Data Augmentation: Improving your Datasets

Bruno Padovese
MERIDIAN, Institute for Big Data Analytics,
Dalhousie University, Halifax, Canada

Passive Acoustic Monitoring (PAM)

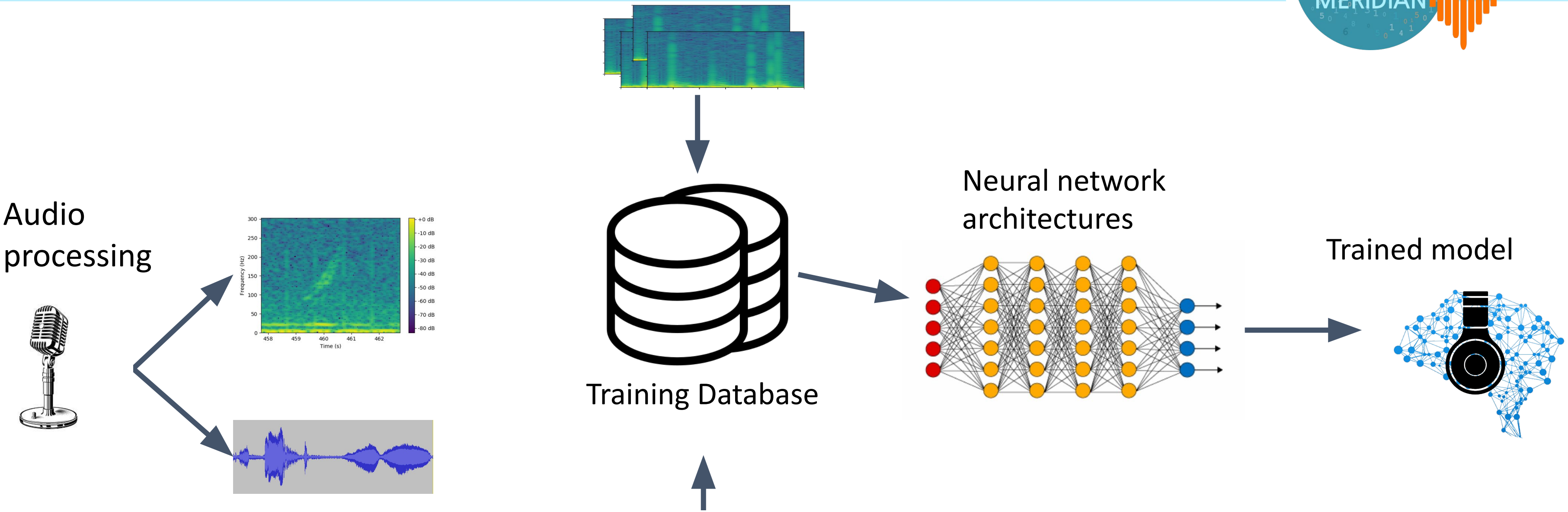
- One of the best ways to monitor marine mammals
  - Capable of months of uninterrupted data collection
  - Efficient way of monitoring large remote areas

- Massive amounts of data generated
  - Easily exceeds capacity for manual labeling

- Automated sound detection and classification systems can help mitigate this problem

Machine learning can help us build these systems

Audio Processing Pipeline



filename	sel_id	label	start	end
NOPP6_EST_20090329_084500.wav	0	1	890.100436	893.100436
NOPP6_EST_20090329_090000.wav	0	1	51.413506	54.413506
	1	1	41.592974	44.592974
	2	1	97.386199	100.386199
	3	1	115.234384	118.234384
	4	1	288.680821	291.680821

Annotation tables

DNN

DNN are particularly known for requiring large amounts of data

- Costly to build a dataset a well annotated dataset
- Domain specific data may require input from experts

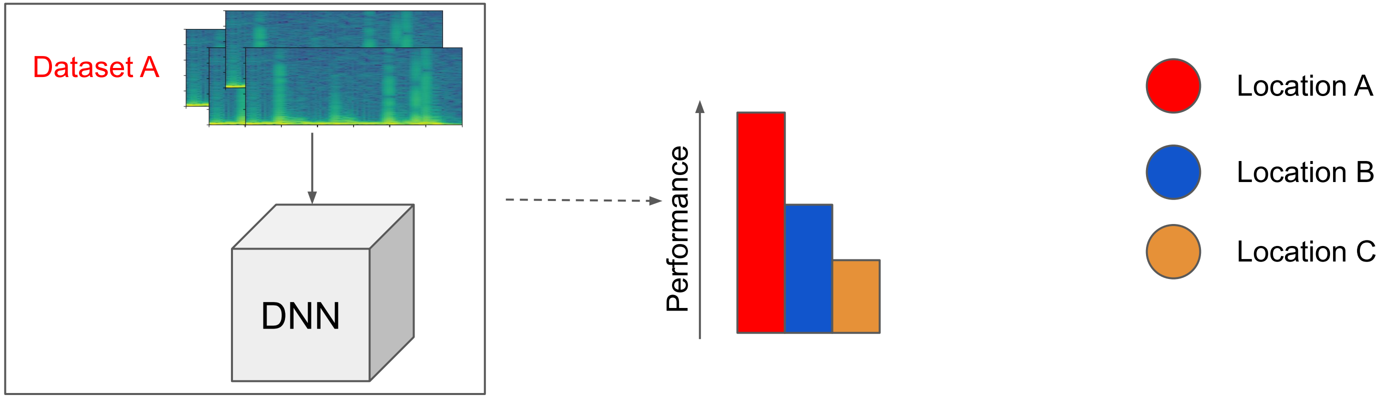
		label	start	end
filename	sel_id			
NOPP6_EST_20090329_084500.wav	0	1	890.100436	893.100436
NOPP6_EST_20090329_090000.wav	0	1	51.413506	54.413506
	1	1	41.592974	44.592974
	2	1	97.386199	100.386199
	3	1	115.234384	118.234384
	4	1	288.680821	291.680821

DNNs problems in Underwater Acoustics

DNNs may be sensitive to:

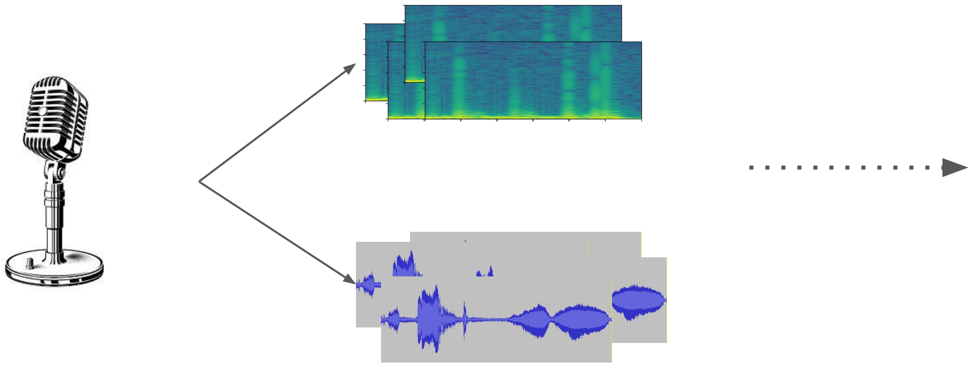
- Changes in amplitude
- Different types of hydrophones
- Distinct geographic locations

Thus, a model trained in one Dataset may perform poorly when tested in another location



Solutions…?

- More data…
  - It would always be helpful to simply have access to more data from all sources
    - Would require a lot of effort towards collecting and annotating more data



filename	sel_id	label	start	end
NOPP6_EST_20090329_084500.wav	0	1	890.100436	893.100436
NOPP6_EST_20090329_090000.wav	0	1	51.413506	54.413506
	1	1	41.592974	44.592974
	2	1	97.386199	100.386199
	3	1	115.234384	118.234384
	4	1	288.680821	291.680821

- What if we could artificially inflate the size of our dataset?

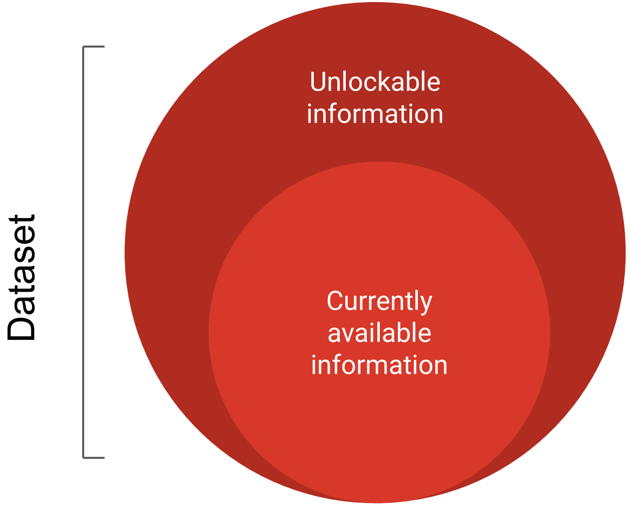
What is data augmentation?

Data Augmentation is a data-space solution to the problem of data limitation

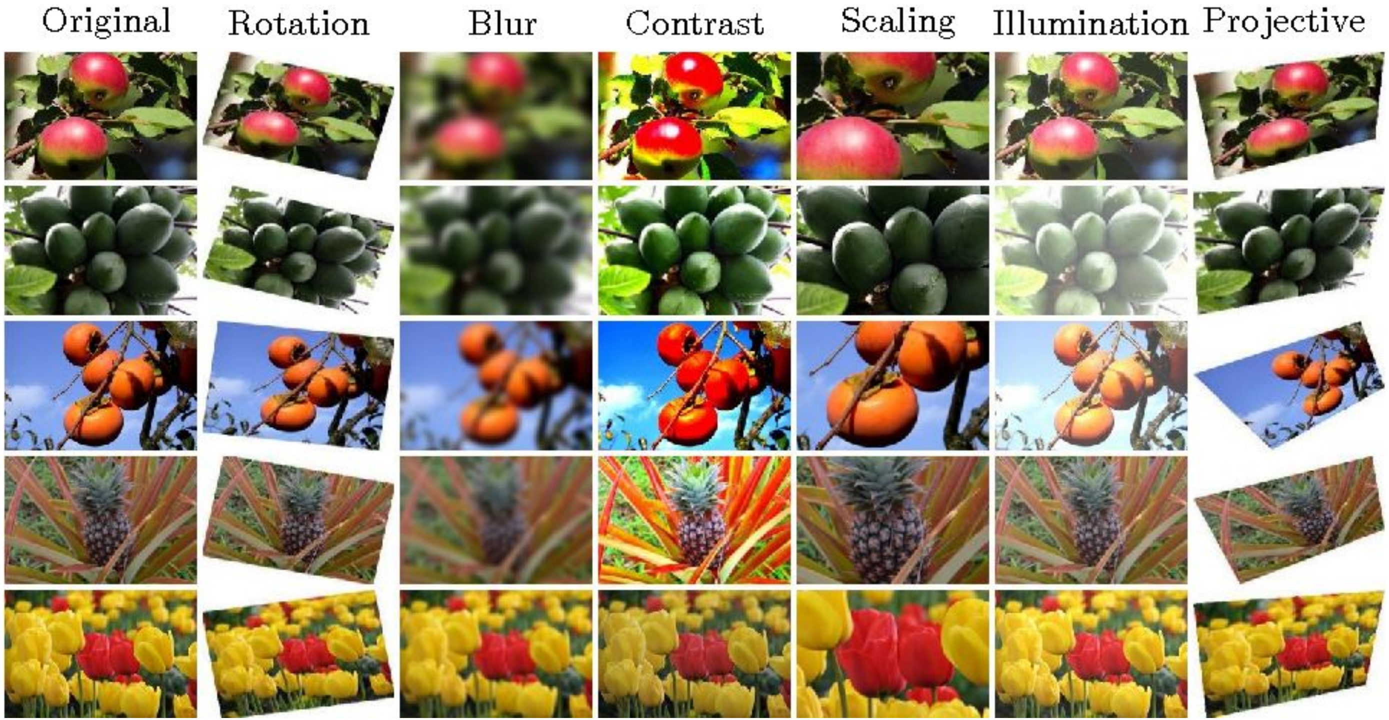
- Suit of techniques that enhance the size and quality of training datasets
- Inexpensive way to acquire more labeled data

Possible techniques include:

- Geometric transformations
- Color space augmentations
- Mixing
- Random erasing
- Deep learning based methods



Simple Data Augmentation

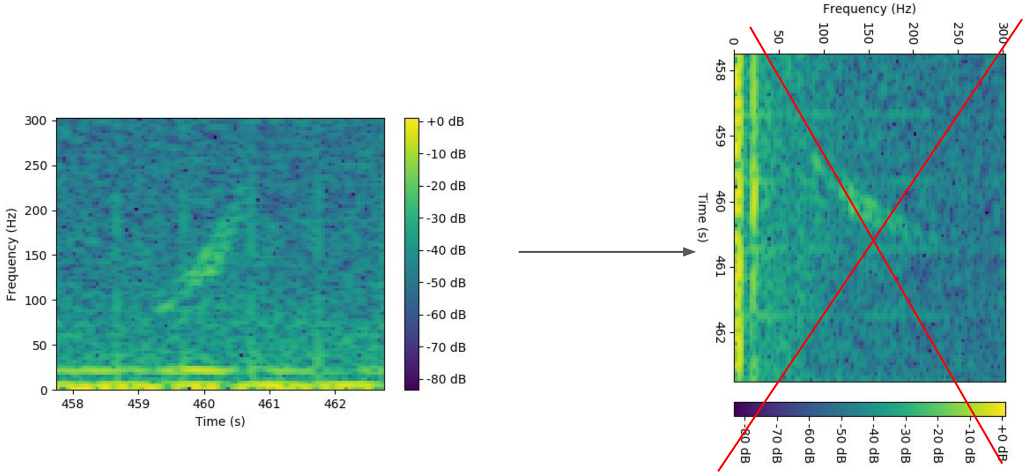


Data Augmentation In acoustics

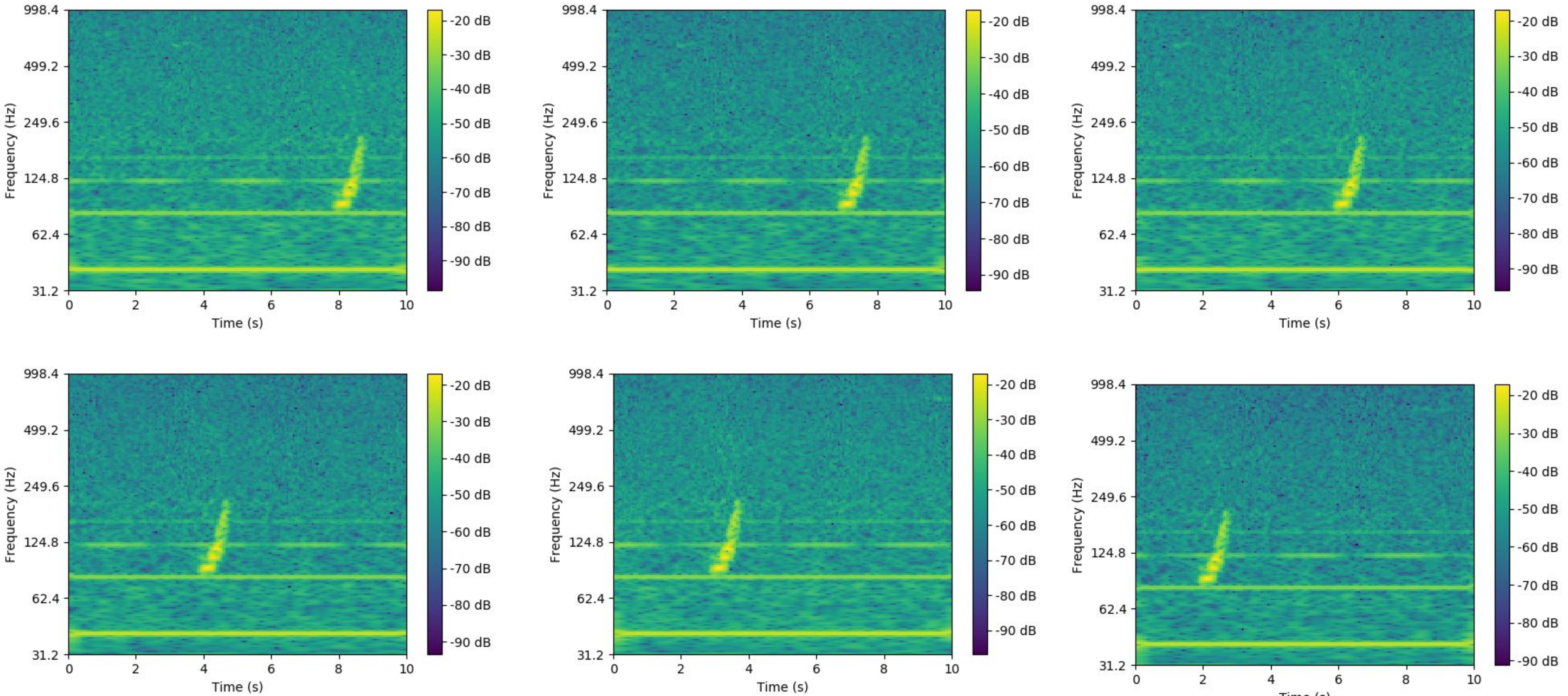
- The type of Augmentation will depend on your data
  - Is the label preserved post-transformation?

Rotations and flips

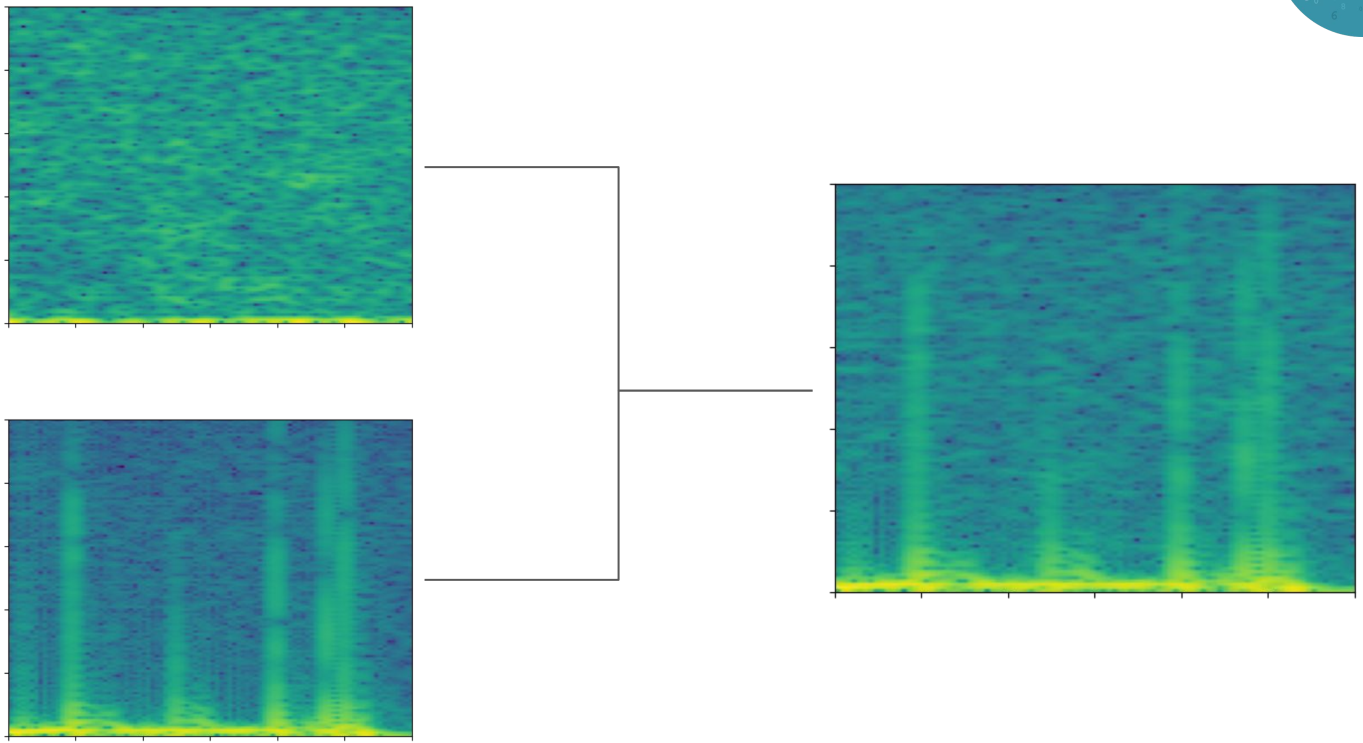
- Generally safe on ImageNet
- Problematic on spectrograms



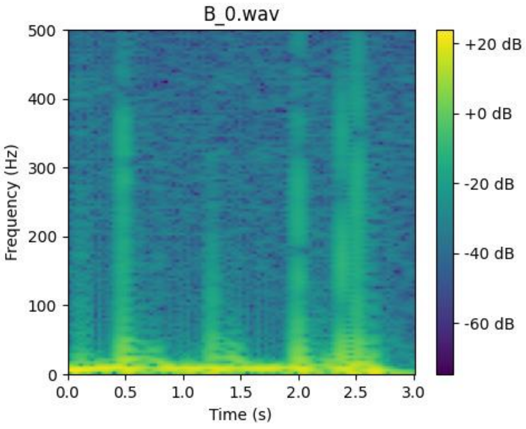
Temporal Shifting



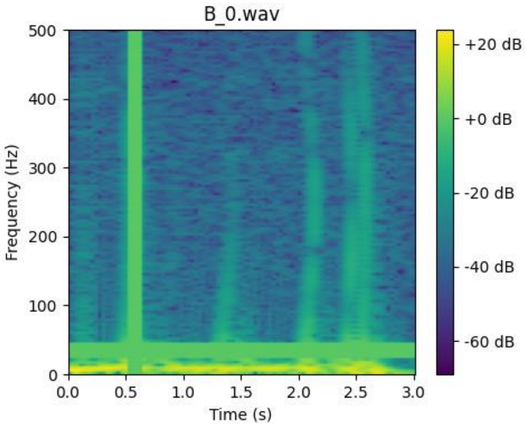
Mixup



SpecAugment

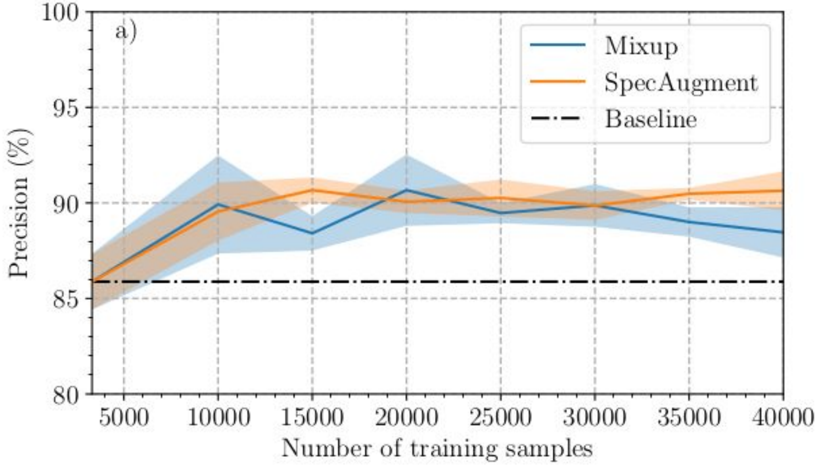


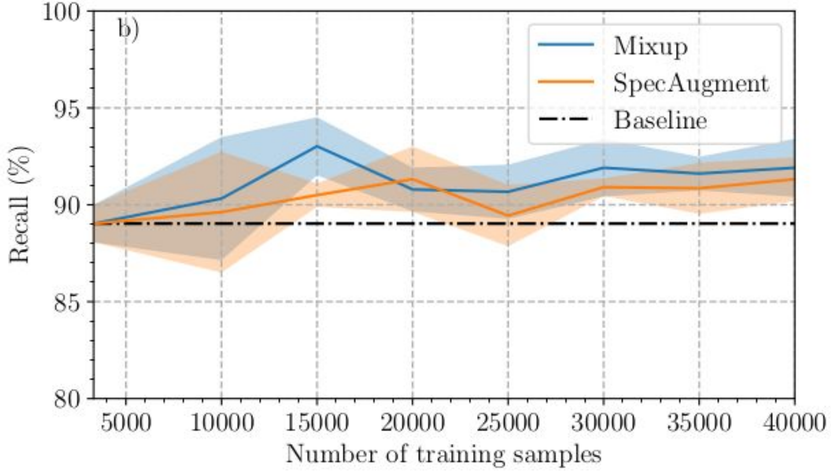
Masking + warping



Benefits of data Augmentation

- Original training dataset: 3309 samples





Limitations of Data Augmentation

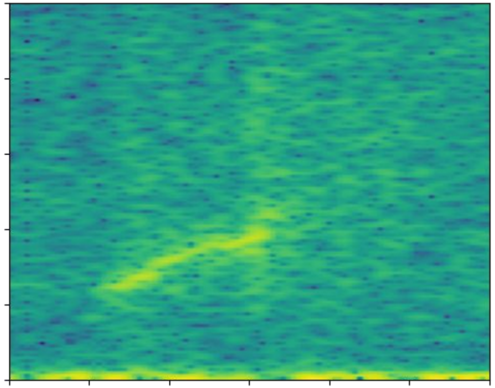
- Simpler augmentation methods generate new samples in a very specific manner
  - Limited by the type of transformation
  - Won't generate completely new data
  - Generative methods based on deep learning are capable of modeling raw audio

- Limited by the data already available
  - Data augmentation does not replace real data

Deep Generative Models

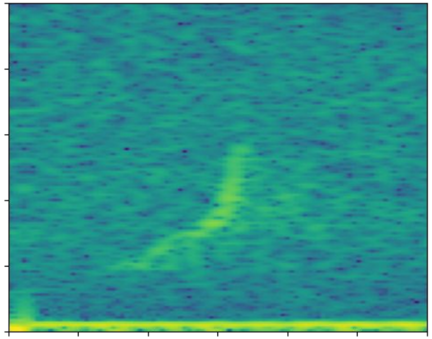
Generative Models - Acoustic

NARW upcall

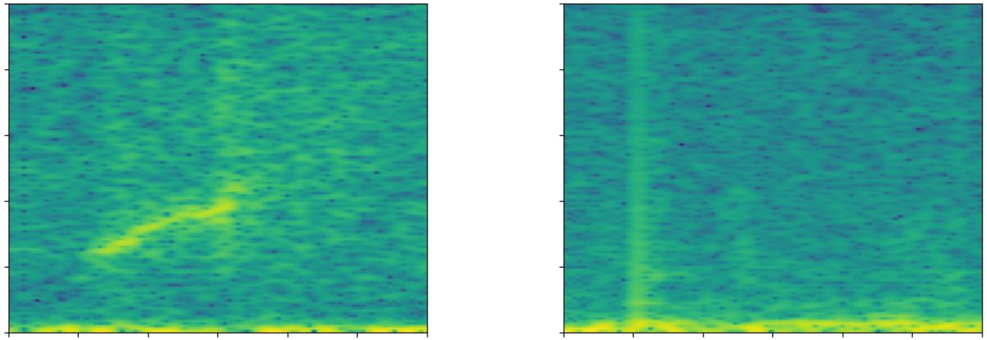


SampleRNN - Quality of generations

Original



Generated



So... what is good enough?

These generative methods will generate data that can vary in quality

Are our generated samples good enough to be included in our dataset?

- Problem: Feeding our model bad quality samples to our training procedure might harm the model's ability

We can evaluate our generated samples through an active learning approach

- Classify each generation and ask a human analyst to oversee the most uncertain samples