

DATA ANALYTICS USING DEEP LEARNING

GT 8803 // FALL 2018 //

JACOB LOGAS

EFFICIENT CONTENT-BASED AUDIO RETRIEVAL

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

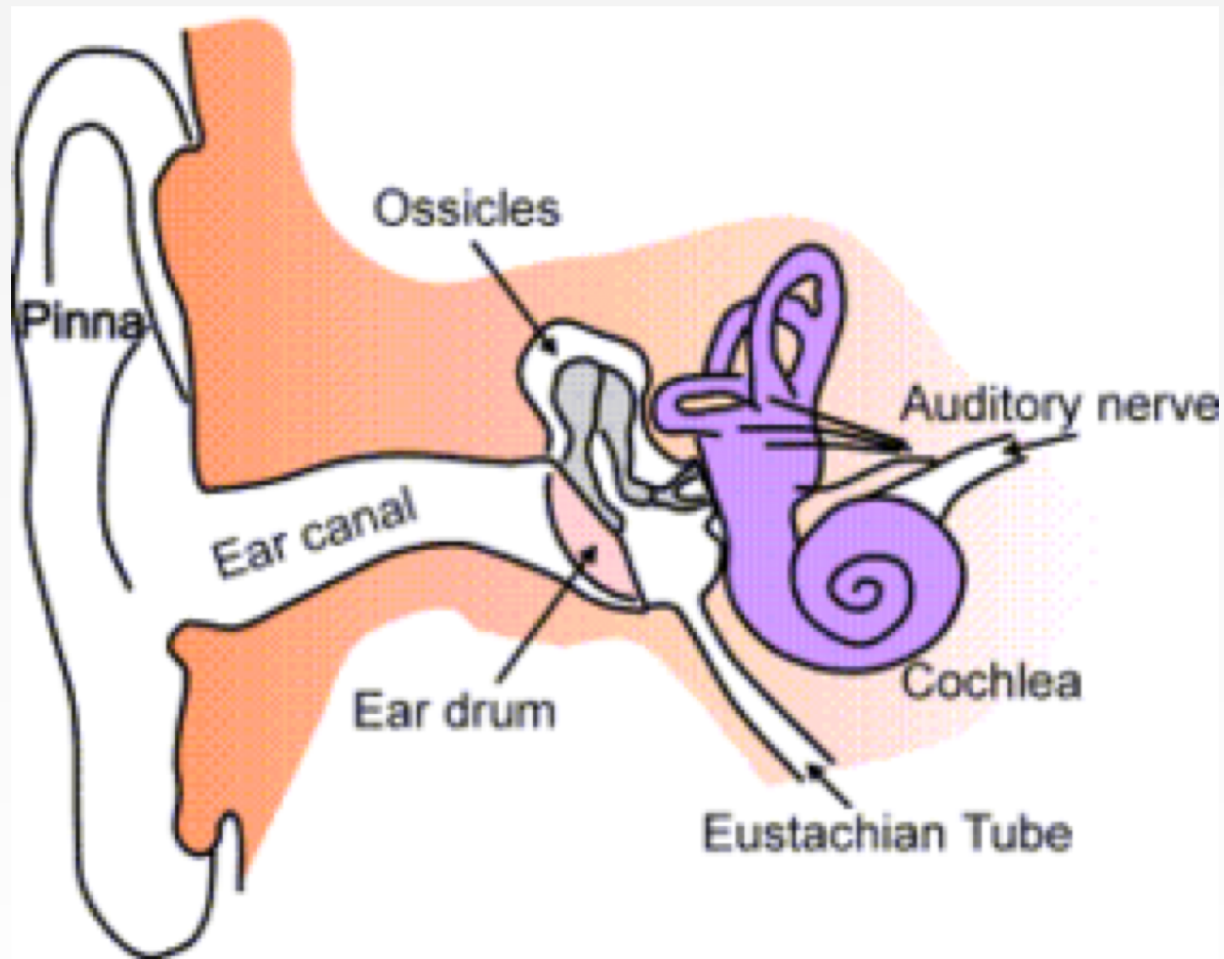
RESEARCH QUESTIONS

- How does a hierarchical approach to audio retrieval affect 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?

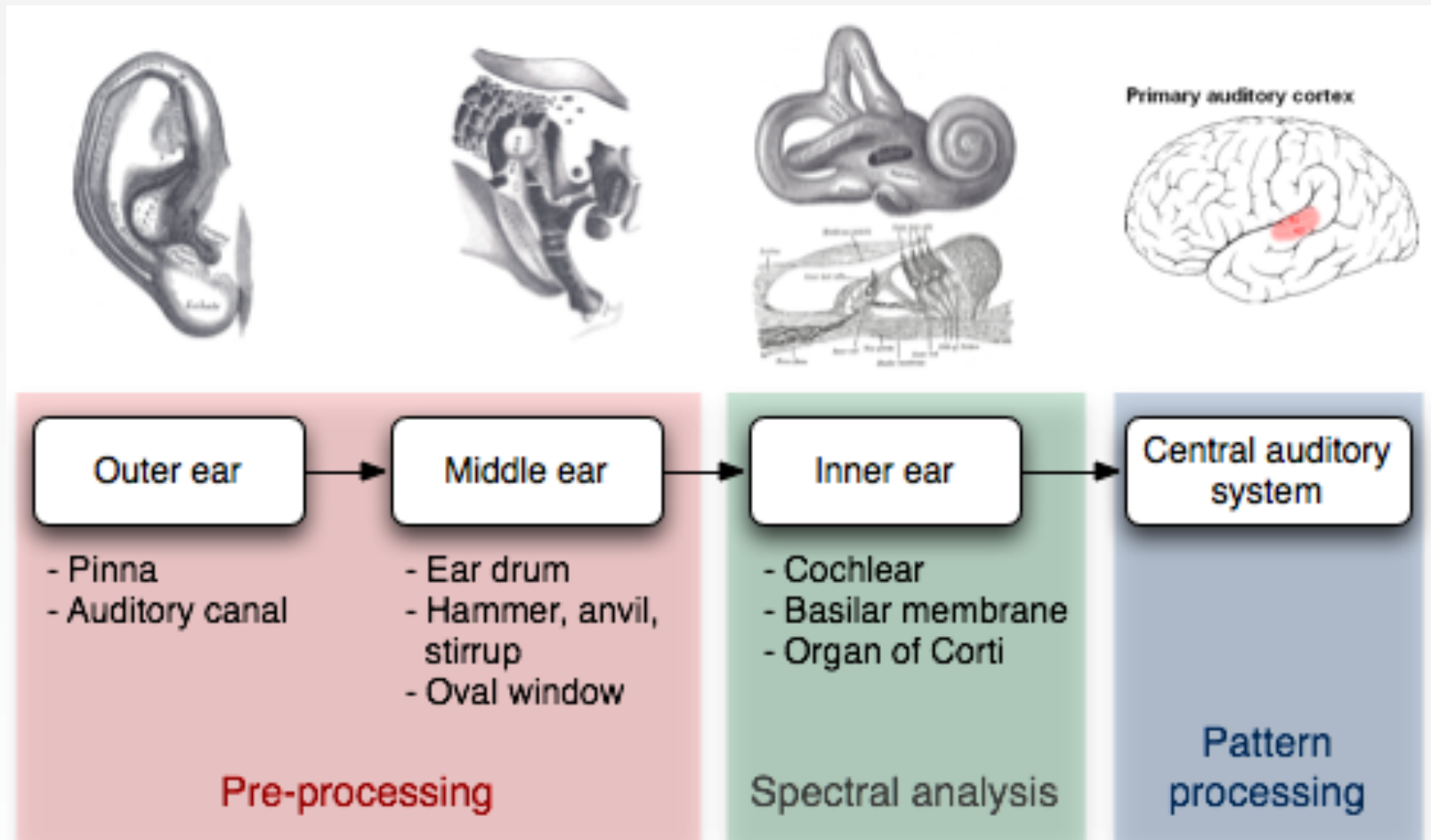
GOALS

- 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

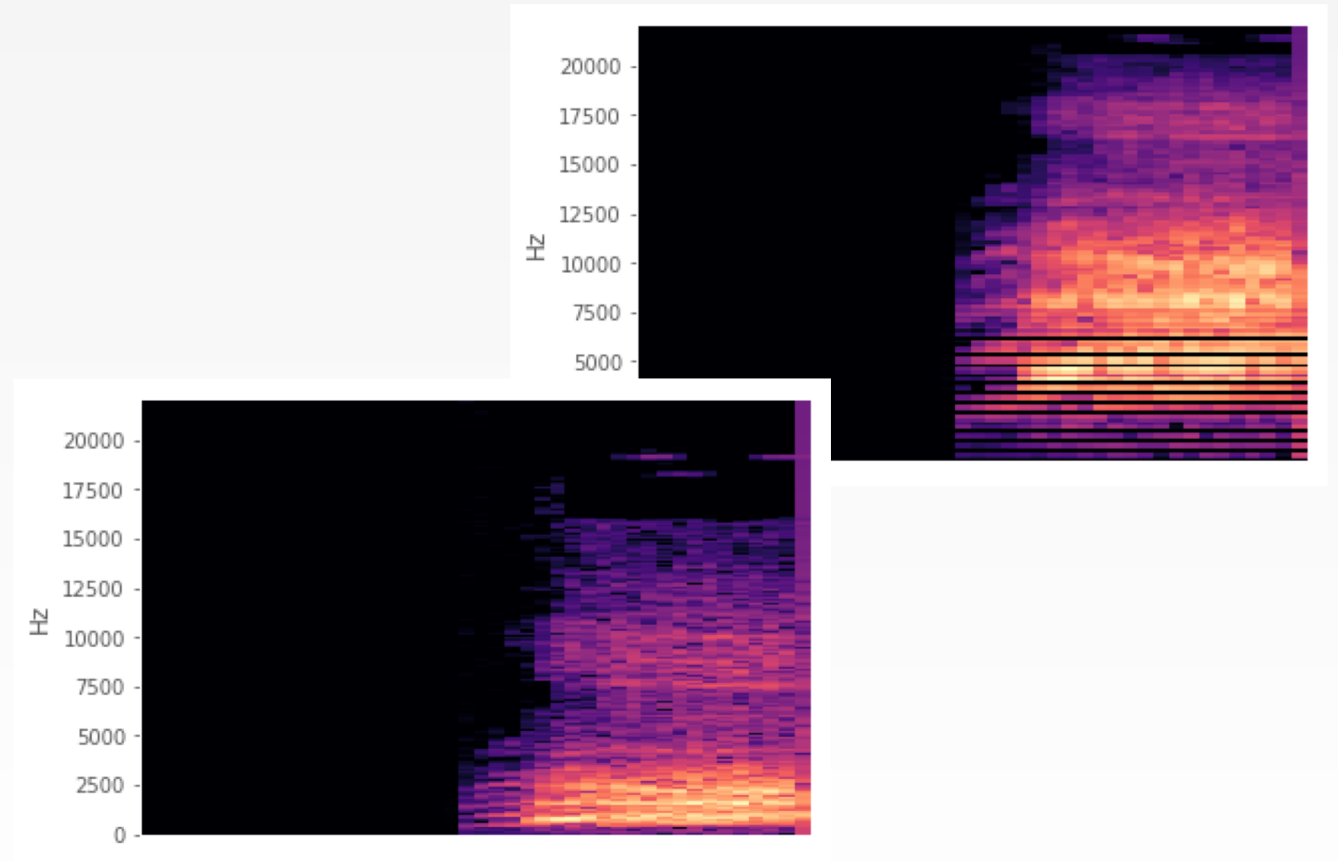


WHY NOT JUST USE CNNs?



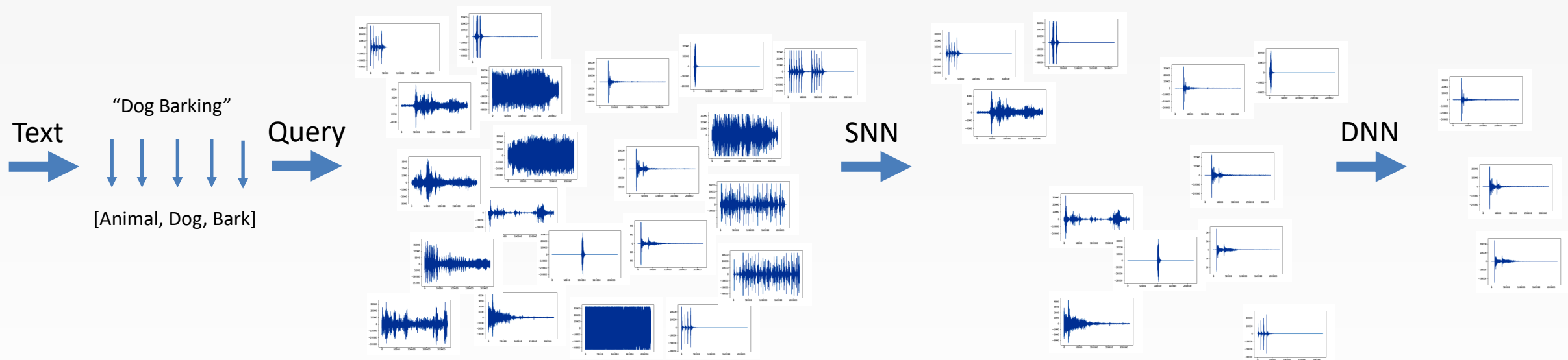
REPRESENTATION

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



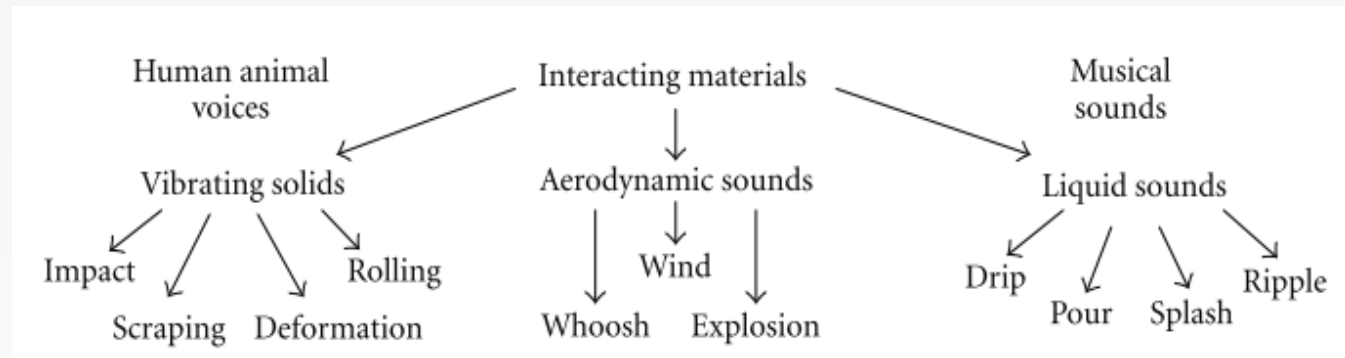
PROPOSED APPROACH

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

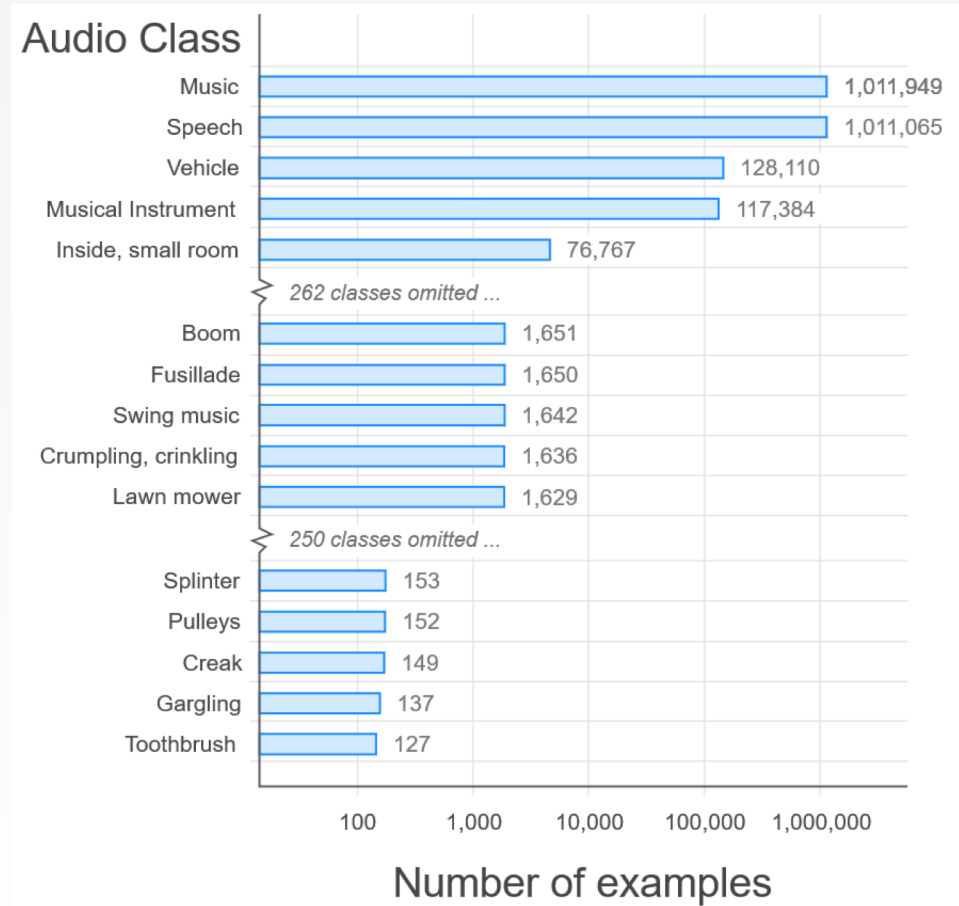
DEVELOPMENT DATASET

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

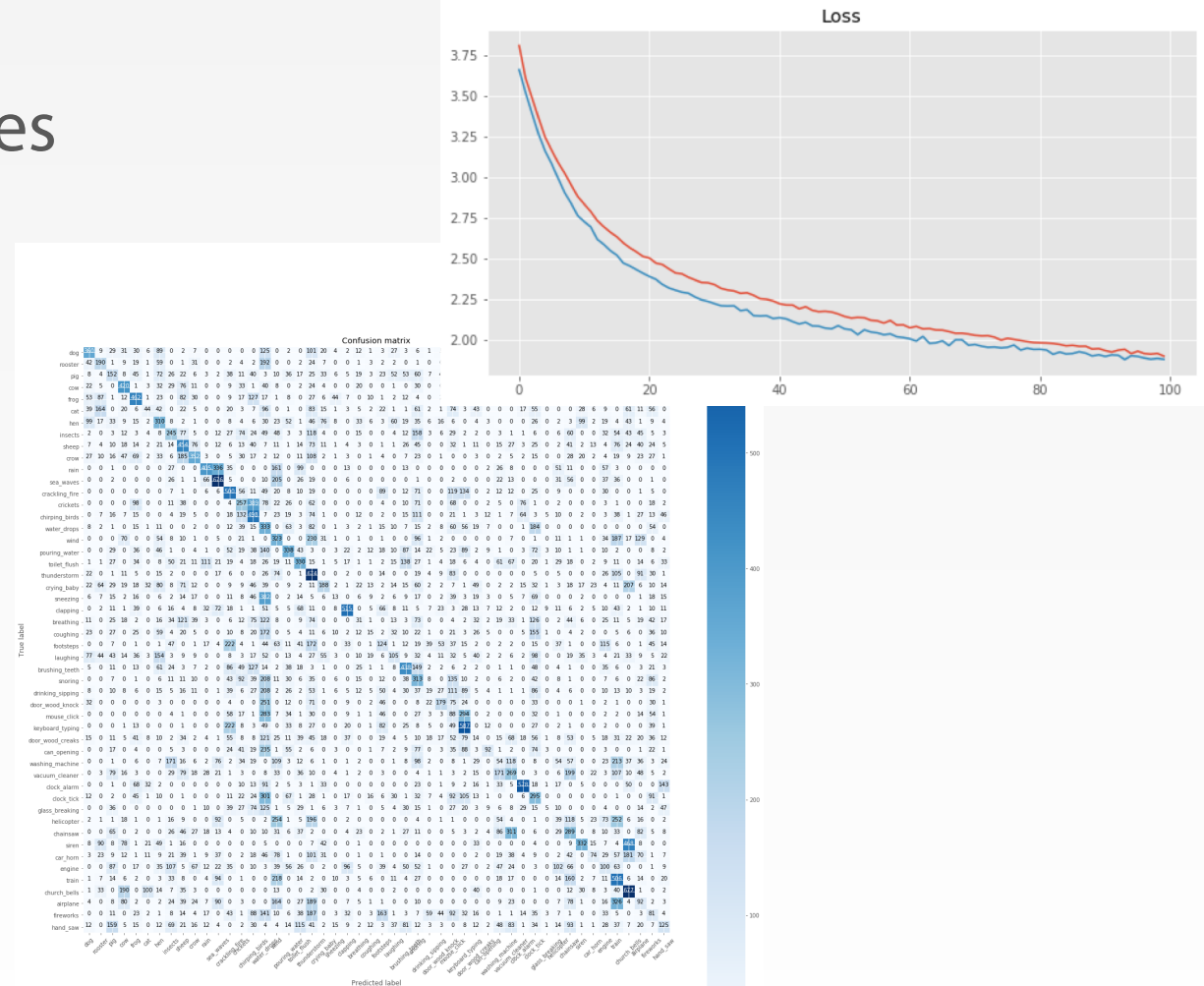
TARGET DATASET



EXPERIMENTS

DEFAULT CNN

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

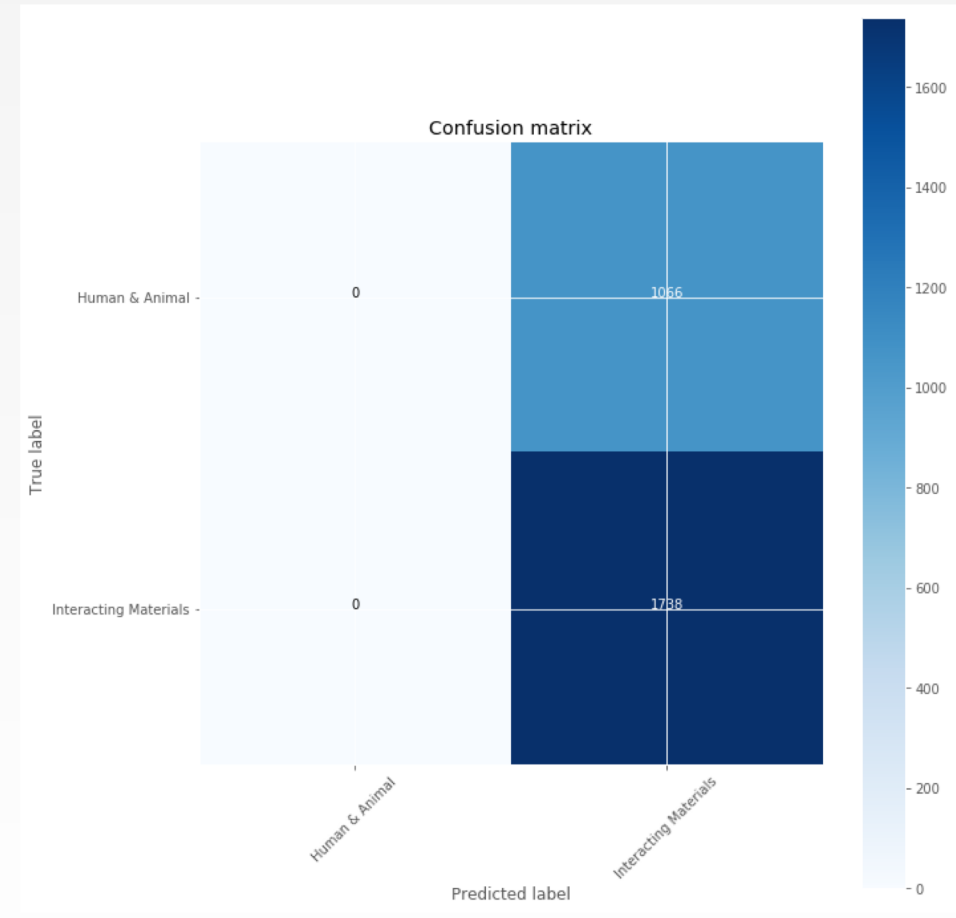


TOP LEVEL EXPERIMENTS

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

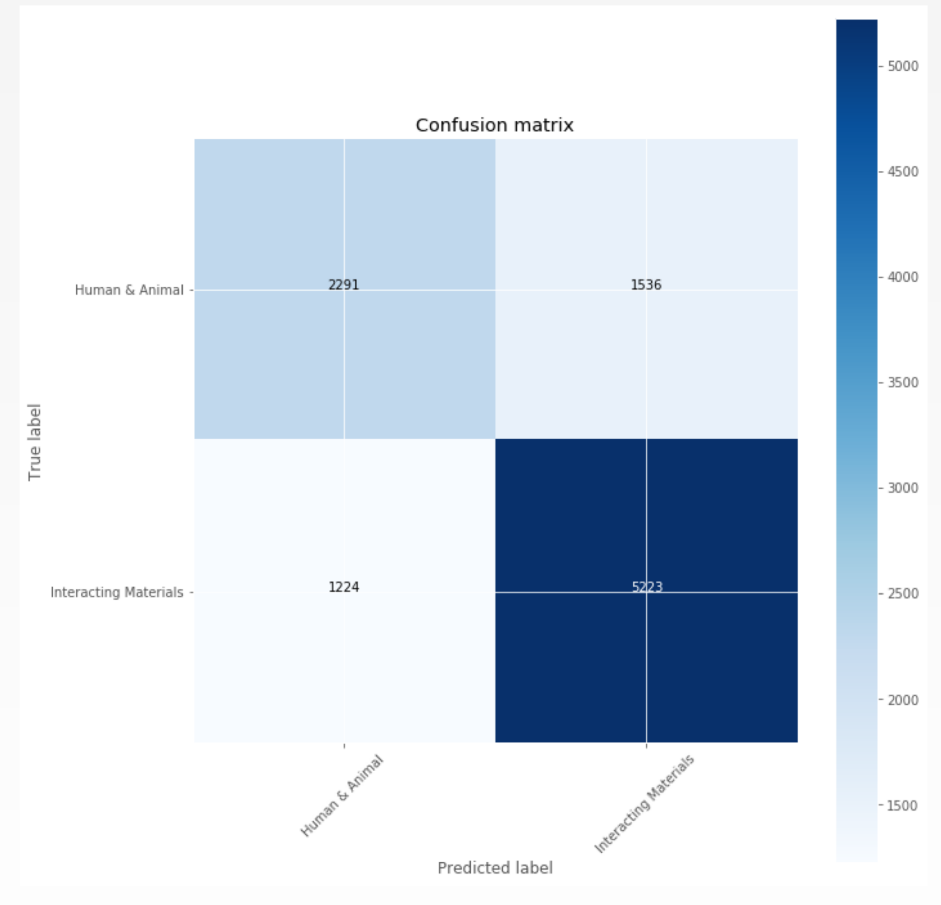
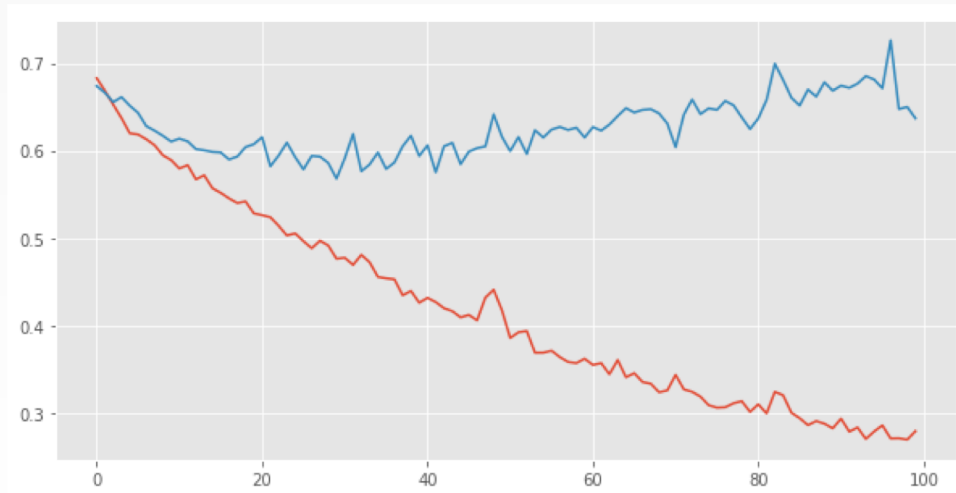
TOP LEVEL EXPERIMENTS

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



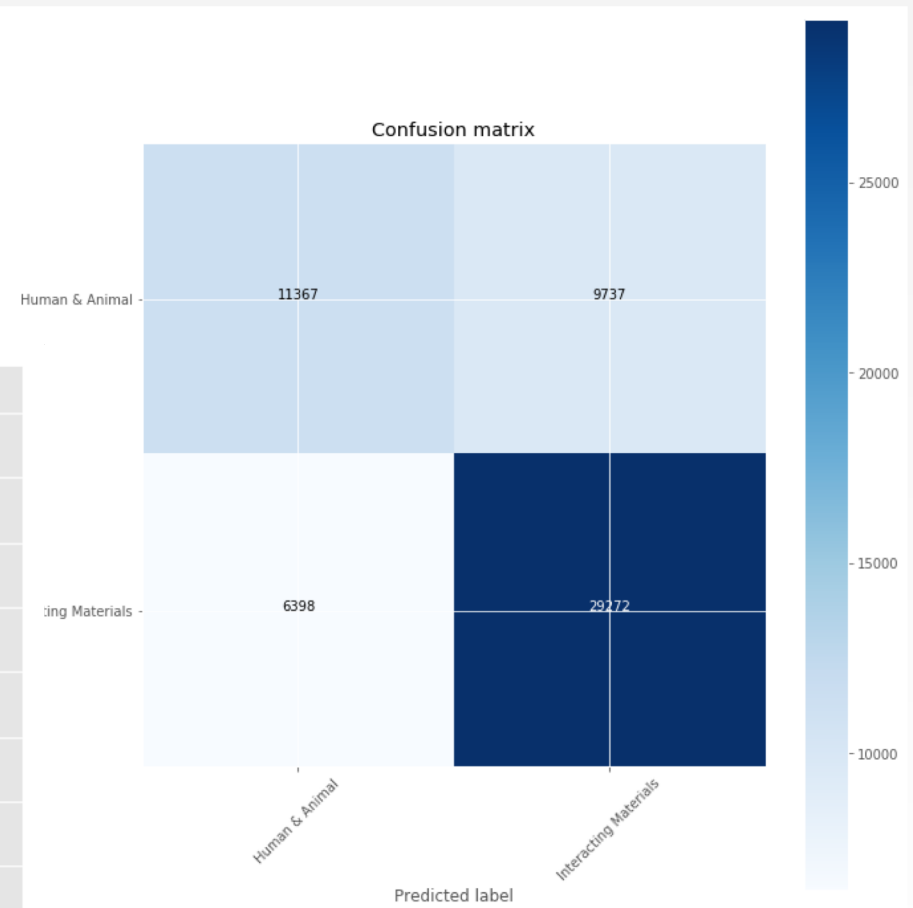
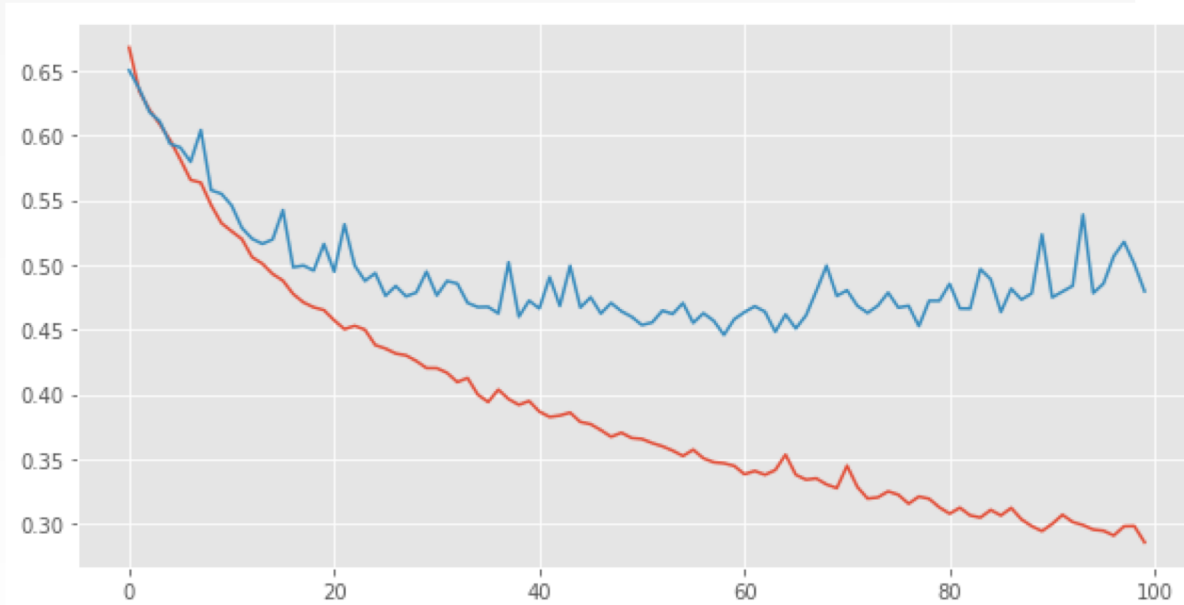
TOP LEVEL EXPERIMENTS

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



HIERARCHY TOP LEVEL

- 250ms Window
– 72% Testing Accuracy

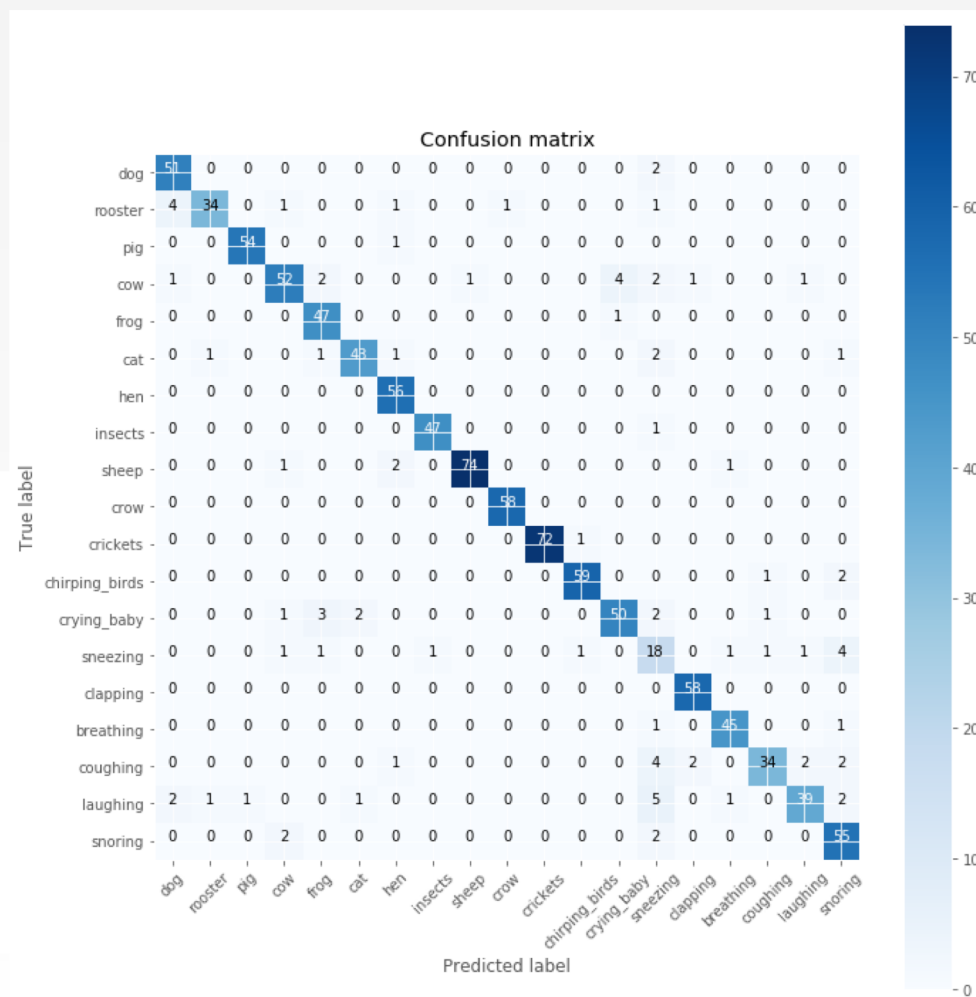
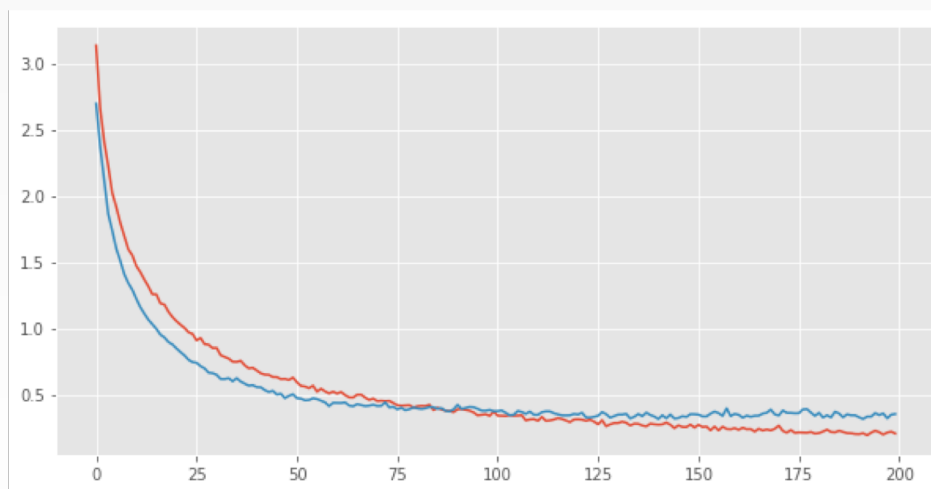


LOWER LEVEL EXPERIMENTS

- 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

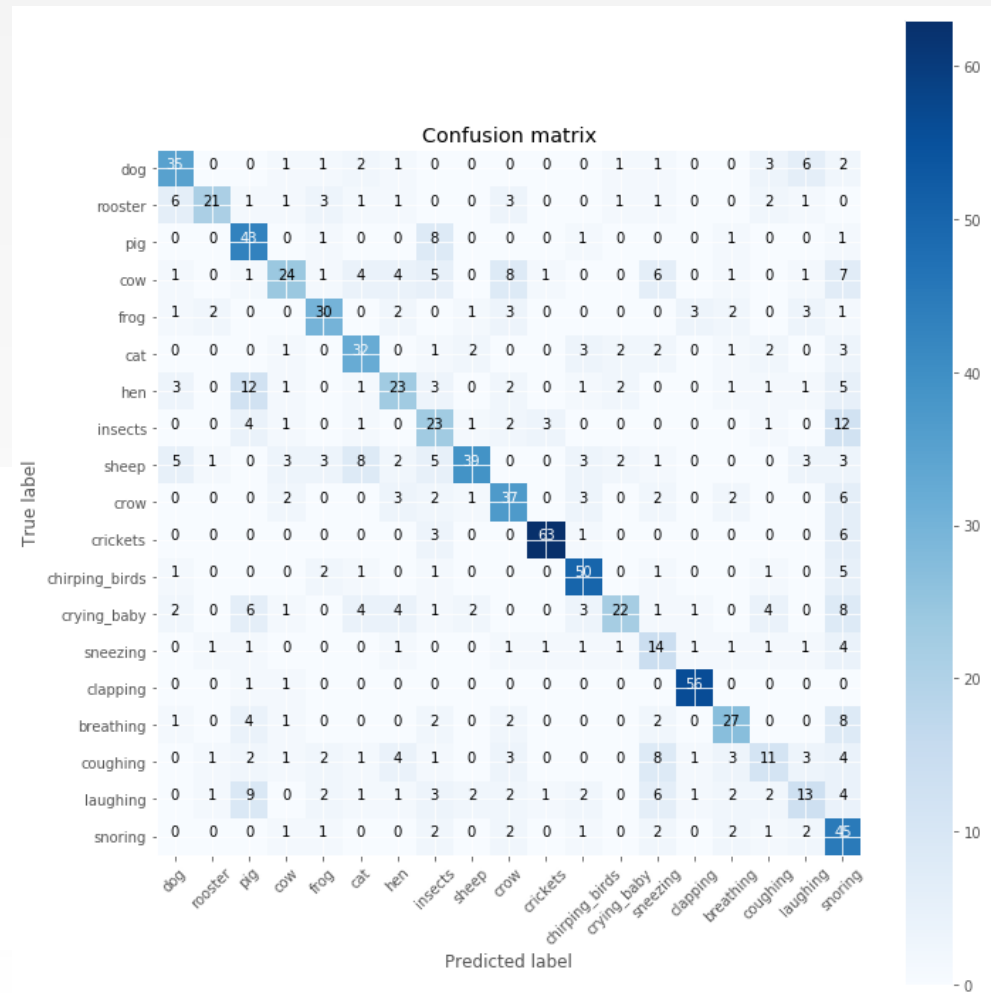
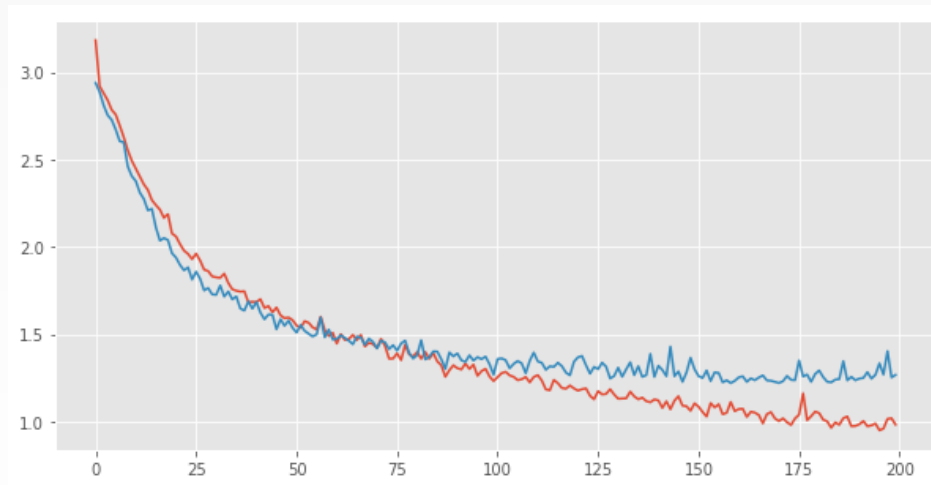
- Actual performance much worse ~30%
- Confusion between breath sounds

3	5	139	144	101
	16	103	101	67
9	5	115	164	22
	35	118	69	407
ing	breathing	coughing	laughing	snoring

dog	561	30	0	34	6	32	76	1	7	21	4	18	89	113	9
rooster	177	264	1	58	25	13	17	6	12	67	7	6	24	135	12
pig	8	6	290	48	5	80	42	18	53	30	8	166	48	16	6
cow	107	2	23	244	105	16	151	23	156	14	5	7	39	43	13
frog	98	24	2	9	625	46	12	0	48	13	0	42	129	30	14
cat	167	55	18	48	61	35	59	12	48	91	2	17	94	61	32
hen	145	10	20	44	16	48	285	103	20	2	52	27	151	30	18
insects	47	13	14	56	3	2	49	402	25	5	94	61	6	30	1
sheep	91	48	68	24	28	79	94	15	405	26	7	26	43	50	70
crow	110	39	16	65	143	48	45	67	50	271	5	20	22	7	96
crickets	0	17	2	25	1	25	0	36	3	63	260	383	76	196	0
chirping_birds	62	48	30	14	23	33	38	68	24	25	96	416	149	57	17
crying_baby	41	45	5	17	13	43	15	42	34	23	24	141	312	213	91
sneezing	47	6	14	46	12	16	17	16	30	9	3	42	25	296	4
dapping	47	51	22	27	7	53	25	22	27	95	1	51	22	96	564

ANIMAL SOUNDS EXPERIMENT

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



ANIMAL SOUNDS EXPERIMENT

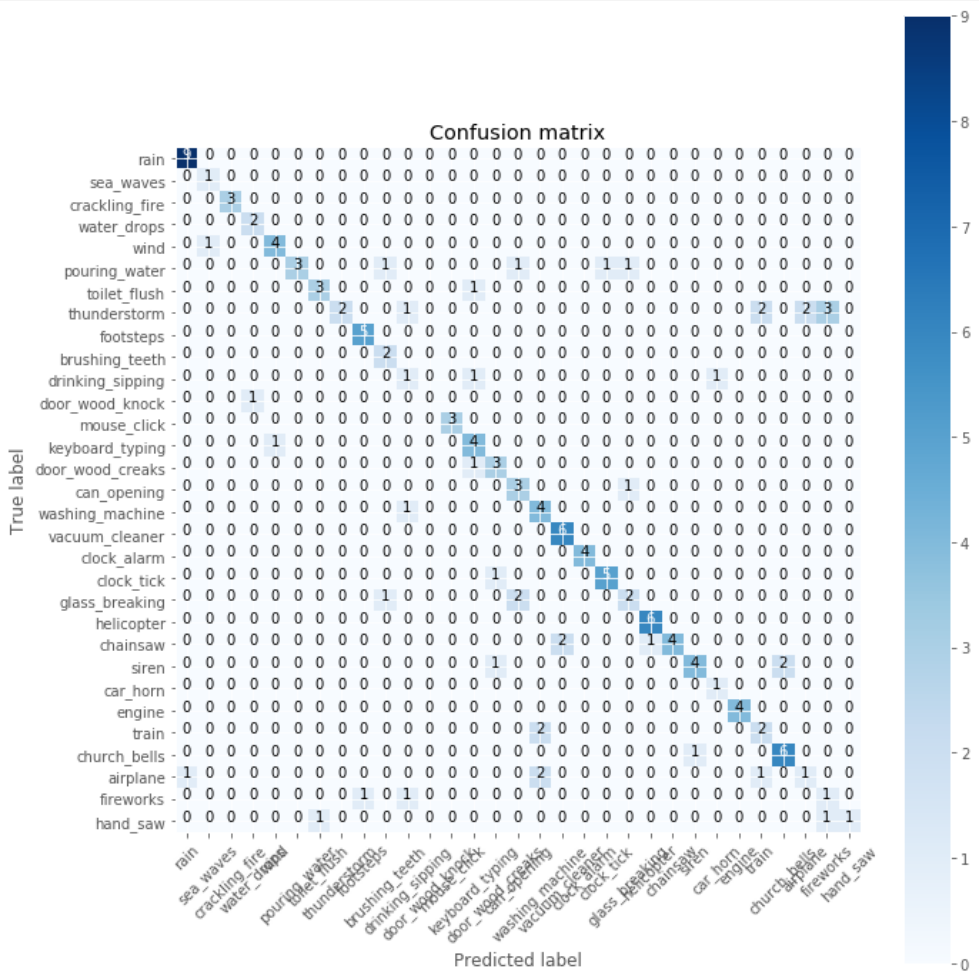
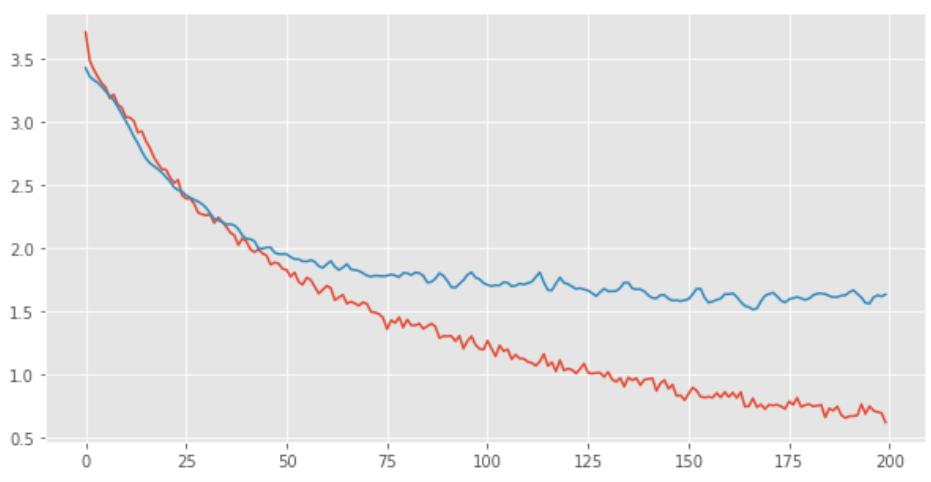
- Actual performance close to feature vector ~30%
- Confusion more spread
- Likely more random guessing

A confusion matrix showing the relationship between actual and predicted classes for 20 different animal sounds. The classes are: dog, rooster, pig, cow, frog, cat, hen, insects, sheep, crow, crickets, chirping_birds, crying_baby, sneezing, clapping, breathing, coughing, laughing, and snoring. The diagonal elements represent correct classifications, and the off-diagonal elements represent misclassifications. The matrix is symmetric, indicating that the model's performance is consistent regardless of the direction of the confusion.

dog	459	9	40	43	13	34	152	3	6	14	0	5	40	83	33	24	47	41	61
rooster	67	277	15	37	23	47	40	1	4	59	9	7	45	124	1	7	34	27	61
pig	7	8	352	38	10	5	40	69	14	52	17	15	19	22	146	19	63	20	160
cow	107	12	88	306	6	24	219	51	33	19	18	10	42	21	5	36	21	14	81
frog	88	87	66	27	308	3	11	8	87	58	6	93	16	30	202	72	30	28	24
cat	103	115	4	80	14	138	52	8	17	29	9	20	77	66	9	9	45	36	198
hen	134	15	99	29	9	25	315	52	13	17	29	10	140	8	9	70	62	39	78
insects	28	1	145	61	39	29	32	491	26	29	22	48	13	6	61	12	21	1	146
sheep	46	15	100	78	56	36	47	121	309	112	8	62	25	11	68	46	19	13	62
crow	54	7	101	82	102	5	37	81	53	410	15	19	0	22	70	22	8	5	44
crickets	21	0	34	6	1	8	1	182	1	59	190	469	1	36	0	2	2	2	233
chirping_birds	39	5	111	16	16	17	22	88	35	11	106	583	5	6	7	8	6	15	149
crying_baby	14	70	52	28	21	121	79	30	17	36	24	53	366	72	15	60	46	20	99
sneezing	26	12	28	7	52	29	6	5	7	46	9	39	10	277	19	13	81	31	107
clapping	9	2	38	15	10	0	2	24	1	23	0	0	3	44	888	1	85	13	28
breathing	91	1	130	28	17	28	35	63	35	118	9	106	7	138	50	55	67	52	56
coughing	56	4	62	14	20	37	44	2	20	37	14	15	19	100	42	24	120	77	77
laughing	107	23	150	31	42	30	172	23	10	50	7	27	73	36	14	47	82	100	84
snoring	20	0	22	28	9	50	15	41	0	55	49	85	7	64	82	14	39	19	632

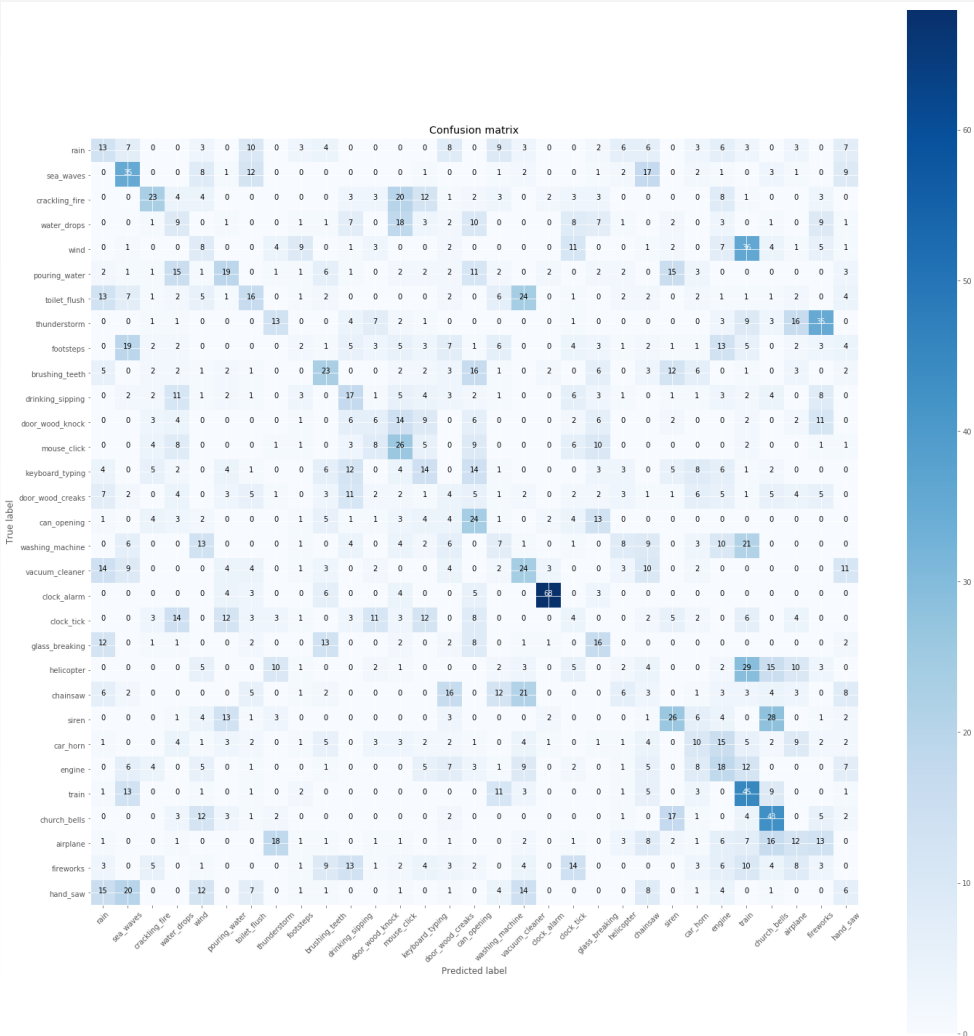
INTERACTING MATERIALS EXPERIMENT

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



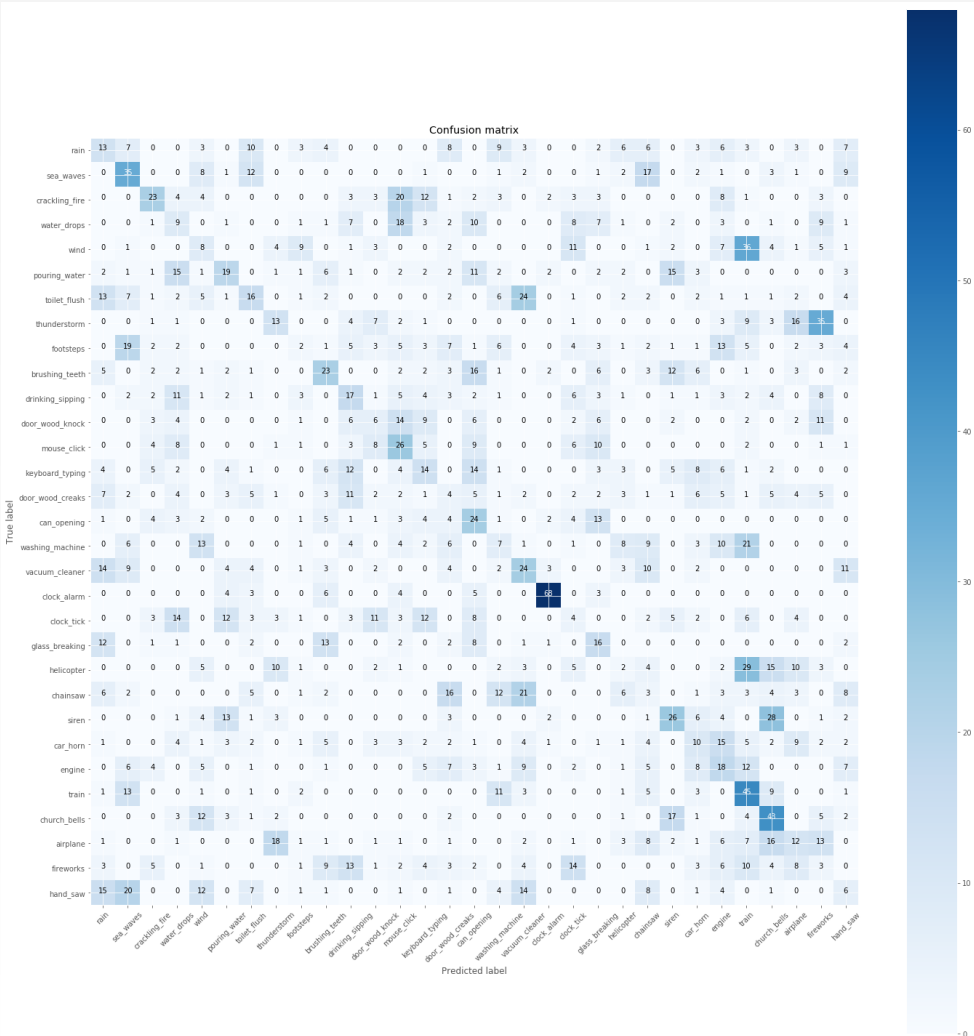
INTERACTING MATERIALS EXPERIMENT

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



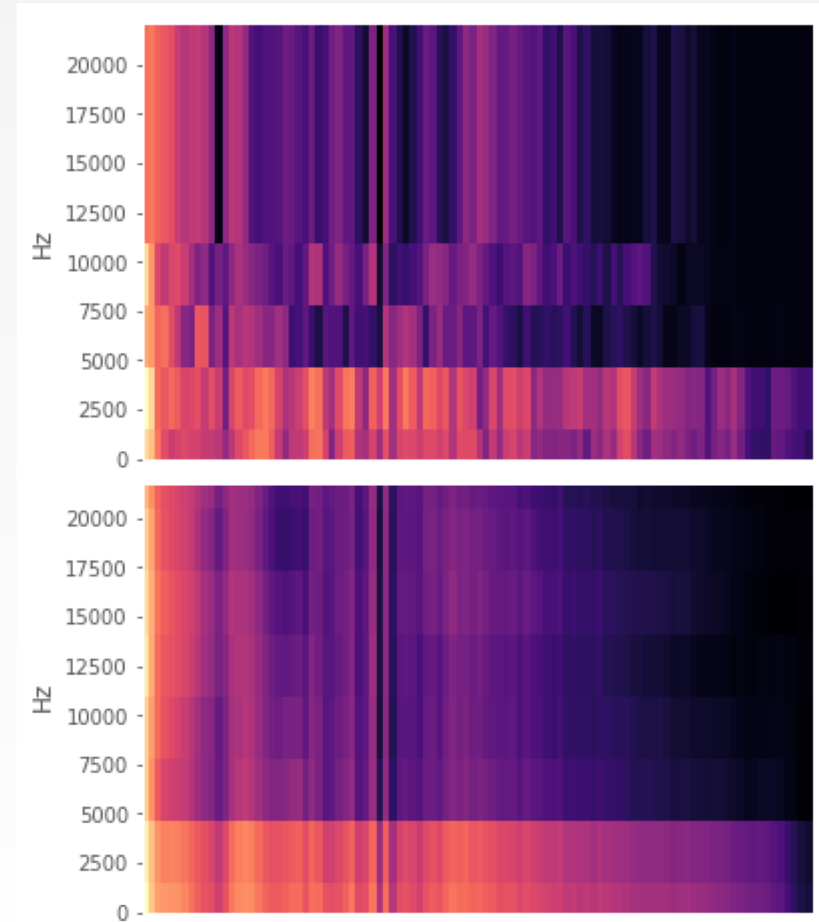
INTERACTING MATERIALS EXPERIMENT

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



FORAY INTO AUTOENCODING

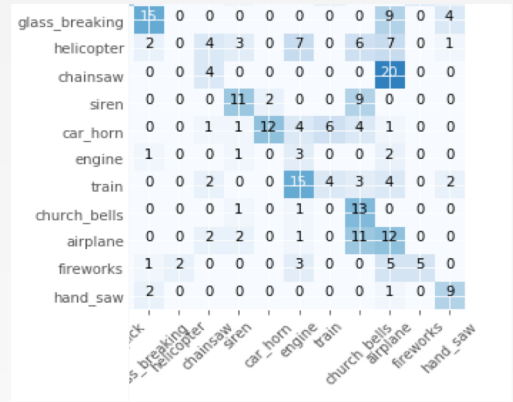
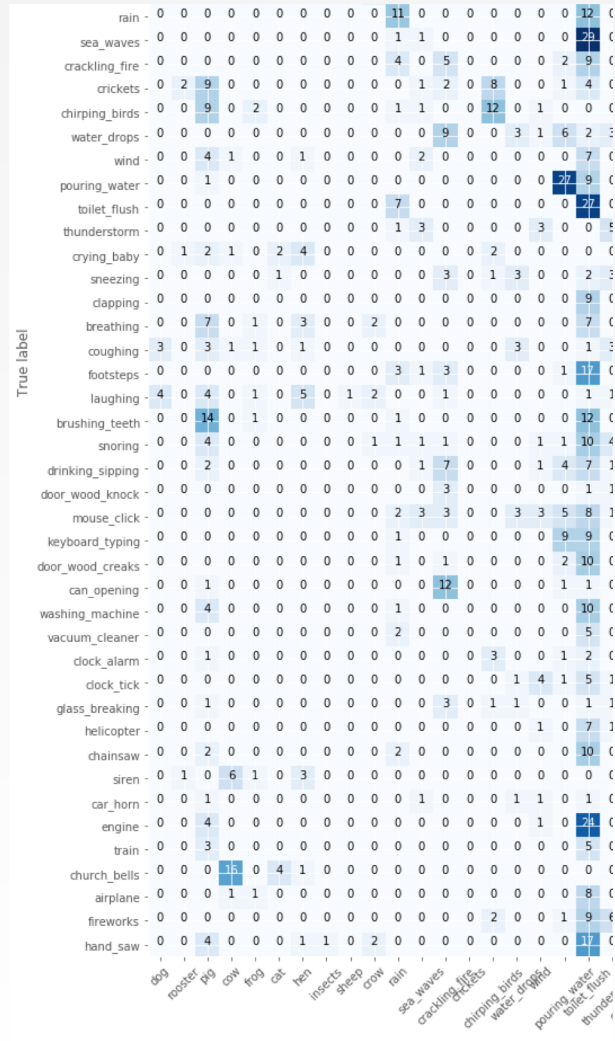
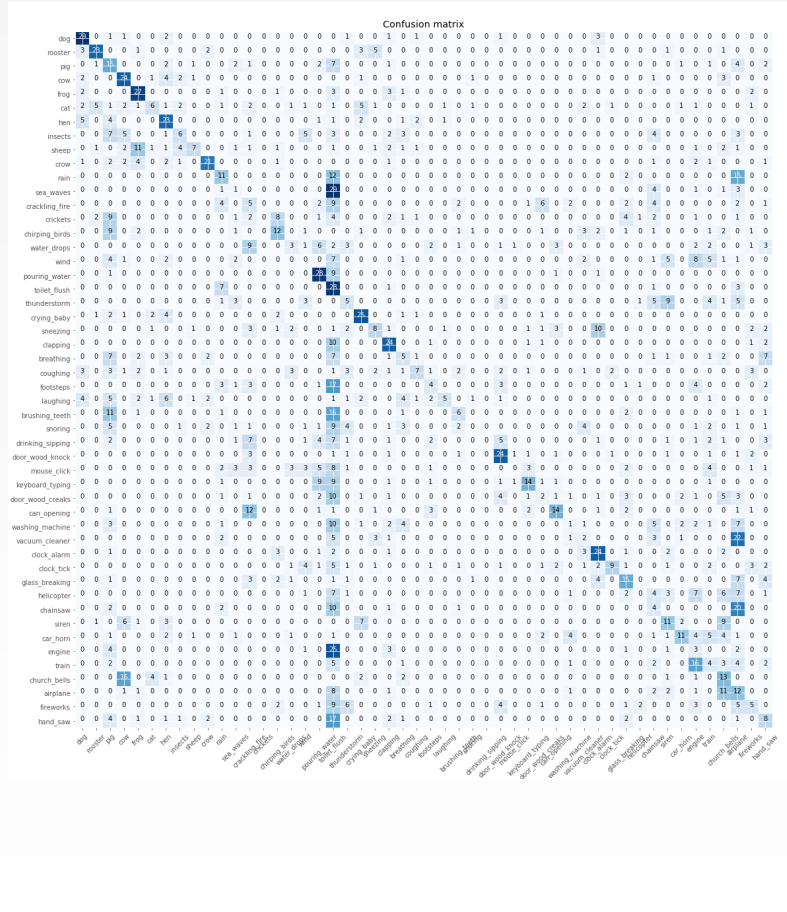
- Autoencoder using NN
- Use dilated audio frames
 - Attempt at temporal feature
- No improvement
- CNN provided no improvement



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

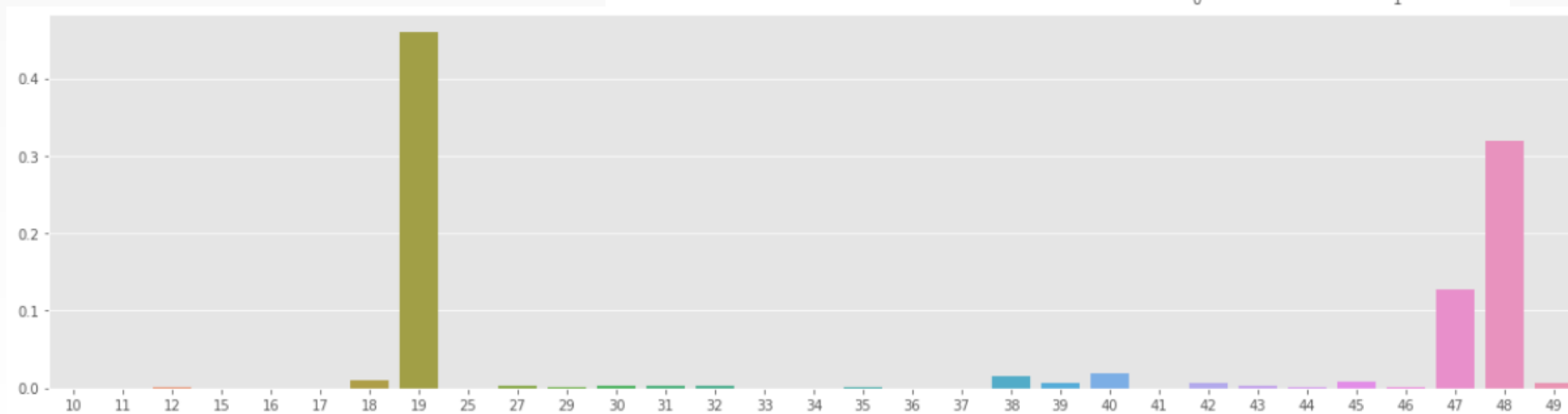
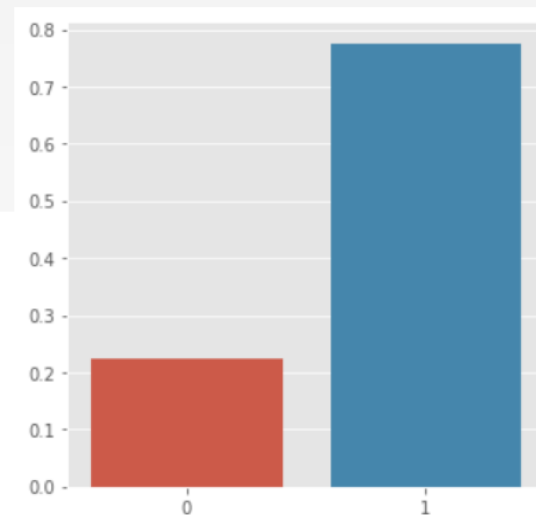
OVERALL PERFORMANCE



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?
