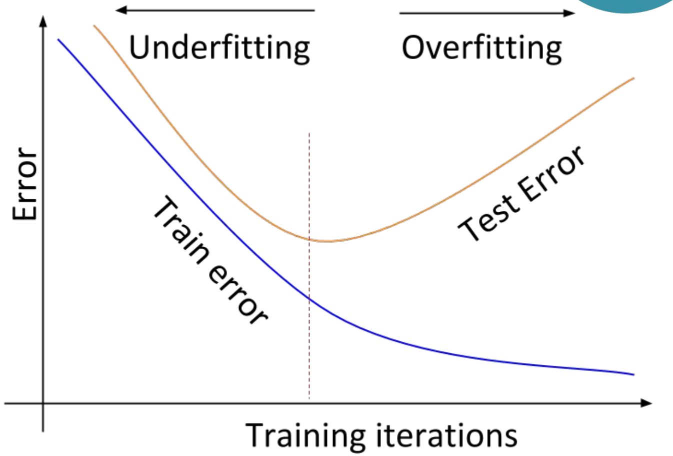
Data Augmentation: Improving your Datasets

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

DNNs are smart, but not always…

- Tend to overfit the training set
- Can easily inherit and perpetuate biases



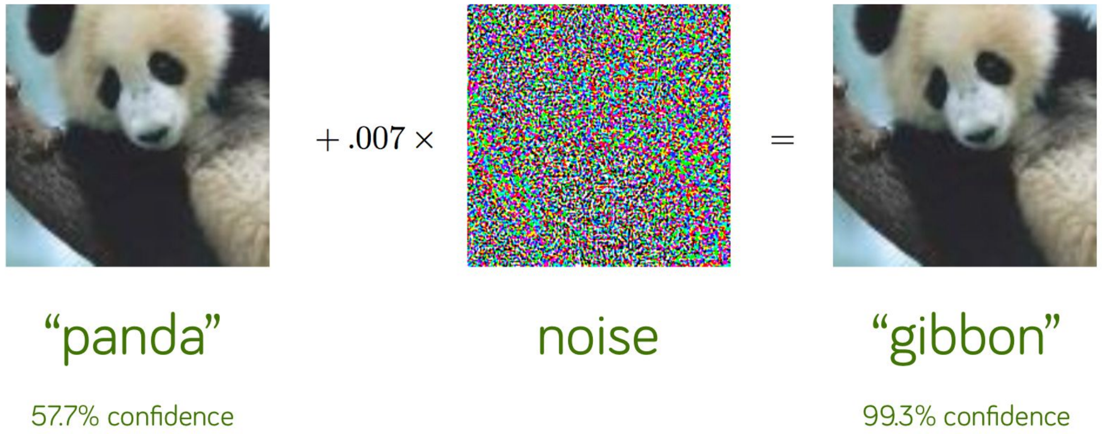
filename	sel_id	label	start	end
NOPP6_EST_20090329_084500.wav	0	1	890.100436	893.100436
NOPP6_EST_20090329_090000.wav	0	1	51.413506	54.413506
	1	1	41.592974	44.592974
	2	1	97.386199	100.386199
	3	1	115.234384	118.234384
	4	1	288.680821	291.680821

Benefit from large amounts of labeled data

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

DNNs classifiers can often be sensitive to very minor changes

- Ex: Adversarial attacks

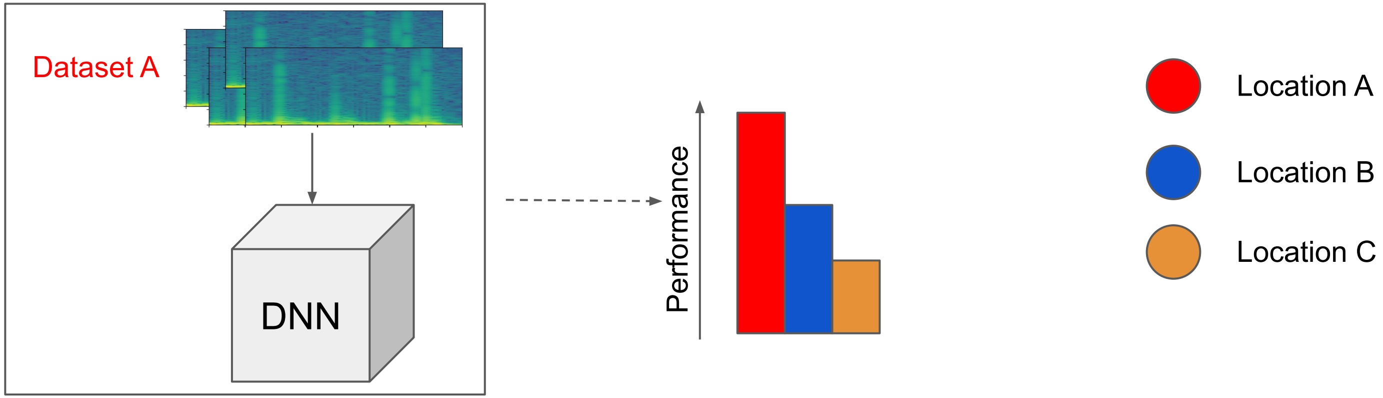


DNNs problems in Underwater Acoustics

It may happen naturally as well:

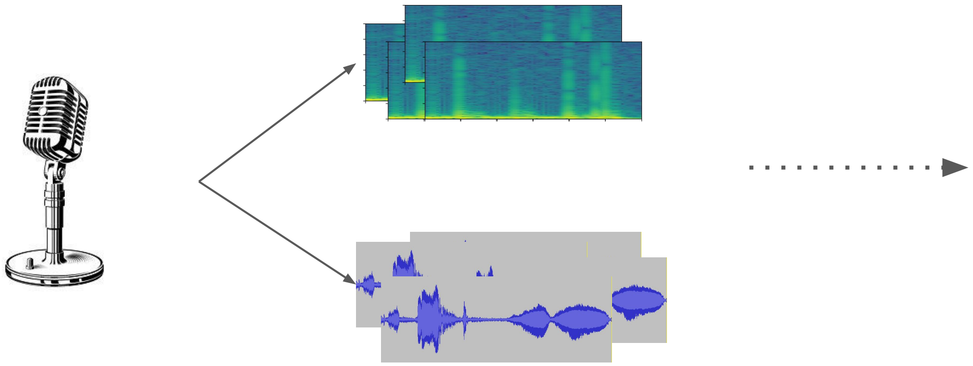
- Changes in amplitude
- Different types of hydrophones
- Distinct geographic locations

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



		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

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

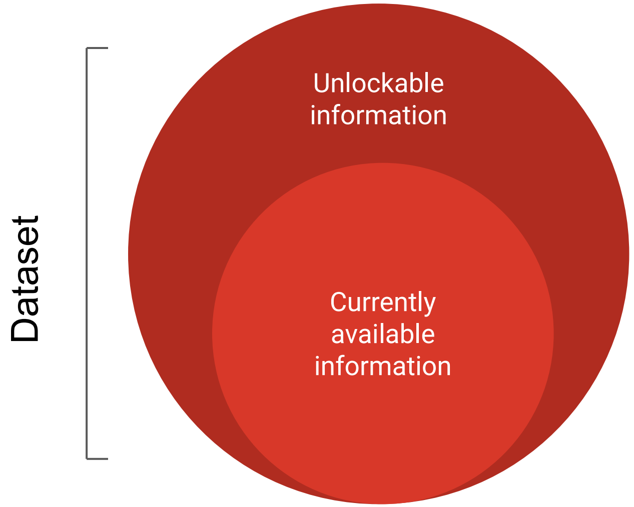
What is data augmentation?

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

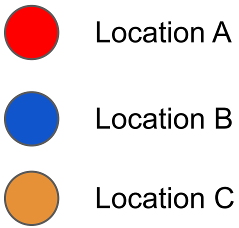
Possible techniques include:

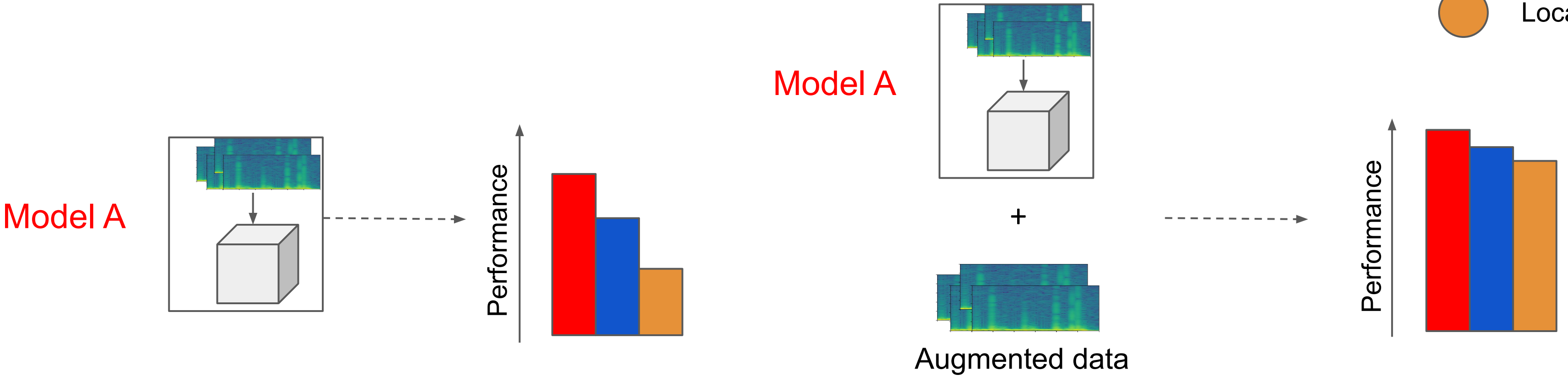
- Geometric transformations
- Color space augmentations
- Mixing
- Random erasing
- Deep learning based methods



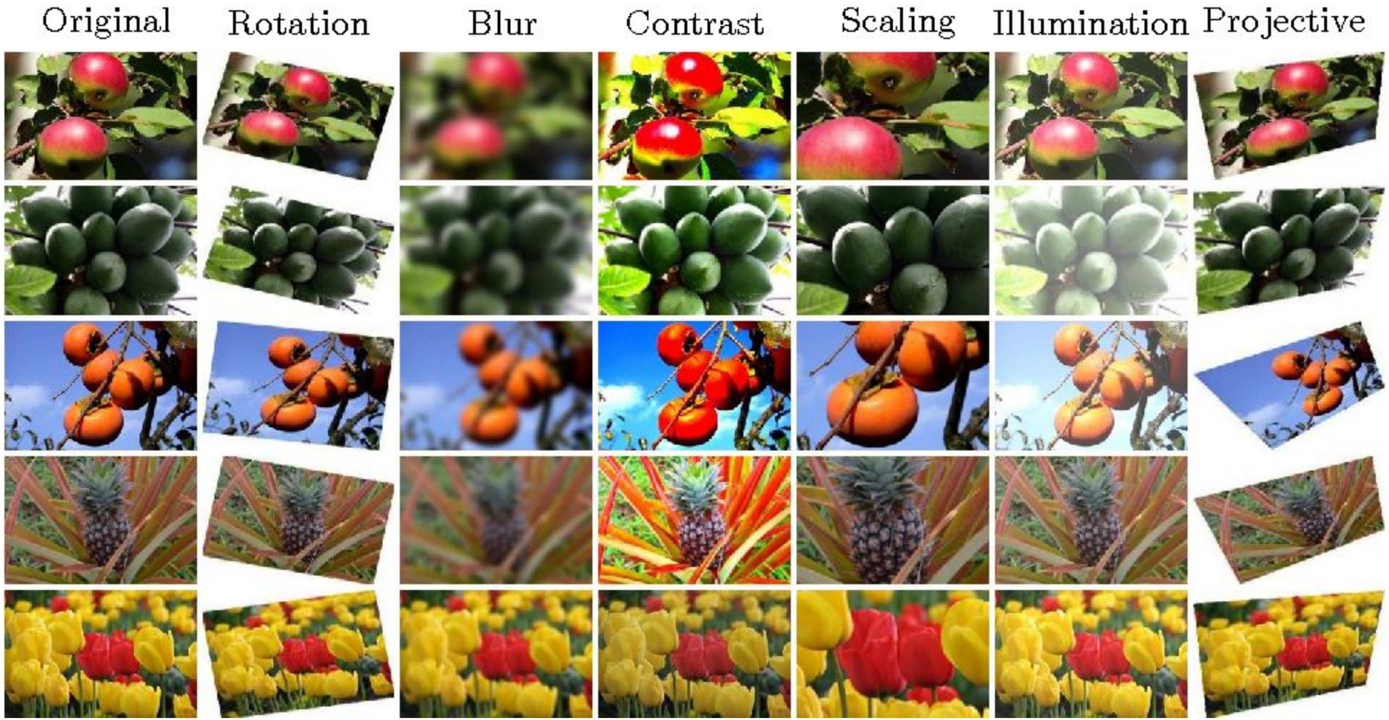
What is data augmentation?

- Data Augmentation can:
  - Enhance the quality of training sets
    - Leading to better classifiers
  - Expands model adaptability to other contexts (e.g. Other geographic locations)





Simple Data Augmentation



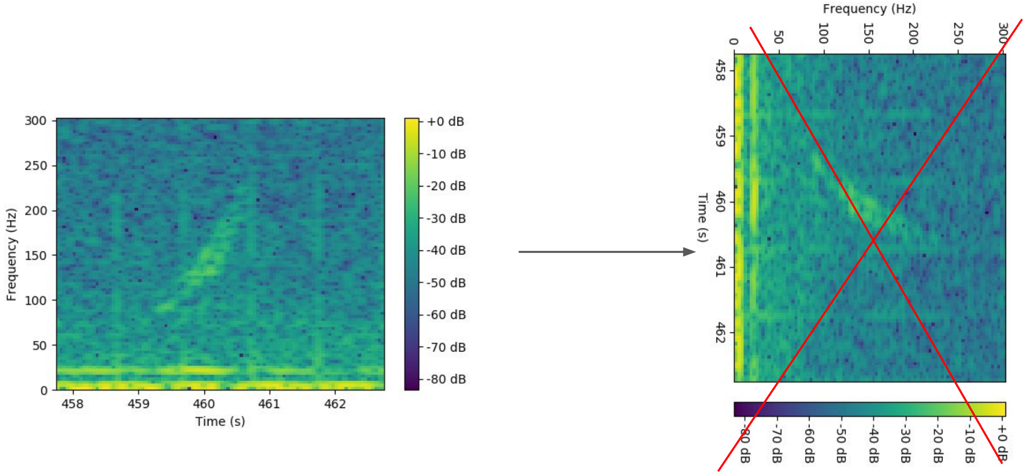
Source: Data Augmentation for Plant Classification

‘Safety’ of Data Augmentation

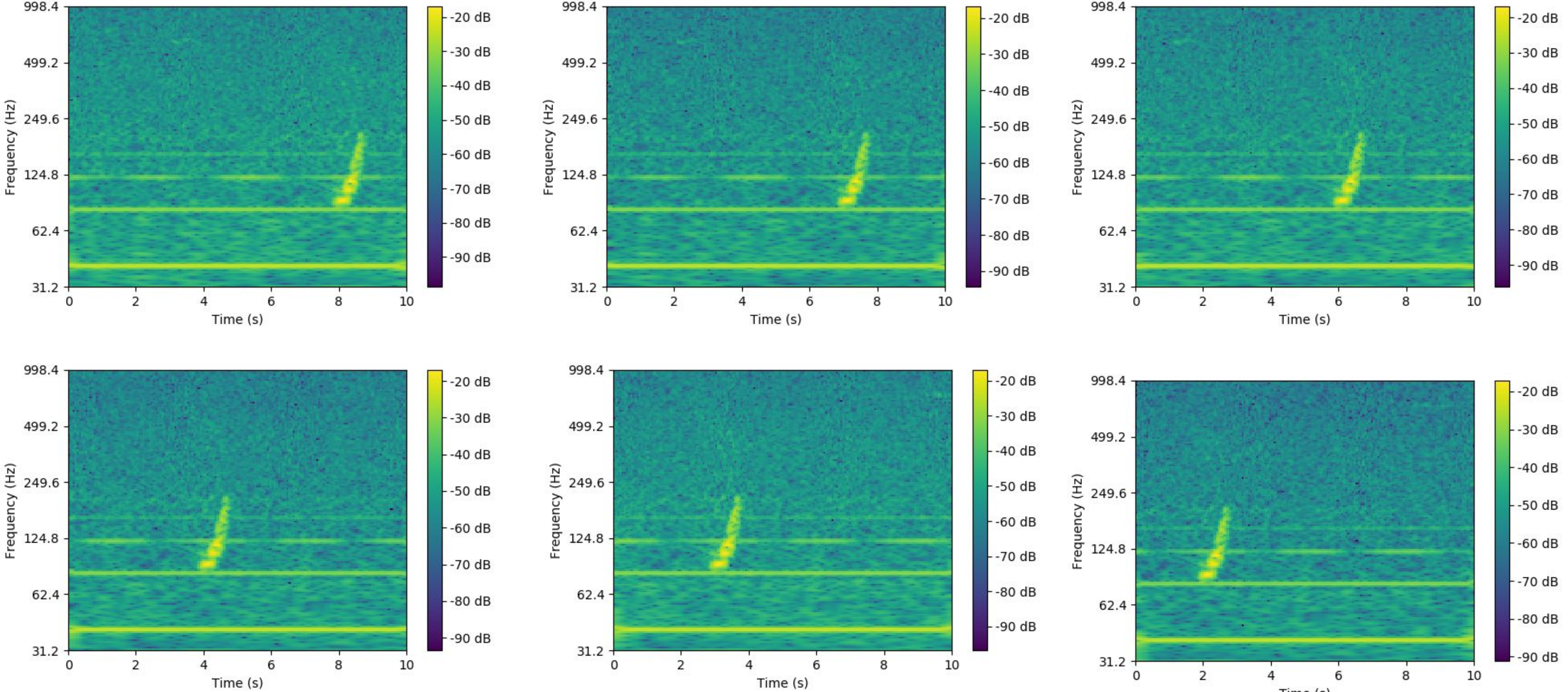
- The type of Augmentation will depend on your data
  - Is the label preserved post-transformation?
  - ‘Unsafe’ transformations are those that do not preserve the label

Rotations and flips

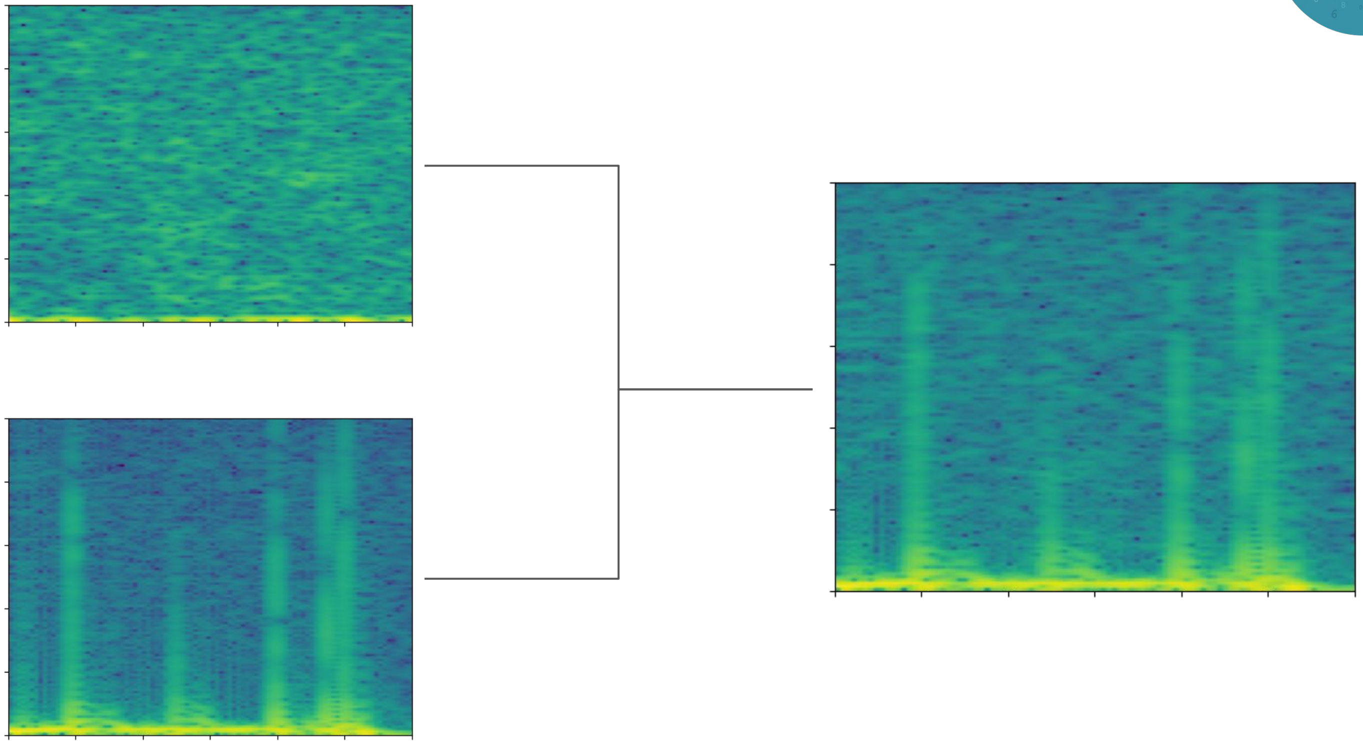
- Generally safe on ImageNet
- Problematic on spectrograms



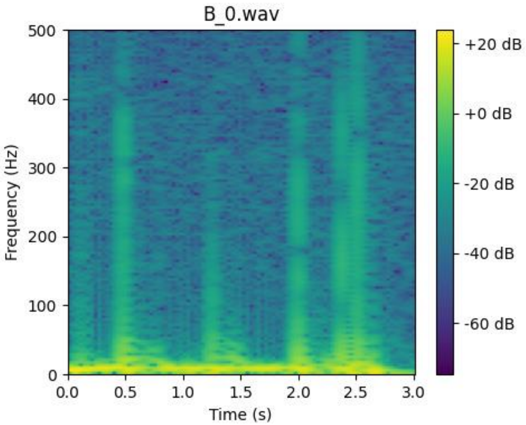
Temporal Shifting



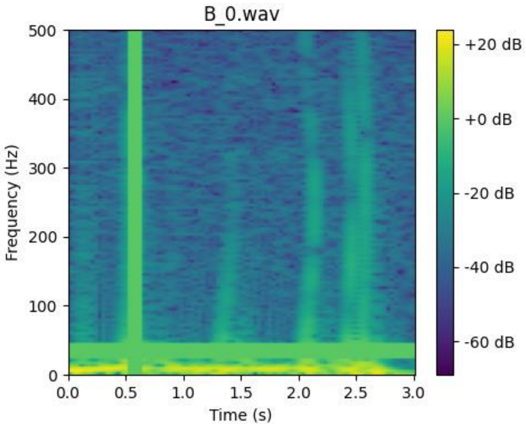
Mixup



SpecAugment

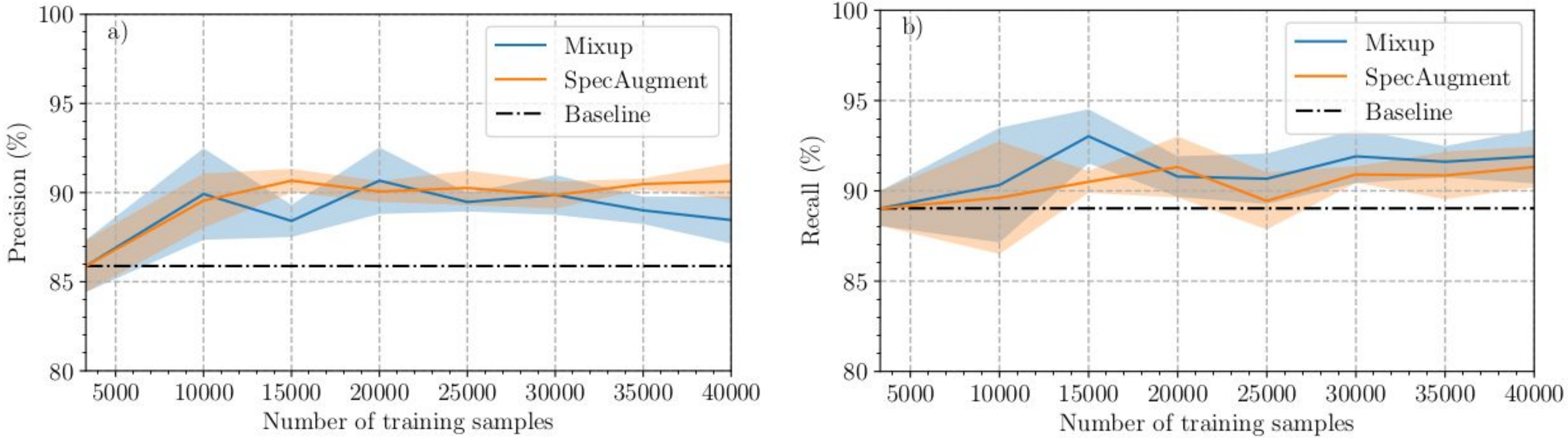


Masking + warping



Benefits of data Augmentation

- Original training dataset: 3309 samples



Kayra: A a python package that provides implementation of several data augmentation techniques

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

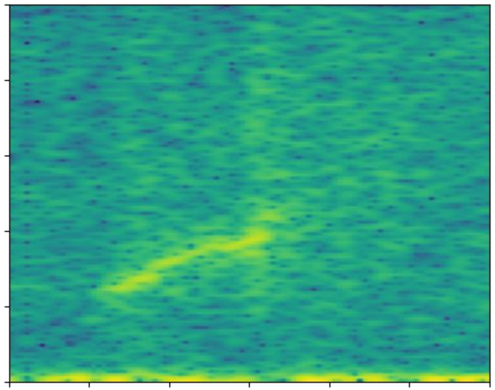
- Limited by the data already available
  - Data augmentation does not replace real data

Deep Generative Models

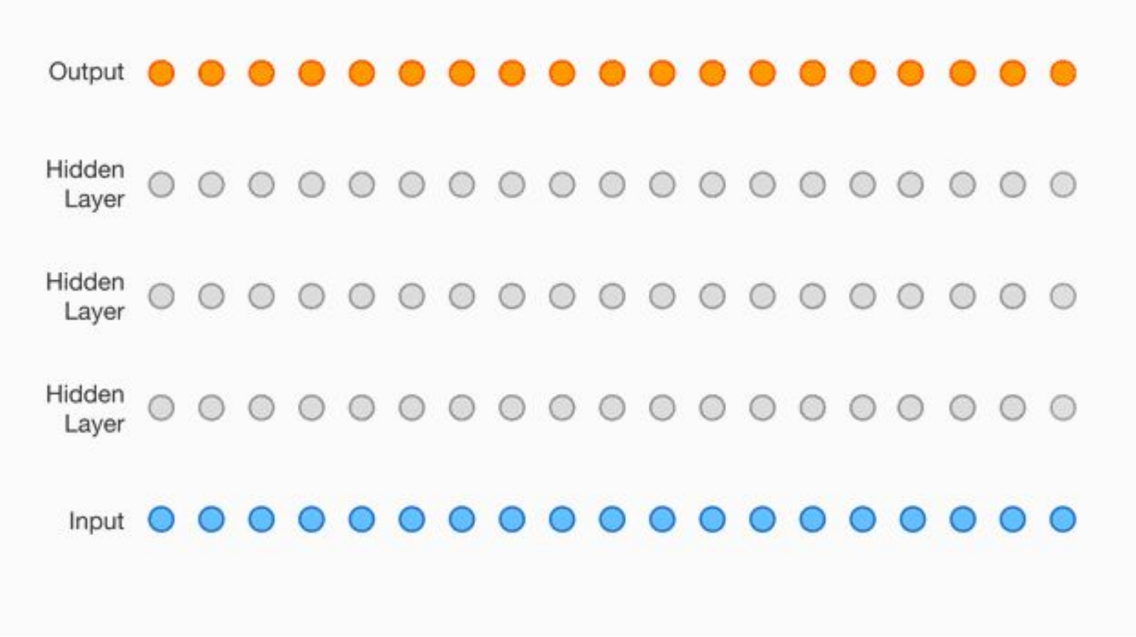
Generative Models - Audio

Human Speech

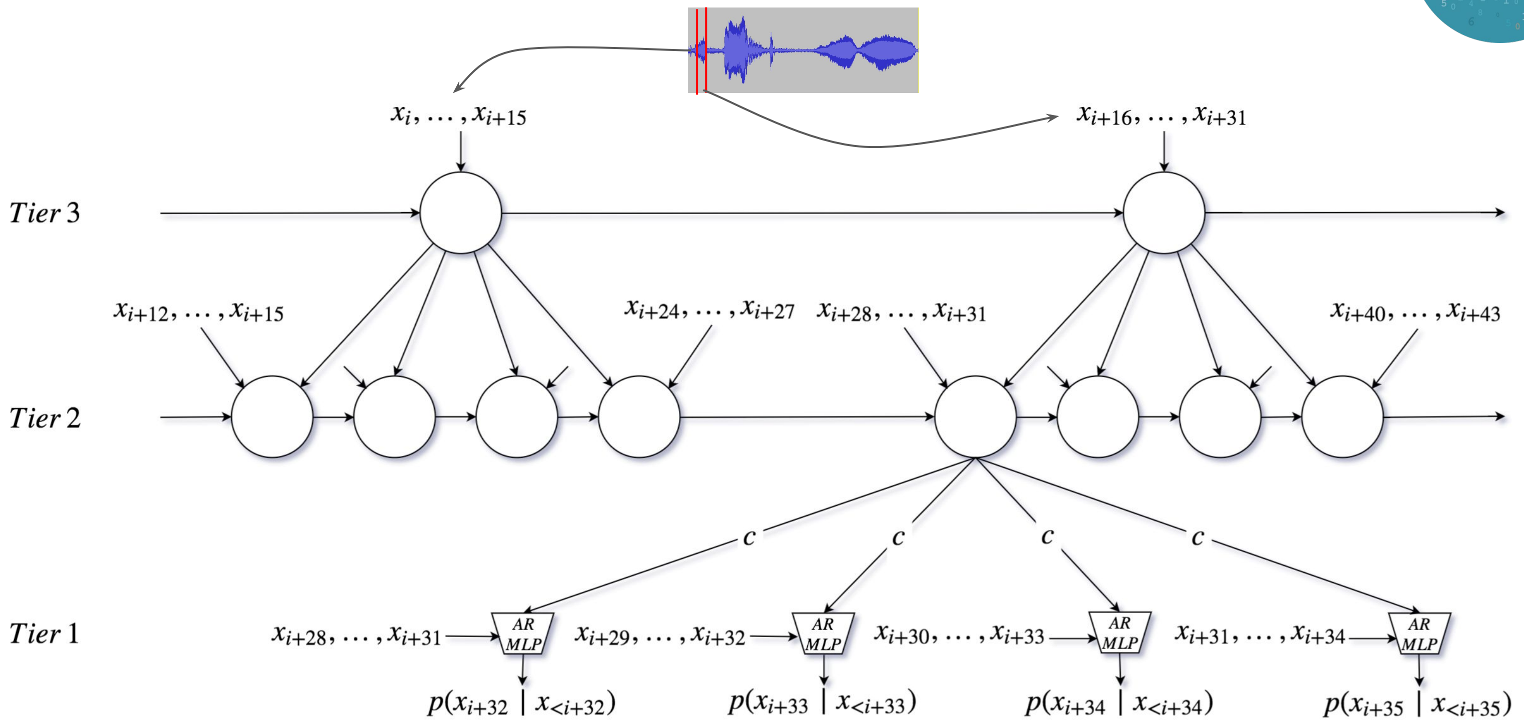
NARW upcall



Generative Models - WaveNets

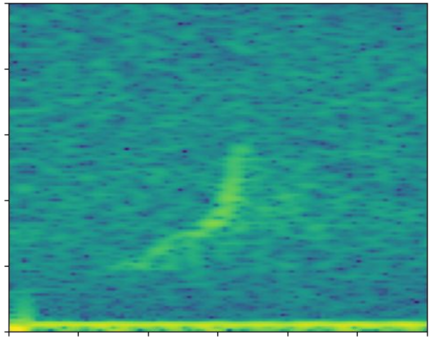


SampleRNN

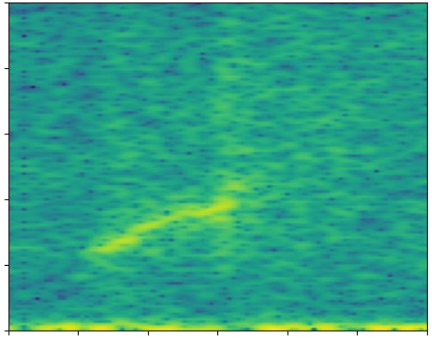

$x_i, \ldots, x_{i+15}$
$x_{i+16}, \ldots, x_{i+31}$
Tier 3
$x_{i+12}, \ldots, x_{i+15}$
$x_{i+24}, \ldots, x_{i+27}$
$x_{i+28}, \ldots, x_{i+31}$
$x_{i+40}, \ldots, x_{i+43}$
Tier 2
c
c
c
c
Tier 1
$x_{i+28}, \ldots, x_{i+31}$
AR MLP
$x_{i+29}, \ldots, x_{i+32}$
AR MLP
$x_{i+30}, \ldots, x_{i+33}$
AR MLP
$x_{i+31}, \ldots, x_{i+34}$
AR MLP
$p(x_{i+32} \mid x_{<i+32})$
$p(x_{i+33} \mid x_{<i+33})$
$p(x_{i+34} \mid x_{<i+34})$
$p(x_{i+35} \mid x_{<i+35})$

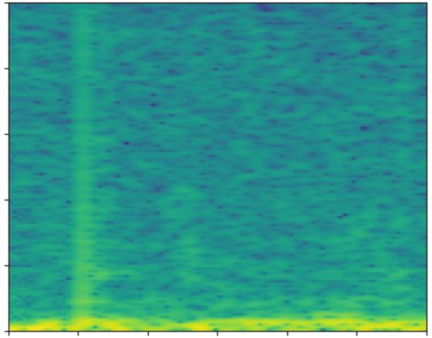
SampleRNN - Quality of generations

Original



Generated





So… what is good enough?

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

- Problem: labeling a new sample as positive when there is no vocalization
- How do we evaluate our generated samples?

There are several possibilities with trade-offs

- Conduct a manual labelling process of the generated samples
  - Would produce an accurate augmented set
  - Expensive
- Use a pre-trained model to classify each generation
  - Inexpensive but less accurate
  - Could inherit model bias

Generated samples evaluation

- A more balanced approach can be considered
  - Manually label some generated samples
  - Use these samples in conjunction with a pre-trained model to label the remaining generations

- Clustering methods can be used to group generated samples into classes
  - With a visualization tool, we can then ask for the user to label only the samples that the method is not confident about