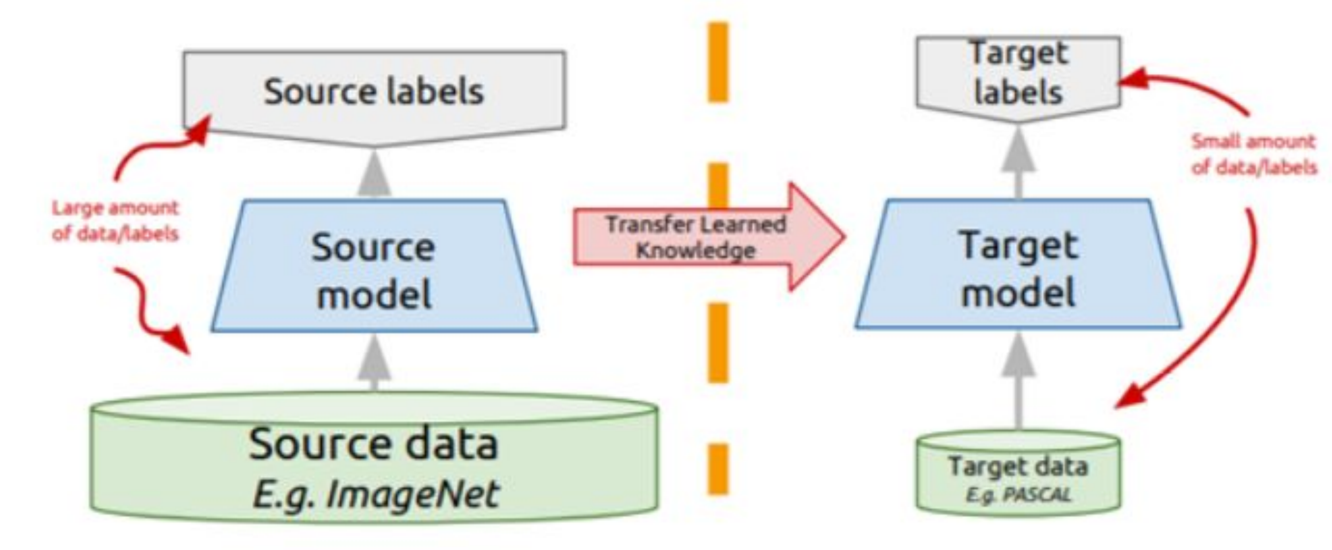
Atelier Machine Learning

Transfer apprentissage pour détection de scènes acoustiques biologiques sous-marine

Paul Nguyen et Dorian Cazau

Transfer learning

"the situation where what has been learned in one setting is exploited to improve generalization in another setting"



Useful when

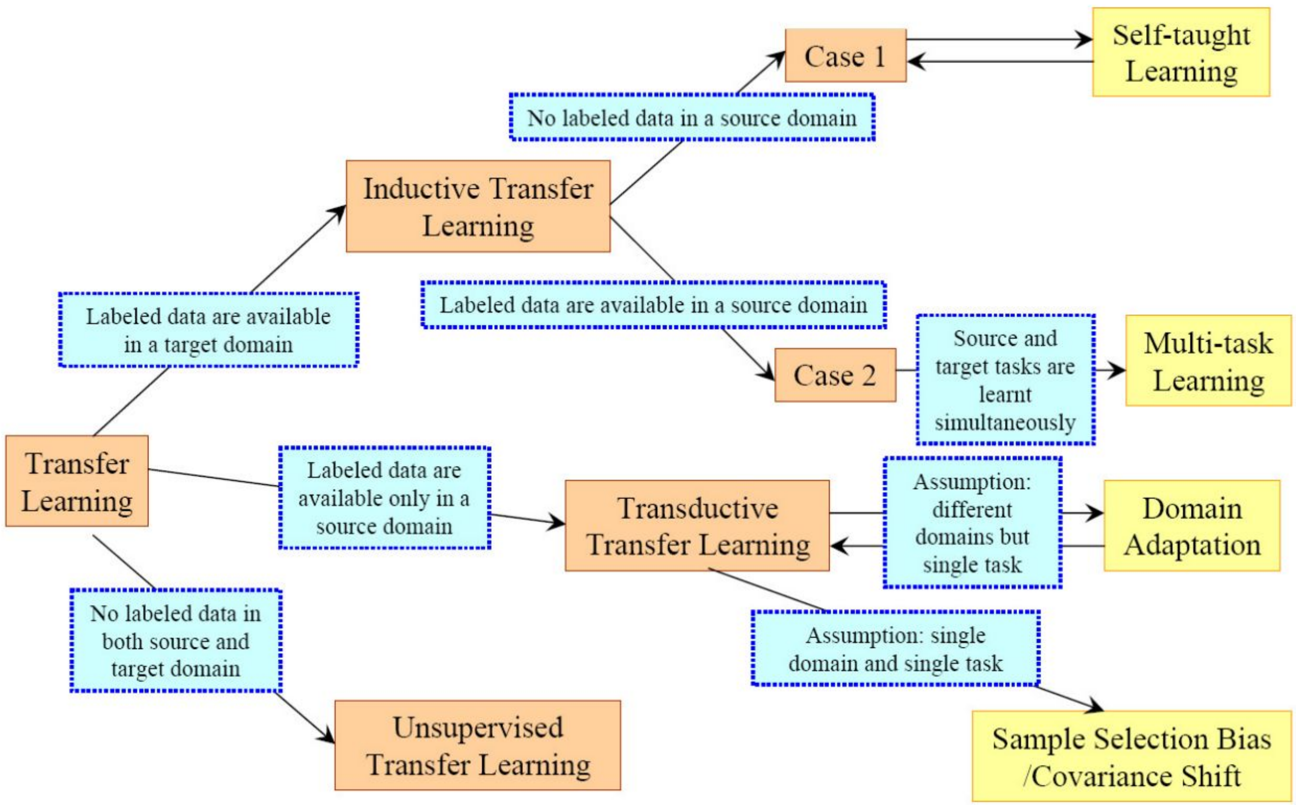
- small annotated datasets
- events are inherently rare

Pan and Yang (2010) "A survey on transfer learning," IEEE Transactions on knowledge and data engineering

Different settings of transfer

inductive transfer learning (task adaptation)

transductive transfer learning (domain adaptation)



Pan and Yang (2010) “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering

Google Audio set

Big but “weakly labeled” SOURCE audio dataset

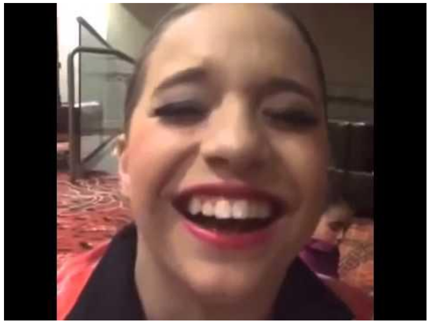
Weakly labeled data (= “somewhere within this temporal region there is a sound of interest occuring”)

2,084,320 YouTube videos over 527 classes Cs

- **Human sounds**
  - *Human voice*
  - *Whistling*
  - *Respiratory sounds*
  - *Human locomotion*
  - *Digestive*
  - *Hands*
  - *Heart sounds, heartbeat*
  - *Otoacoustic emission*
  - *Human group actions*
- **Animal sounds**
  - *Domestic animals, pets*
  - *Livestock, farm animals, working animals*
  - *Wild animals*
- **Natural sounds**
  - *Wind*
  - *Thunderstorm*
  - *Water*
  - *Fire*
- **Music**
  - *Musical instrument*
  - *Music genre*
  - *Musical concepts*
  - *Music role*
  - *Music mood*
- **Sounds of things**
  - *Vehicle*
  - *Engine*
  - *Domestic sounds, home sounds*
  - *Bell*
  - *Alarm*
  - *Mechanisms*
  - *Tools*
  - *Explosion*
  - *Wood*
  - *Glass*
  - *Liquid*
  - *Miscellaneous sources*
  - *Specific impact sounds*
- **Source-ambiguous sounds**
  - *Generic impact sounds*
  - *Surface contact*
  - *Deformable shell*
  - *Onomatopoeia*
  - *Silence*
  - *Other sourceless*
- **Channel, environment and background**
  - *Acoustic environment*
  - *Noise*
  - *Sound reproduction*

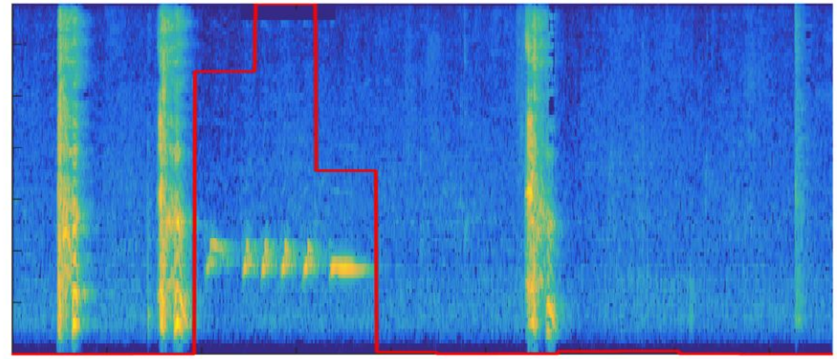
Cs = Bird vocalization, Insect, Chirp, Cricket

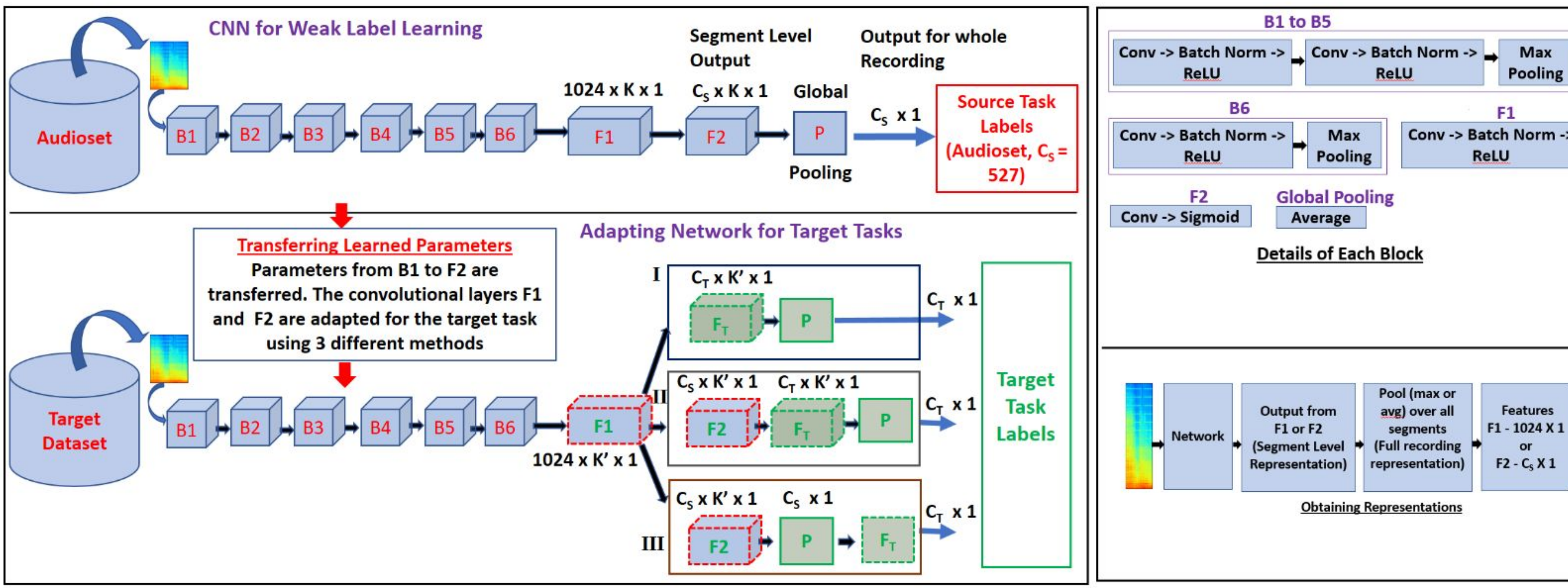
Cs = Music, Speech, Female singing, Child singing



Gemmeke et al. (2017) “Audio set: An ontology and human-labeled dataset for audio events,” in IEEE ICASSP

Kumar2018: State-of-the-art in DCASE





Number of filters: {B1 : 16, B2 : 32, B3 : 64, B4 : 128, B5 : 256, B6 : 512}

Kumar et al. (2018) “Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes”, arxiv

Results

MAUC		MAP	
$\mathcal{N}_{\mathrm{S}}^{\mathrm{slat}}$	$\mathcal{N}_{\mathrm{S}}$	$\mathcal{N}_{\mathrm{S}}^{\mathrm{slat}}$	$\mathcal{N}_{\mathrm{S}}$
0.915	0.927 (**+1.3%**)	0.167	0.213 (**+27.5%**)

Simple trick,
huge improvement !

(slat = strong label assumption training)

Methods	Mean Accuracy
Piczak [20]	64.5 %
Tokozume [30]	71.0 %
Aytar [16]	74.2 %
Proposed (F1)	**83.5 %**

Network	F1		F2	
	$max()$	$avg()$	$max()$	$avg()$
$\mathcal{N}_{\mathrm{S}}$	**82.8**	81.6	65.5	64.8
$\mathcal{N}_{\mathrm{T}}^{\mathrm{I}}$	**83.5**	81.3	–	–
$\mathcal{N}_{\mathrm{T}}^{\mathrm{II}}$	**83.5**	81.8	81.9	81.5
$\mathcal{N}_{\mathrm{T}}^{\mathrm{III}}$	83.3	82.6	82.6	81.9

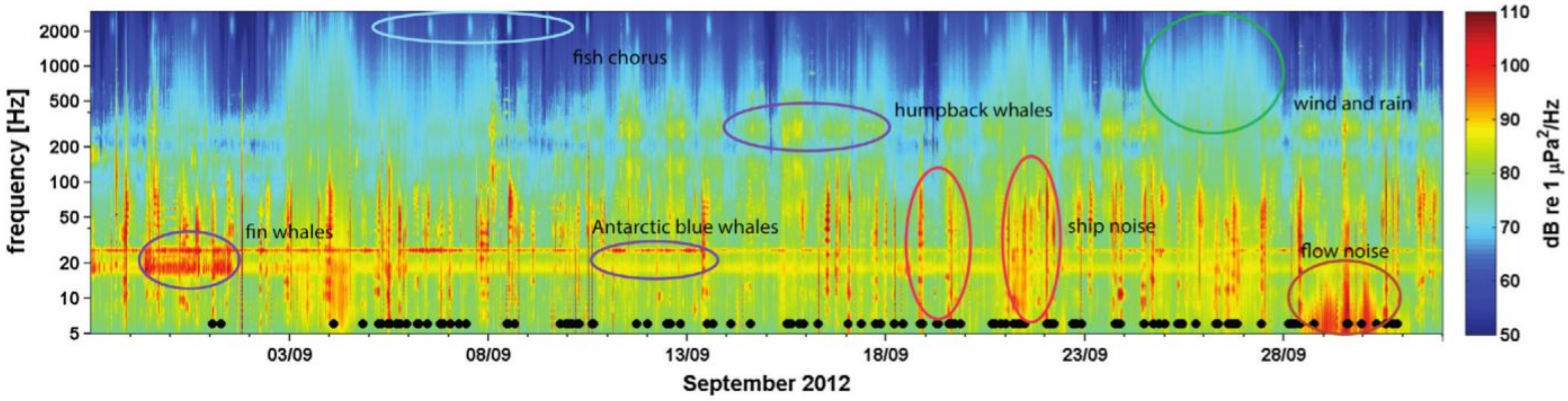
Target tasks

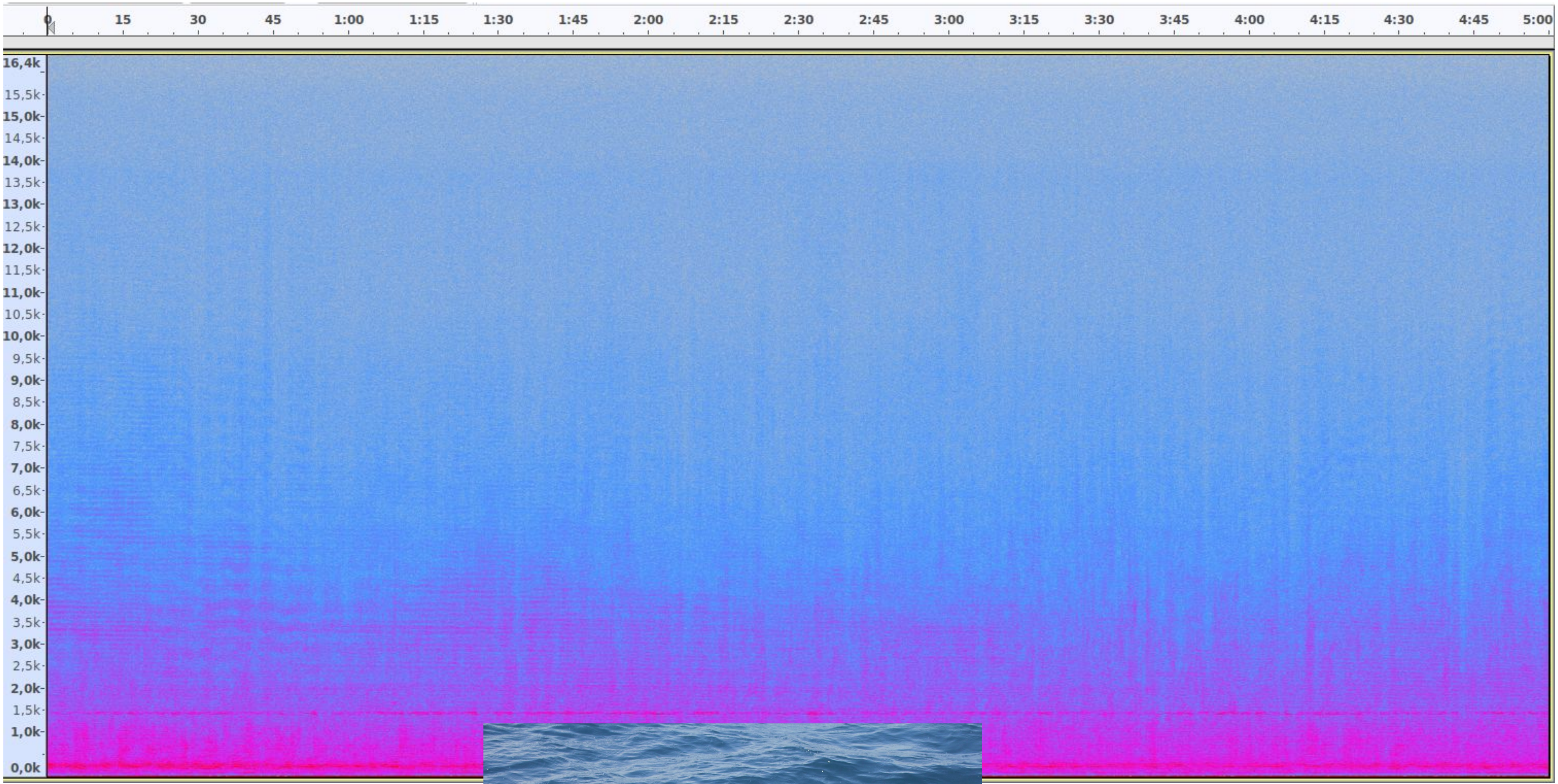
- **Acoustic Event Classification** (ESC-50 dataset: 50 class events from broad categories, Animals , Natural Soundscapes and Water Sounds, Human Non Speech..)
- **Acoustic Scene Classification** (DCASE2016 dataset: 30 seconds examples for 15 acoustic scenes)

Network	F1		F2		Network	F1		F2	
	$max()$	$avg()$	$max()$	$avg()$		$max()$	$avg()$	$max()$	$avg()$
$\mathcal{N}_{\mathrm{S}}$	72.2	69.8	59.1	60.4	$\mathcal{N}_{\mathrm{T}}^{\mathrm{II}}$	75.5	73.0	73.8	73.9
$\mathcal{N}_{\mathrm{T}}^{\mathrm{I}}$	75.2	73.7	–	–	$\mathcal{N}_{\mathrm{T}}^{\mathrm{III}}$	76.6	73.7	72.5	73.3

Kumar et al. (2018) “Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes”, arxiv

A weakly labeled underwater acoustic scene



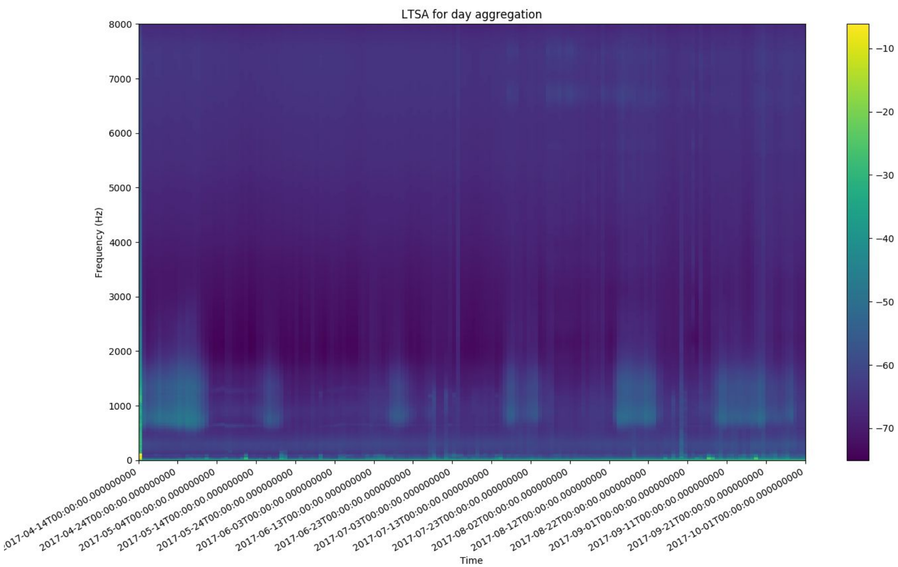

15
30
45
1:00
1:15
1:30
1:45
2:00
2:15
2:30
2:45
3:00
3:15
3:30
3:45
4:00
4:15
4:30
4:45
5:00
16,4k
15,5k
15,0k
14,5k
14,0k
13,5k
13,0k
12,5k
12,0k
11,5k
11,0k
10,5k
10,0k
9,5k
9,0k
8,5k
8,0k
7,5k
7,0k
6,5k
6,0k
5,5k
5,0k
4,5k
4,0k
3,5k
3,0k
2,5k
2,0k
1,5k
1,0k
0,0k

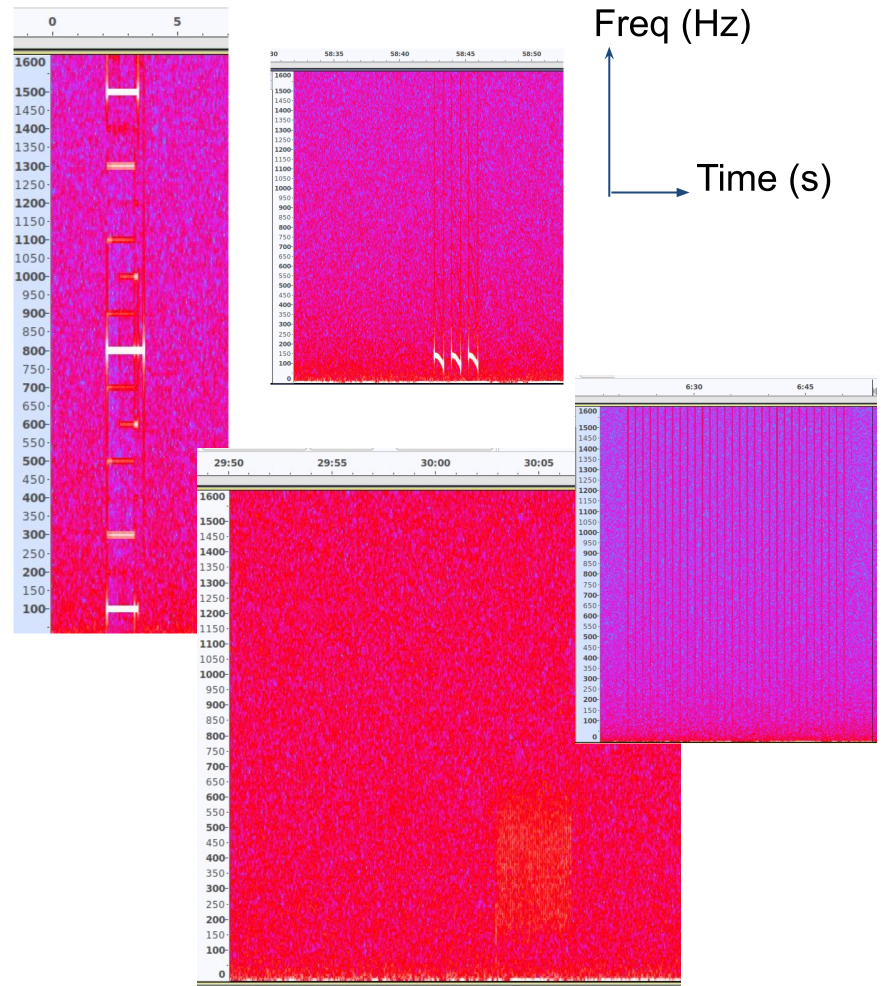

© Fabio Cassiano

??

Datasets	Chagos	Synthetic
Sampling rate	16000Hz	3200 Hz
WAV file duration / recording campaign duration	11min09s / ~6mois	1h / 1 year

Long term spectrogram: having an overview of acoustic activity (spectrum daily-averaged)





Experiments (domain adaptation):

- ➔ Who’s the best?
  - Thresholding (rule-based decision)
  - Supervised machine learning algorithm: Support Vector Machine (SVM) [1]
  - Unsupervised machine learning algorithm: Kmeans
  - Supervised Deep Learning algorithm: Convolutional Recurrent Neural Network (CRNN)

- ➔ Real world VS Synthetic audio data?

WHY DO WE NEED SYNTHETIC DATA

- small annotated datasets
- events are inherently rare

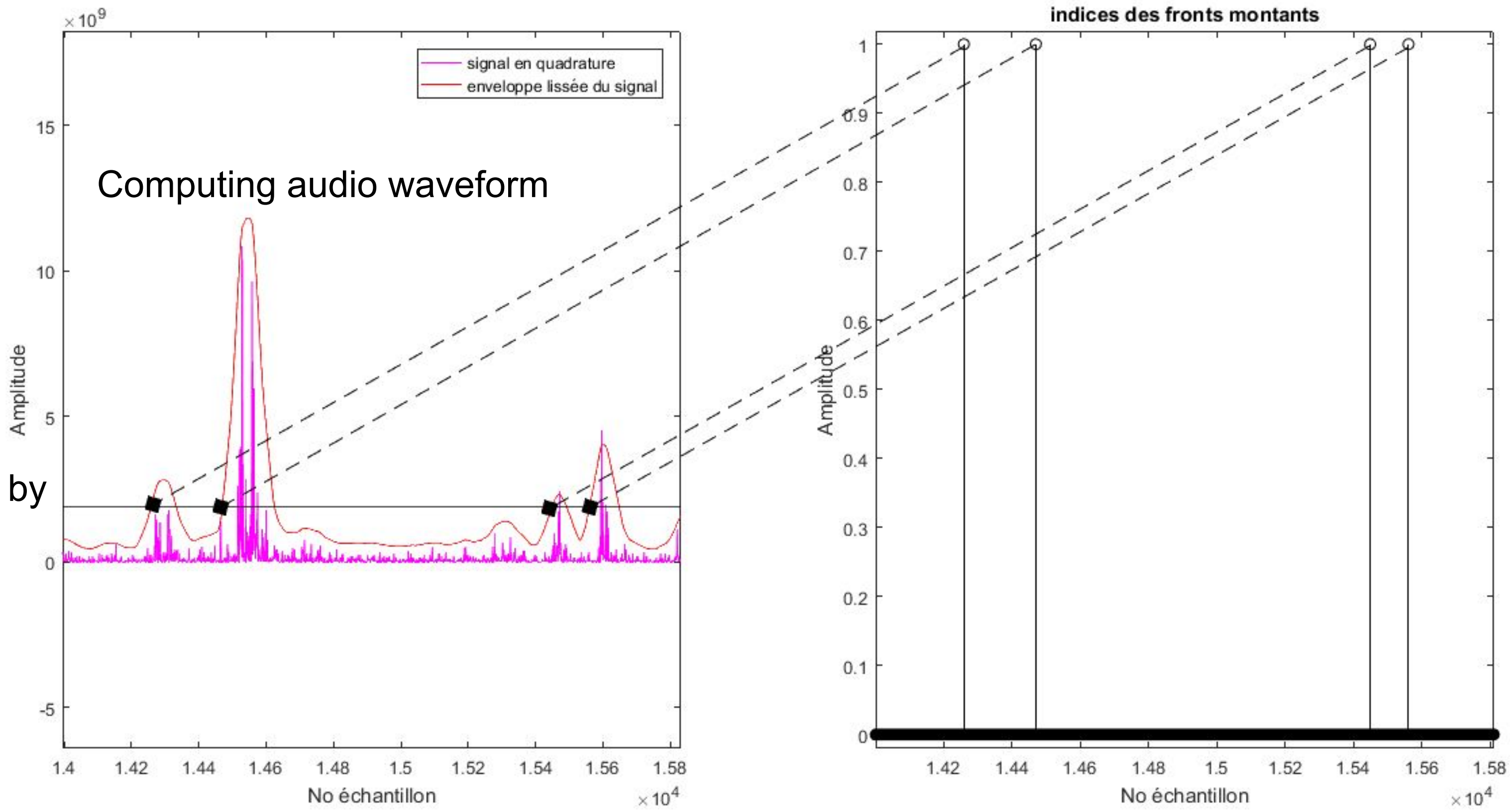
BUT…

- Difficulty in generating synthetic data
- Quality of the data model?
- Inconsistencies when trying to replicate complexities within original datasets
- Difficulty in tracking all necessary features required to replicate the data
- The presence of bias within the synthetic data
- May require validation against real world data
- Simplified representations within datasets can have hidden effects on the performance of an algorithm when used in a real world setting

[1] https://github.com/amlb/amlb.github.io/blob/master/2019-03-12_SVM/presentation.pdf Atelier ML de Clément Dechesne

Thresholding (baseline)

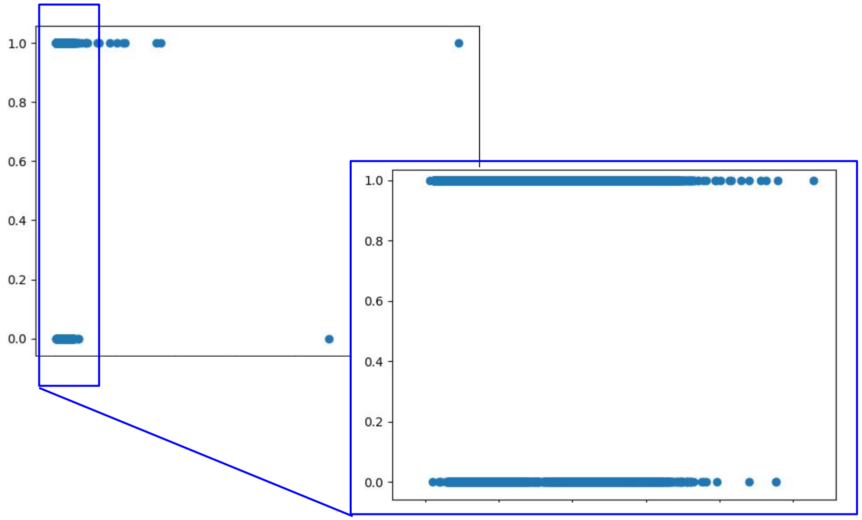
Thresholding by amplitude



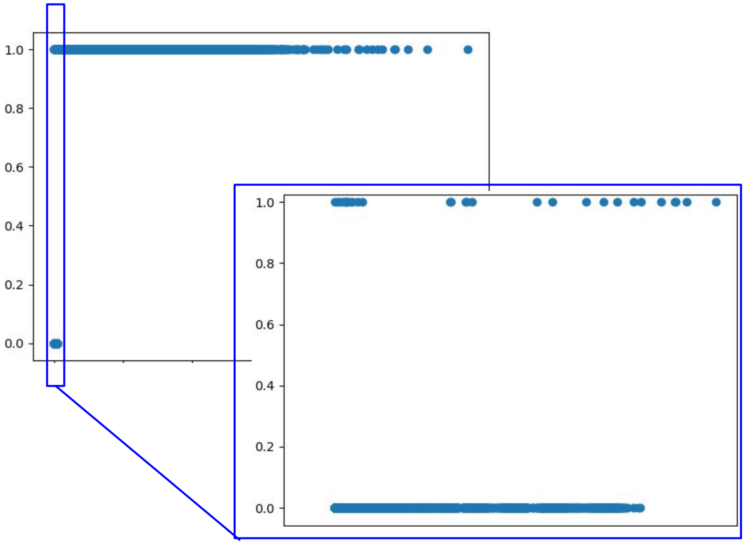
SVM [1]

Computing energy on the whole audio file
Divide presence/absence groups only with the energy

Chagos



Synthetic

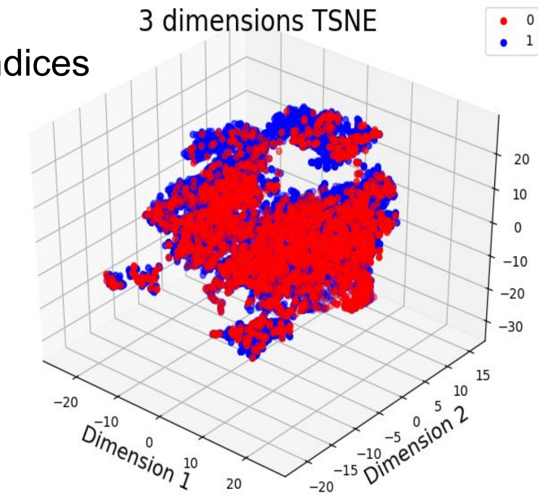


[1] https://github.com/amlb/amlb.github.io/blob/master/2019-03-12_SVM/presentation.pdf Atelier ML de Clément Dechesne

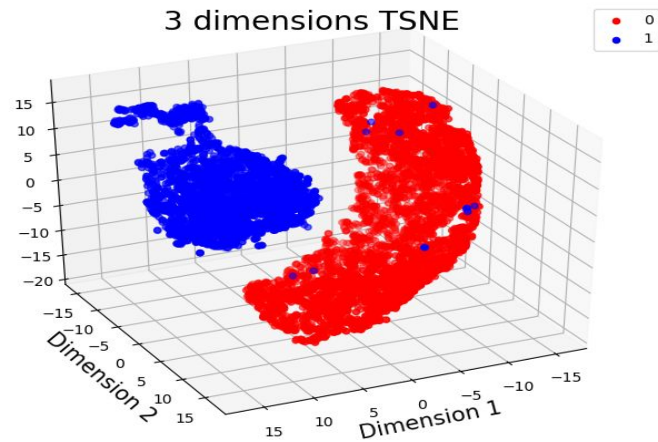
K-means

Computing ecoacoustic indices

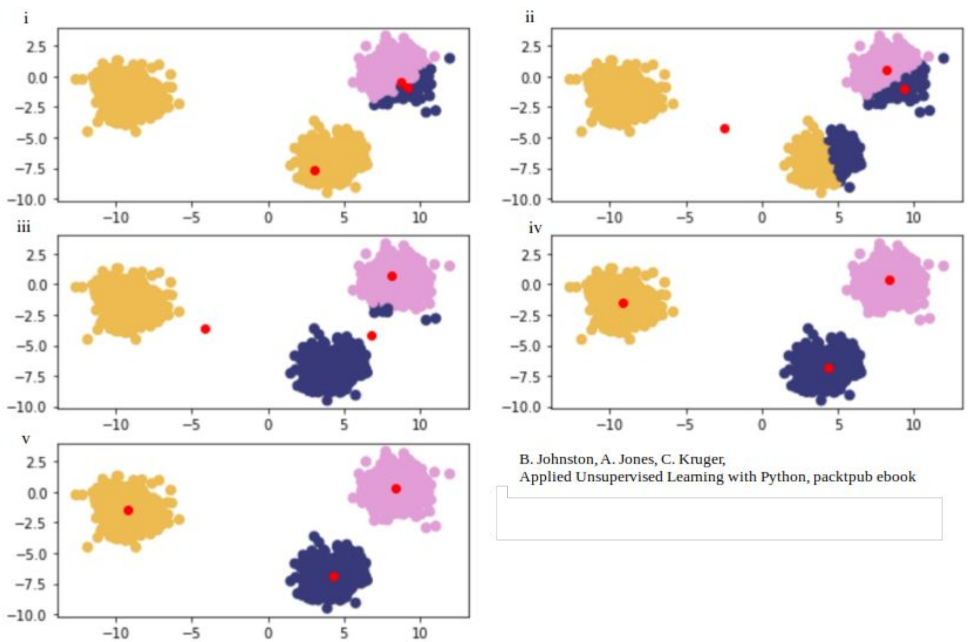
Chagos



Synthetic



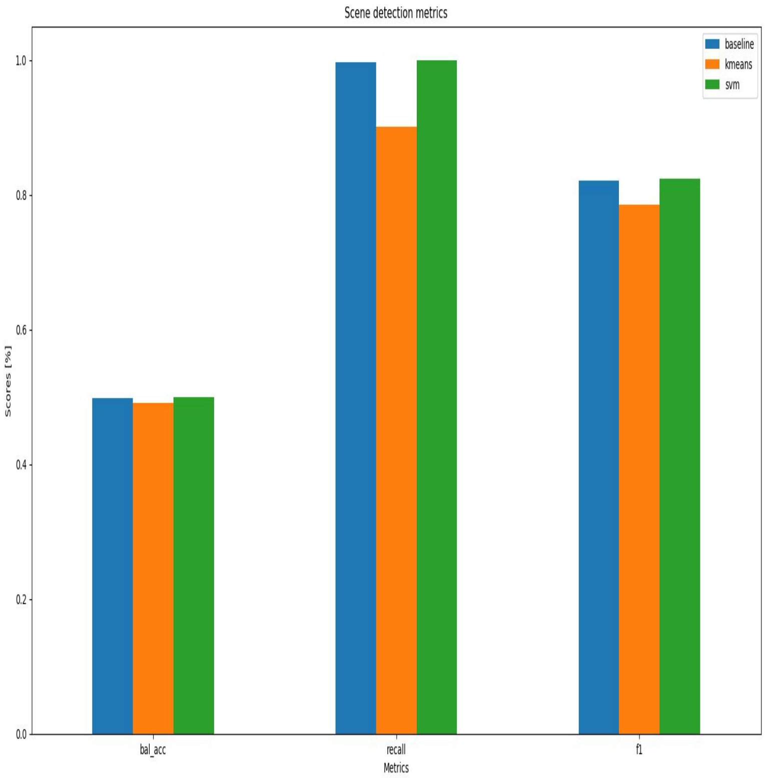
Clustering samples



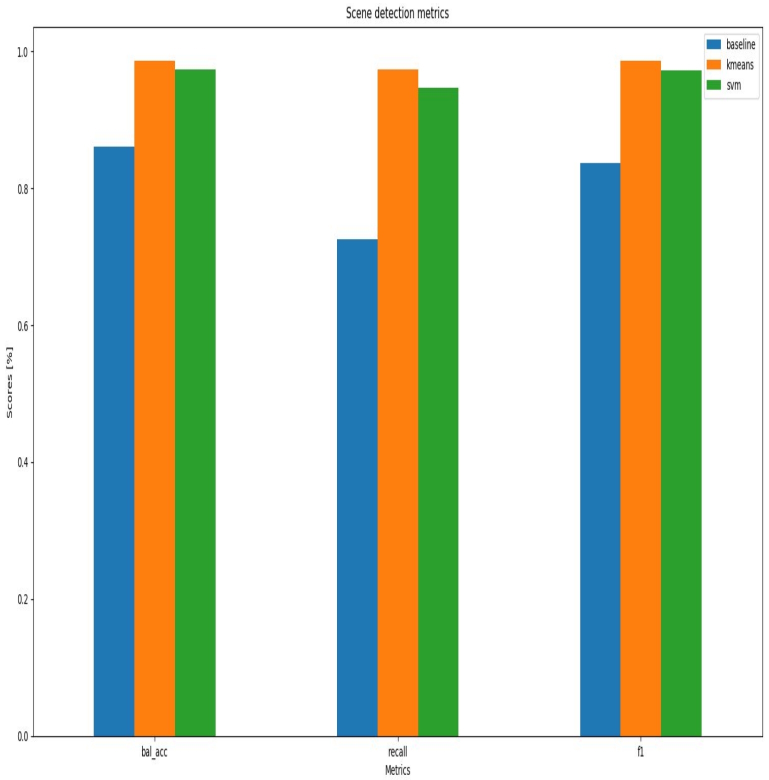
B. Johnston, A. Jones, C. Kruger,
Applied Unsupervised Learning with Python, packtpub ebook

Benchmark

Chagos



Synthetic

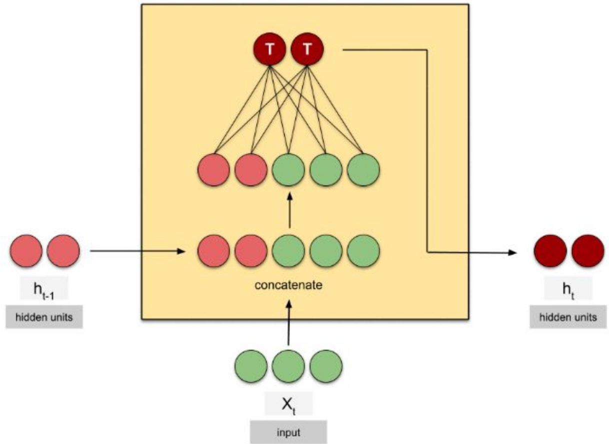


Synthetic -> Chagos: predictions: only zeroes...

CRNN

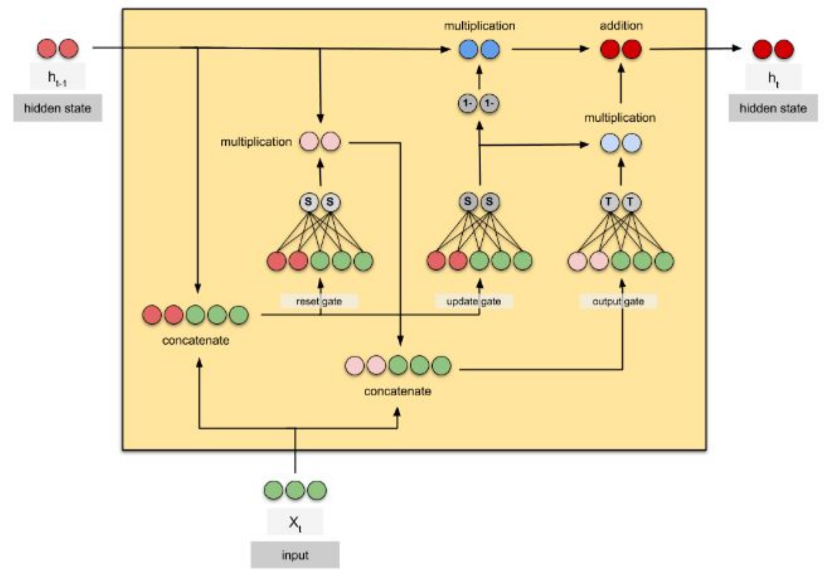
"Capacity to learn the acoustic units of the events with CNN, while the specific temporal order within the events is captured by the following recurrent layers." [2]
Event based but applied with sliding window for scene detection

RNN Layer



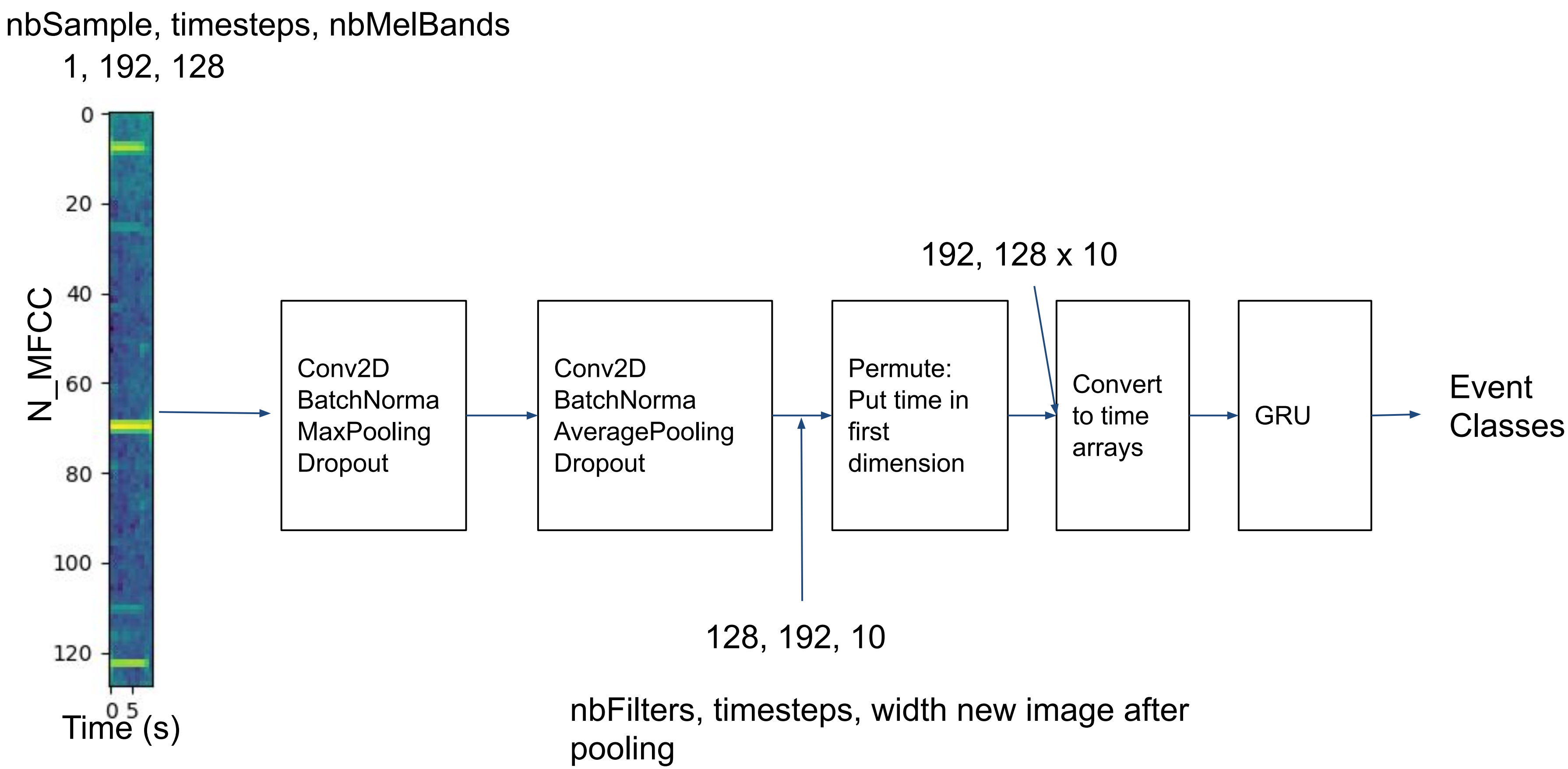
GRU Layer:
3 gates:
- update: how much of the past information should be propagated in the future.
- reset: how much of the past information should be omitted
- output: new internal memory state

Pink Multiplication: store the relevant information from the past.



[2] P. Arora et al., 2017, A Study on Transfer Learning for Acoustic Event Detection in a Real Life Scenario, 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)
https://towardsdatascience.com/animated-rnn-lstm-and-gru-ef124d06cf45

CRNN



[2] P. Arora et al., 2017, A Study on Transfer Learning for Acoustic Event Detection in a Real Life Scenario, 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)

Some issues with underwater audio classification

- Sampling rate
- Duration
- Annotation
- Synthetic data too easy to classify
- Difficult to model underwater environments and sounds