DATA ANALYTICS USING DEEP LEARNING

GT 8803 // FALL 2018 //

JACOB LOGAS

EFFICIENT CONTENT-BASED AUDIO RETRIEVAL

CREATING THE NEXT®

AUDIO RETRIEVAL

- Querying continues to be an open question
- Naïve approach
  - Costly similarity functions
  - Query-by-example
  - Heuristics
- Common Approaches
  - Gaussian Mixture Models
  - Support Vector Machines

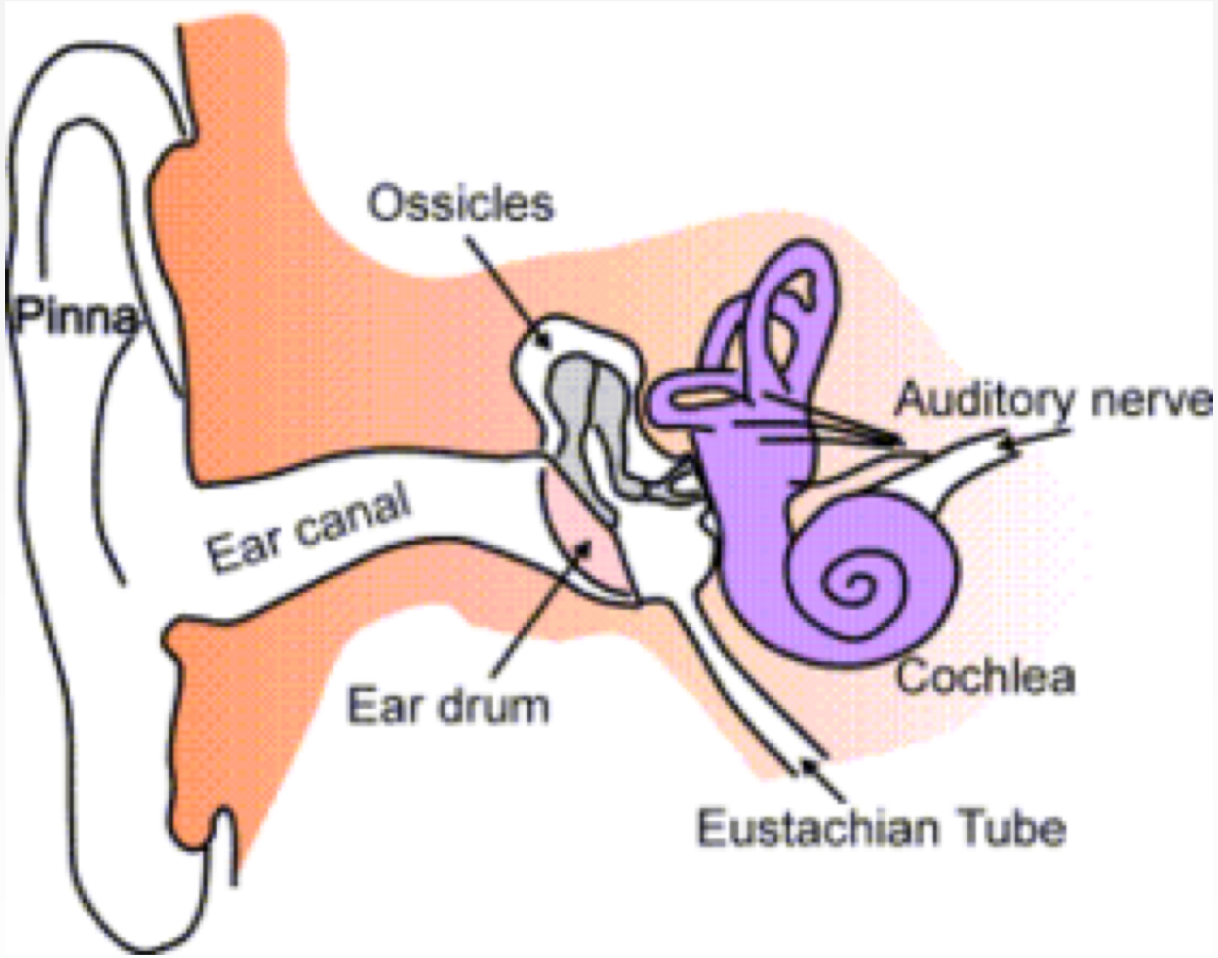
RESEARCH QUESTIONS

- How does a hierarchical approach to audio retrieval effect performance?
- Will deep learning provide a significant improvement over naïve methods?
- Can unstructured audio datasets be queried with complex content-based text queries?
- Is it feasible to query audio datasets for audio based on context?

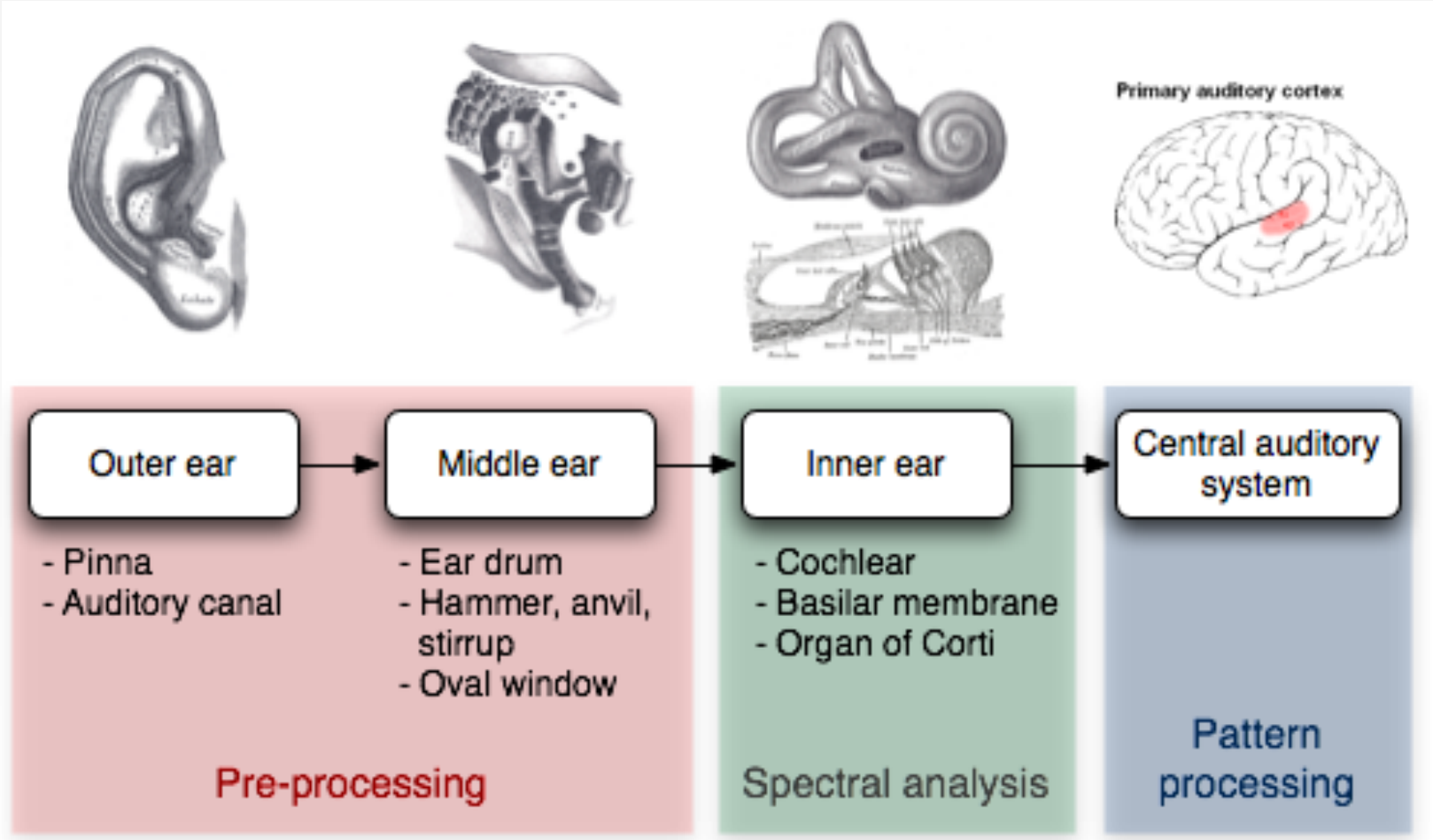
GOALS

- Query an unstructured dataset based on a hierarchical taxonomy of sound
- Build a system that uses deep learning to query unstructured audio datasets
- Have text queries and allow for complex queries with unrestricted vocabulary
- Query audio dataset for a contextual event
- Make faster and at least as accurate as other querying systems

HUMAN PERCEPTION



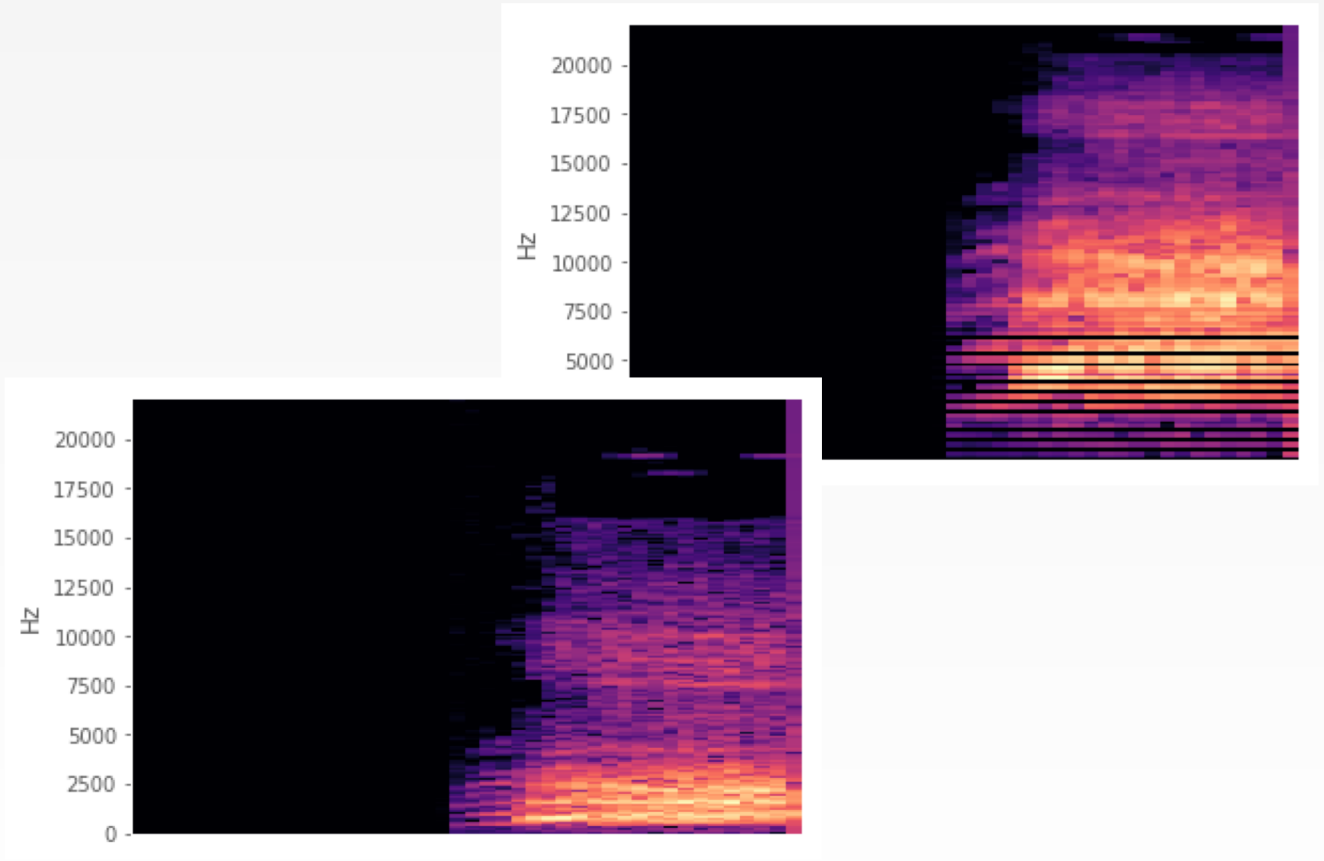
WHY NOT JUST USE CNNS?



Information flow in the auditory system (photo credit: Universität Zu Lübeck)

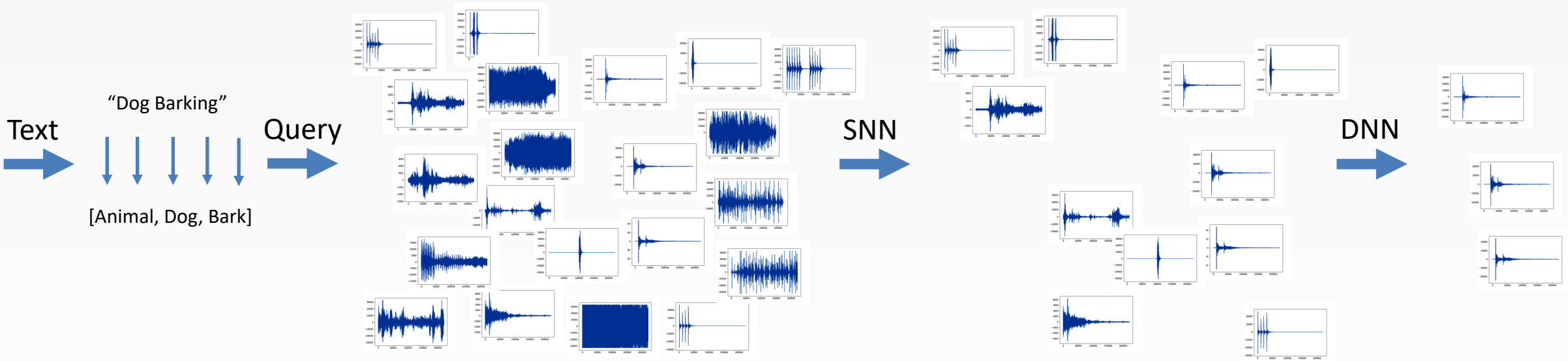
REPRESENTATION

- Spectral Properties
  - MFCCs
- Intensity Properties
  - Maximum frequency
- Statistical Properties
  - Zero Crossing Rate
- Temporal Properties
- Cognitive Properties



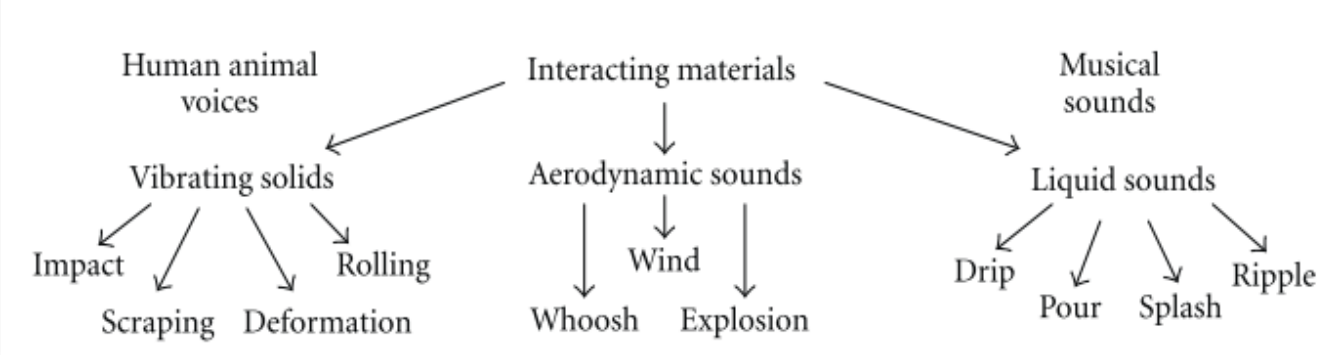
PROPOSED APPROACH

OVERVIEW



GAVER TAXONOMY

- Taxonomy of sound
- Based on human perception
- Listen to sounds differently
- Leads to hierarchy of neural networks
- Each network tuned for kind of sound



Gerard Roma, Jordi Janer, Stefan Kersten, Mattia Schirosa, Perfecto Herrera, and Xavier Serra. 2010. Ecological acoustics perspective for content-based retrieval of environmental sounds. Eurasip Journal on Audio, Speech, and Music Processing 2010: 1–11. https://doi.org/10.1155/2010/960863

TAXONOMY LOWER LEVELS

- Much more nuanced discrimination task
- Requires more classification power
- DNN
  - Fully connected layers
  - Many hidden layers
  - Feed-forward algorithm operating on each audio frame separately
  - Can be computationally complex

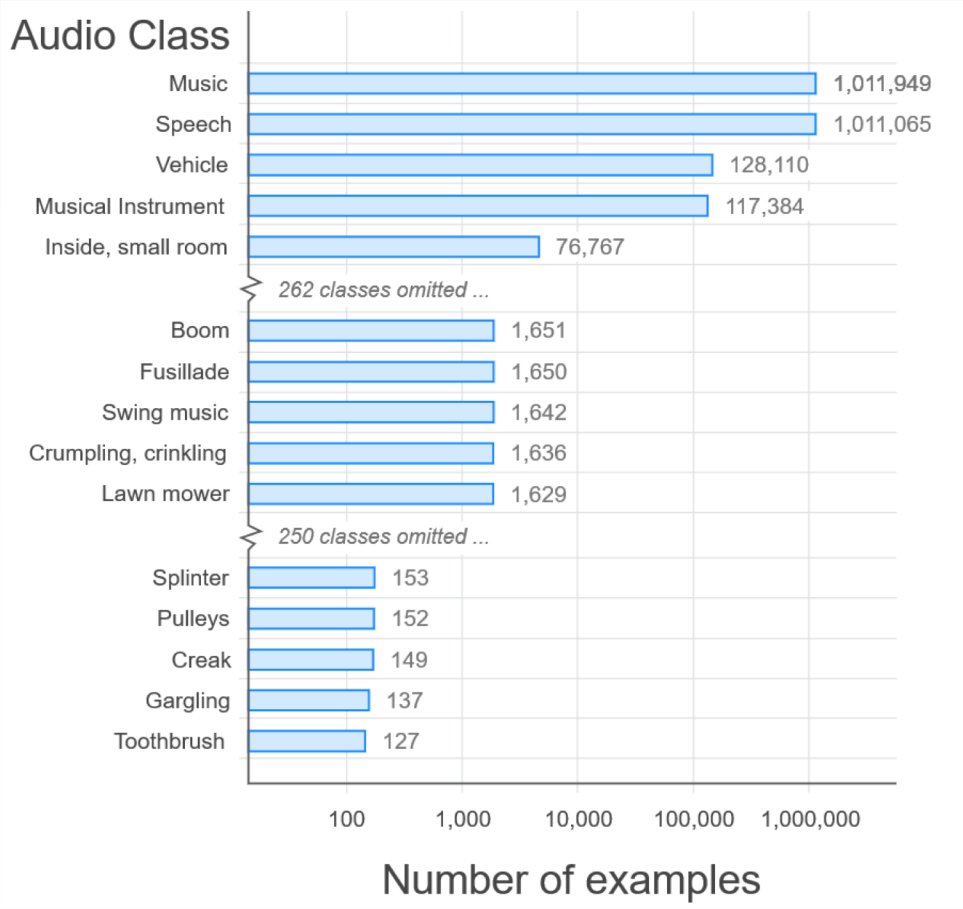
DEVELOPMENT DATASET

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

DEVELOPMENT DATASET (ESC-50)

- Current hierarchy doesn't match Gaver
- Required manual labelling
- 5s audio clips
- Singleton sounds
- Uniformly labelled

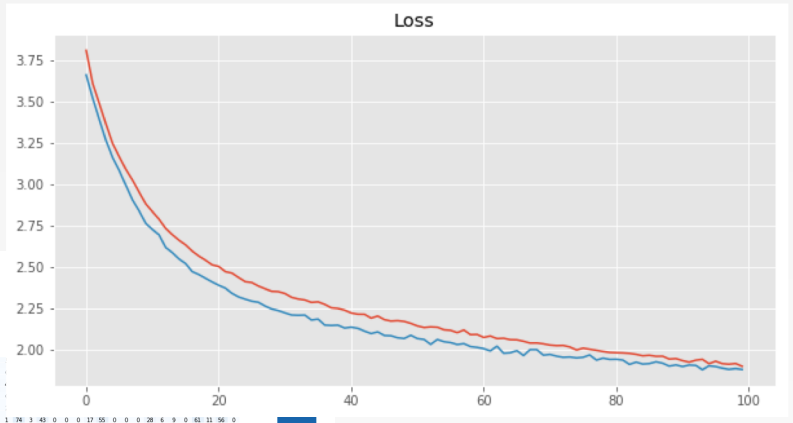
TARGET DATASET

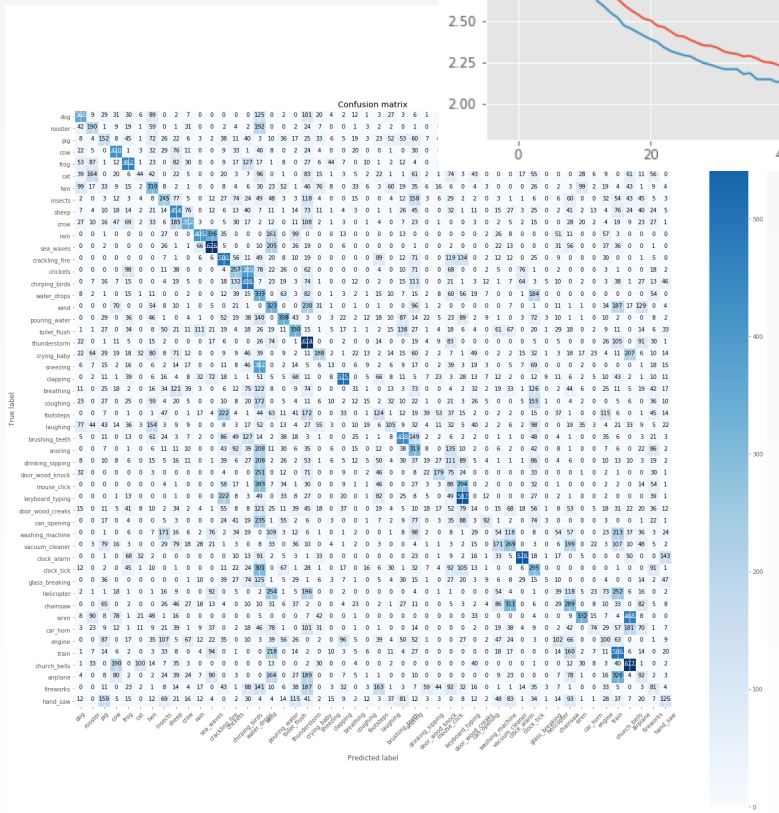


EXPERIMENTS

DEFAULT CNN

- Trained on all 50 classes
- All default settings
- ~24% accuracy
- ~7 minutes to train



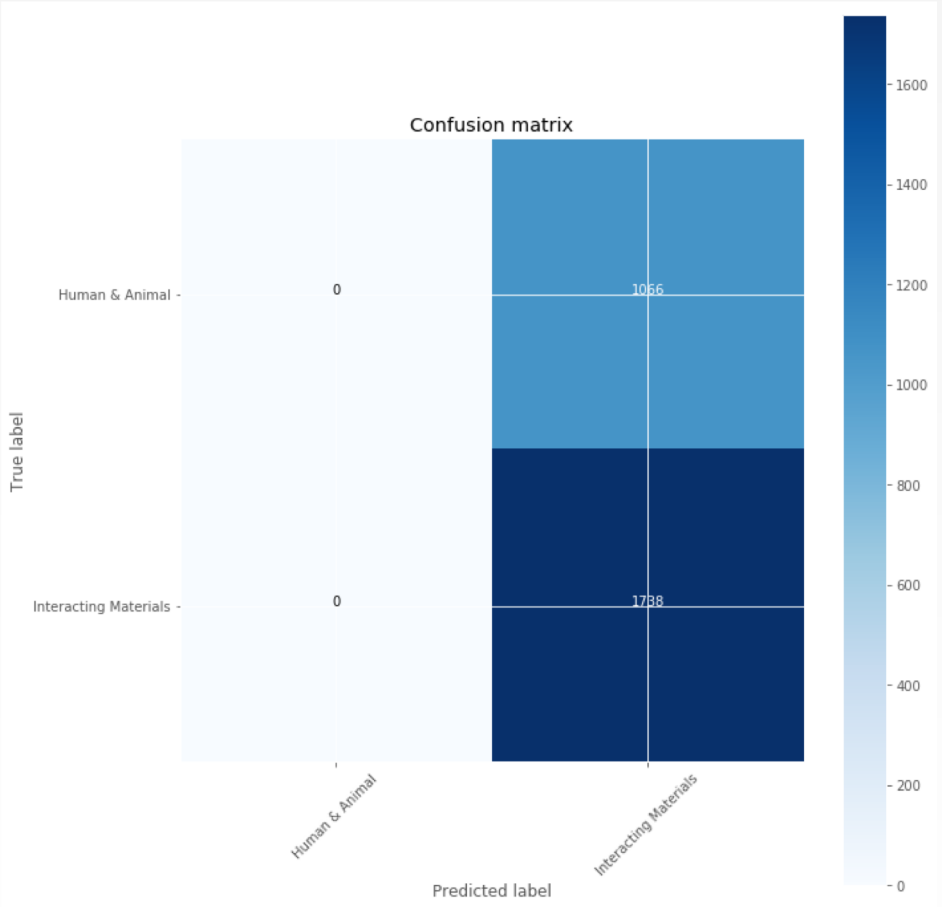


TOP LEVEL EXPERIMENTS

- Vector of features
  - MFCCs
  - First and second derivative
  - Spectral Contrast
- Spectral Image
  - Spectrogram
  - Mel-spectrogram

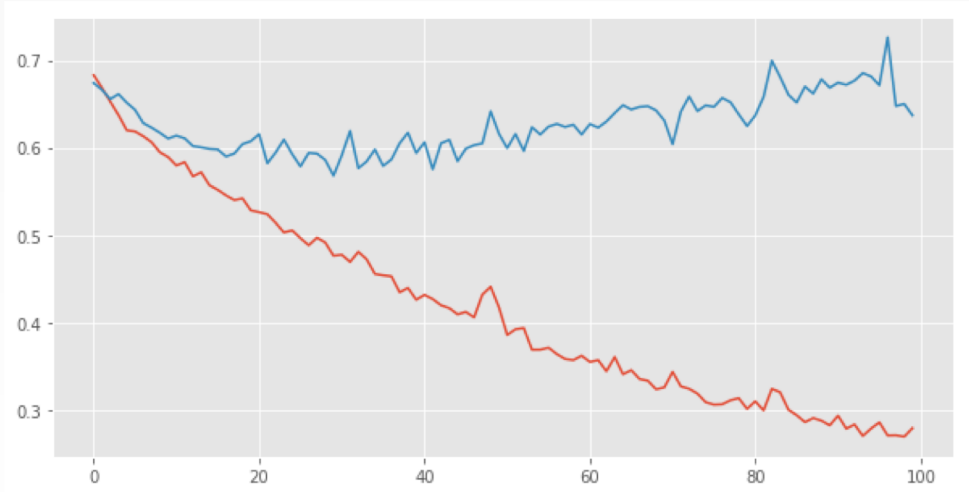
TOP LEVEL EXPERIMENTS

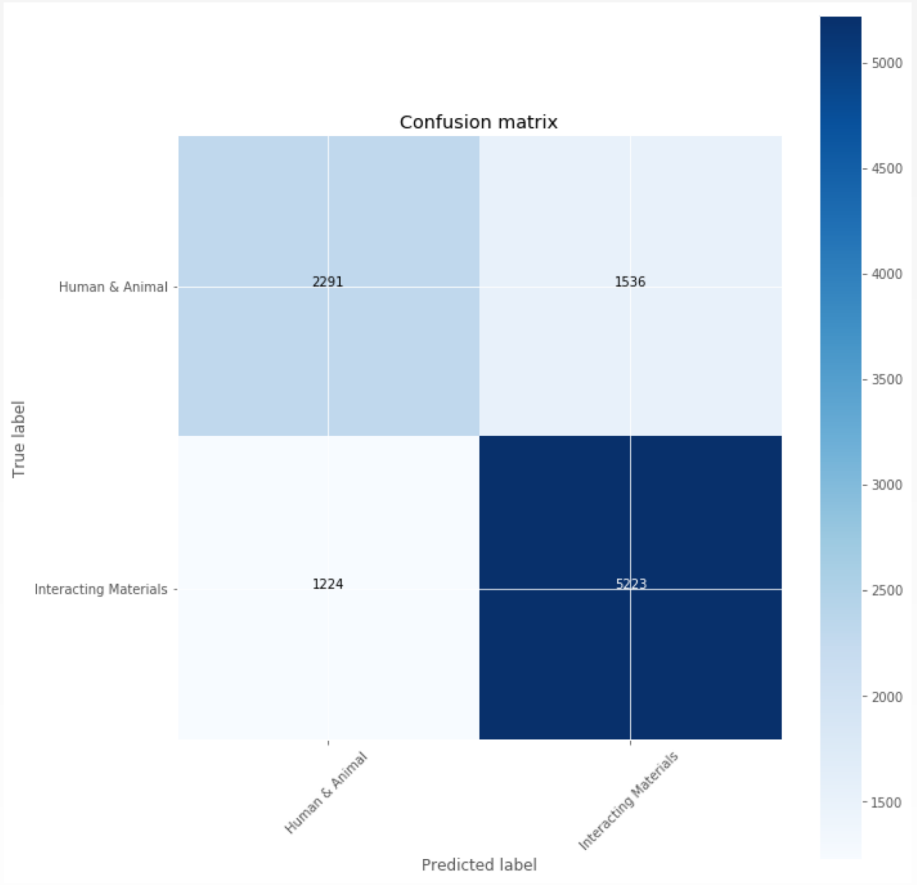
- Feature Vector
  - Loss function unchanged
  - Likely little difference between classes on these axes



TOP LEVEL EXPERIMENTS

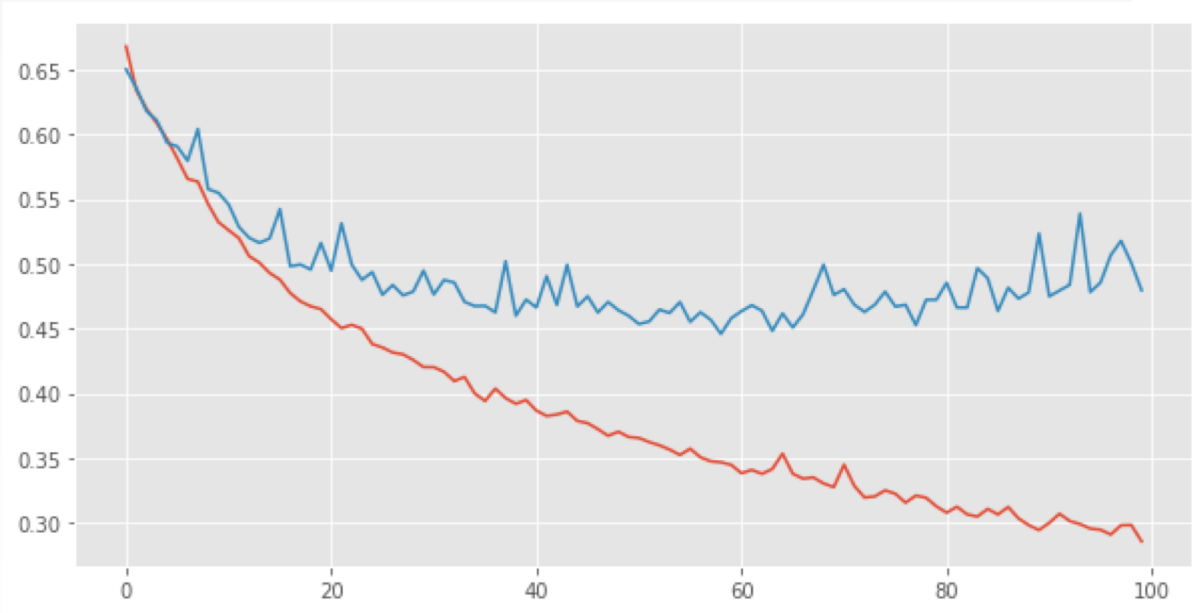
- Spectral Image
  - 250ms and 1s window
  - 125ms and 250ms overlap
- 73% at 1s window

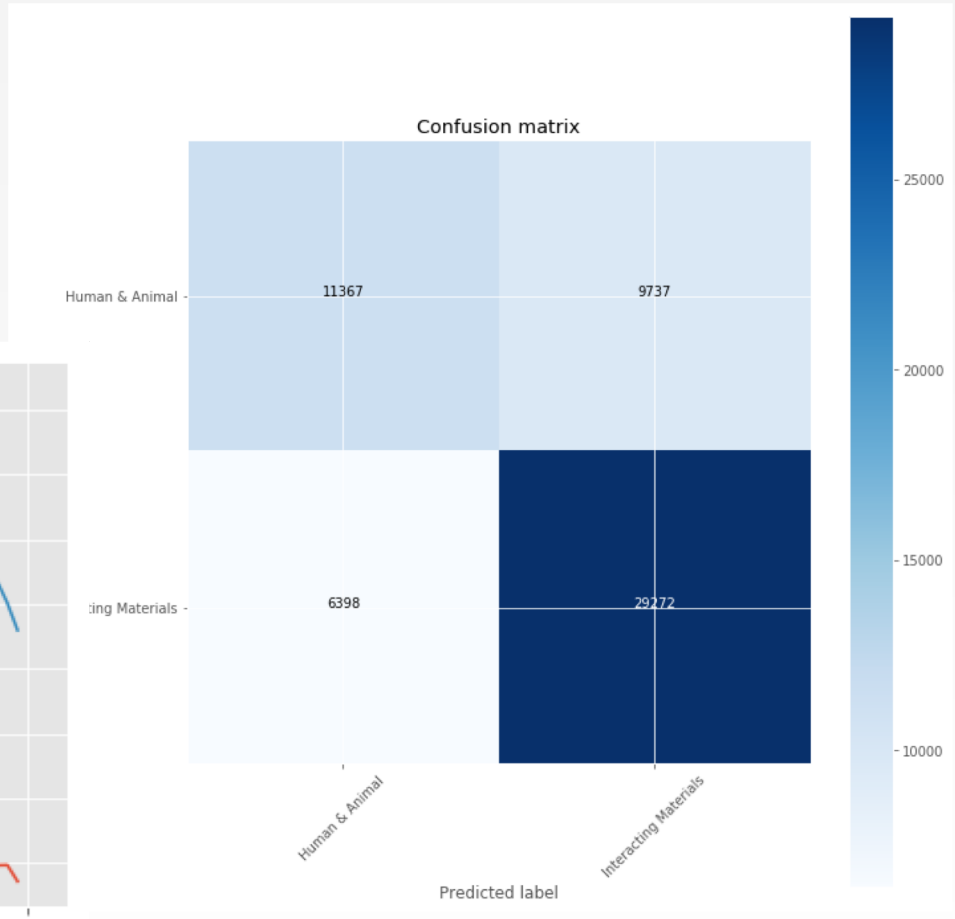




HIERARCHY TOP LEVEL

- 250ms Window
  - 72% Testing Accuracy



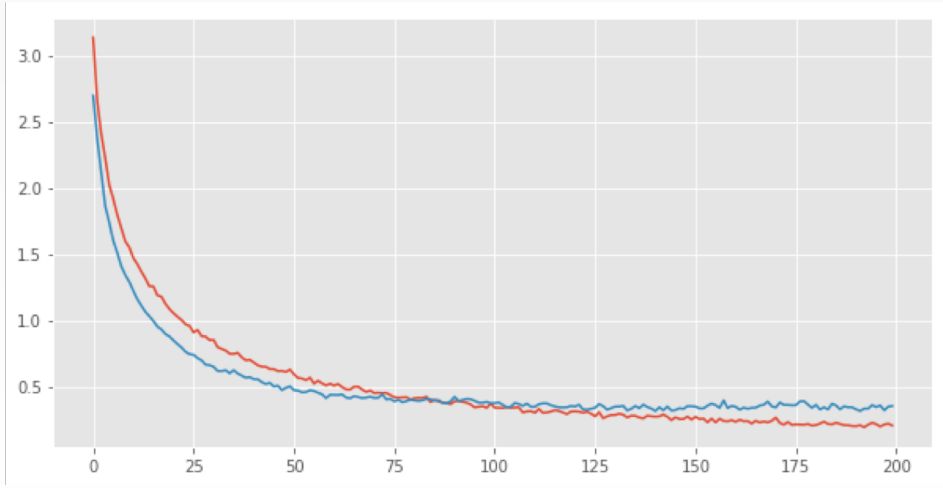


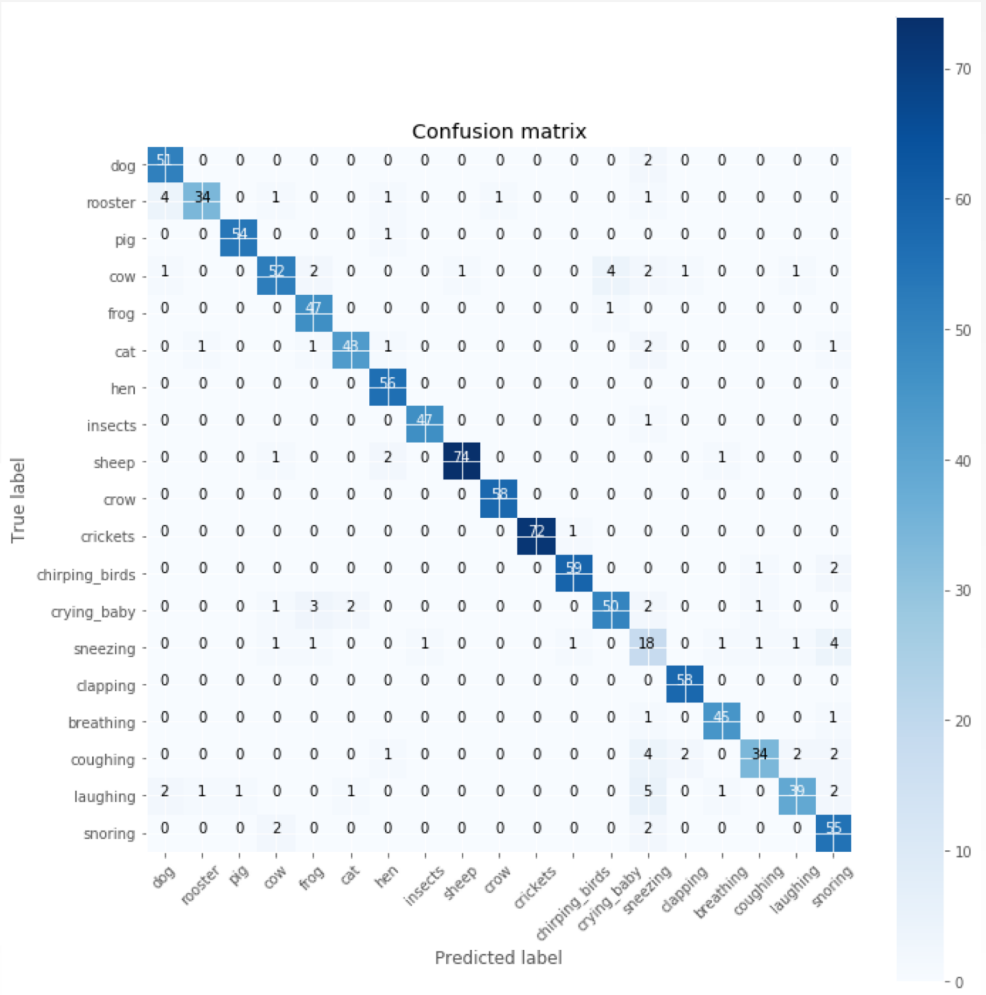
LOWER LEVEL EXPERIMENTS

- Vector of features
  - MFCCs + MFCCs` + MFCCs``
  - Spectral Contrast
- Spectral Image
  - Spectrogram
  - Mel-spectrogram

ANIMAL SOUNDS EXPERIMENT

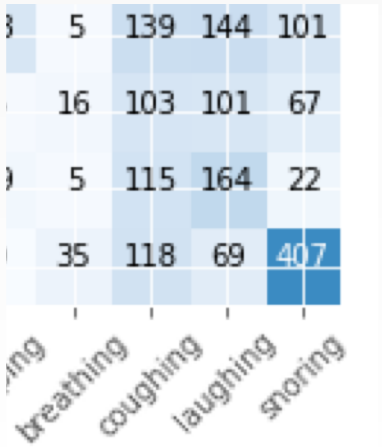
- Feature Vector
  - Learn fold well
  - ~90% accuracy

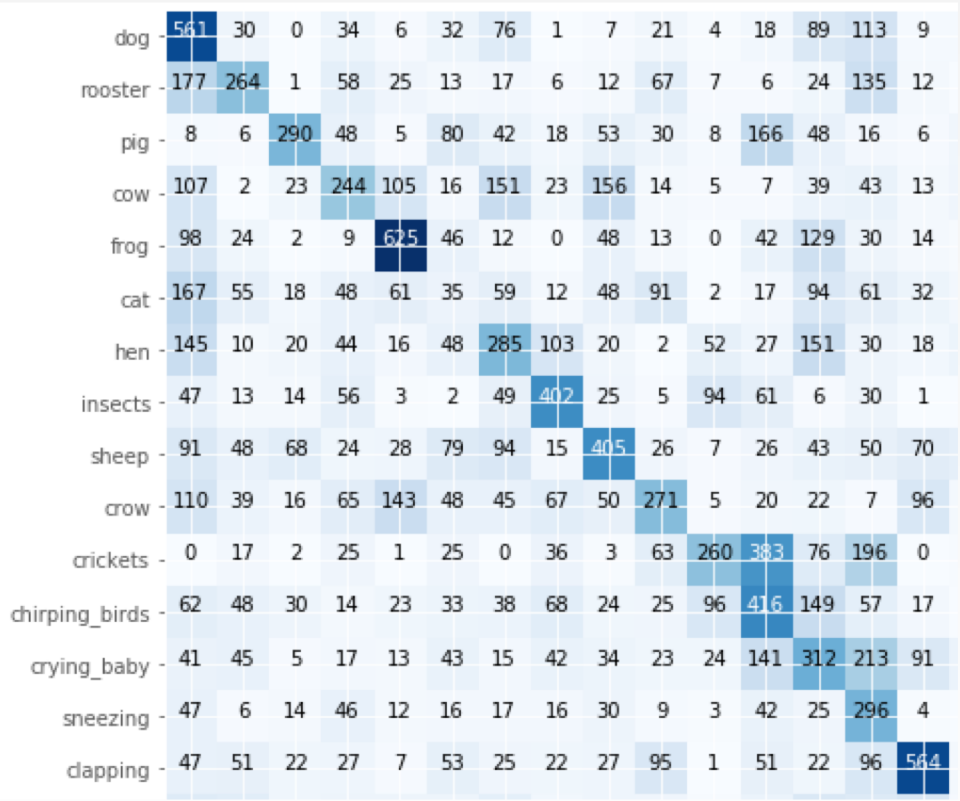




ANIMAL SOUNDS EXPERIMENT

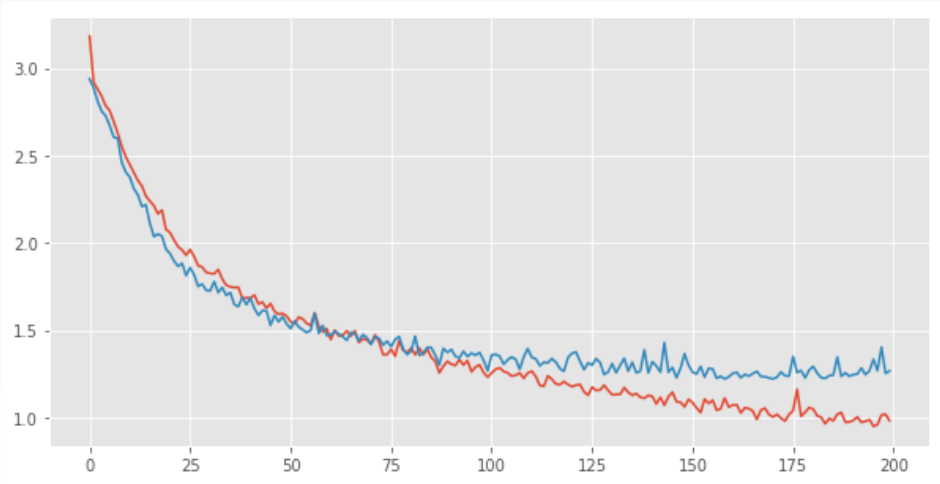
- Actual performance much worse ~30%
- Confusion between breath sounds

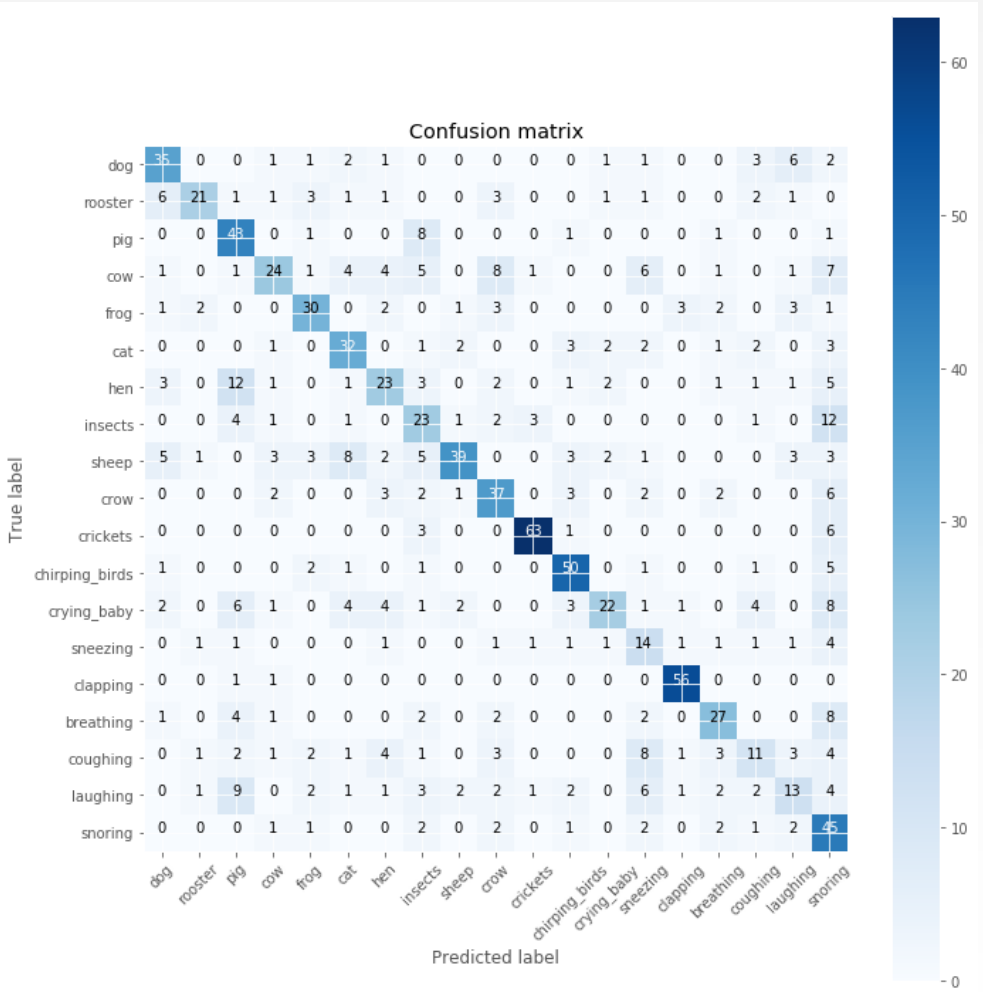




ANIMAL SOUNDS EXPERIMENT

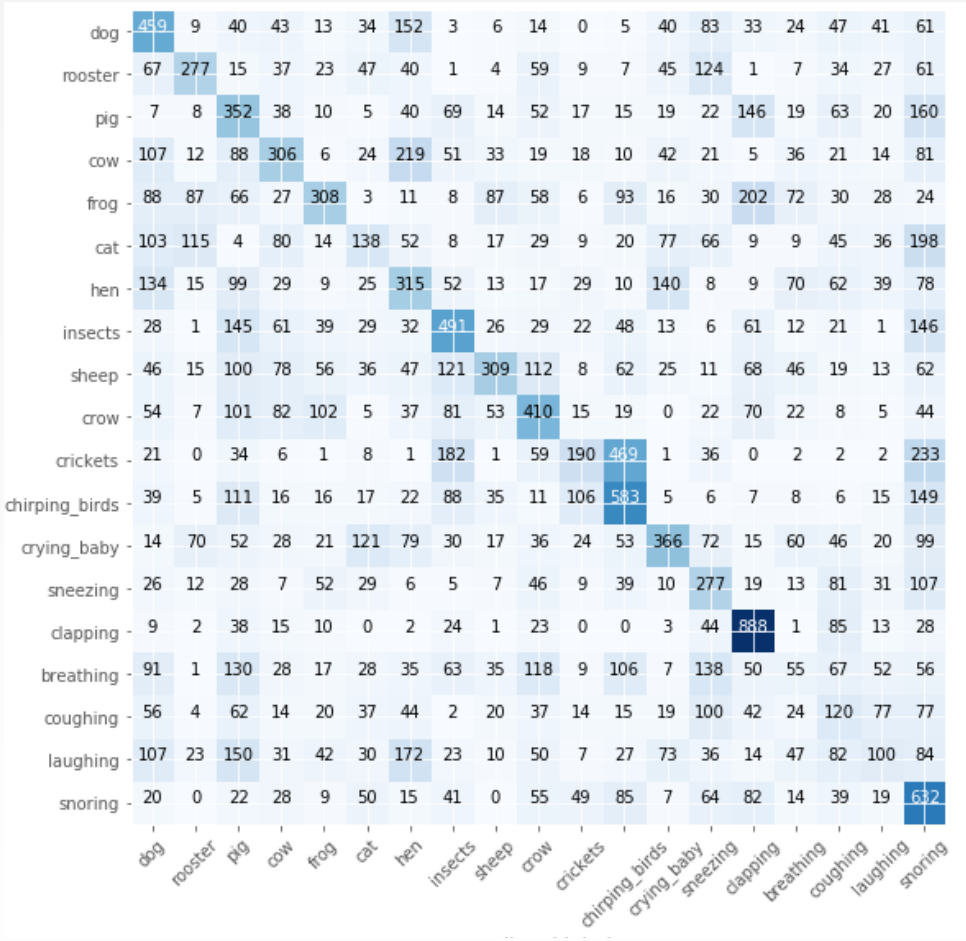
- Spectrogram
  - ~59% accuracy on fold
  - Poor overall performance





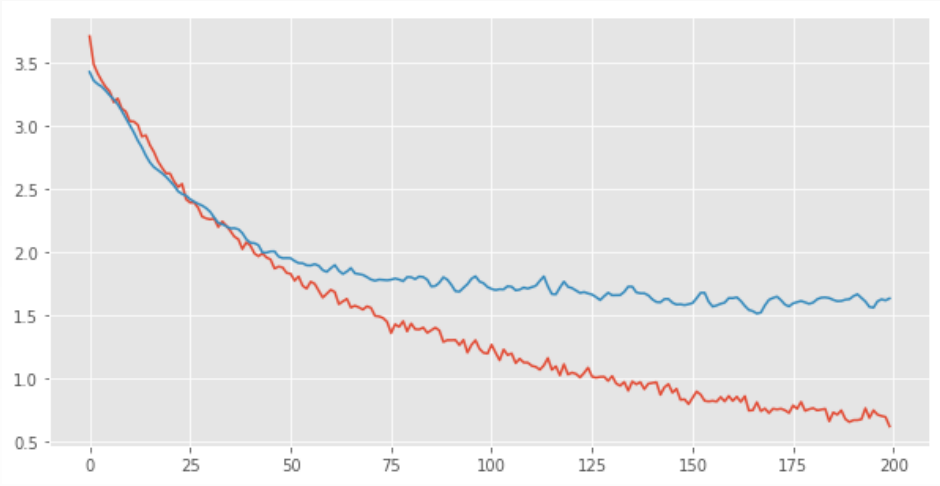
ANIMAL SOUNDS EXPERIMENT

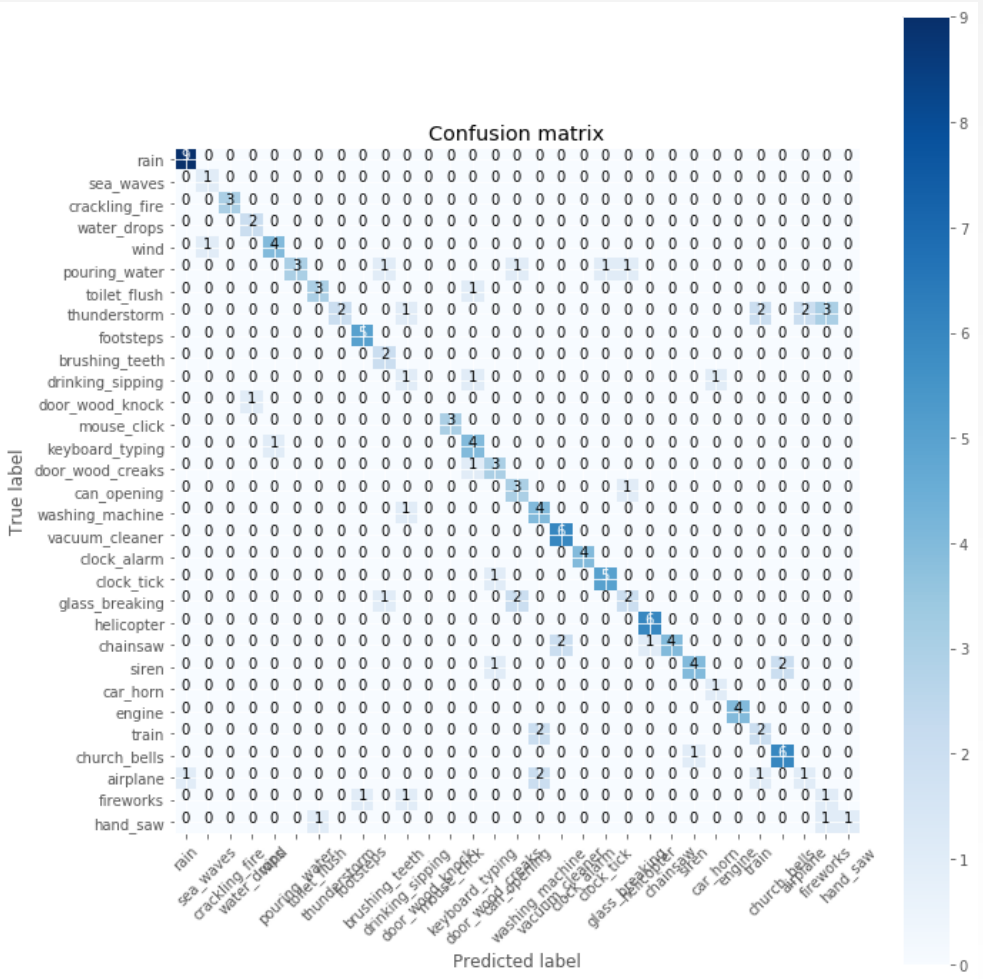
- Actual performance close to feature vector ~30%
- Confusion more spread
- Likely more random guessing



INTERACTING MATERIALS EXPERIMENT

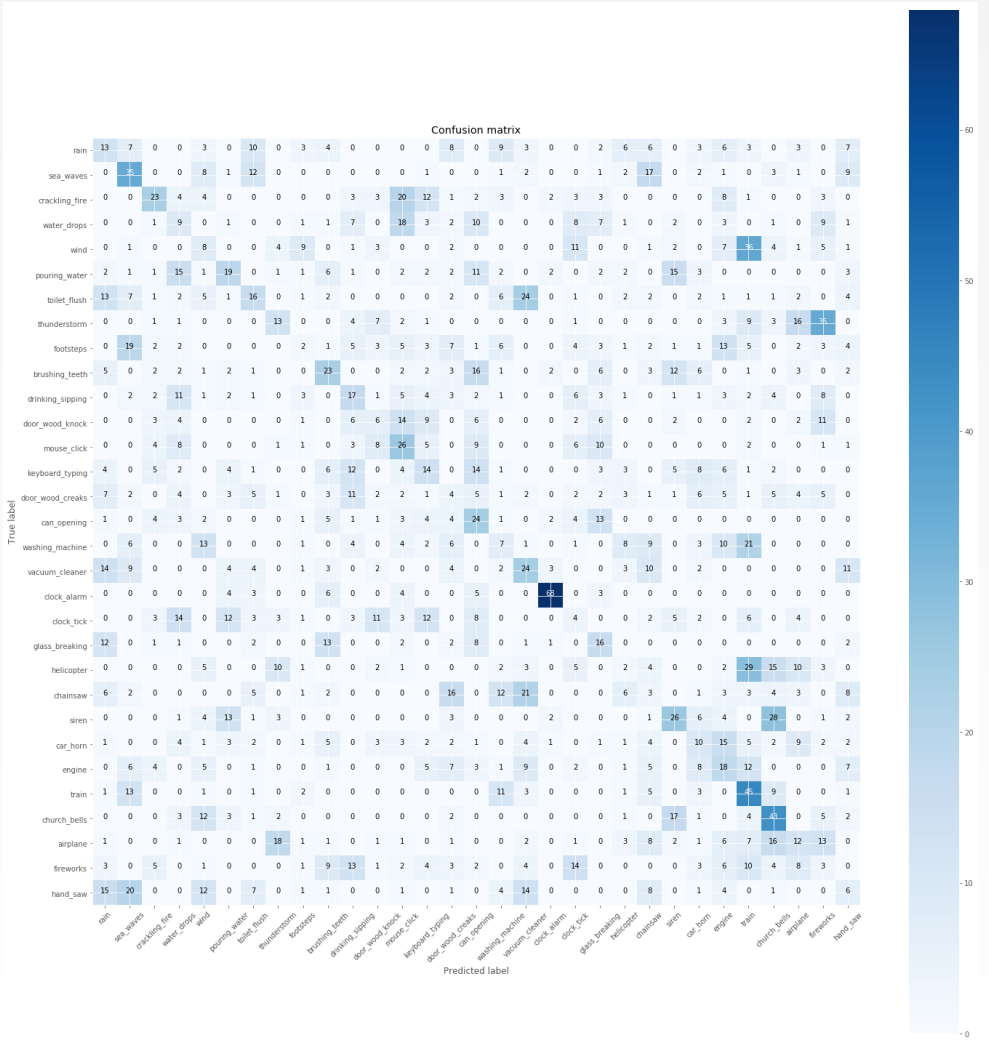
- Feature Vector
  - Fold accuracy 70%
  - Time window has very little effect





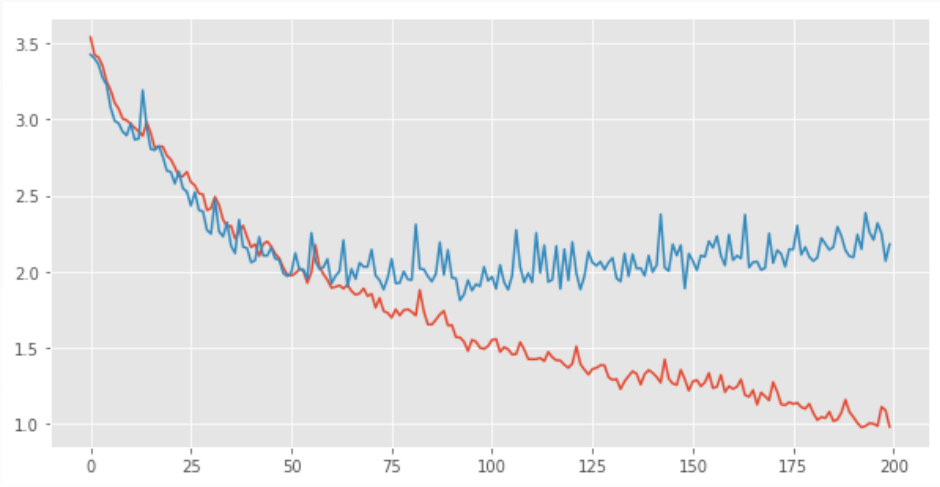
INTERACTING MATERIALS EXPERIMENT

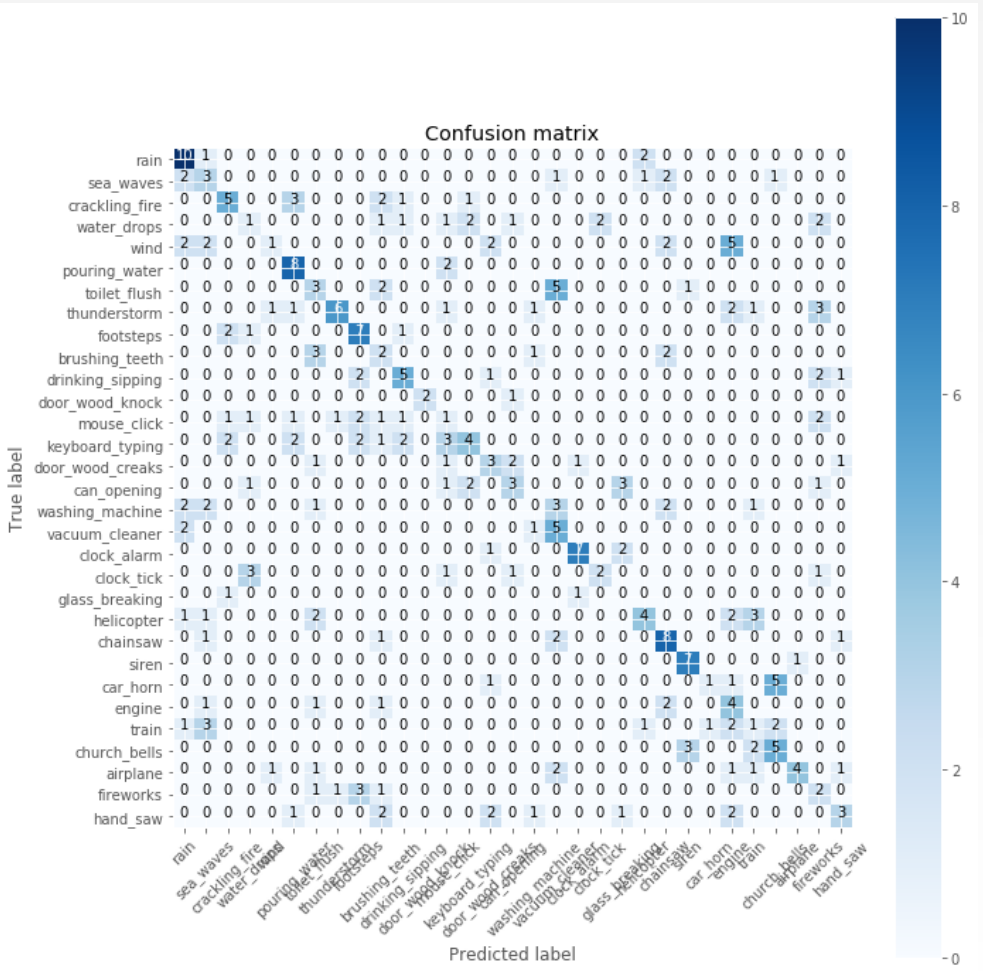
- Actual performance 19%
- Confusion between liquid sounds
  - Dripping, flowing, sipping, etc



INTERACTING MATERIALS EXPERIMENT

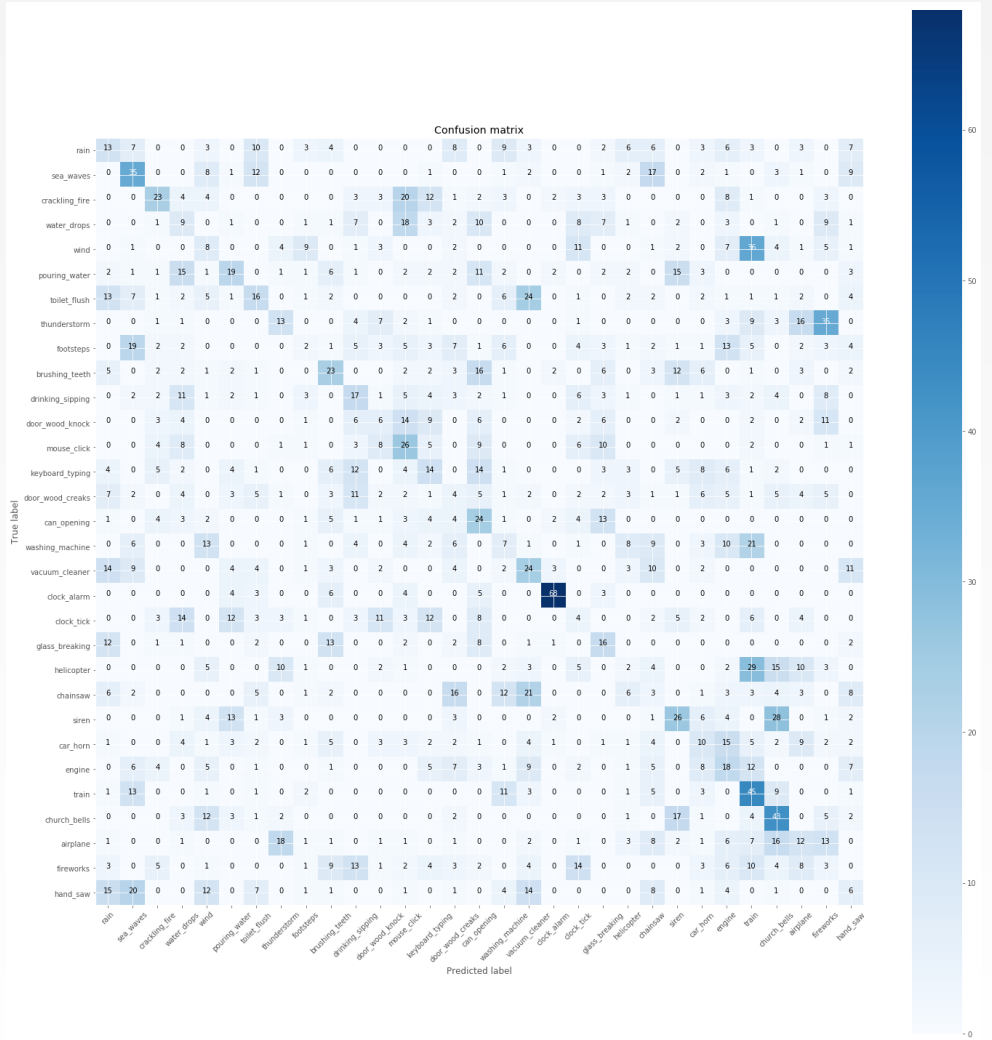
- Spectrogram
  - Fold accuracy 37%
  - Time window has very little effect





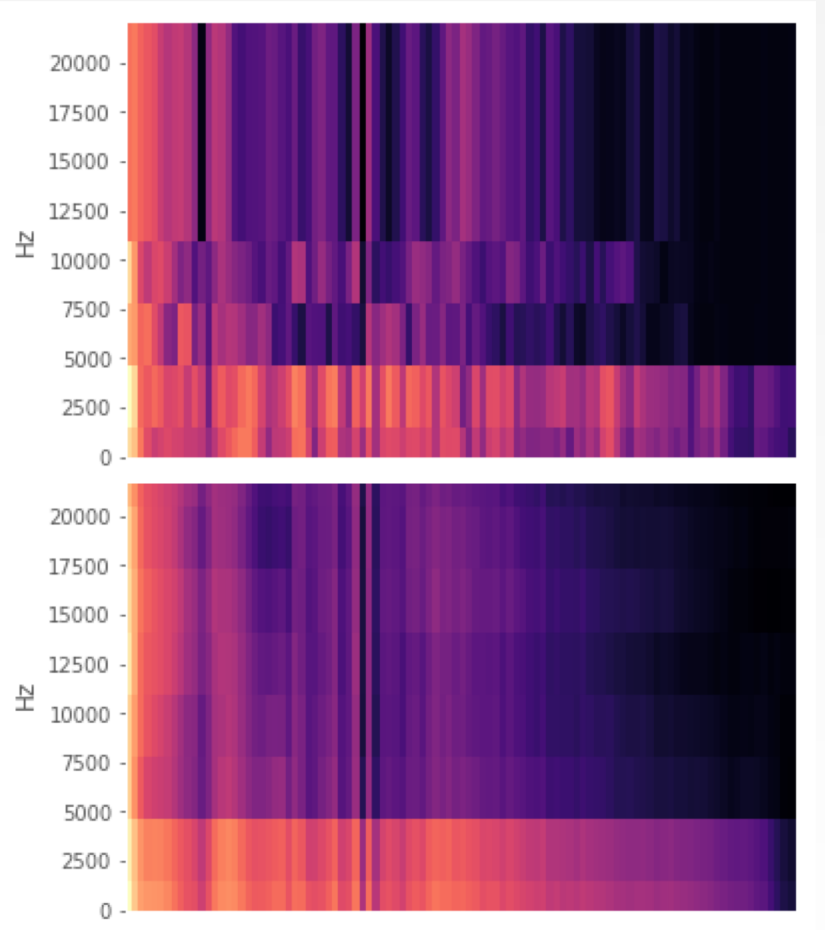
INTERACTING MATERIALS EXPERIMENT

- Actual performance 20%
- Confusion between liquid sounds
  - Dripping, flowing, sipping, etc



FORAY INTO AUTOENCODING

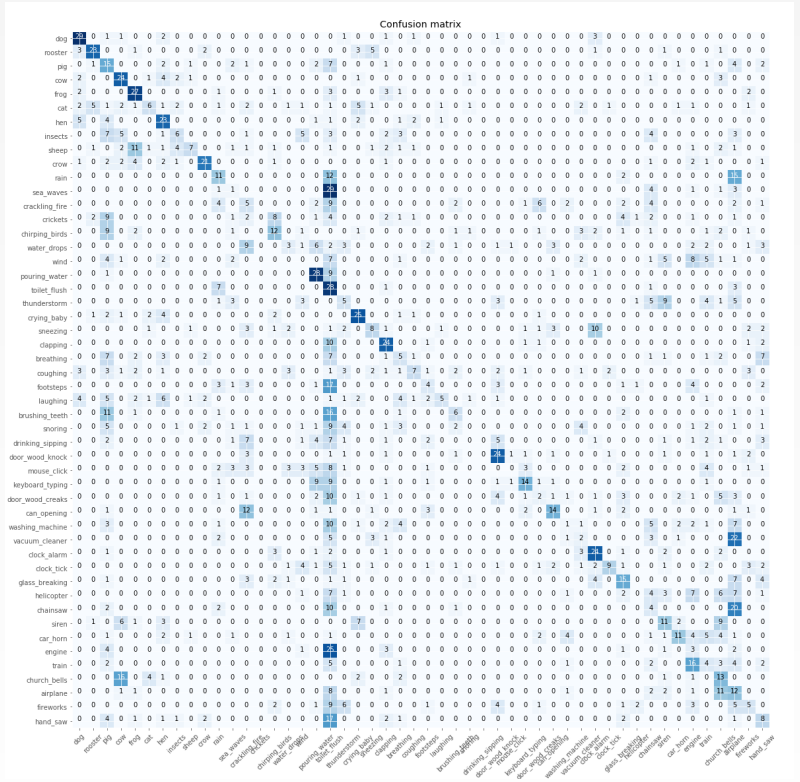
- Autoencoder using NN
- Use dilated audio frames
  - Attempt at temporal feature
- No improvement
- CNN provided no improvement

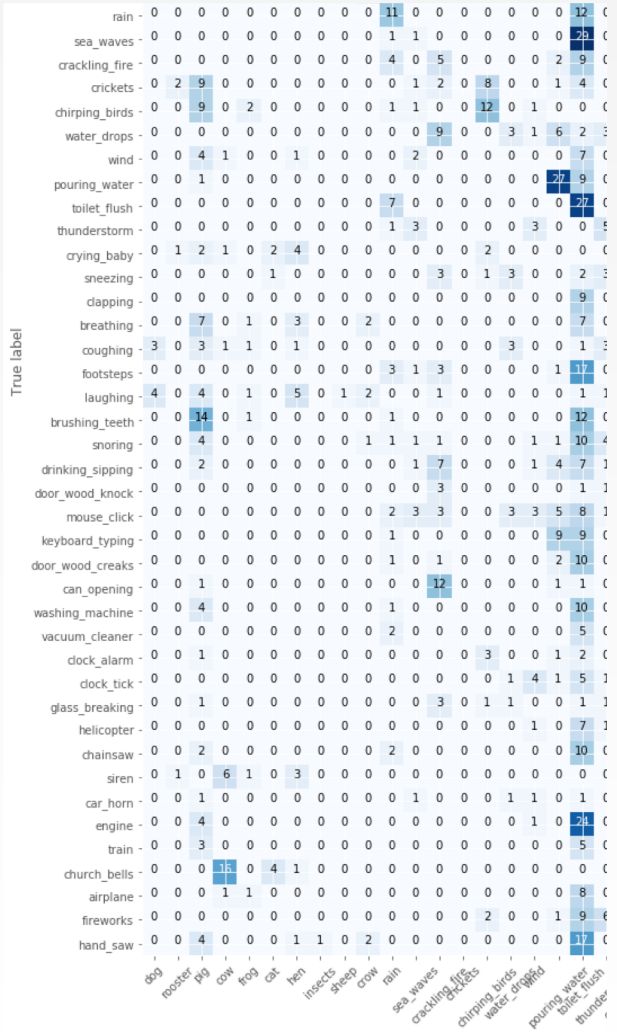


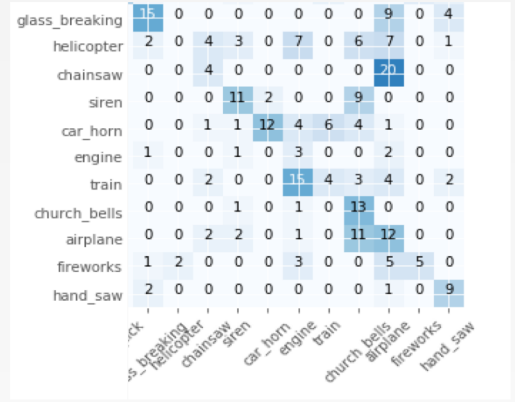
TIME EVALUATION

- Loading folds takes ~6 seconds
- Training Time
  - Top level: ~7 mins
  - Animal sounds: ~7 mins
  - Interacting materials: ~10 mins
- Prediction Time
  - ~11ms/file when predicting classes and probabilities
  - ~8ms/file when only predicting class

OVERALL PERFORMANCE



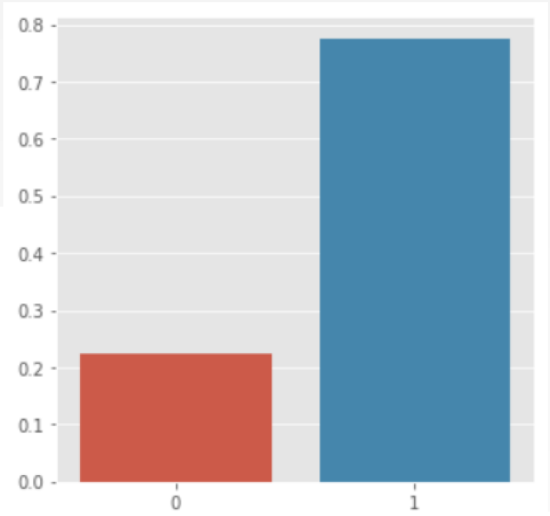


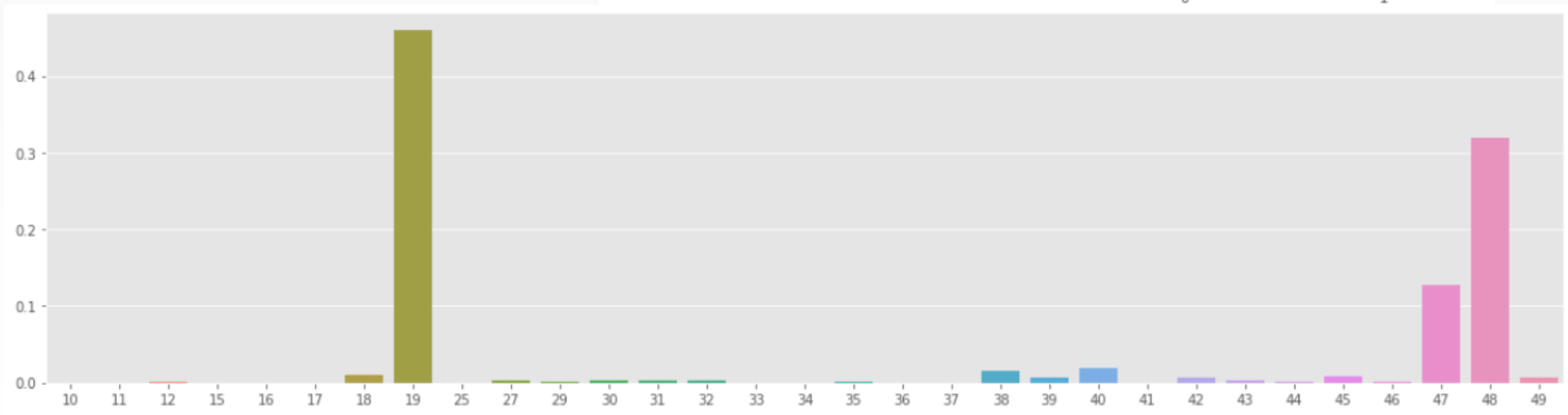


PROBABILISTIC PERFORMANCE

- Not always correct
- Probabilistically Near
- High Level often certain

```
filename        1-115545-C-48.wav
fold                            1
target                         48
category                fireworks
esc10                       False
src_file                   115545
take                            C
h_category                      1
Name: 19, dtype: object
```





GOALS

- Query an unstructured dataset based on a hierarchical taxonomy of sound (95%)
- Build a system that uses deep learning to query unstructured audio datasets (95%)
- Have text queries and allow for complex queries with unrestricted vocabulary (20%)
- Query audio dataset for a contextual event (OOS)
- Make faster and at least as accurate as other querying systems (50%)

FUTURE

- Representation
  - Test with more temporal statistical features
  - Encode features using PCA or other reduction
  - Encode audio directly using LSTM network
- Lower Hierarchy
  - Make NNs totally specialized
- Audio feature extraction
  - Implement on GPU
- Wordnet lowest common denominator

QUESTIONS?