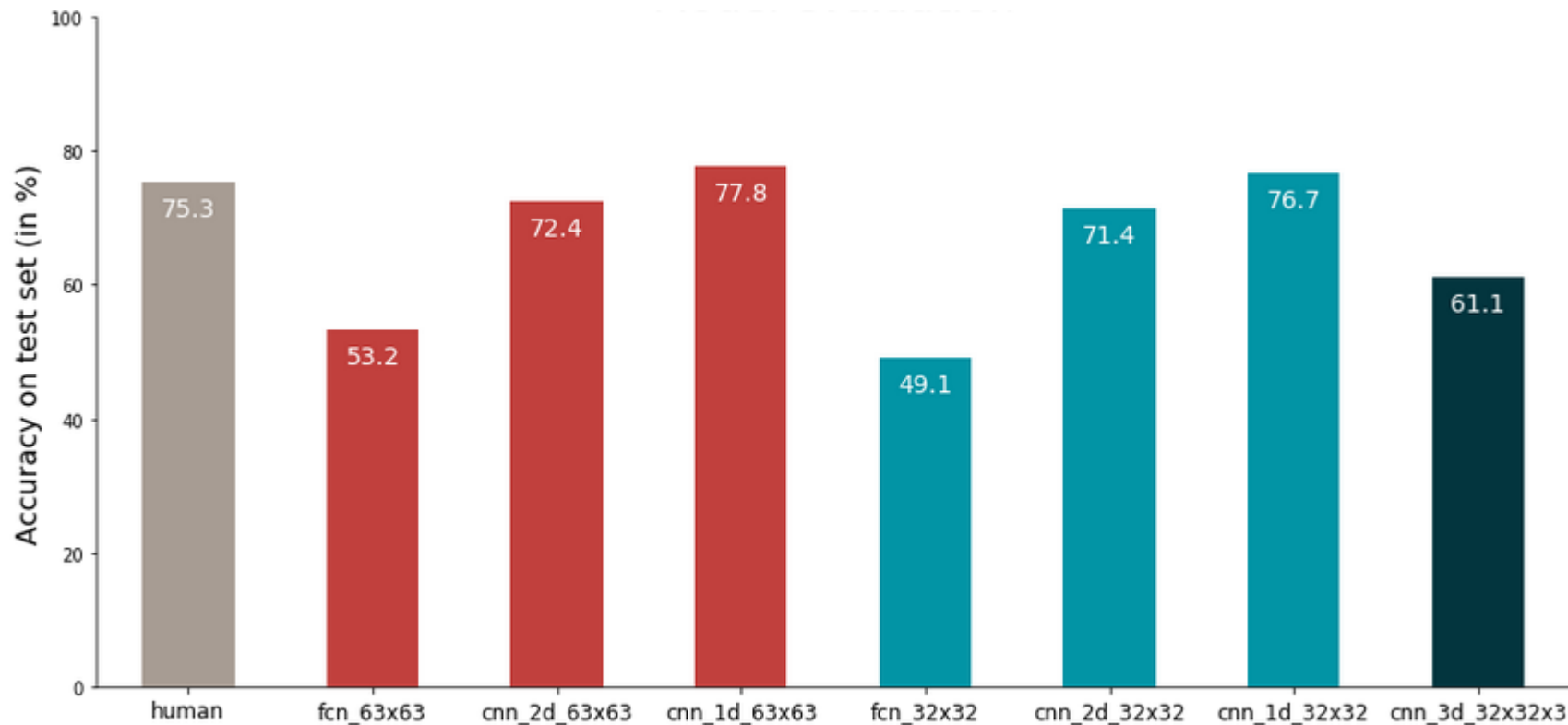
2 convolutional/max pooling layers



dropout - 3 fully connected layers

# | Results with test set

- 1d convolution produces best results
- Reducing feature dimensions has a minor impact on accuracy
- Finer resolution might be needed for a within class classification task



# | References

1. Mijala et al. (2018), Environmental noise monitoring using source classification in sensors. *Applied Acoustics*, Volume 129, Pages 258-267.
2. Redlarski, G., Gradolewski, D., & Palkowski, A. (2014). A System for Heart Sounds Classification. *PLoS ONE*, 9(11), e112673.
3. Nordqvist, P. & Leijon, A. (2004), An efficient robust sound classification algorithm for hearing aids. *J Acoust Soc Am*. Jun;115(6):3033-41.
4. Xie et al. (2018), Acoustic classification of frog within-species and species-specific calls. *Applied Acoustics*, Volume 131, Pages 79-86.
5. Piczak, K. (2015), ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015-1018, ACM.
6. Mesaros, A., Heittola, T., and Virtanen, T. (2018) A multi-device dataset for urban acoustic scene classification. *DCASE2018 Workshop*.
7. Mitrovic, D., Zeppelzauer, M., & Breiteneder, C. (2010), Chapter 3 - Features for Content-Based Audio Retrieval. *Advances in Computers*, Elsevier, Volume 78, Pages 71-150.
8. Hershey et al. (2017), CNN Architectures for Large-Scale Audio Classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 131-135. 10.1109.
9. Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3), 279-283.
10. Boddapati, V. et al. (2017), Classifying environmental sounds using image recognition networks. *Procedia Computer Science*, Volume 112, Pages 2048-2056.