Project CETI

MIT Team Updates

Data

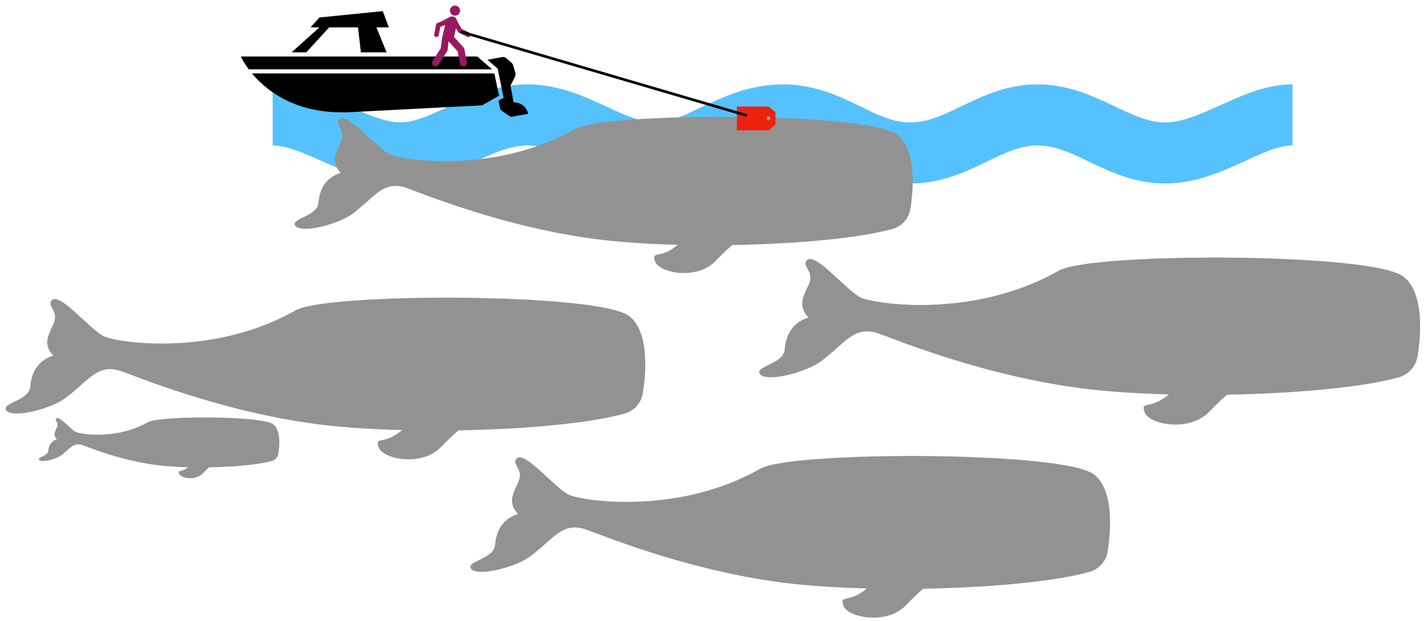
<u>Tagged Data</u>

3950 Codas ~ 22,386 clicks

- Stereo audio recordings

- Rich Annotations

- Gyroscope, Magnetometer, Accelerometer data



Data

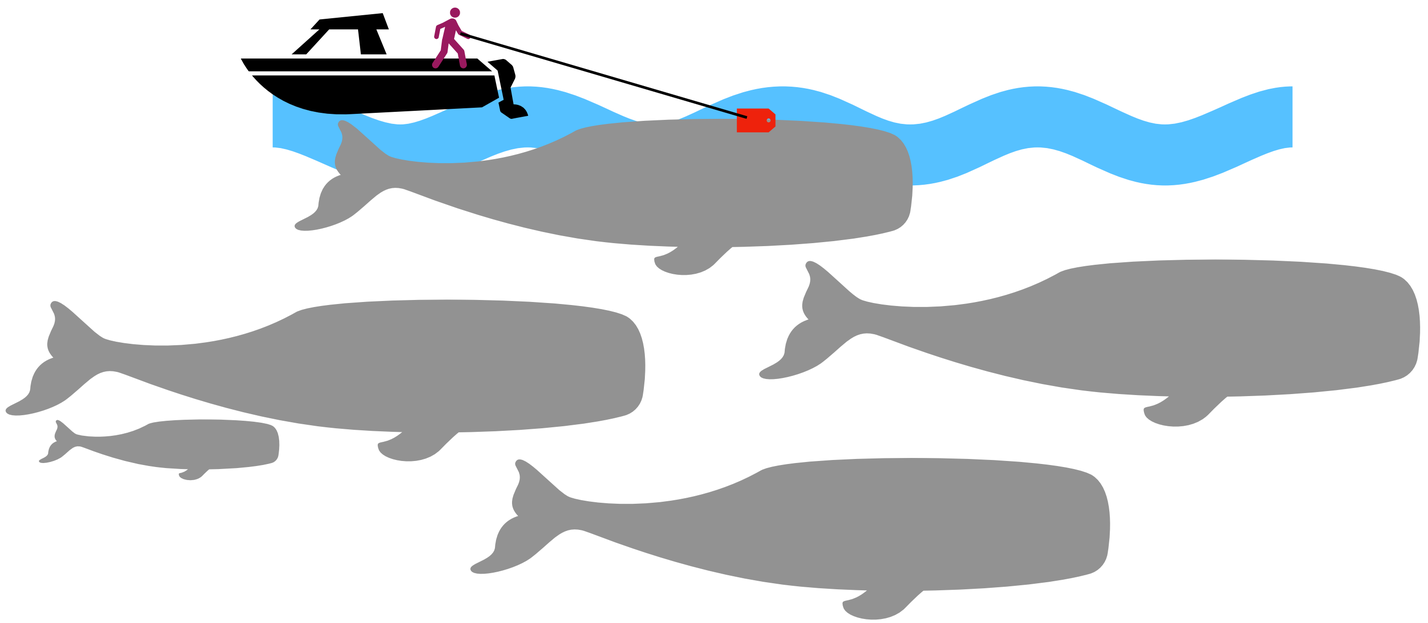
Tagged Data

3950 Codas ~ 22,386 clicks

- Stereo audio recordings

- Rich Annotations

- Gyroscope, Magnetometer, Accelerometer data



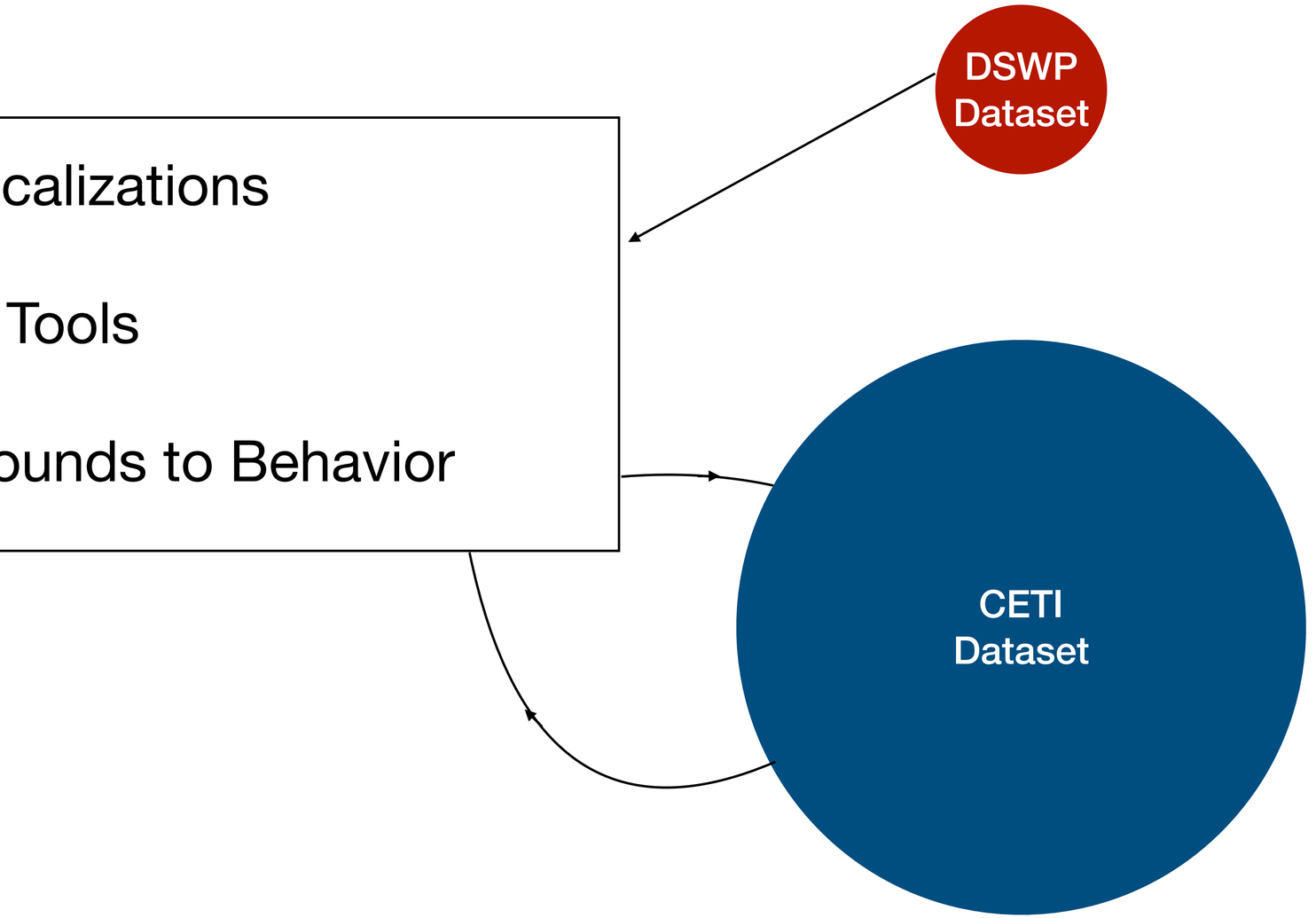
Additional Audio data with no annotations

Outline

1. Finding Structure in the Vocalizations
2. Automatic Annotation and Tools
3. Grounding : Connecting Sounds to Behavior

Outline

1. Finding Structure in the Vocalizations
2. Automatic Annotation and Tools
3. Grounding : Connecting Sounds to Behavior

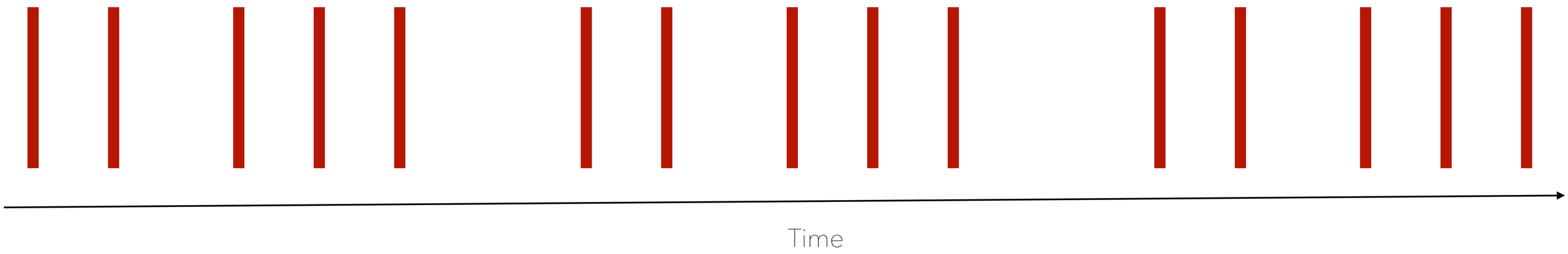


Outline

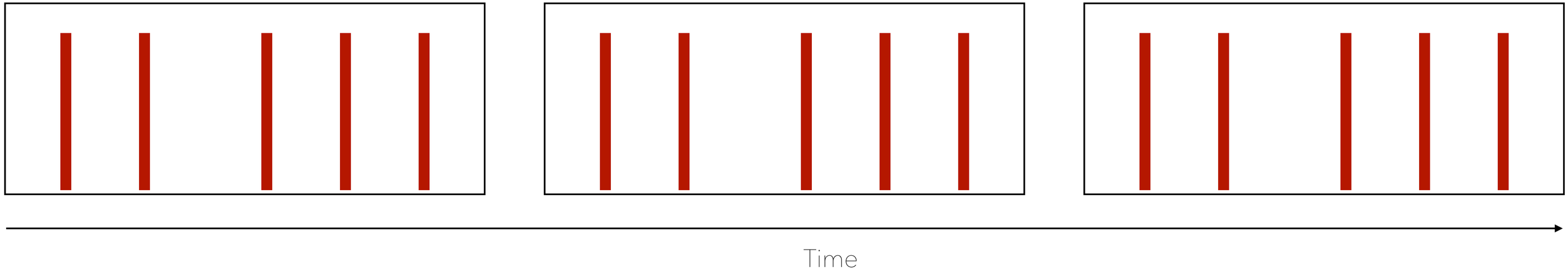
1. Finding Structure in the Vocalizations
2. Automatic Annotation and Tools
3. Grounding : Connecting Sounds to Behavior

Finding Structure in the Sounds

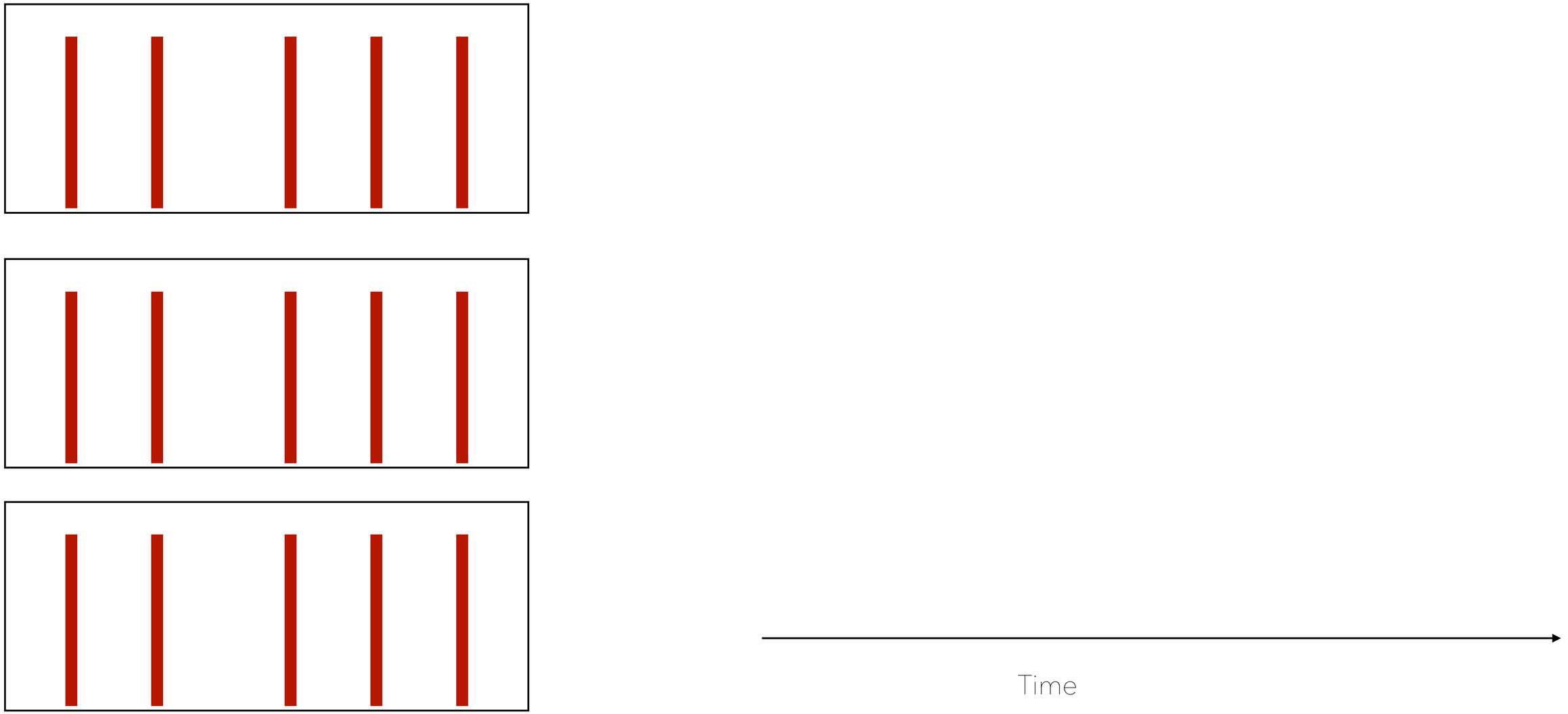
Understanding codas



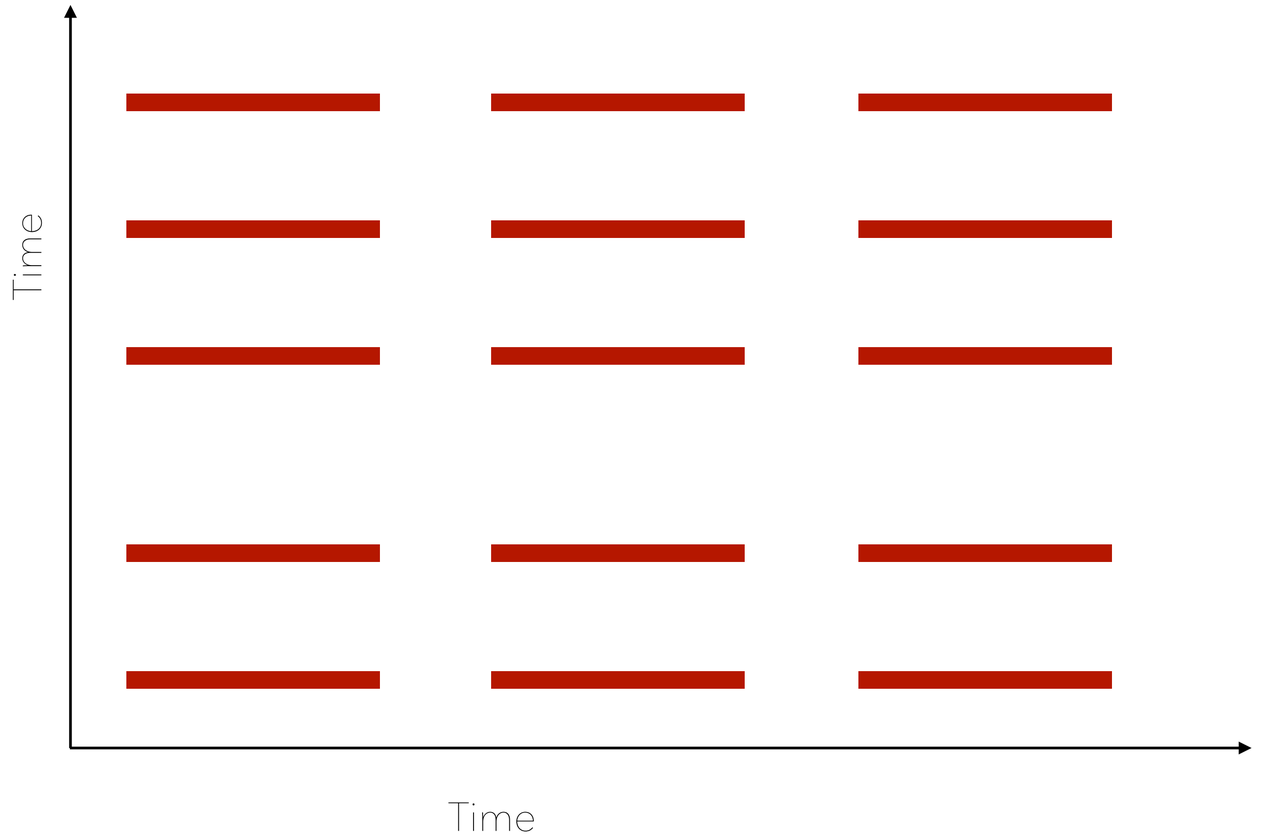
Understanding codas



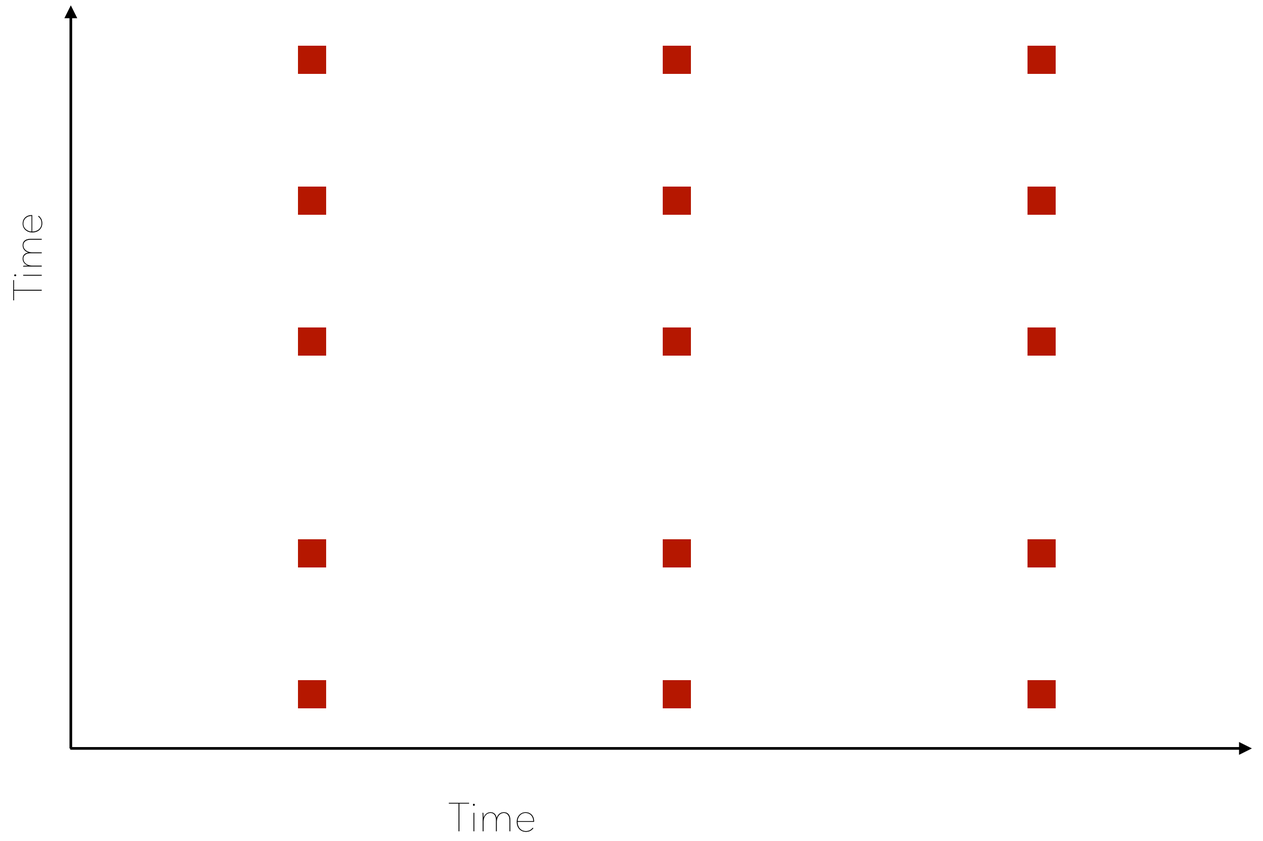
Understanding codas



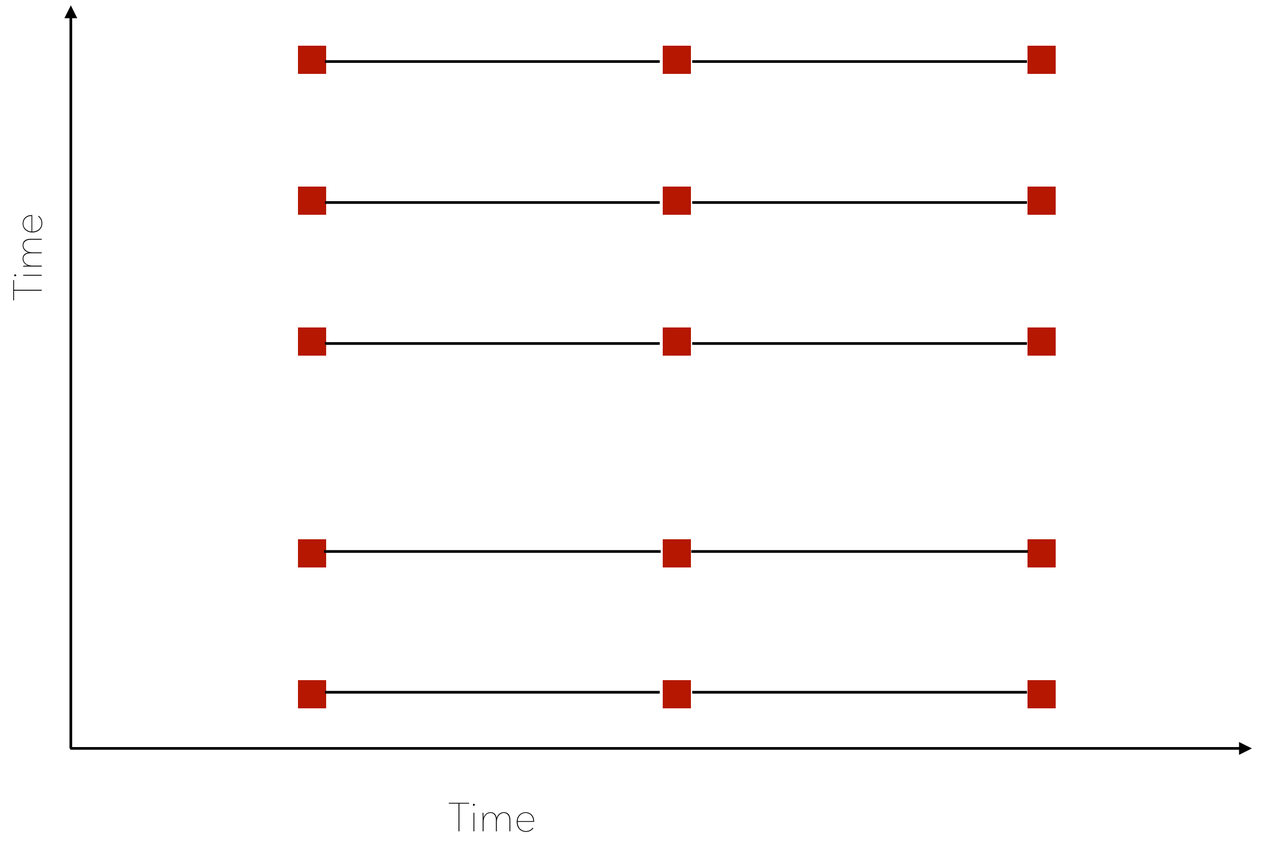
Understanding codas

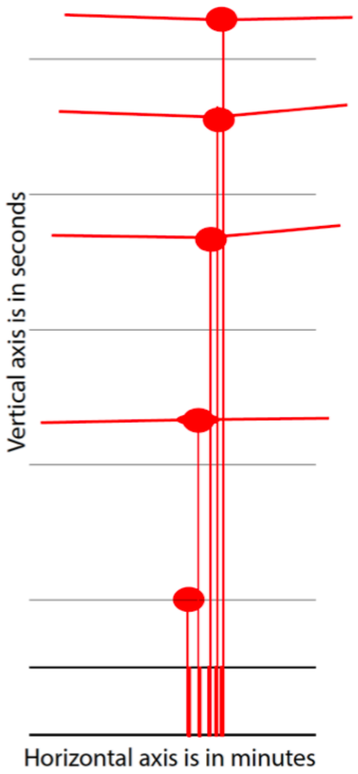


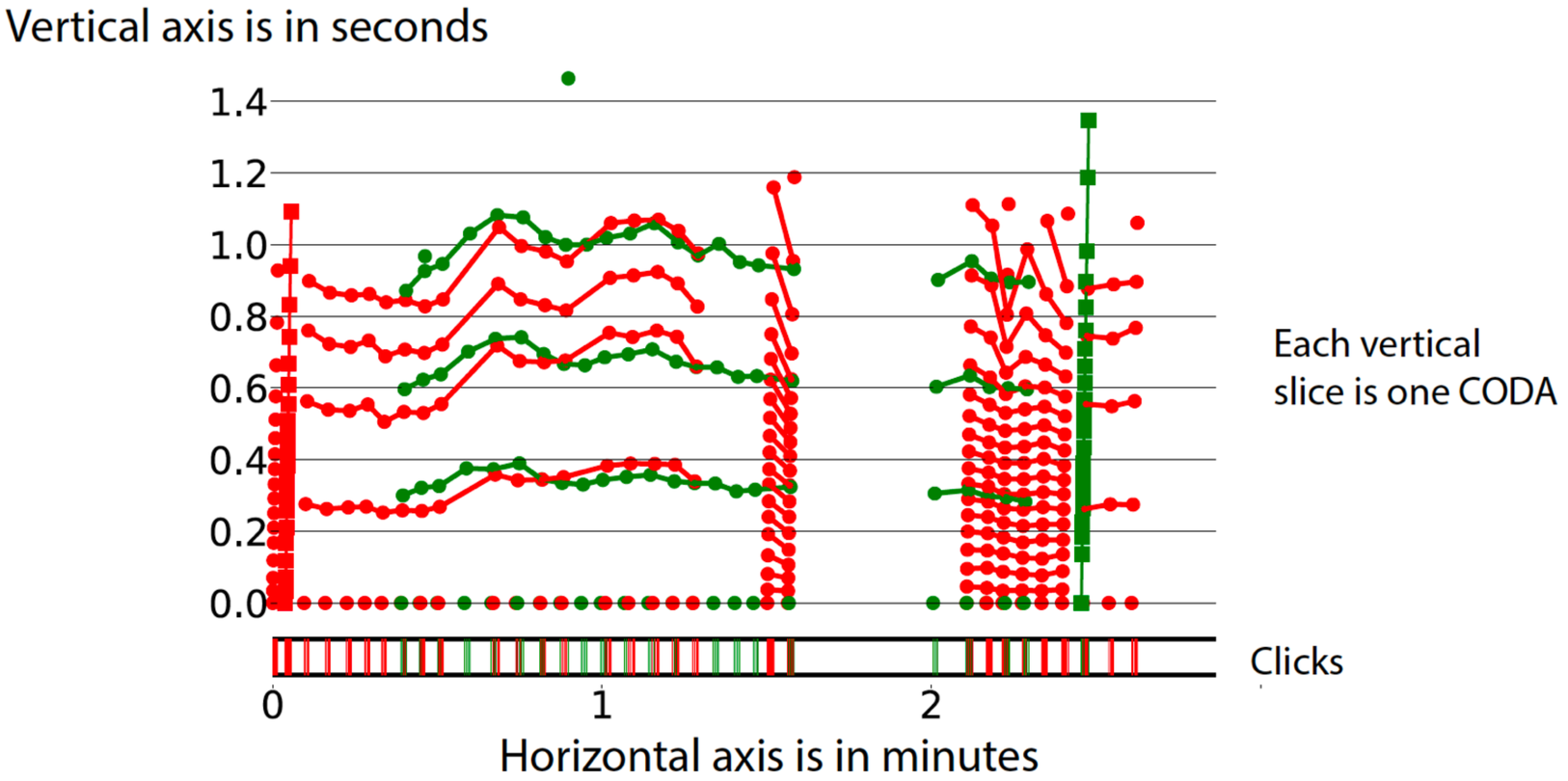
Understanding codas



Understanding codas

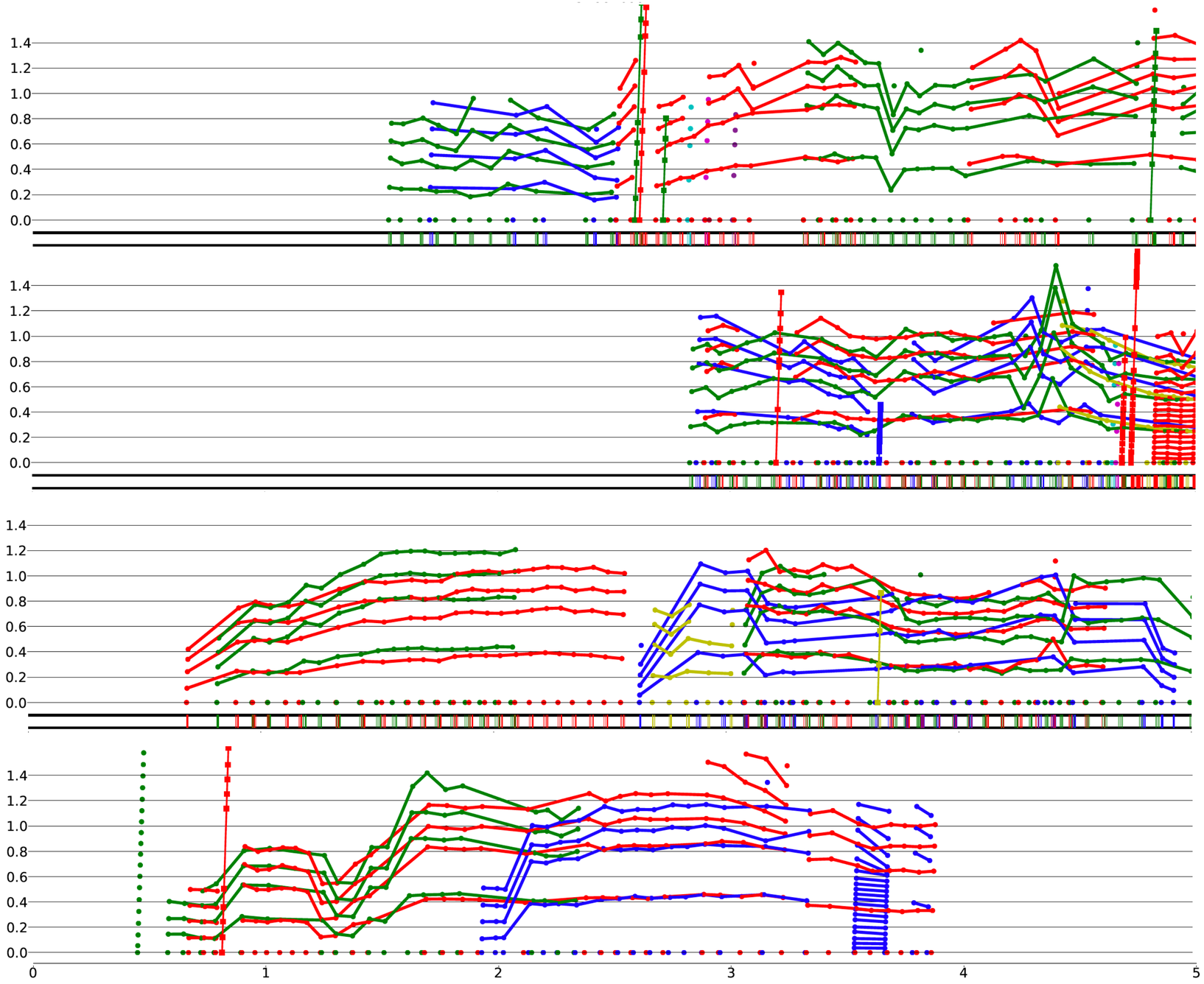


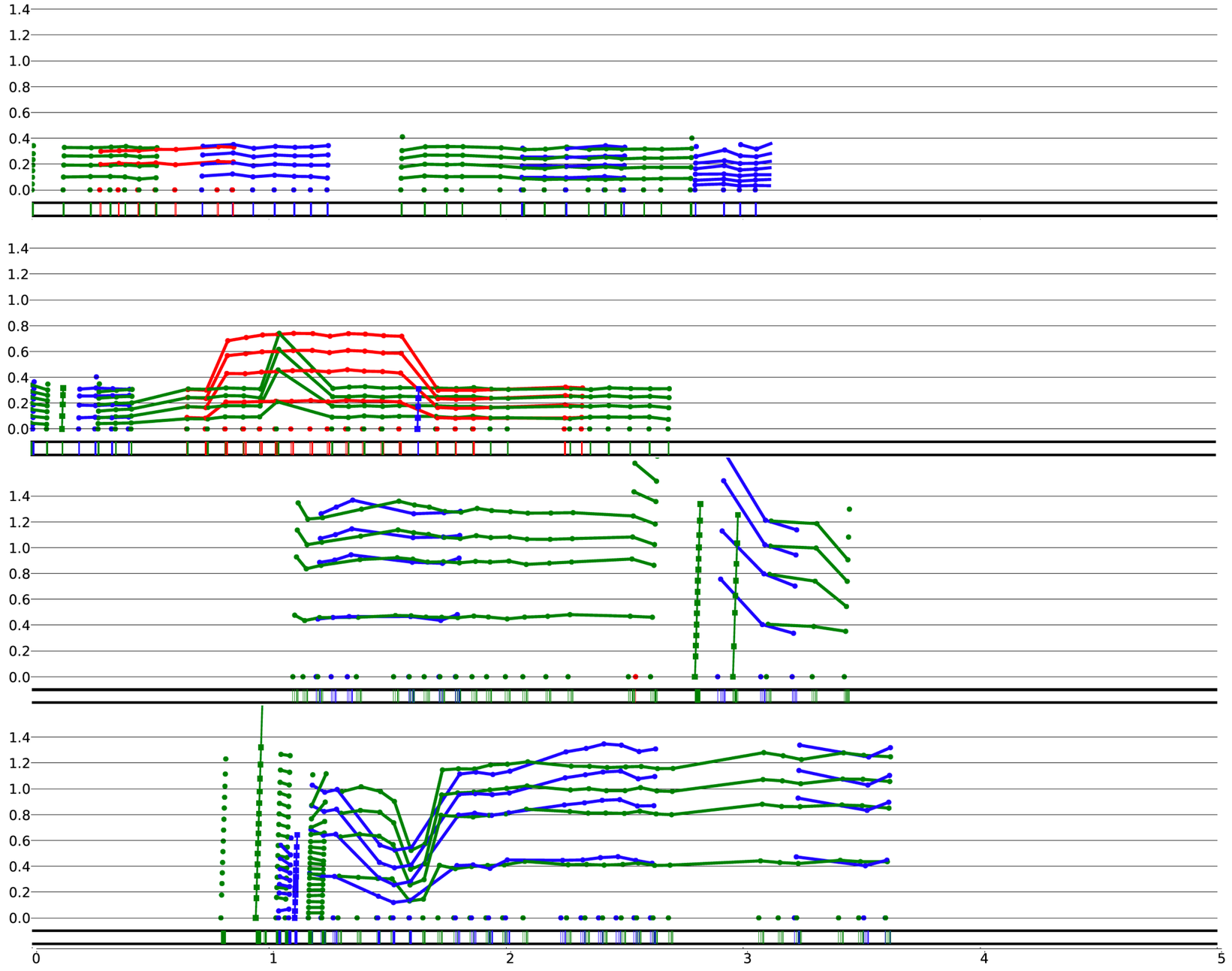




!! Let's say the little amount of variation in the lengths was noise. Then why does the red whale's pattern follow the green whale's pattern?

The book of whales

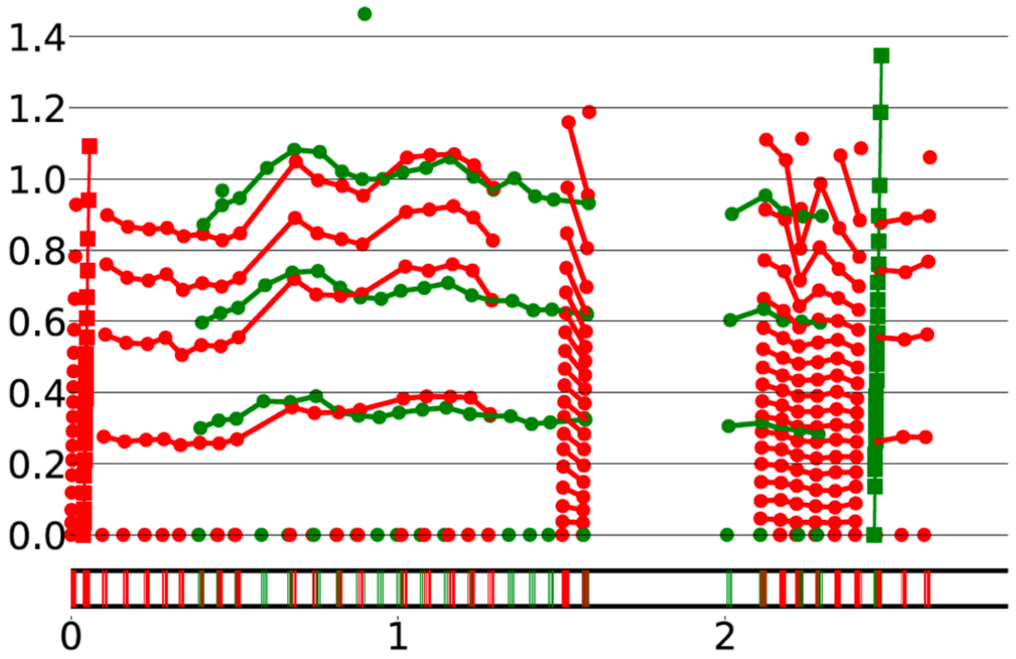

1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5

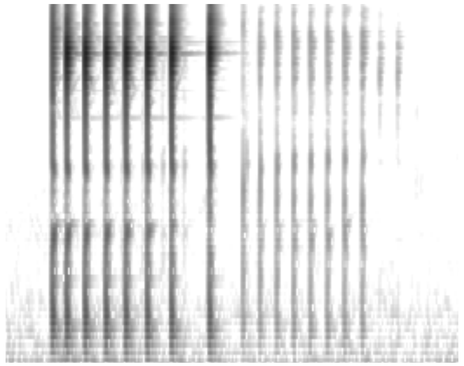
What does it tell us

Predicting the vocalization is a non-trivial problem

1. The coda
2. The speaker
3. The duration
4. The dialogue dynamics!



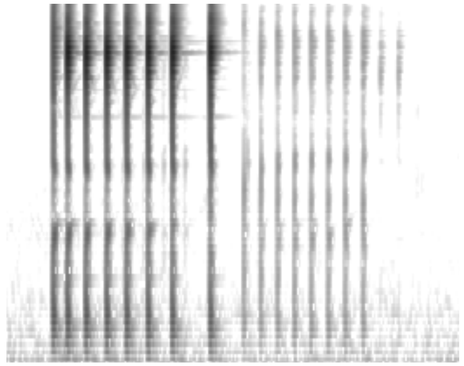
How should we represent the vocalizations?



Attributes:
Which whale?
What coda?
At what depth?
How loud?
At what time?

Representation

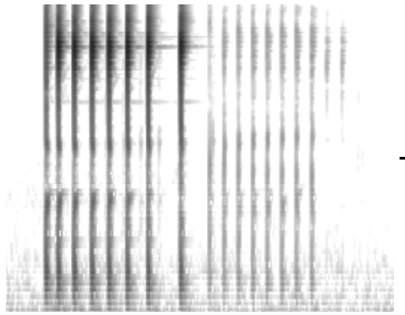
How should we represent the vocalizations?



Attributes: [which whale?,what coda? At what depth?, How loud?, Which direction was it facing?, at what time?]

Continuous like Music?

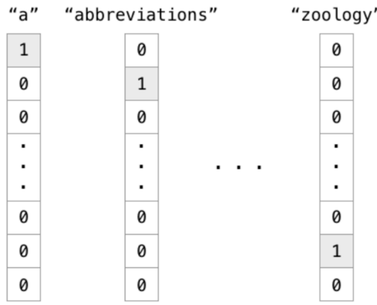
Option1

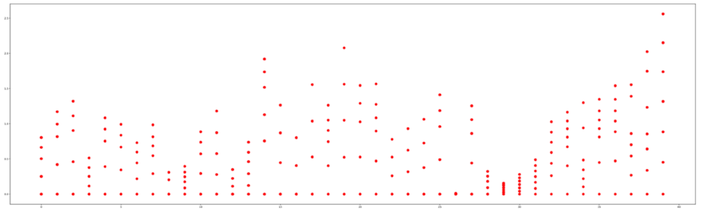


Coda: [8D] :
[ICI1,ICI2,ICI3…,IC7,1/0]

Discrete like Language?

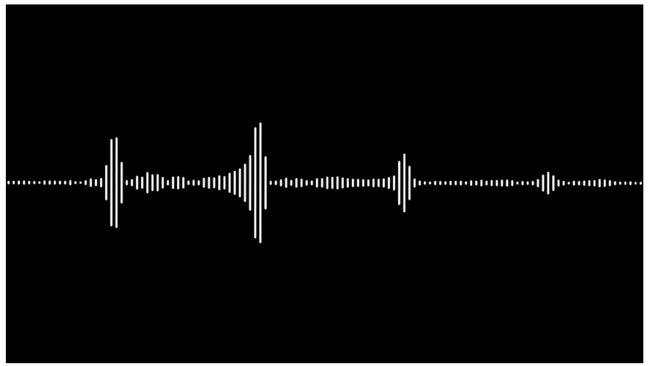
Option2





Cluster: 1 , 2 , 3, 4 …,*n*

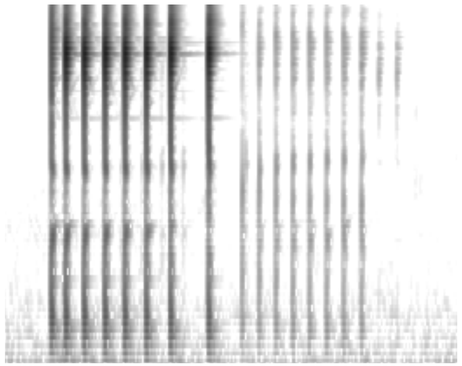
Representation



00010001000010000 1000

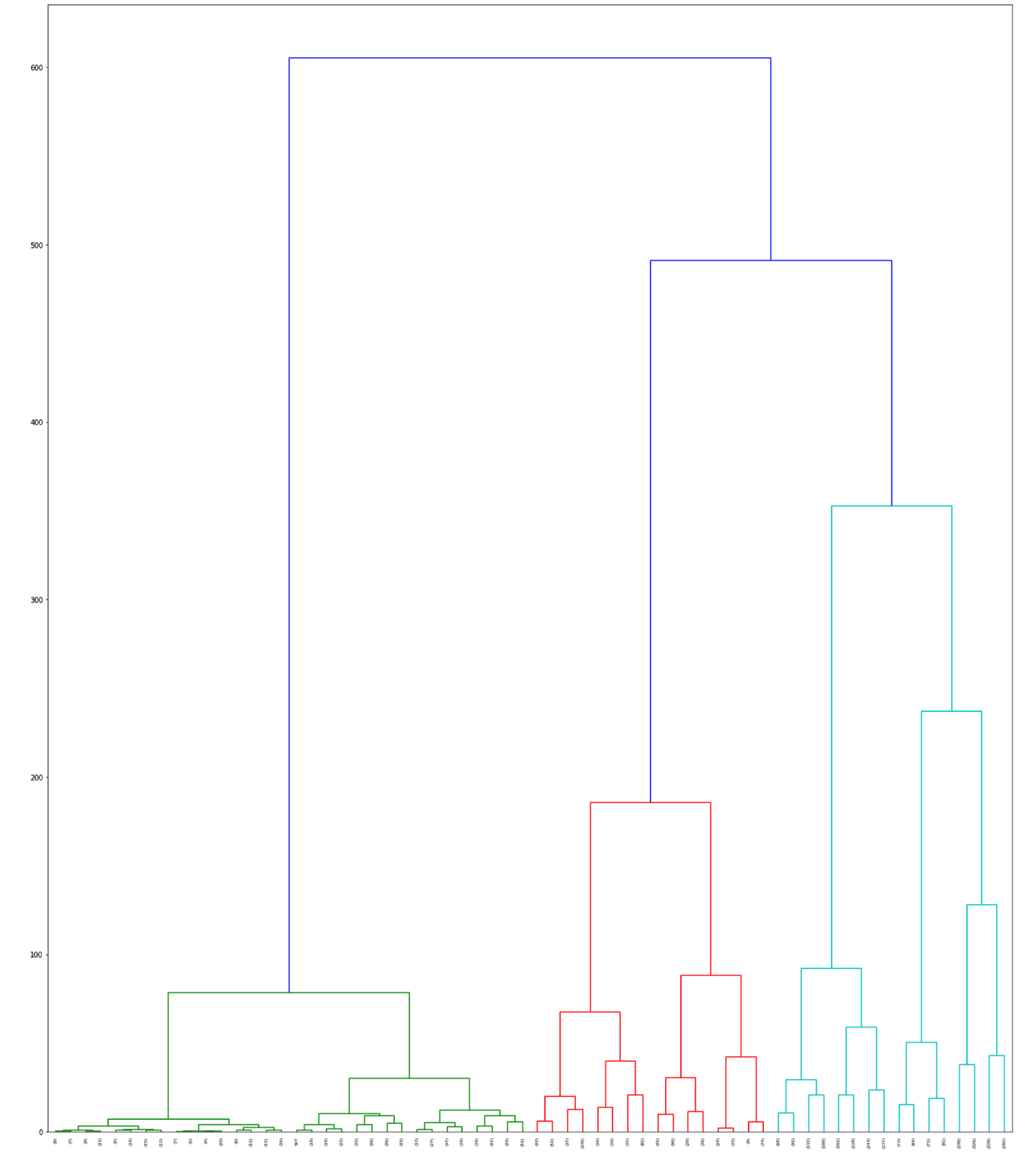
Lower dimensionality better!

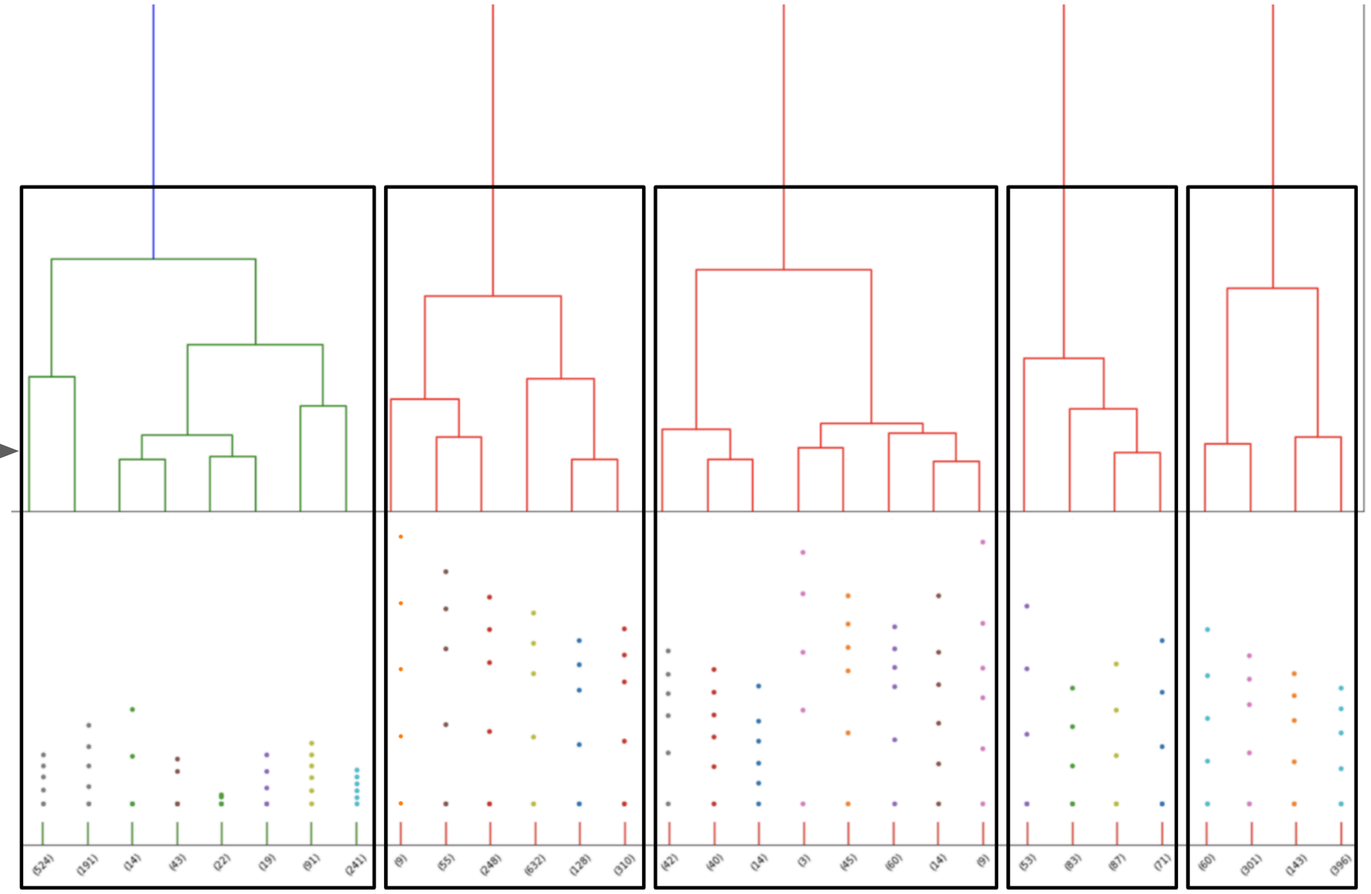
More information better!



Attributes: [which whale?,what coda? At what depth?, How loud?, Which direction was it facing?, at what time?]

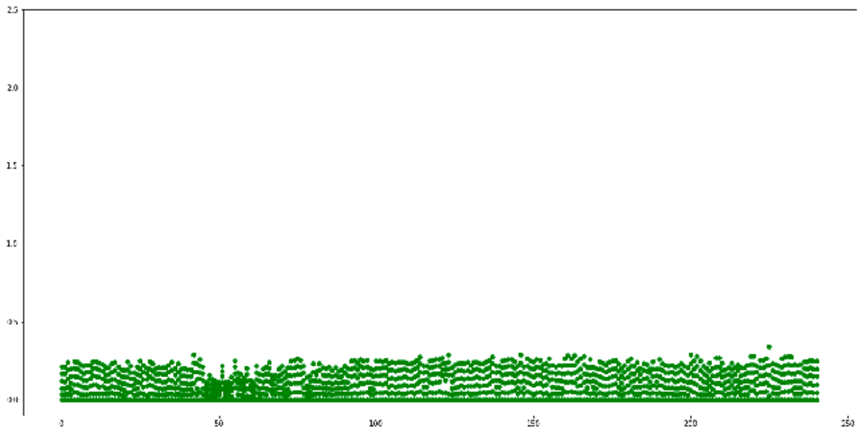
Clustering

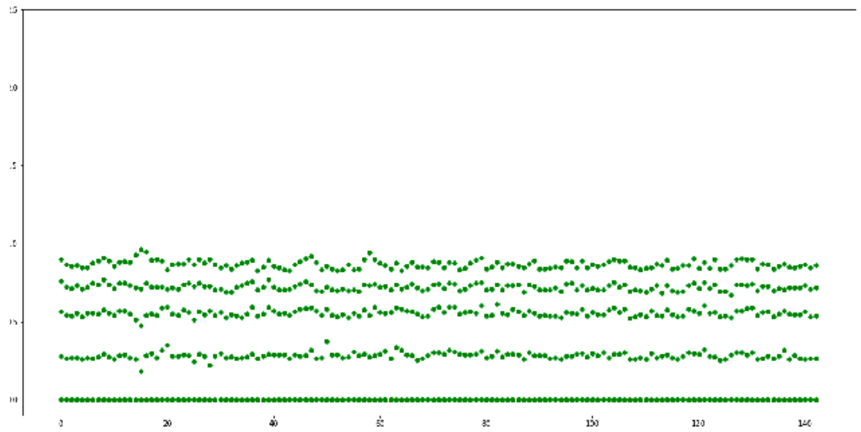


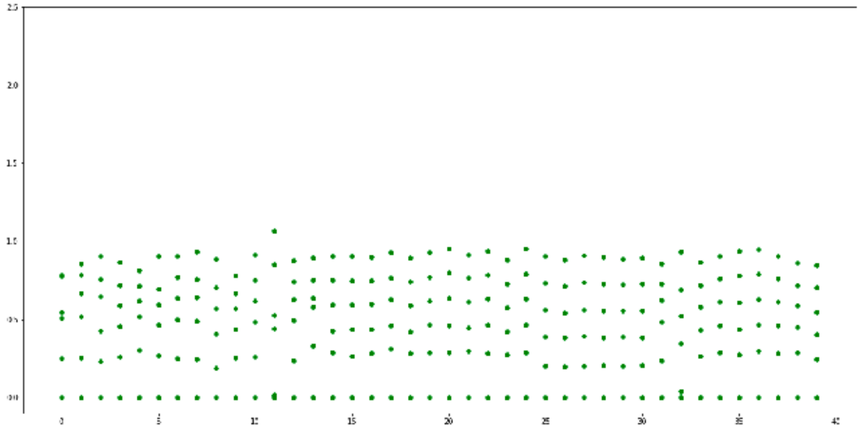


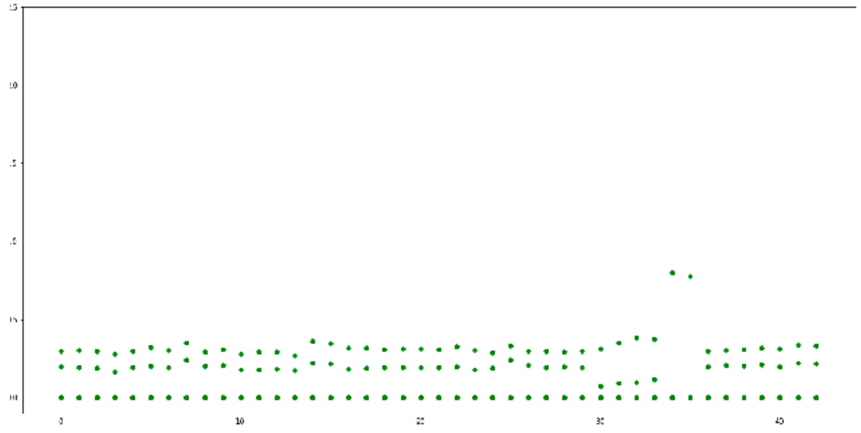
Hierarchical Agglomerative Clusters of CODAs

Variability within clusters



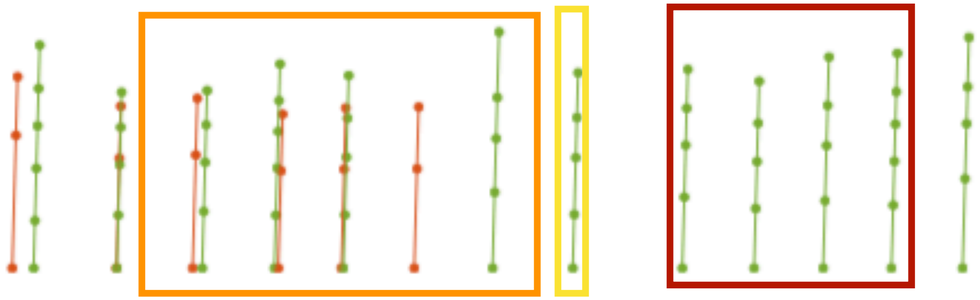


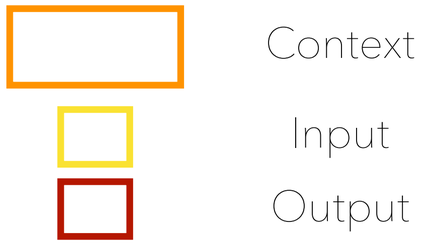




Modeling Whale Conversations

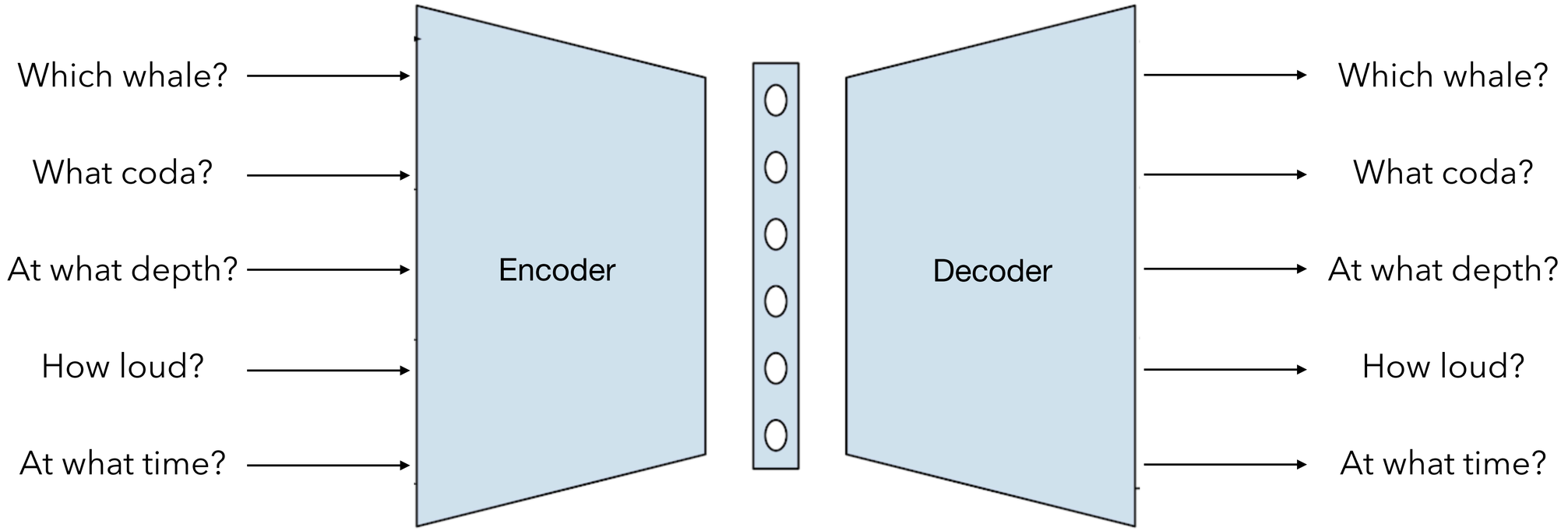
Can we build a model that can predict the vocalizations?





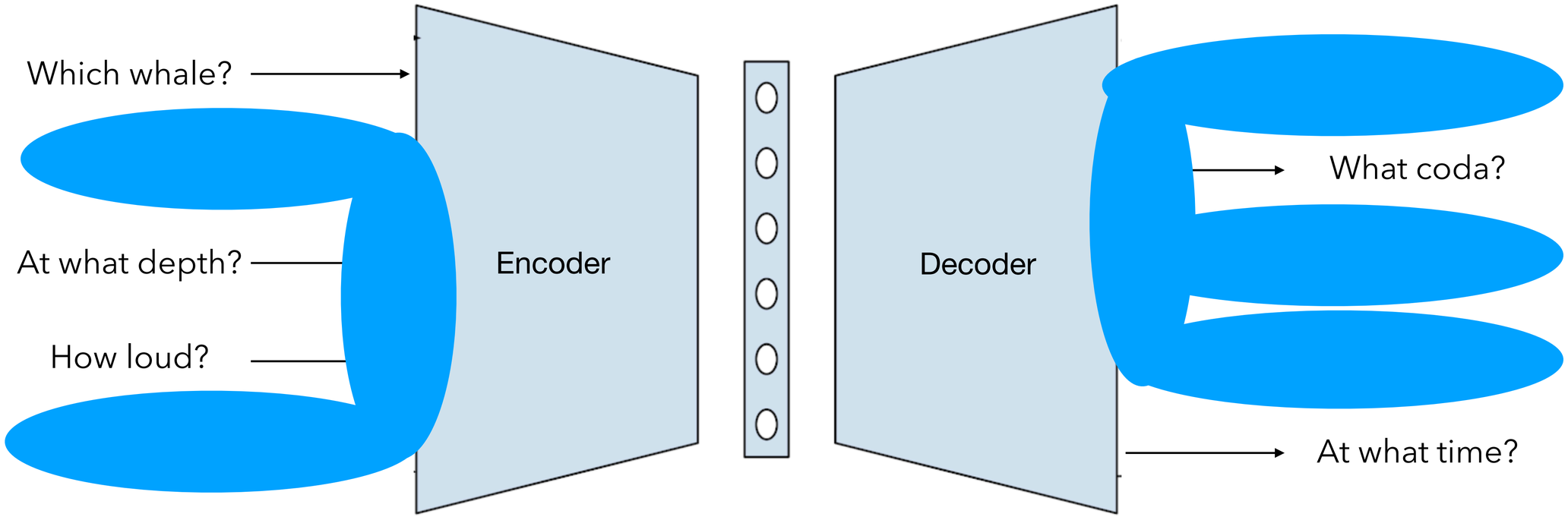
Time = *t, t-1, t-2, t-3, t-4*

Time = *t+1, t+2, t+3, t+4*

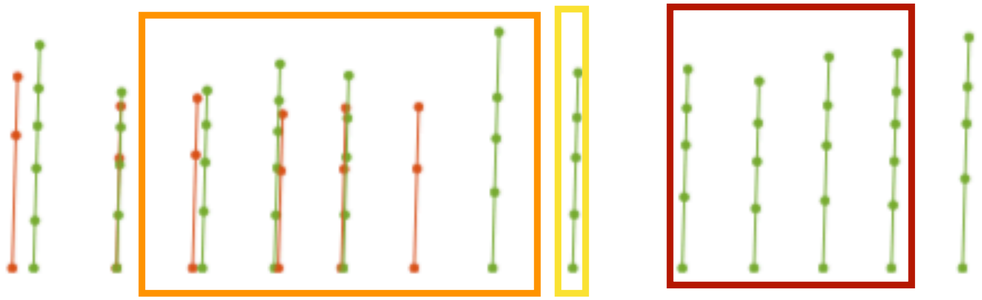


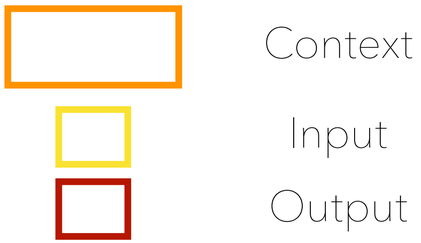
Time = *t, t-1, t-2, t-3, t-4*

Time = *t*



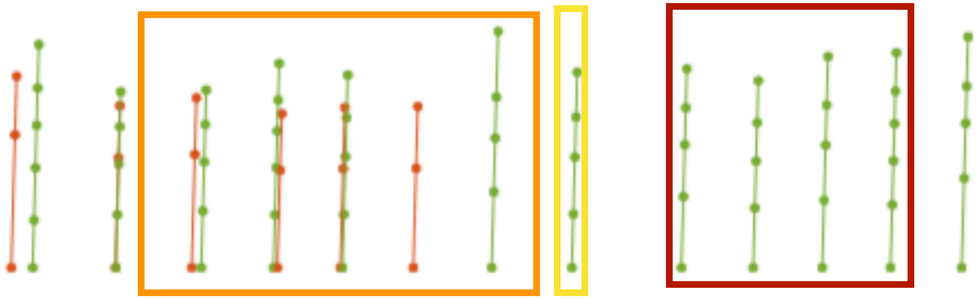
Can we build a model that can predict the vocalizations?

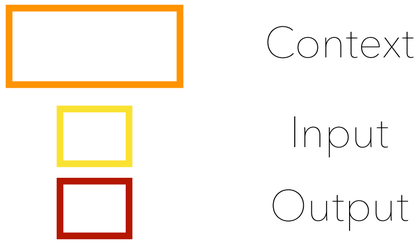


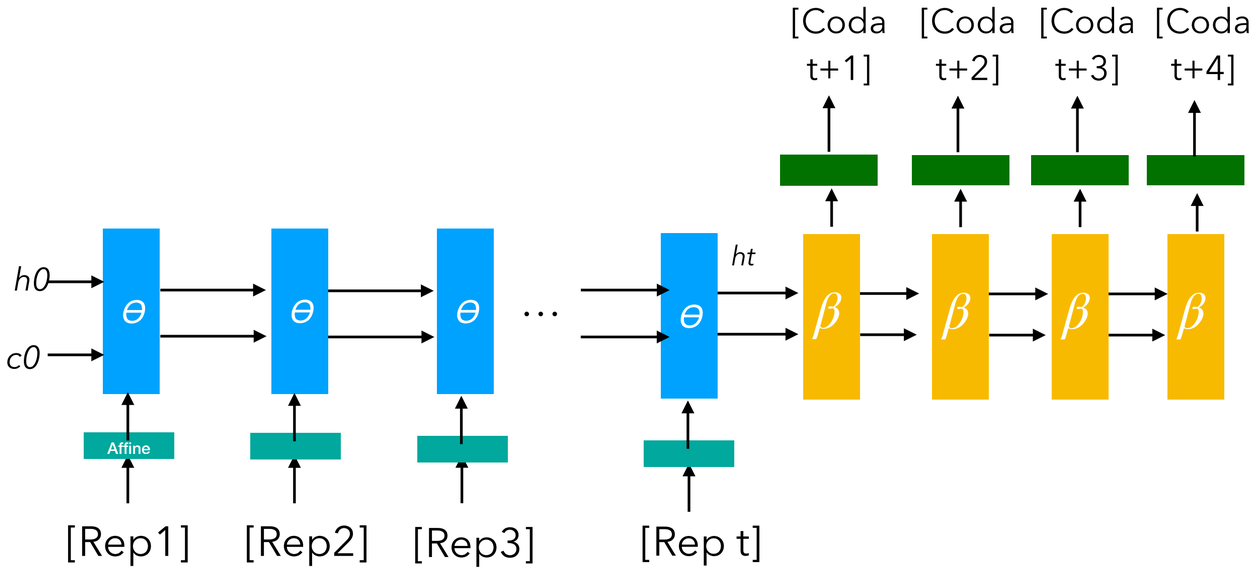


Attributes:
Which whale?
What coda?
At what depth?
How loud?
At what time?

Can we build a model that can predict the vocalizations?







Coda:[Rep]: [Whale ID, start time,ICI1,…,IC7,1/0, Power1,… Power7,Depth]

Predict: The Next Coda

Evaluation

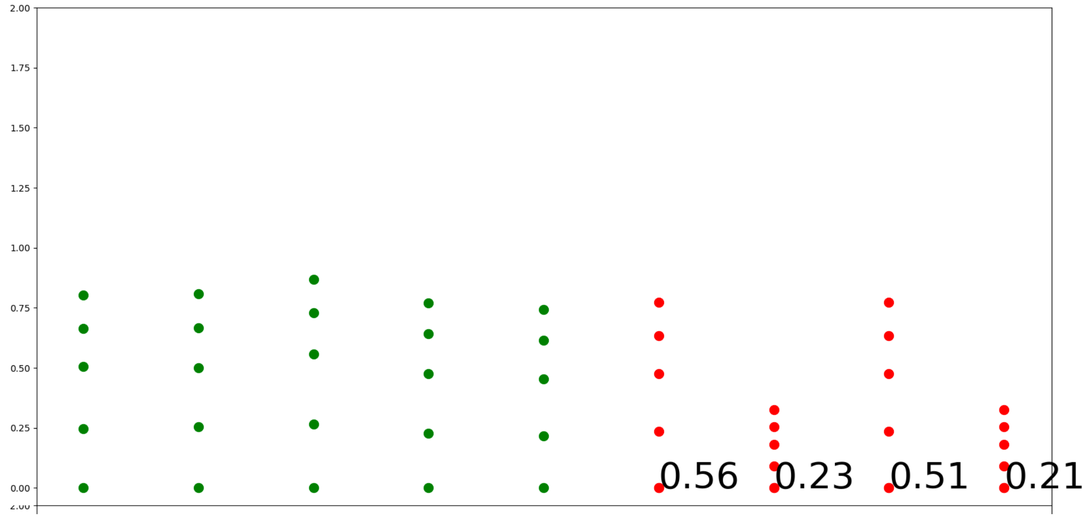
$$PP(X) = 2^{-\frac{1}{n}\log P(x_1,x_2,\ldots,x_n;\theta)}$$

Perplexity: Inverse probability of the test set normalized by the number of words

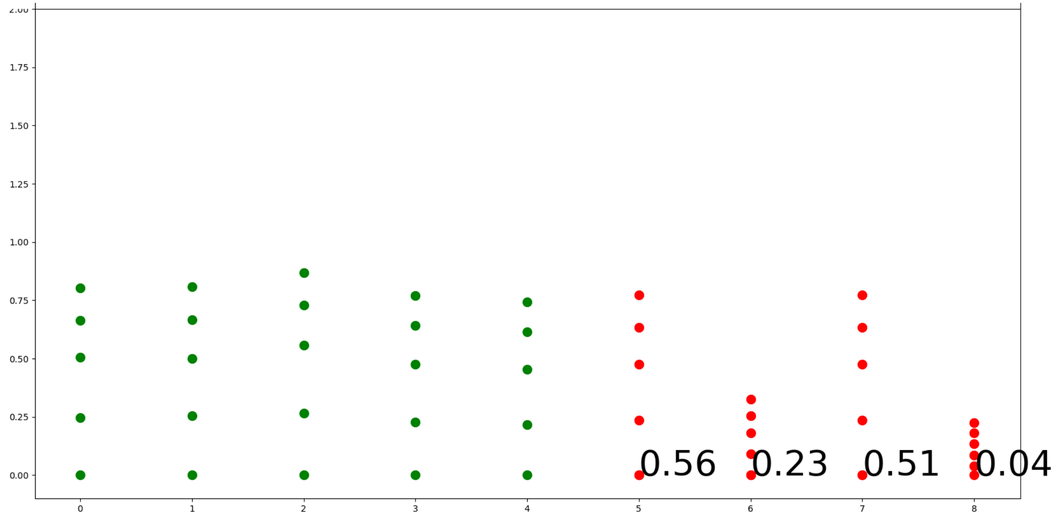
Minimizing perplexity => Maximizing probability

Predictions by the model

Ground Truth

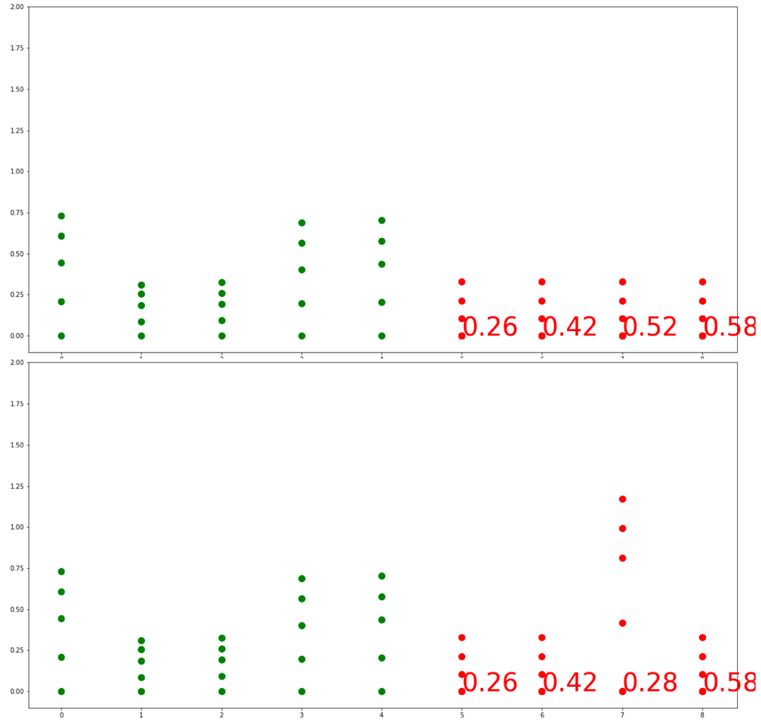


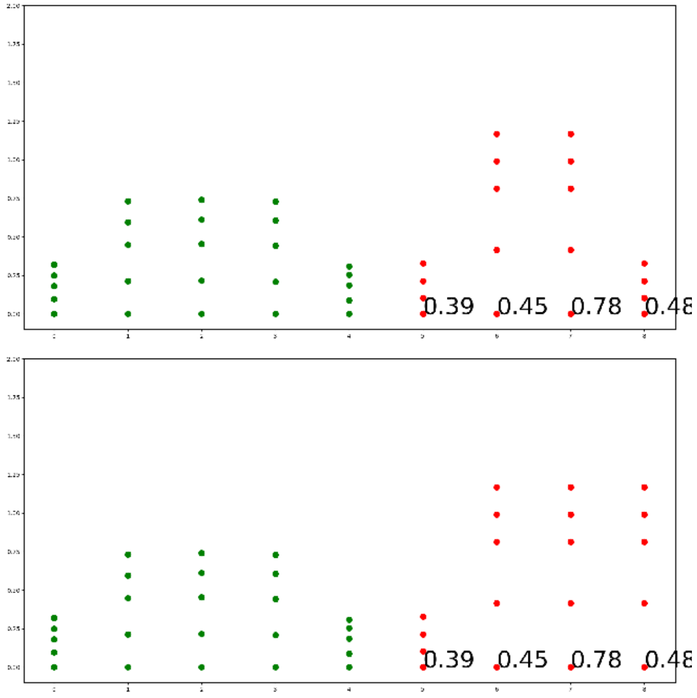
Prediction

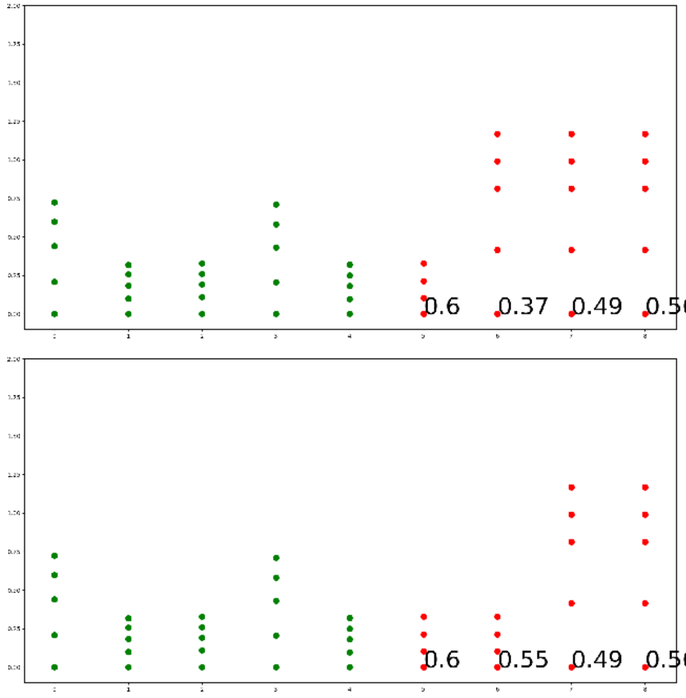


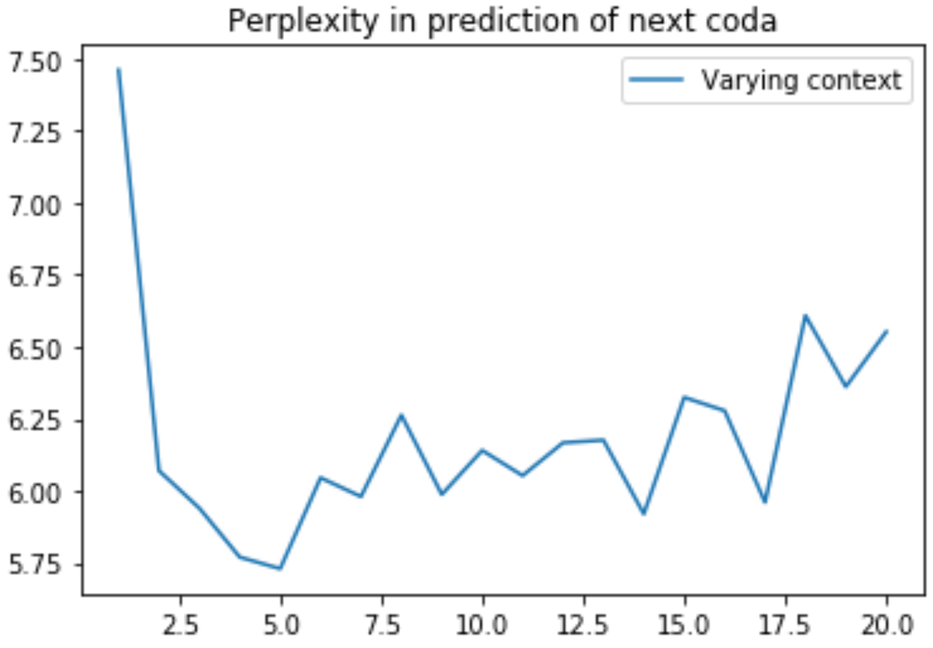
Ground Truth

Prediction

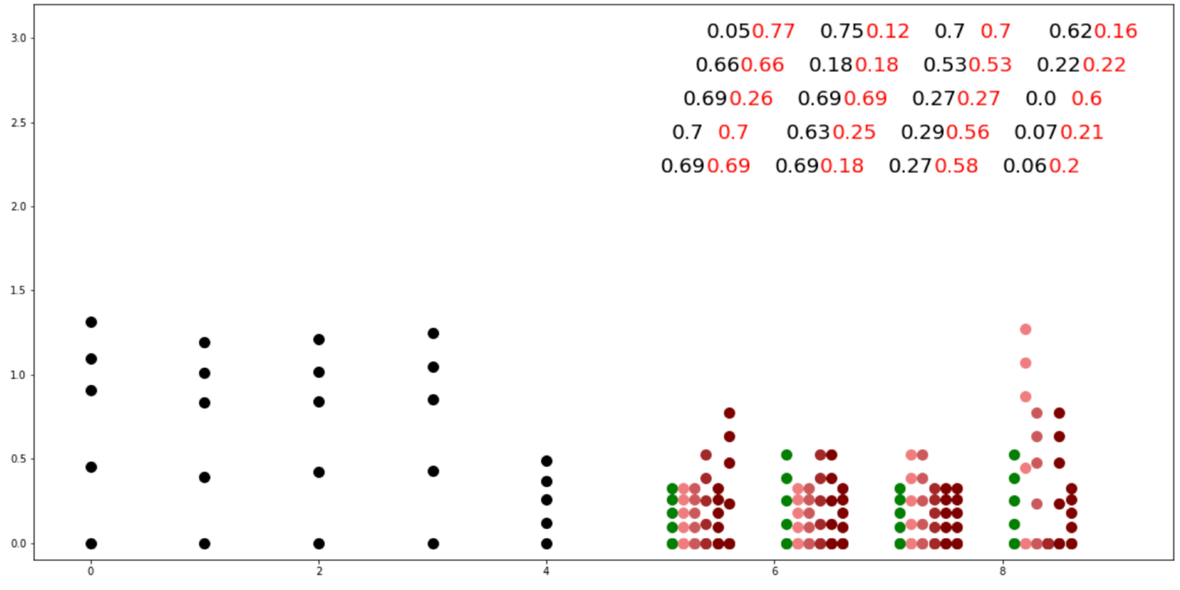


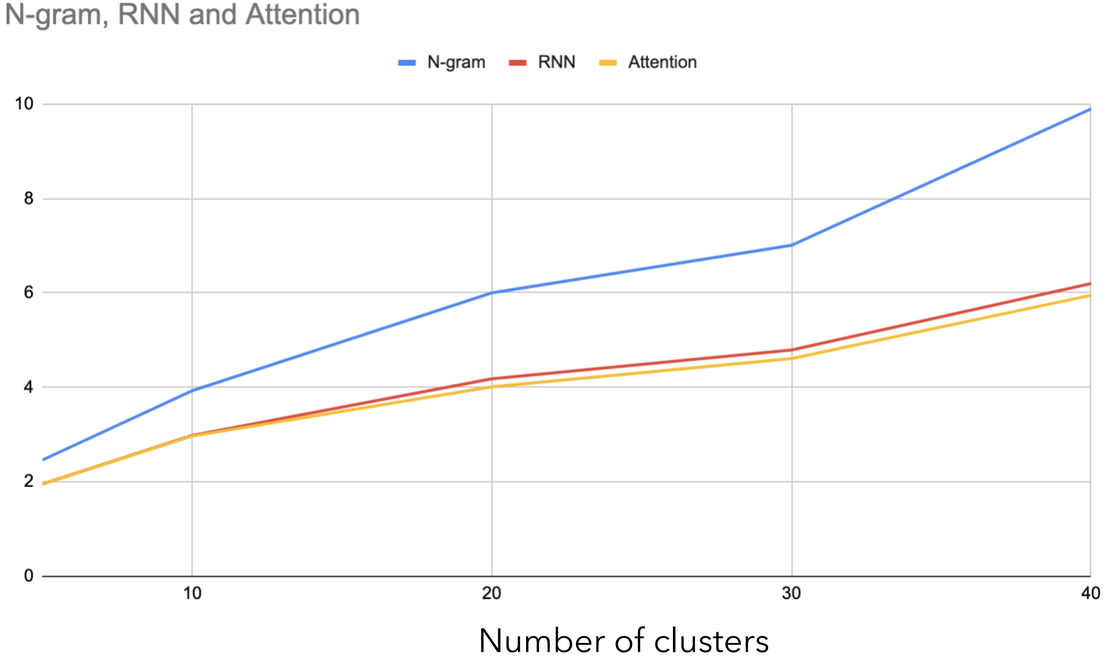




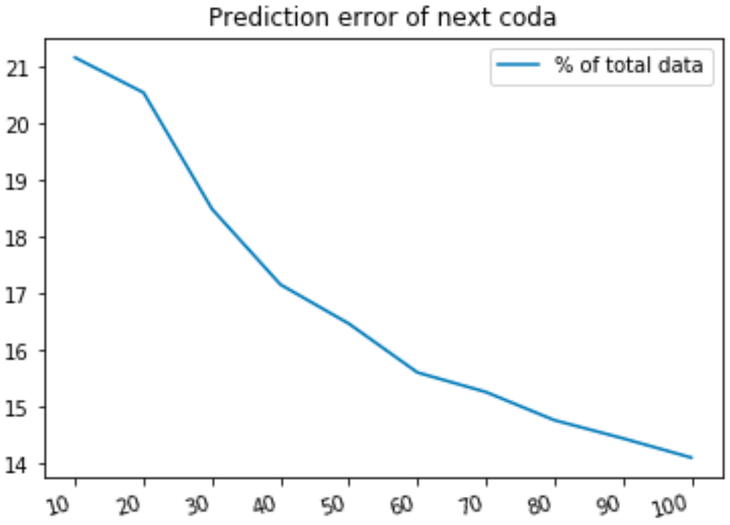


Affect of context size



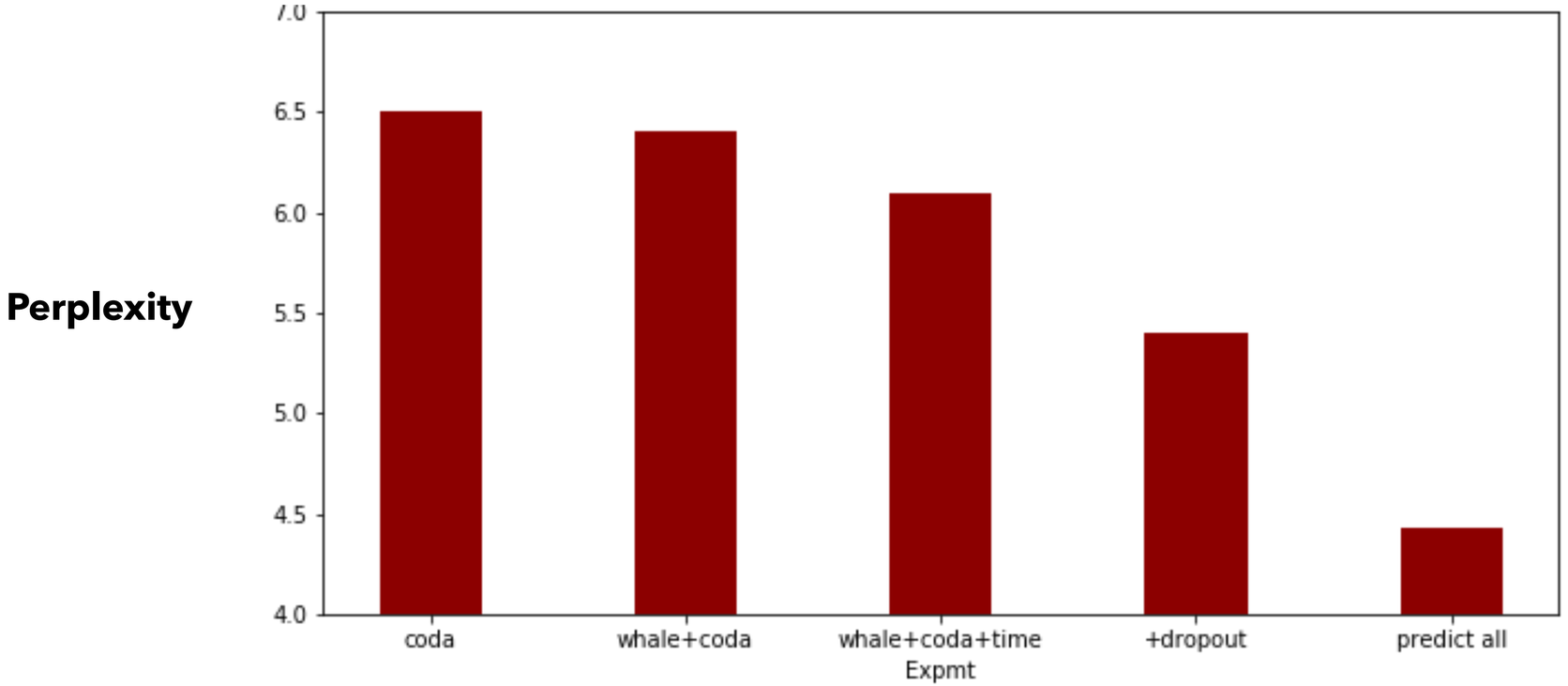


Affect of model complexity



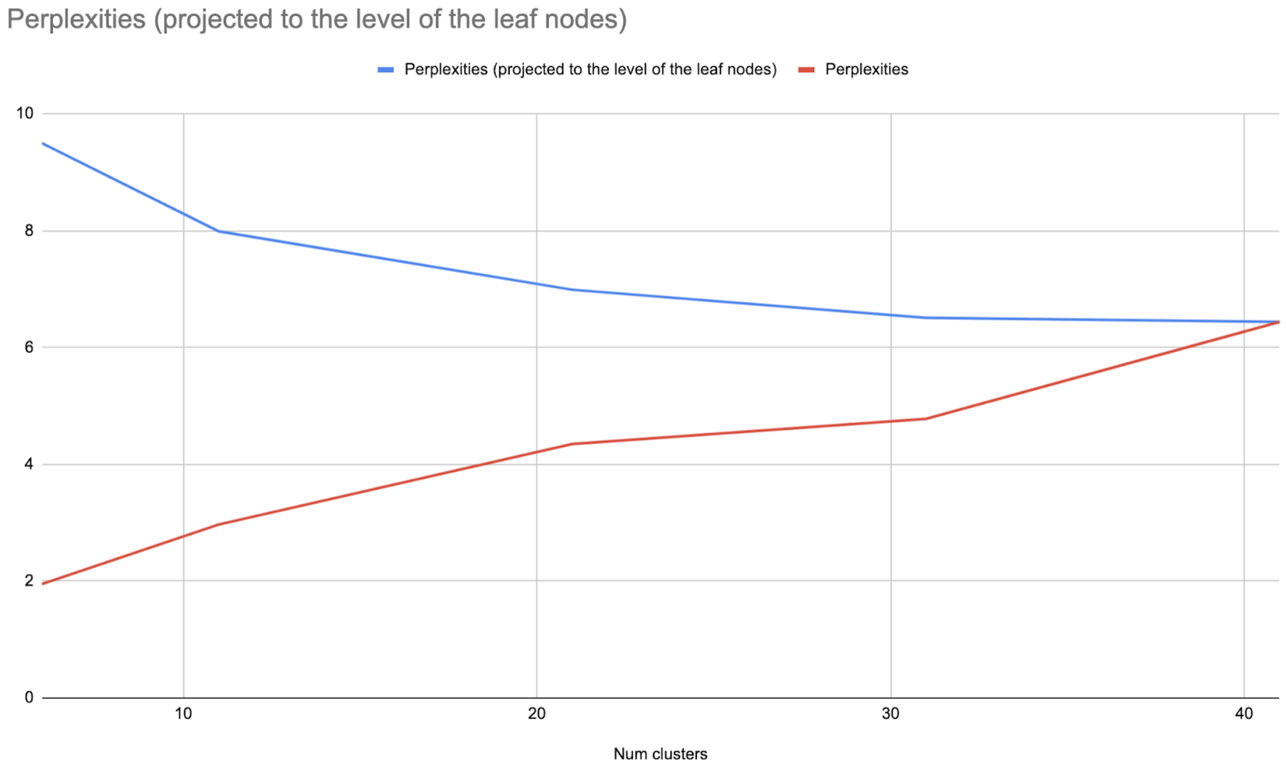
Affect of size of the training data

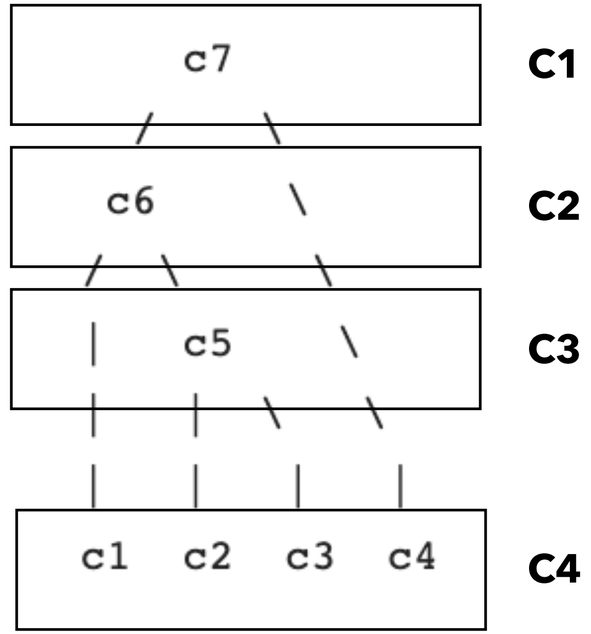
PP of predicting the next coda



What is the smallest # discrete units that may explain the data distribution the best?

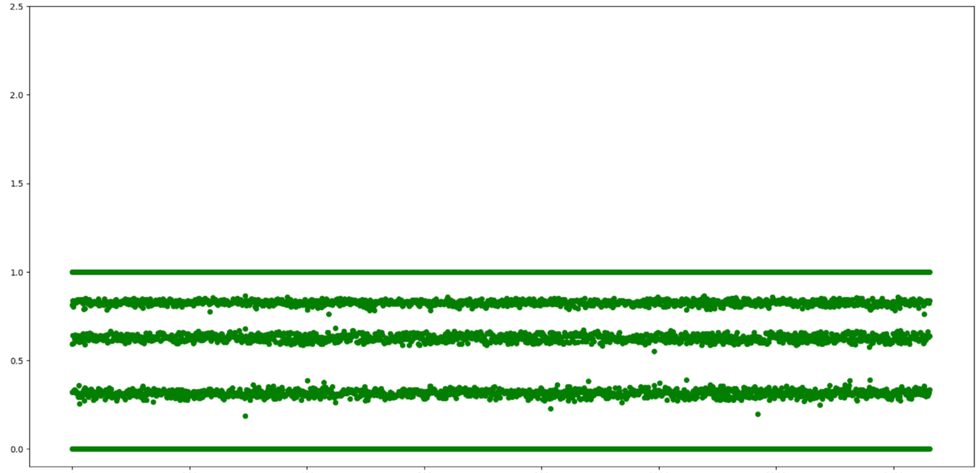
How do we decide how many clusters we should have?

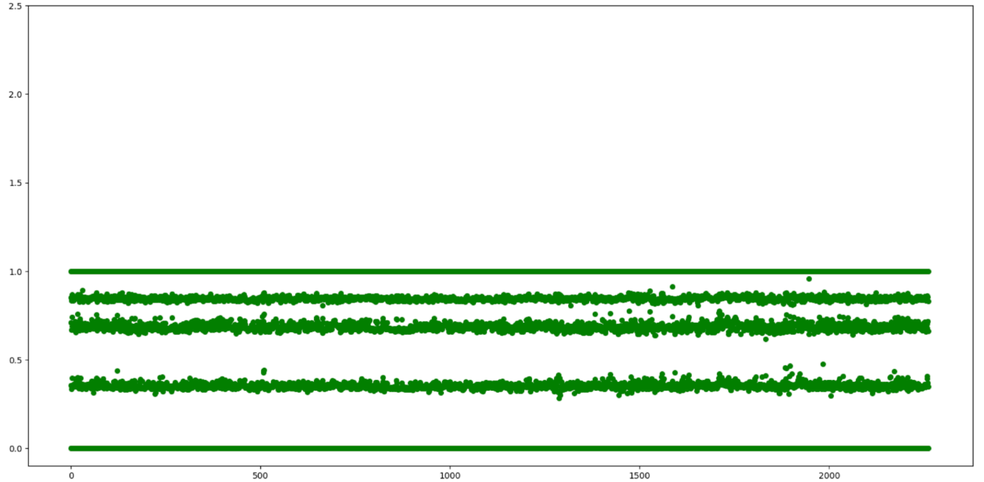


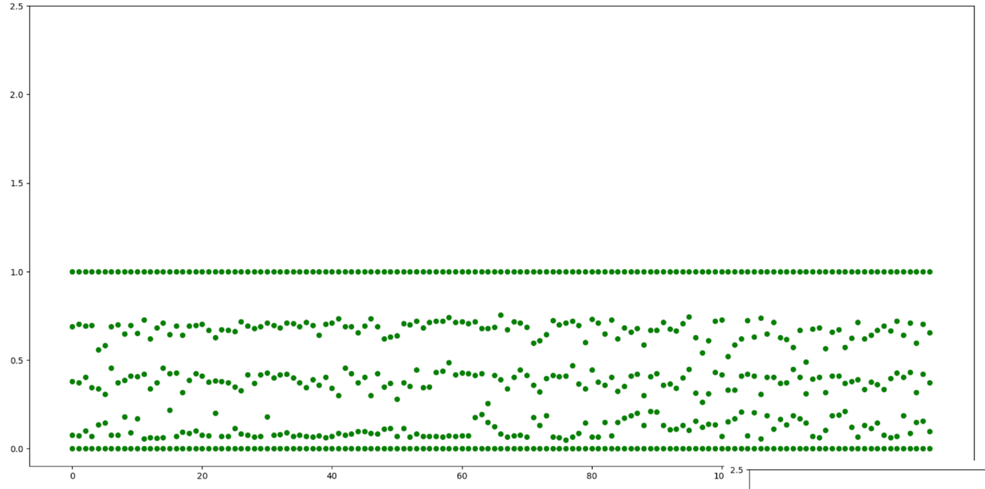


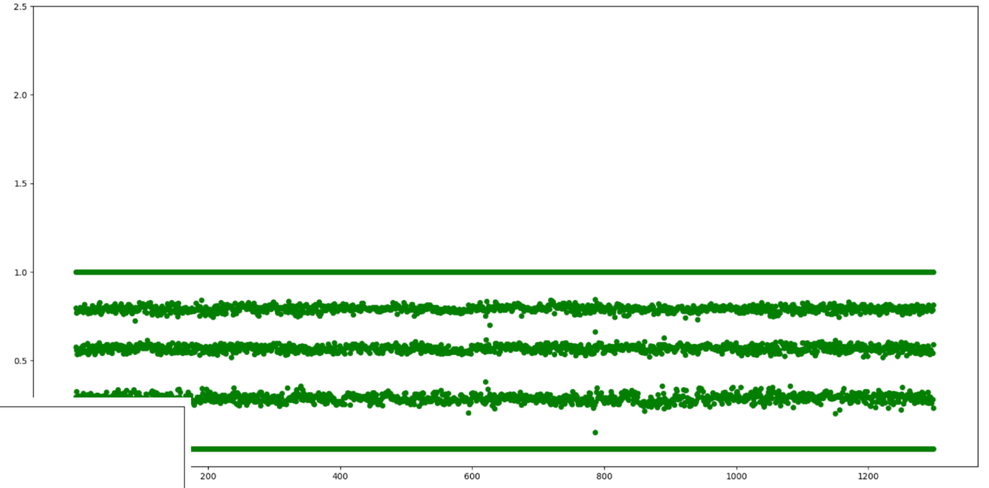
$$p(c3 \mid context) = p(c3 \mid C2) x p_m(c6 \mid context, \theta)$$

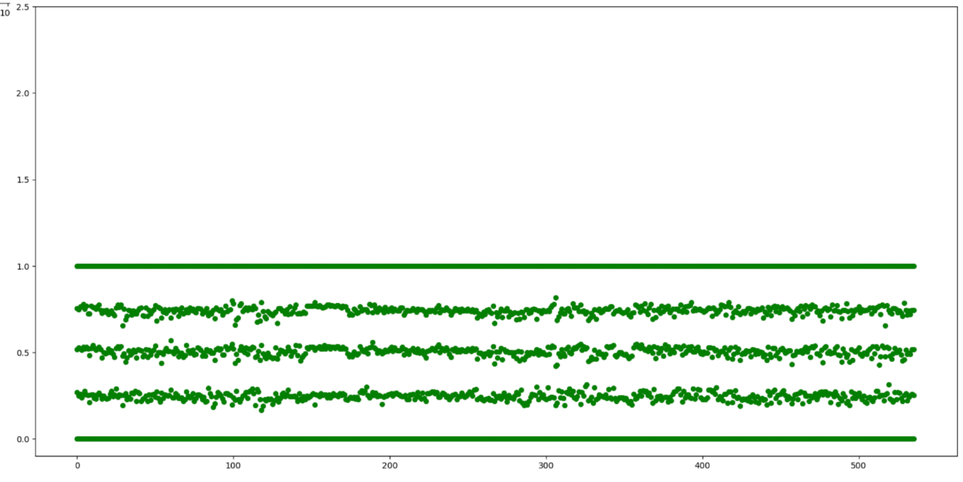
Rhythm and tempo

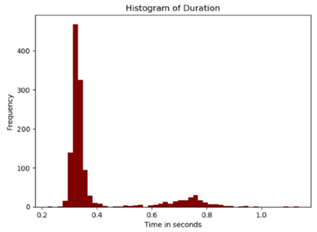

2.5
2.0
1.5
1.0
0.5
0.0
0
200
400
600
800
1000
1200
1400

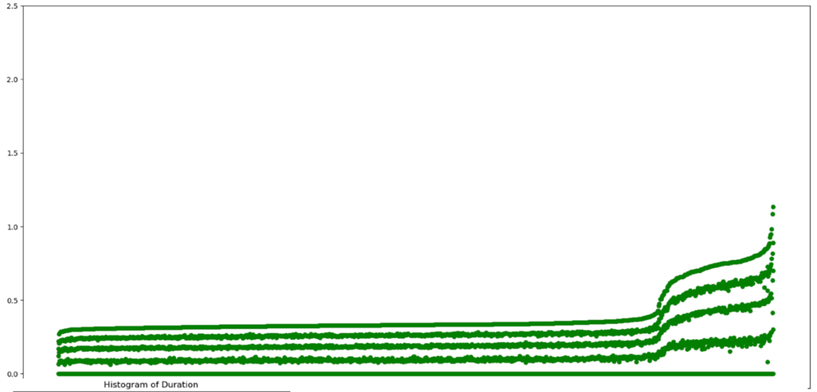

2.5
2.0
1.5
1.0
0.5
0.0
0
500
1000
1500
2000

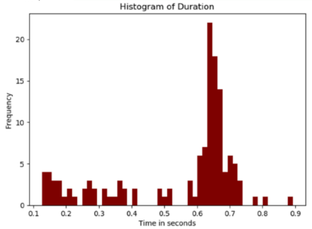

2.5
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80

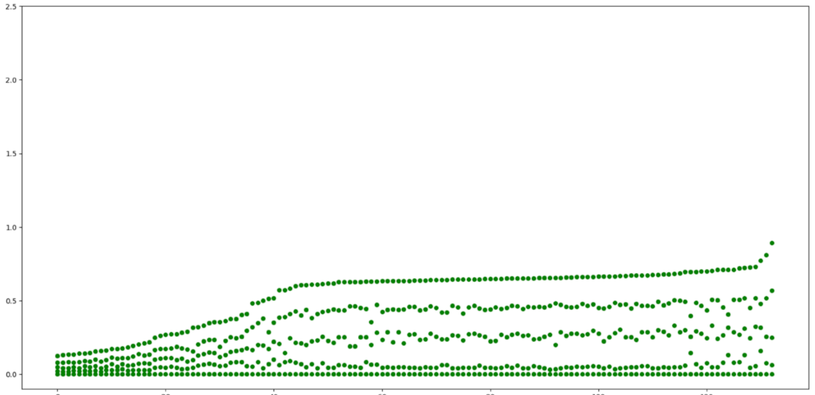

2.5
2.0
1.5
1.0
0.5
200
400
600
800
1000
1200

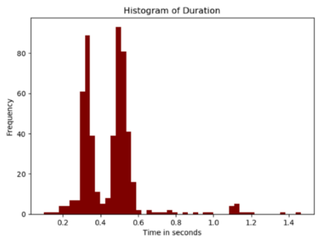

2.5
2.0
1.5
1.0
0.5
0.0
0
100
200
300
400
500

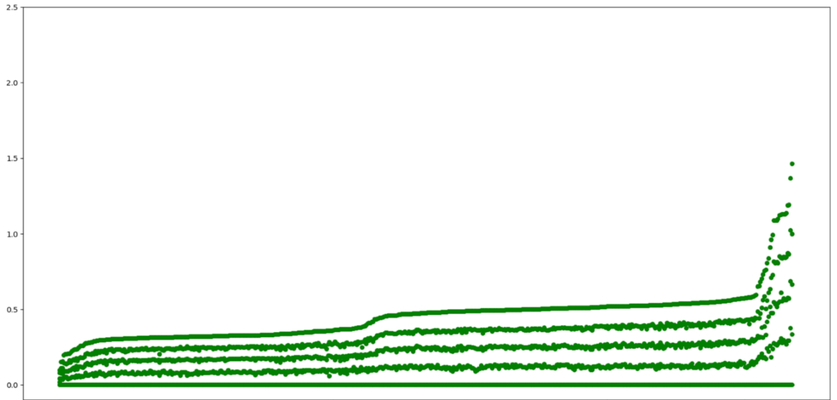

Histogram of Duration
Frequency
Time in seconds

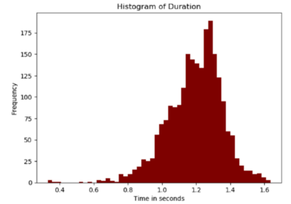


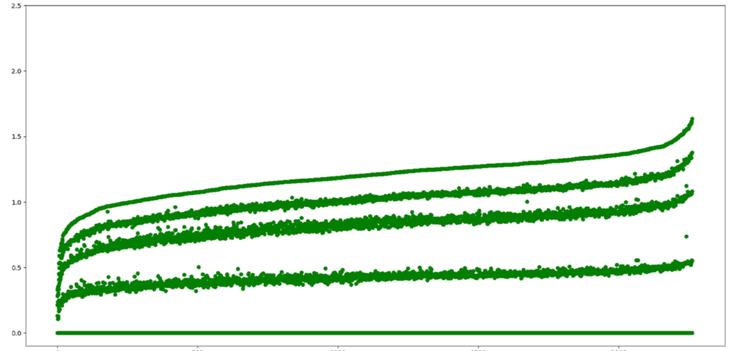

Histogram of Duration
Frequency
Time in seconds

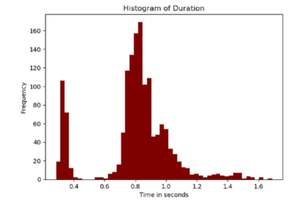


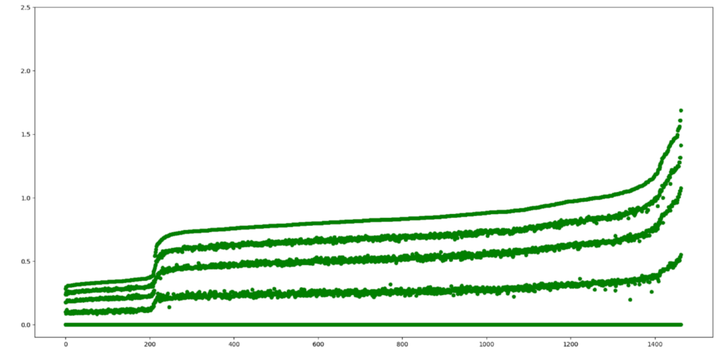

Histogram of Duration
Frequency
Time in seconds

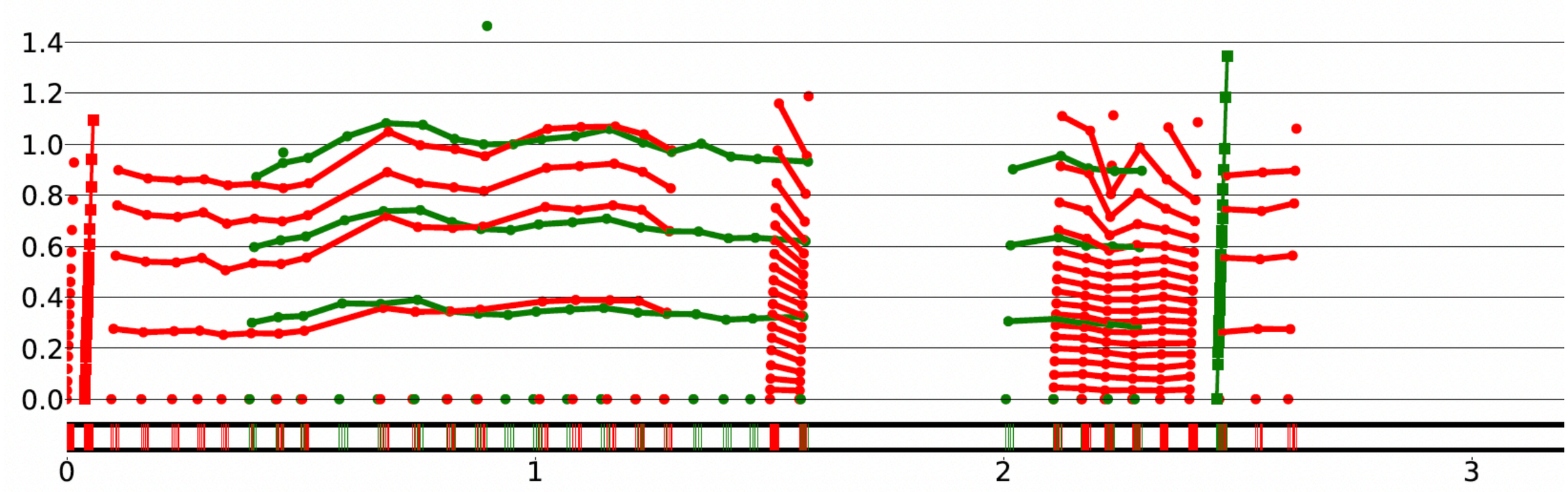



Histogram of Duration
Frequency
Time in seconds

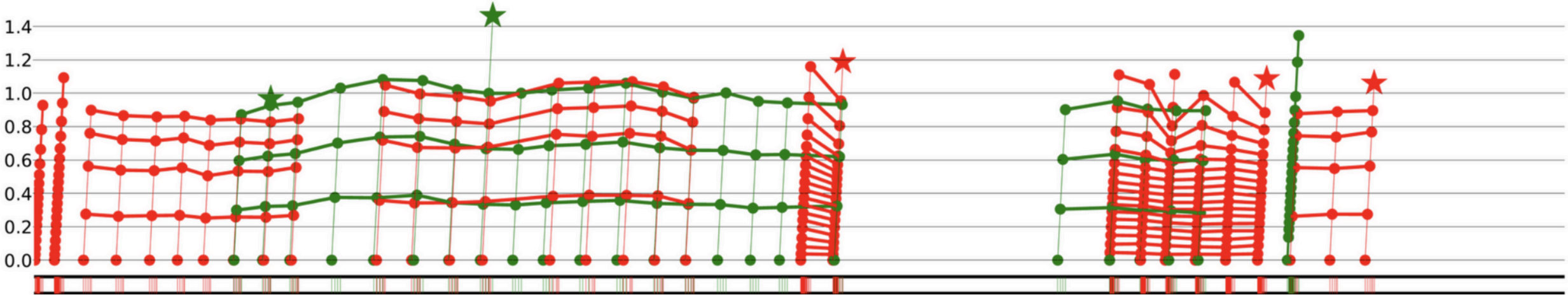



Histogram of Duration
Frequency
Time in seconds



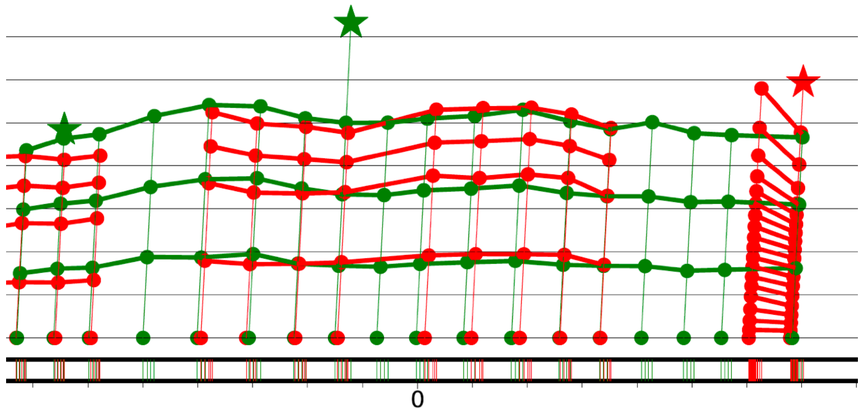

1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3

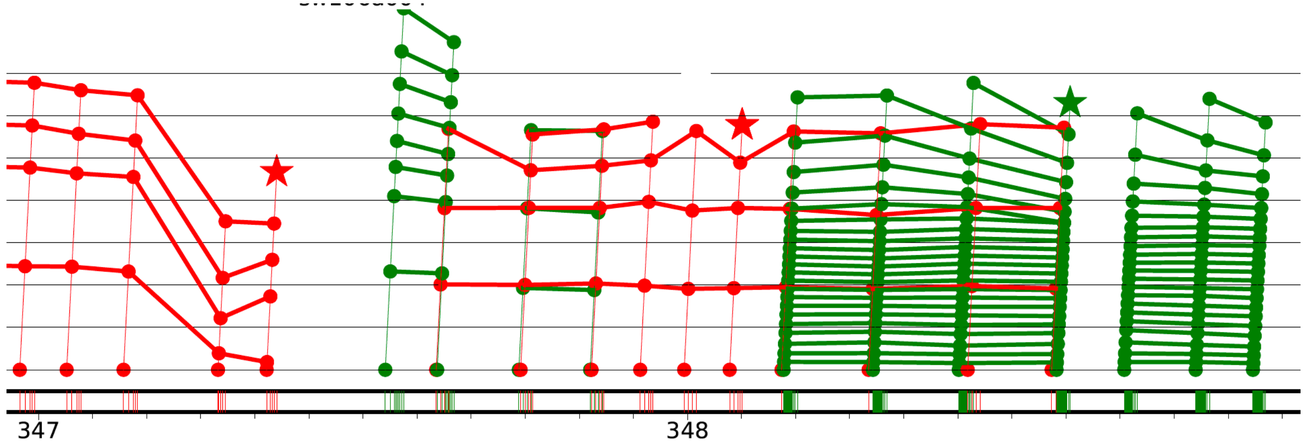
Extra Click


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

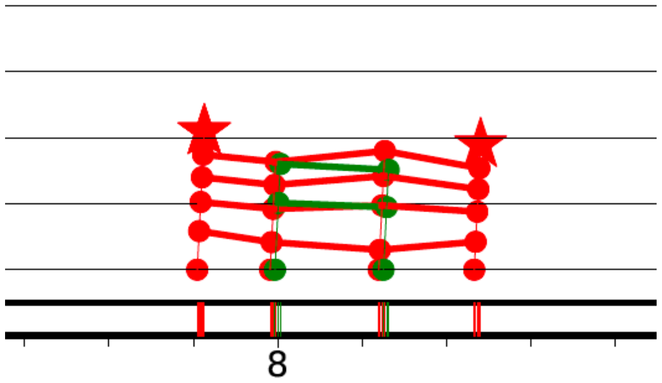
Interruption dropped

End of sentence

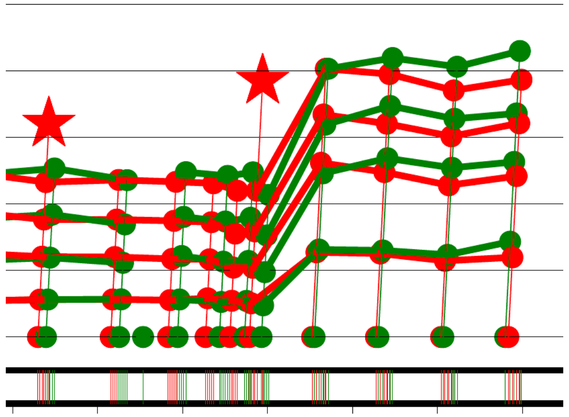




Change of Rhythm

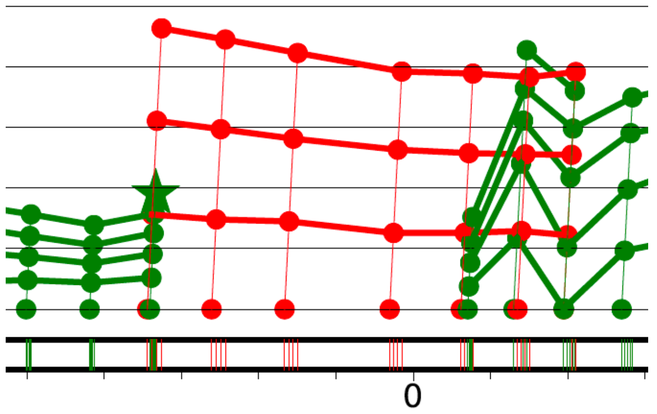


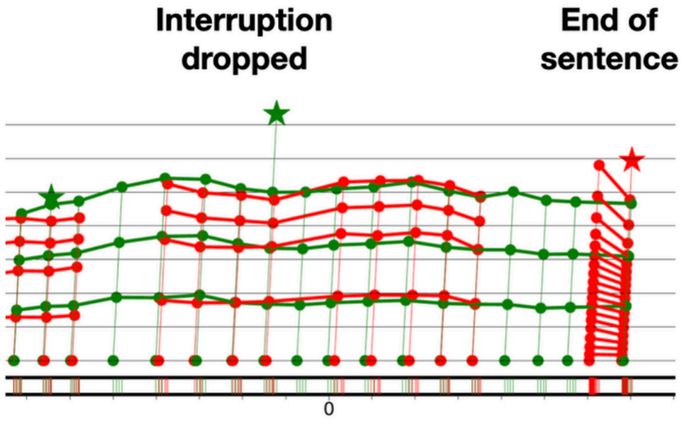
Start of interruption

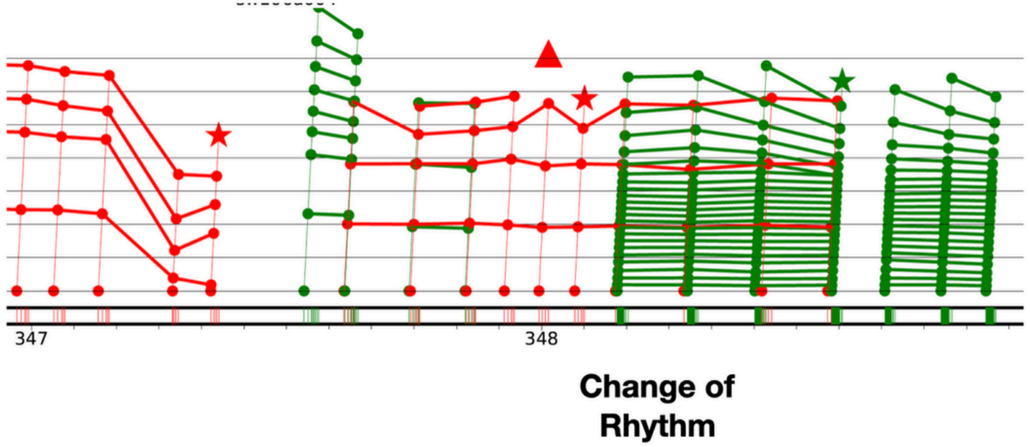


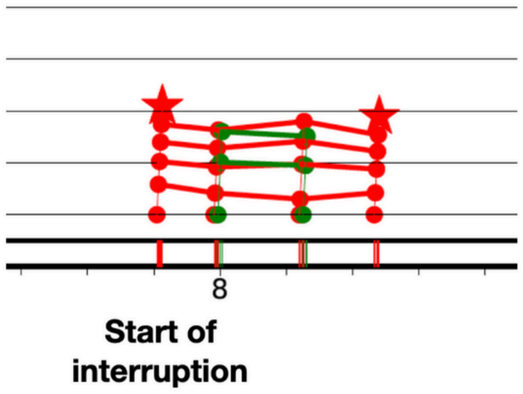
Nothing

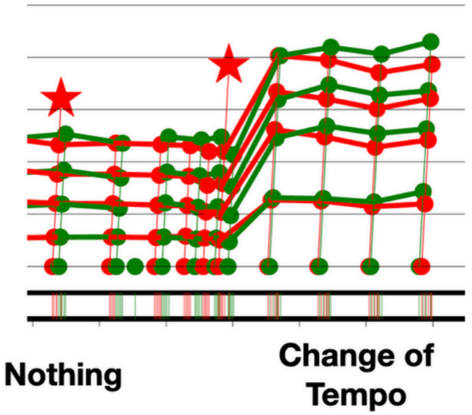
Change of Tempo

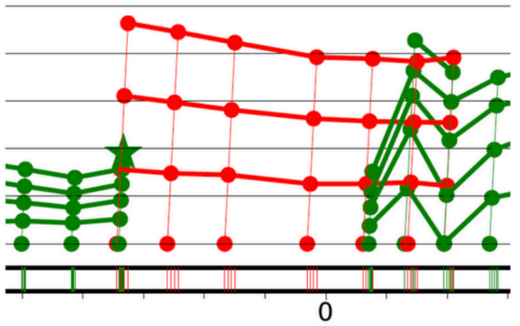








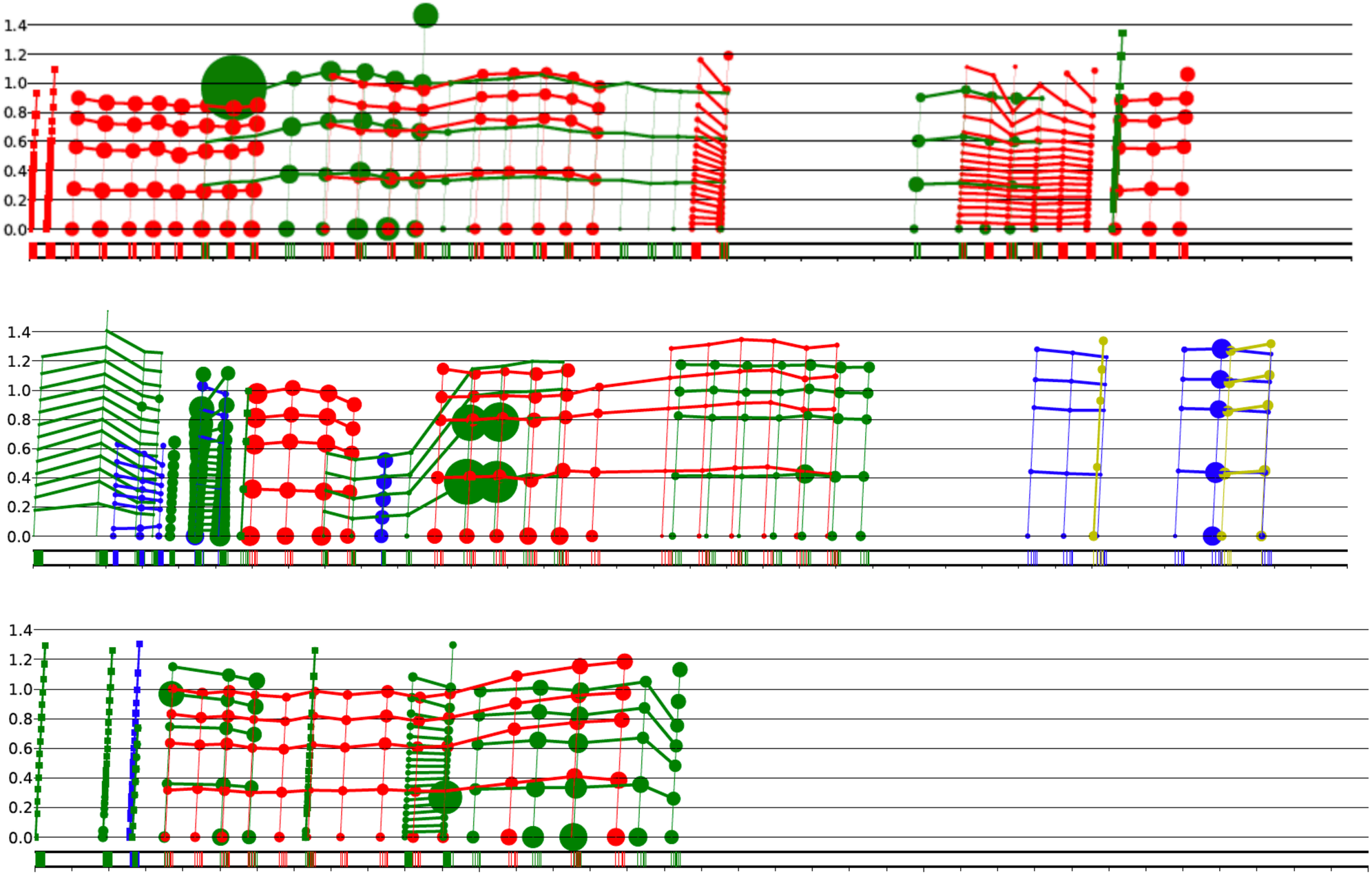




	Predict Interesting behavior
No Extra Click	0.7079533941
Extra Click	0.7619047619

	Predict Change Change in Turn Taking behavior
No Extra Click	0.6712259372
Extra Click	0.7212987013

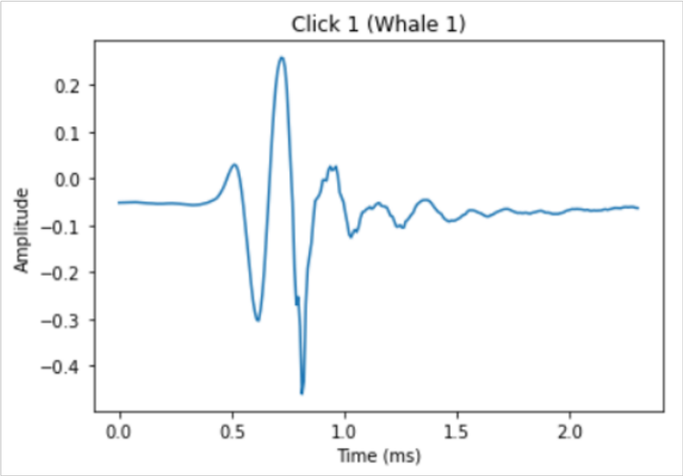
Is when the click starts all that there is to a click?

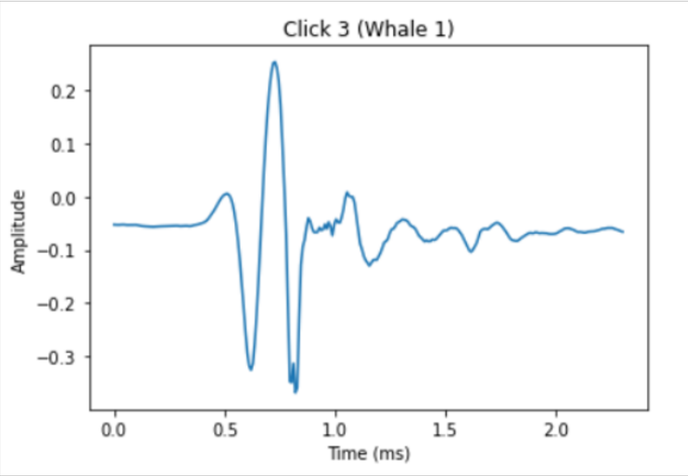


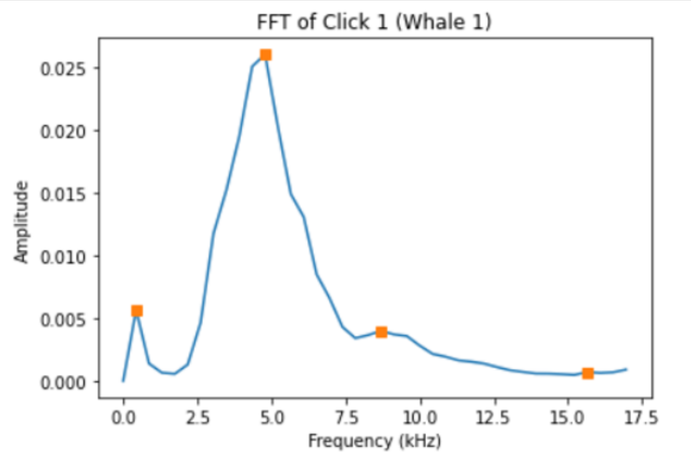
Radius of the circle
∝
power of the click

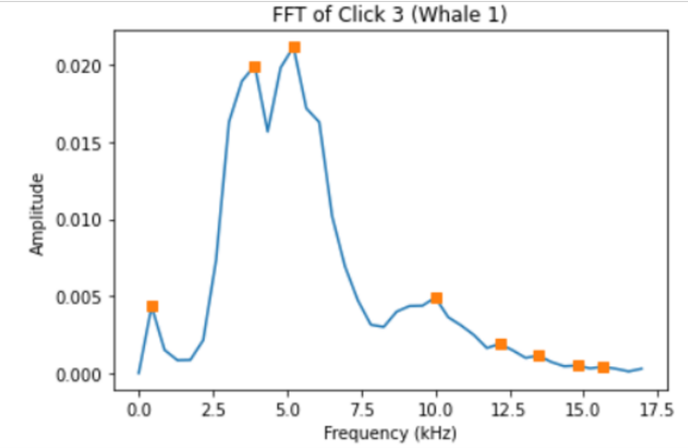
Do sperm whales intentionally control the frequency variation of their clicks?

Miles B. Silva



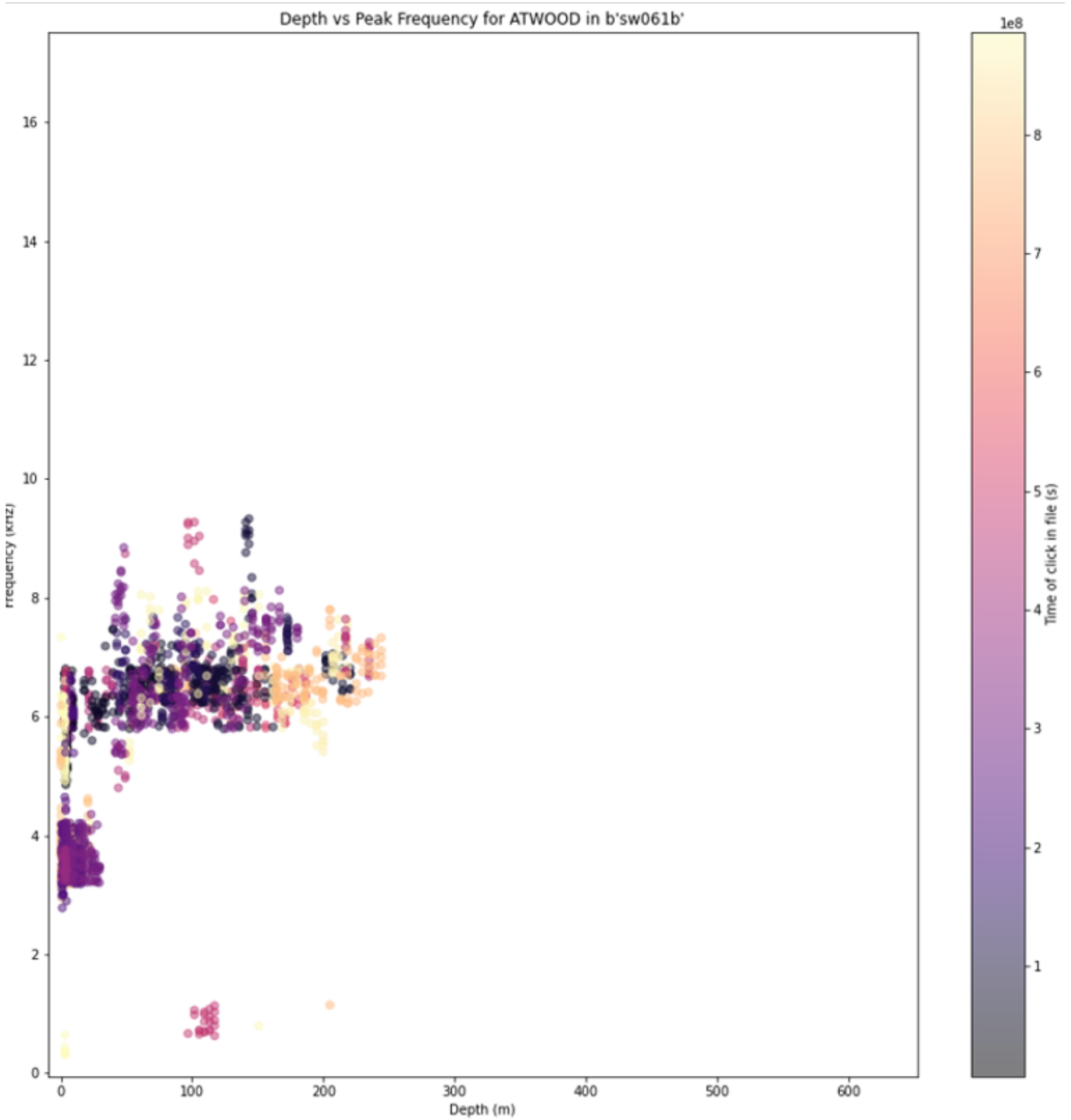




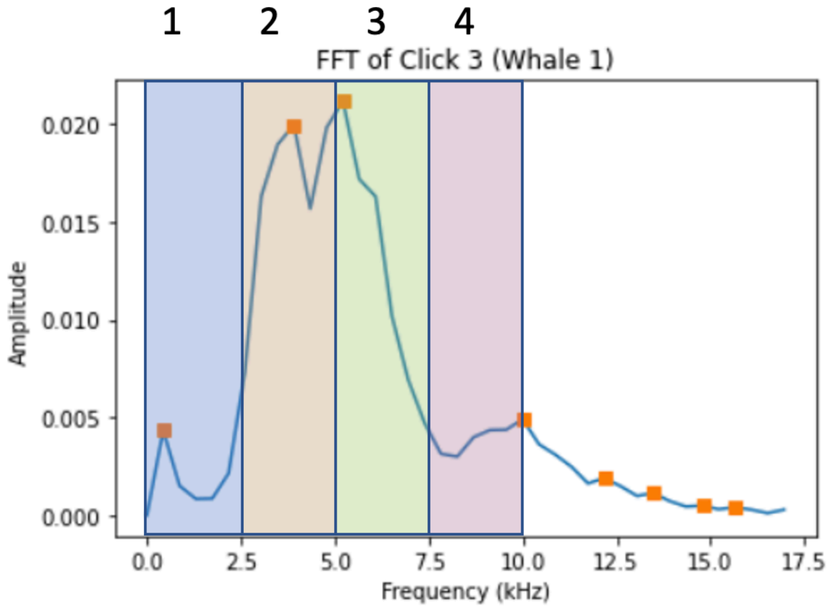


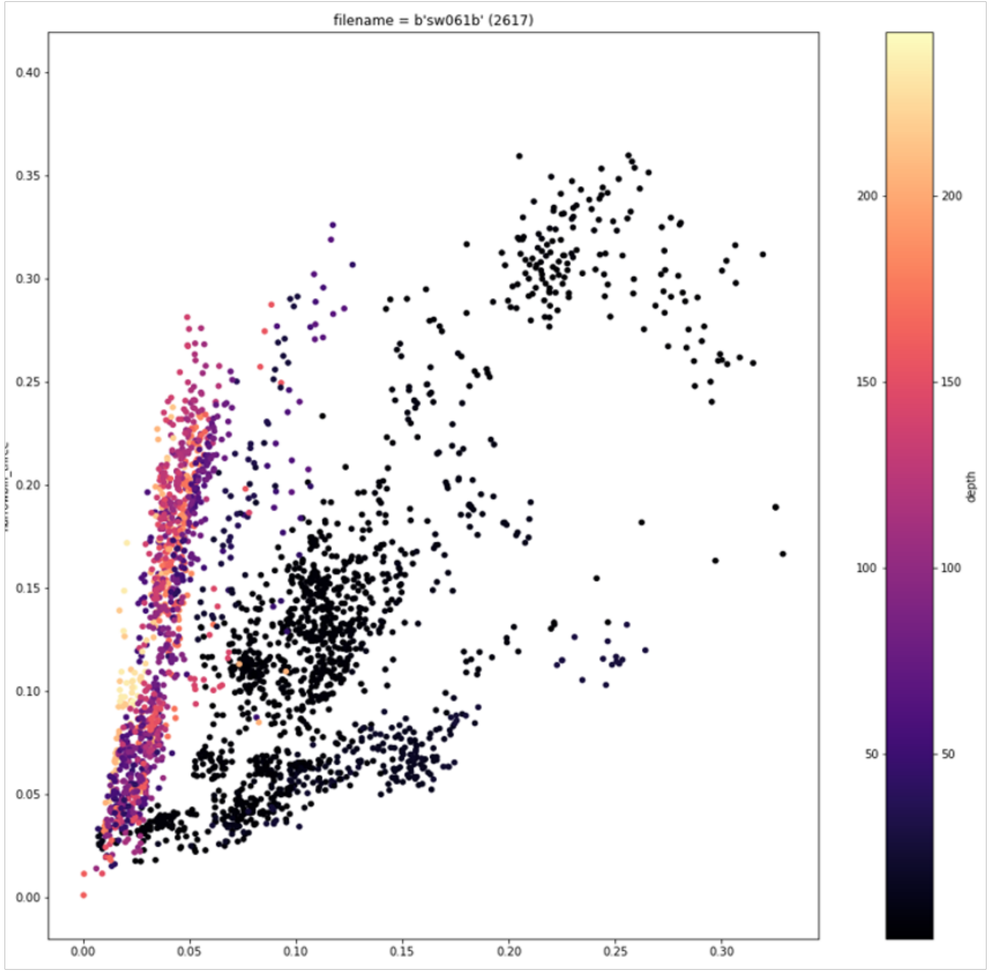
Surface calls : Wider band of frequencies.

Depth : Almost a fixed band of freq

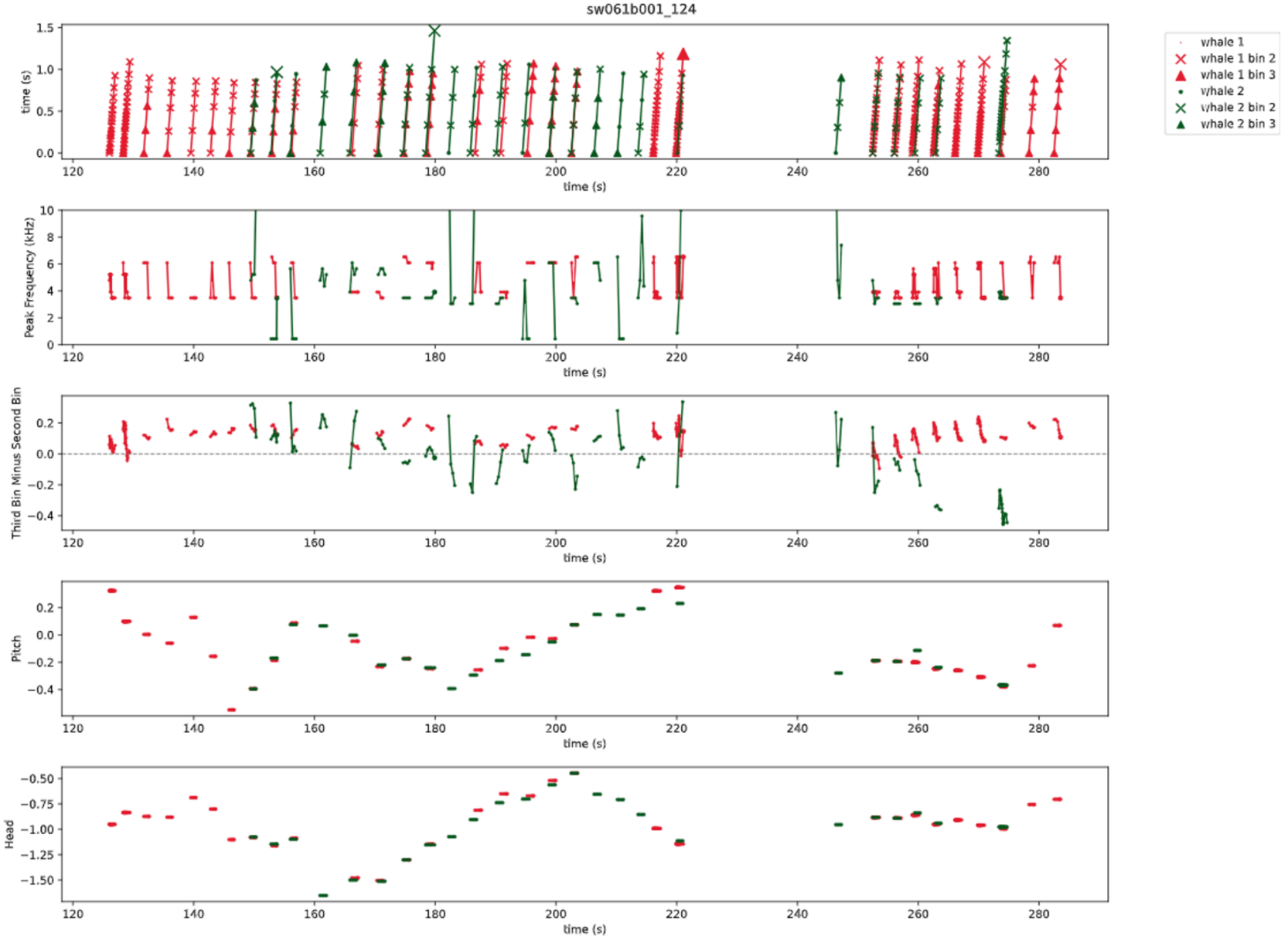


X: Bin 2 vs Y: Bin 3





Surface: Greater variation in emphasis. Depth : More homogenous



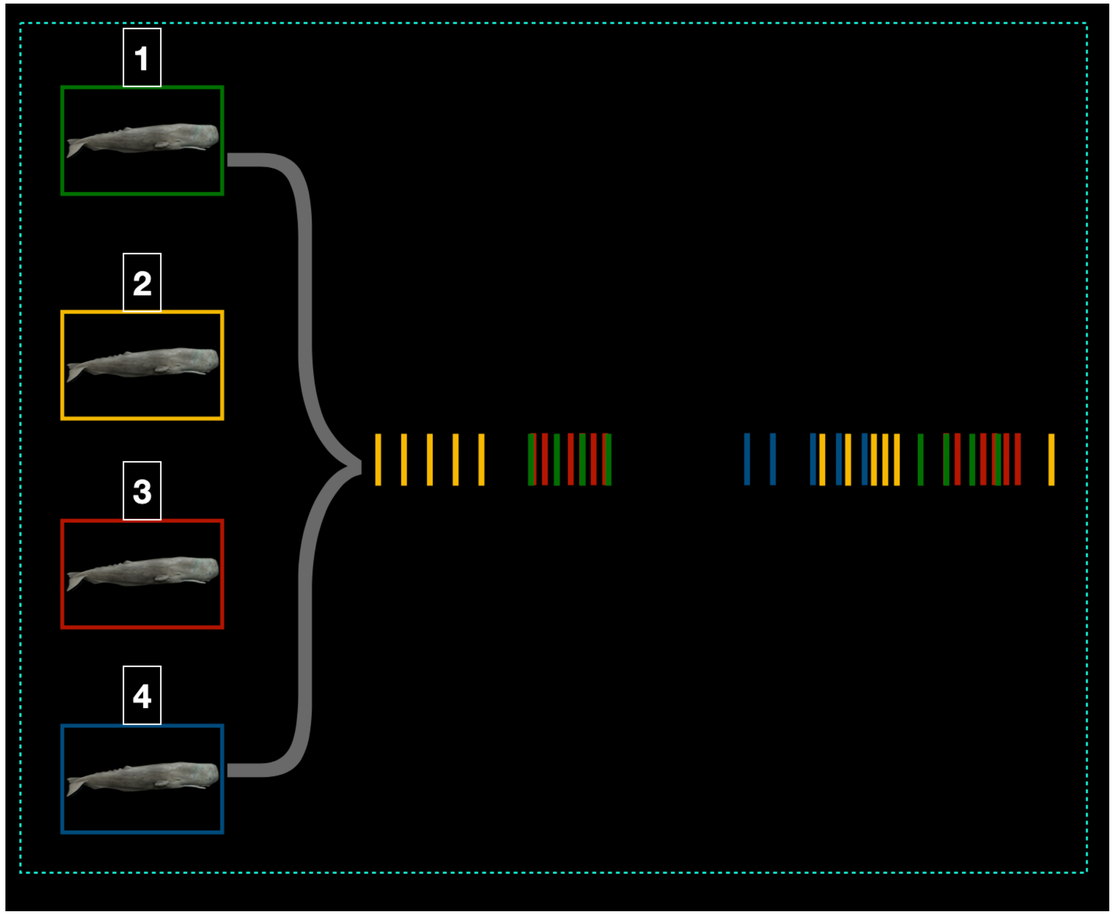
Is this controlled or only a function of behavior?

Summary: Structure in the Sounds

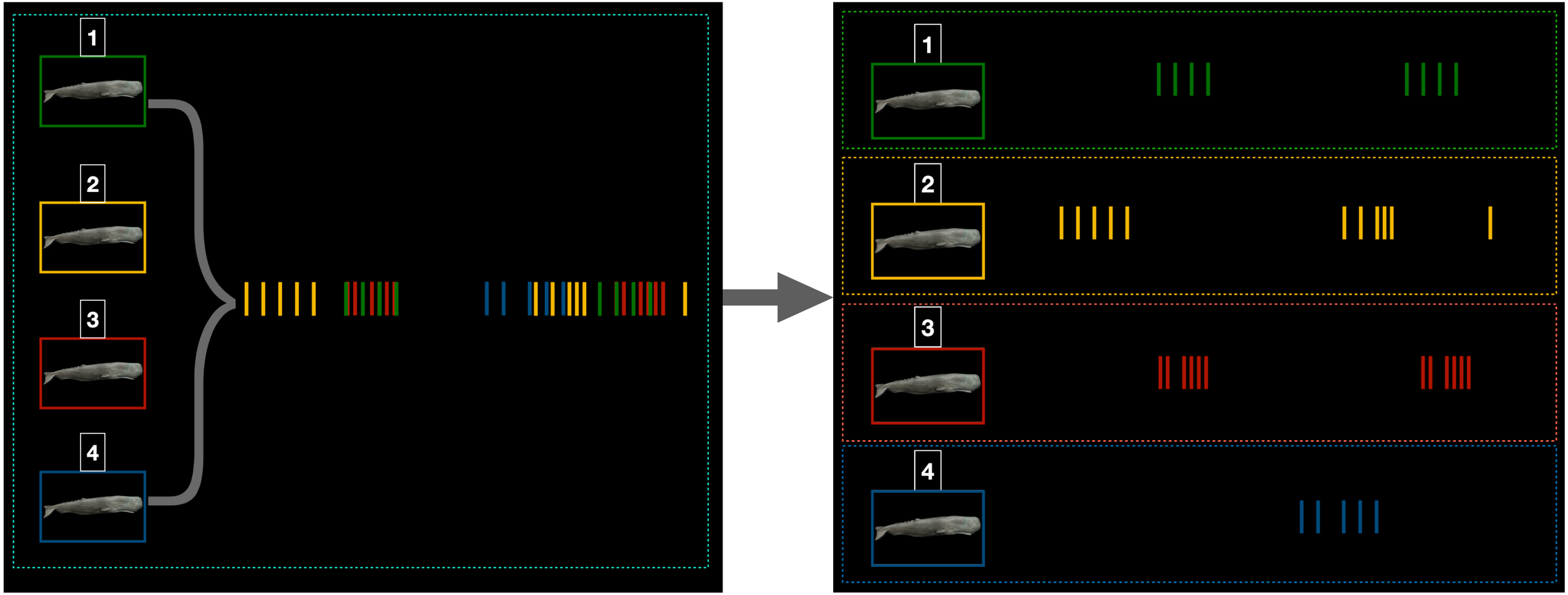
1. Understanding SW Vocalizations : BOW
2. Dialogue: Whale LM (w/ and w/o context) + Protocol
3. What are the smallest units?
4. Deviation from “codas”
    1. Extra click
    2. Looking inside a click

Automatic Annotation

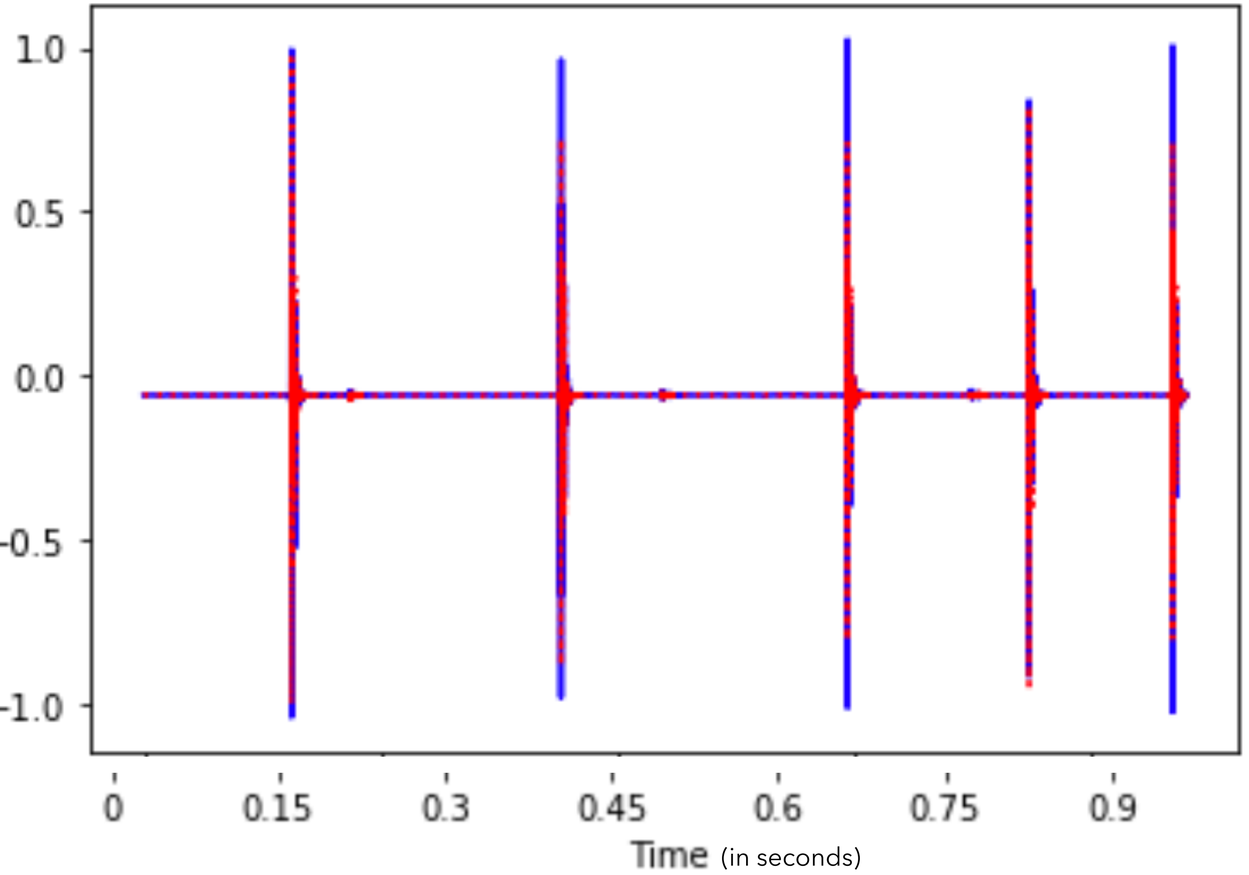
Automatic Annotation

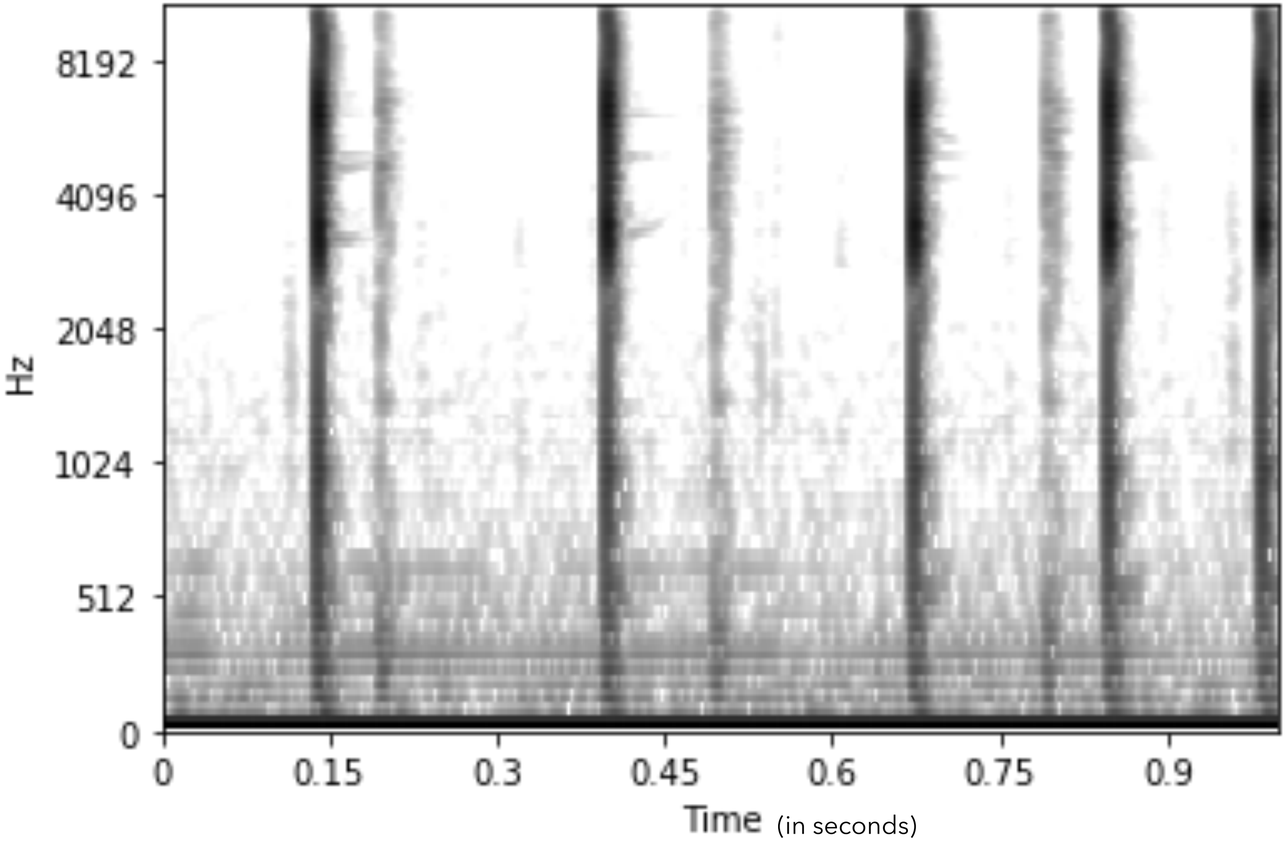


Automatic Annotation

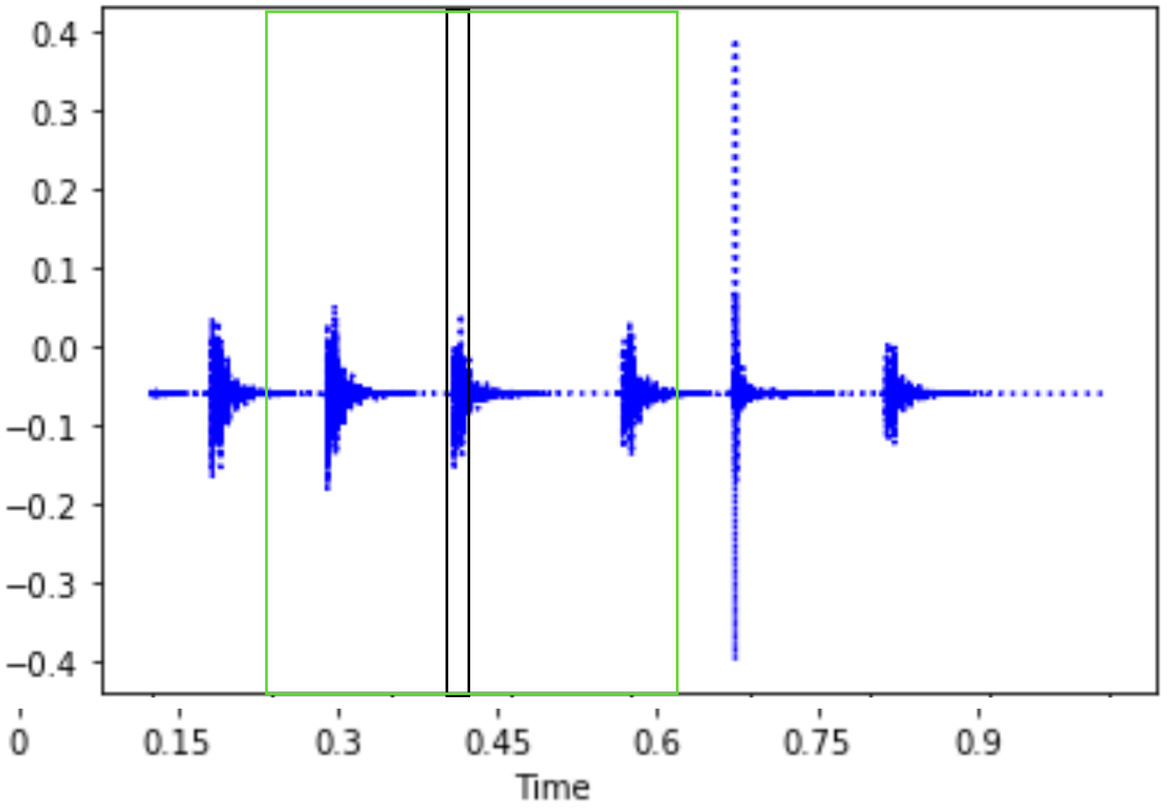


Click Detection



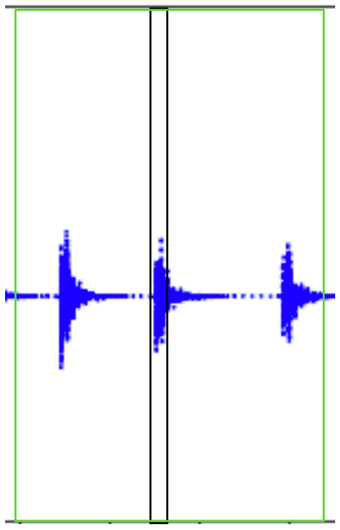


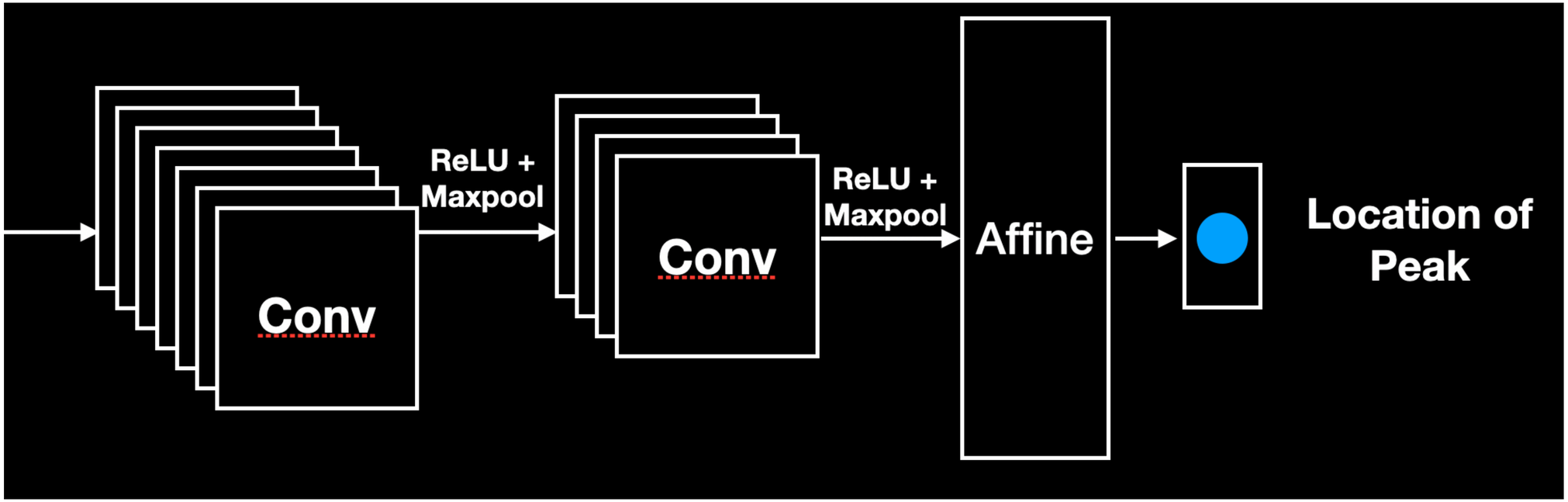
Click Detection



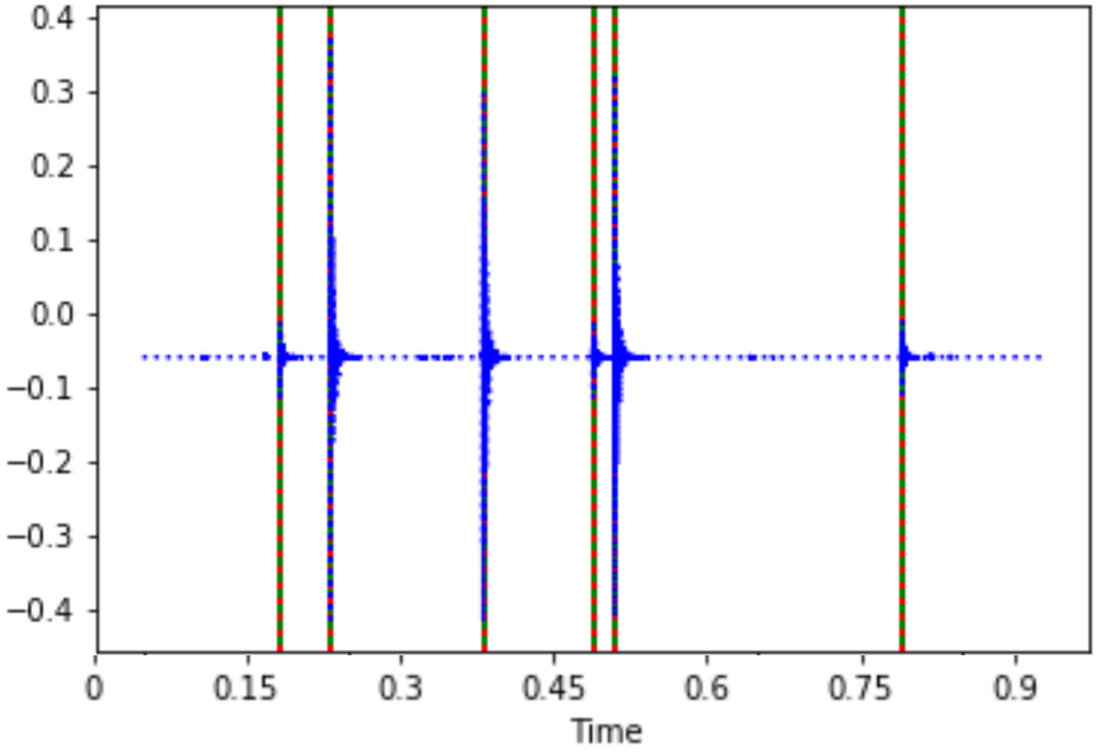
Where is the peak?

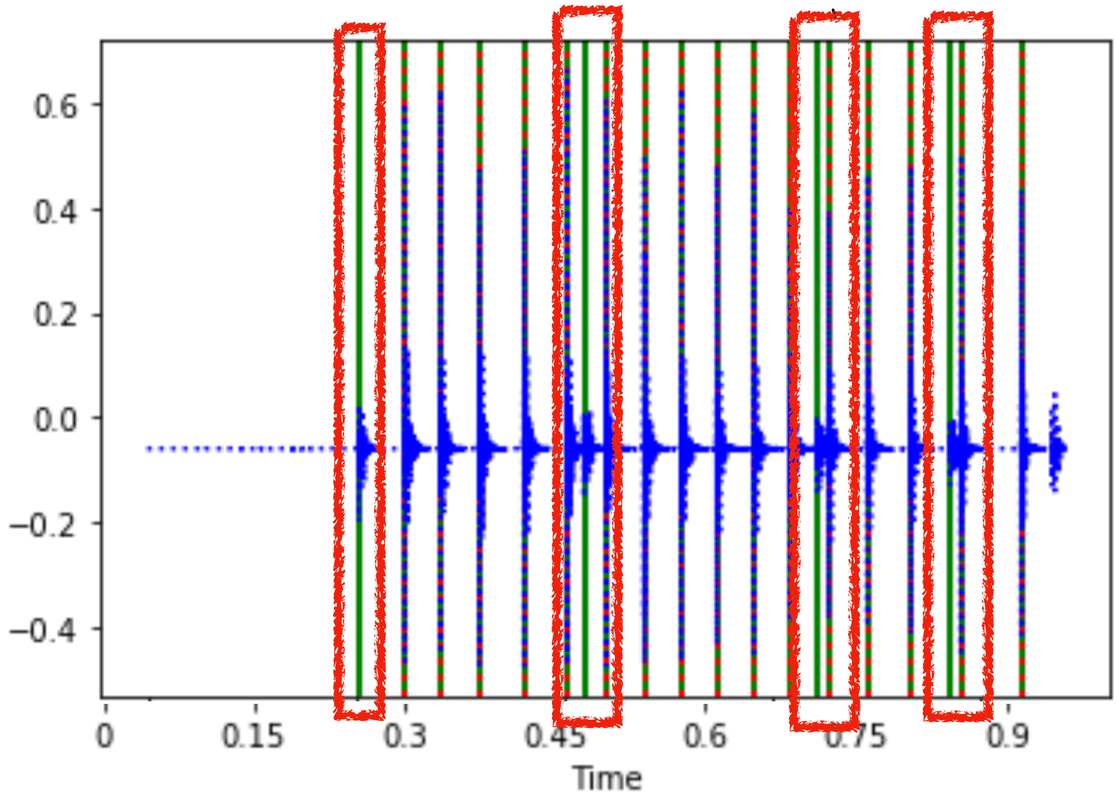
Model




ReLU +
Maxpool
Conv
ReLU +
Maxpool
Conv
Affine
Location of
Peak

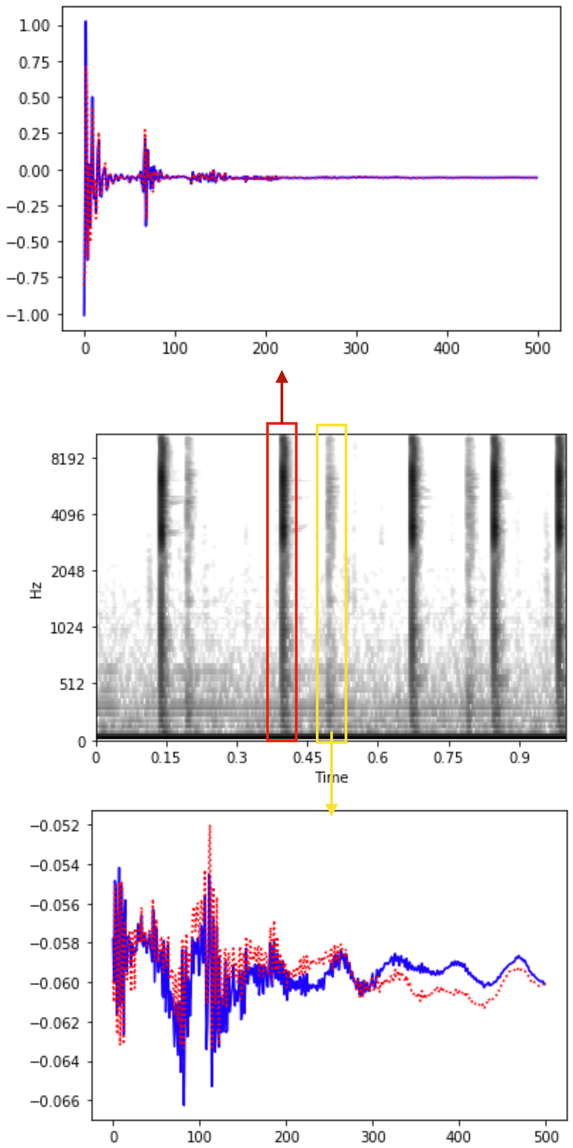
Click Detection





- Can detect the onset of both soft/ loud clicks ~94% accuracy
- Can also recover previously unannotated/unidentified signal!

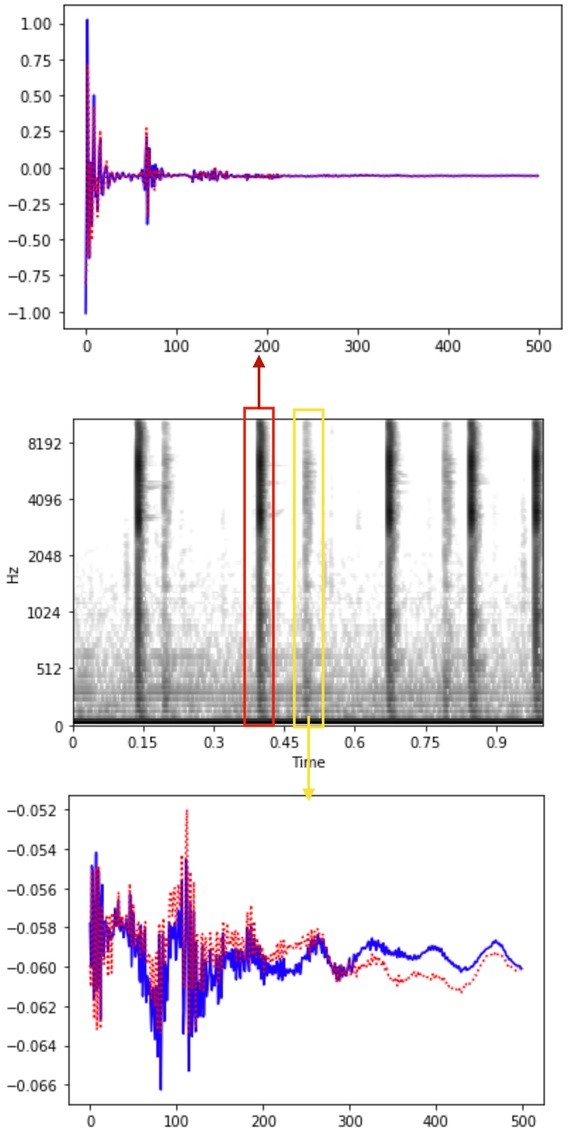
Click Separation



Are the two clicks by the same speaker?

Yes (1) / No (0)

Click Separation



- IPI info + angle of arrival info

Accuracy: 59%

- Raw wav

Accuracy: 69.8%

- Input: Just raw wav of one click: Output: Does the click belong to the whale wearing the mic?

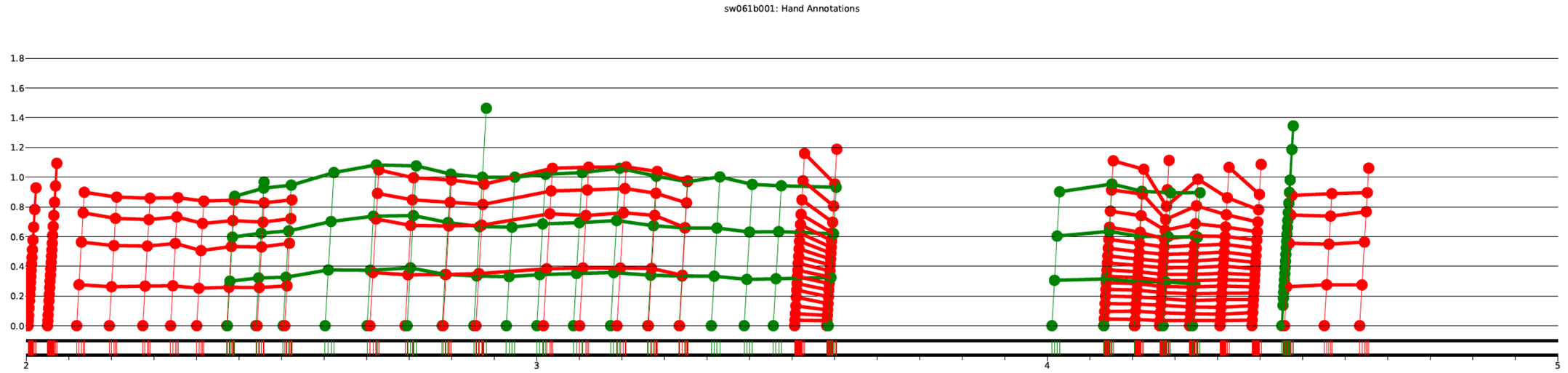
Accuracy: 88%

Automatic Annotation of Unannotated Data

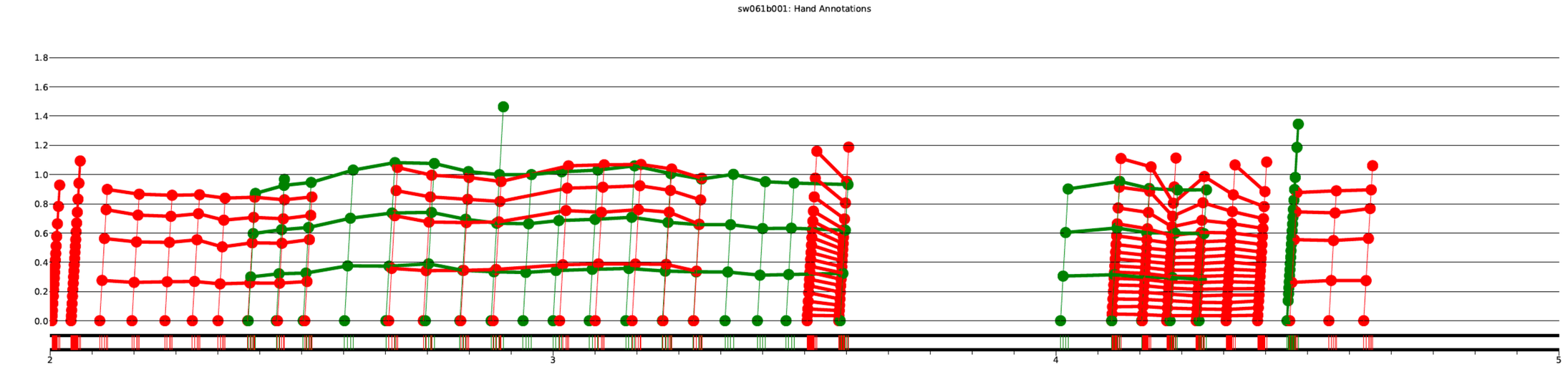
Finnian Jacobson-Schulte

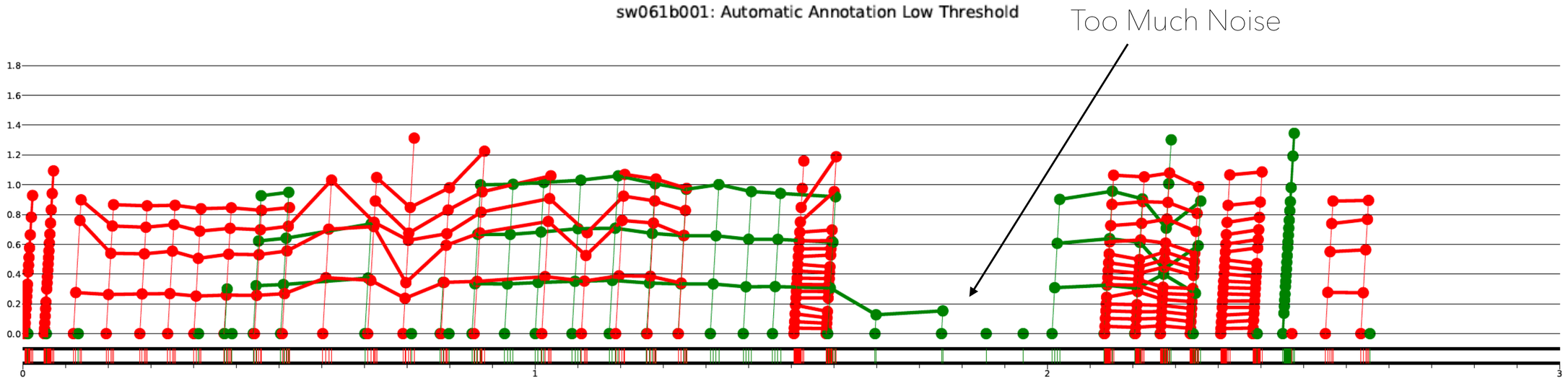
Ioannis Kaklamanis

- Can we recreate the hand annotations using the click detector and click separator models?
- This will allow us to annotate the remaining data, doubling the amount of data we have to work with.

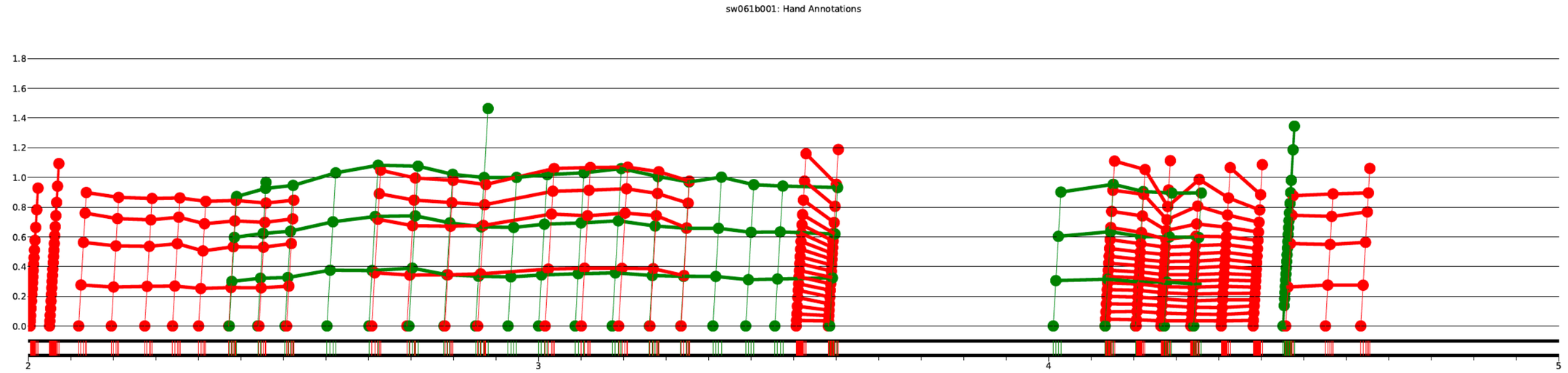


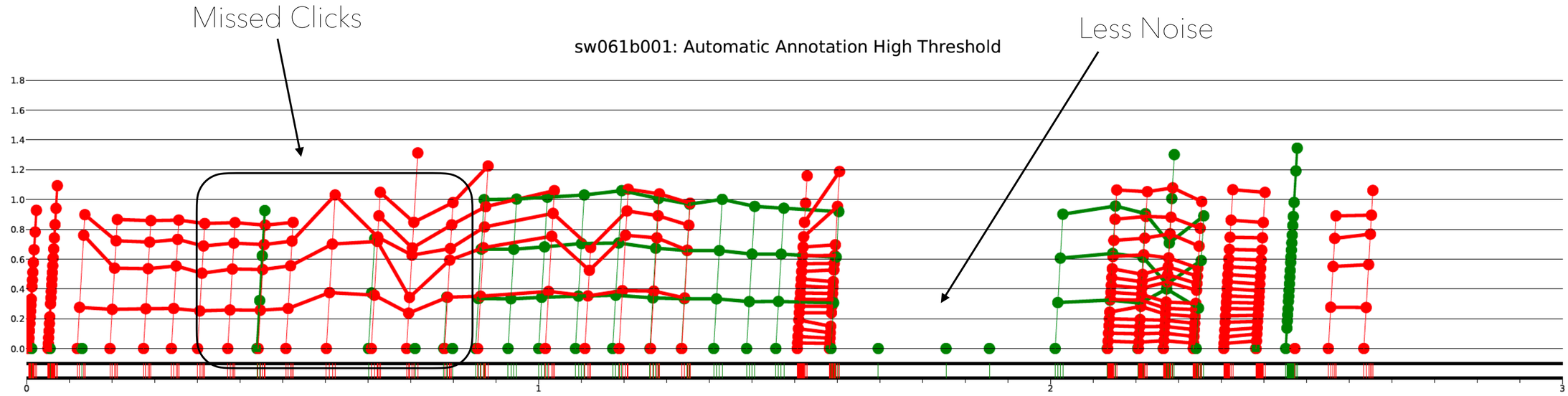
Automatic Annotation - Low Threshold





Automatic Annotation - High Threshold

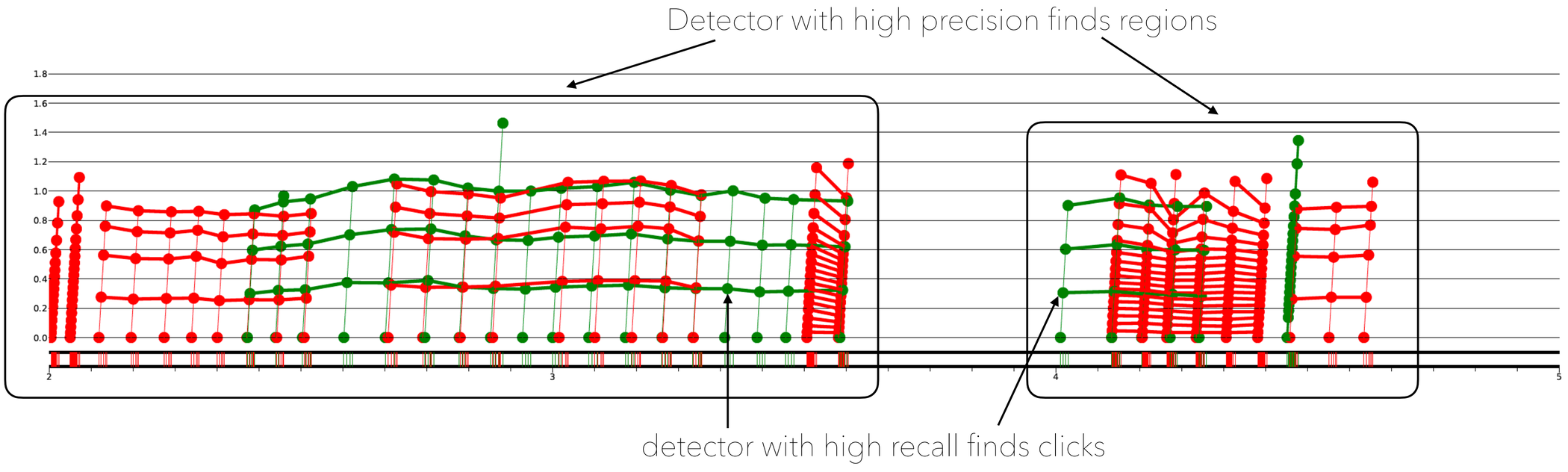




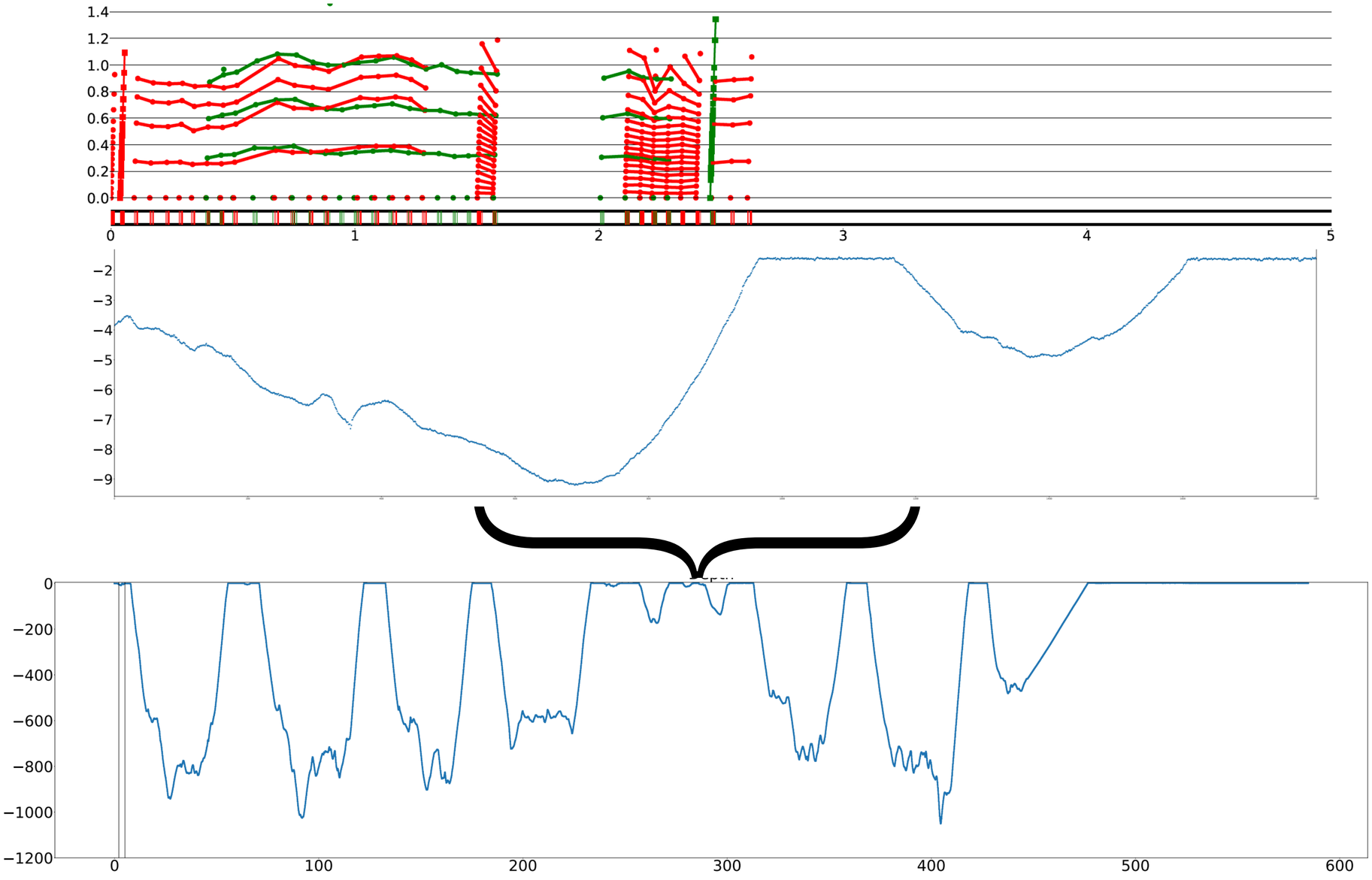
Best of Both Worlds

- Need to detect the soft clicks from non-focal whales, while also not detecting noise as clicks

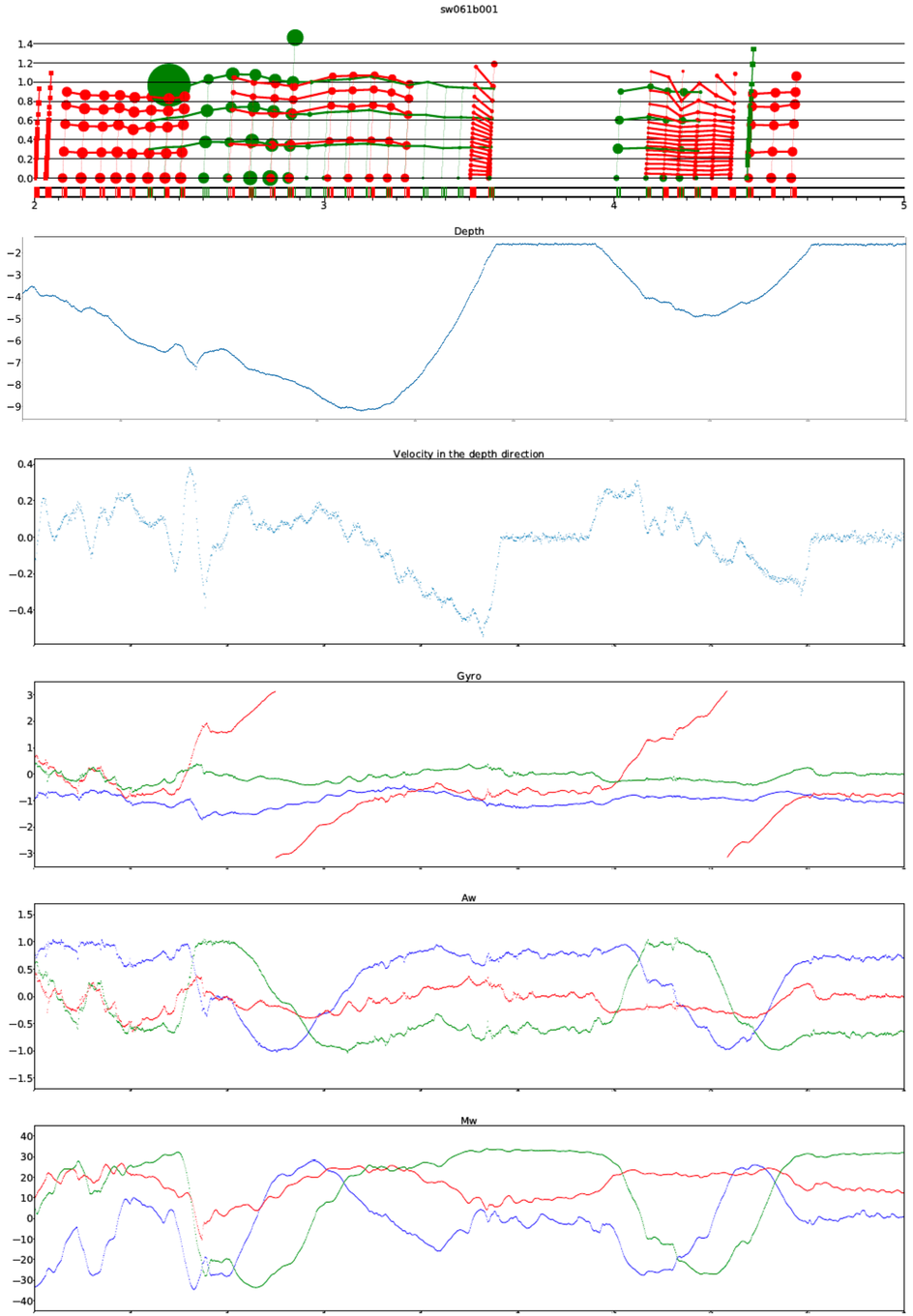
- Use a detector trained with hard noise examples to detect regions where codas are happening, then use a sensitive detector to detect all clicks.



Meta-data


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
−2
−3
−4
−5
−6
−7
−8
−9
0
−200
−400
−600
−800
−1000
−1200
0
100
200
300
400
500
600

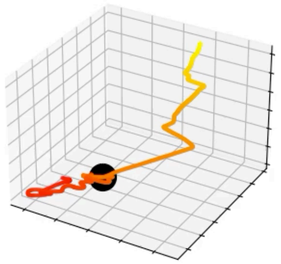
Other Meta-data



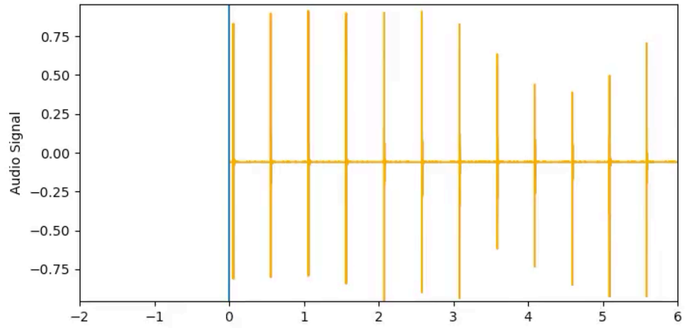
3D Visualization

Finnian Jacobson-Schulte

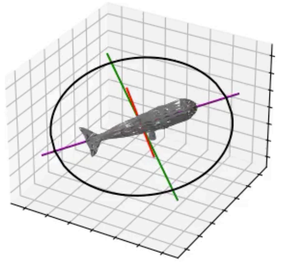
Whale Trajectory



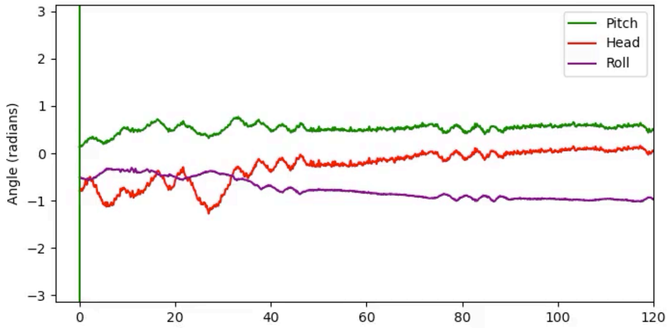
Microphone Audio



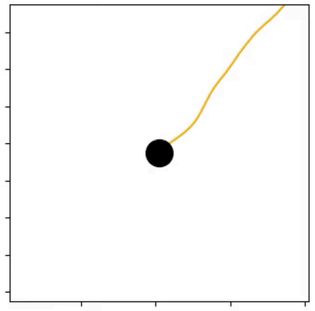
Whale Orientation



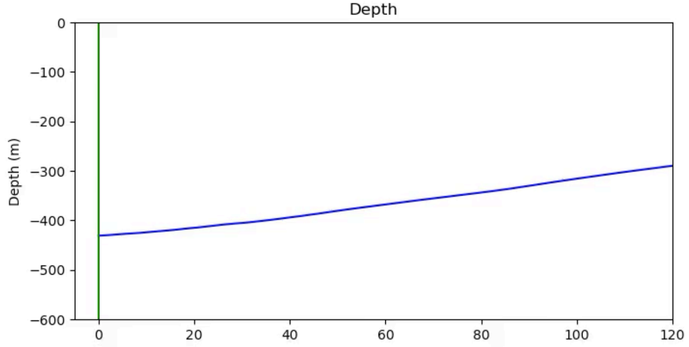
PRH



XY Movement

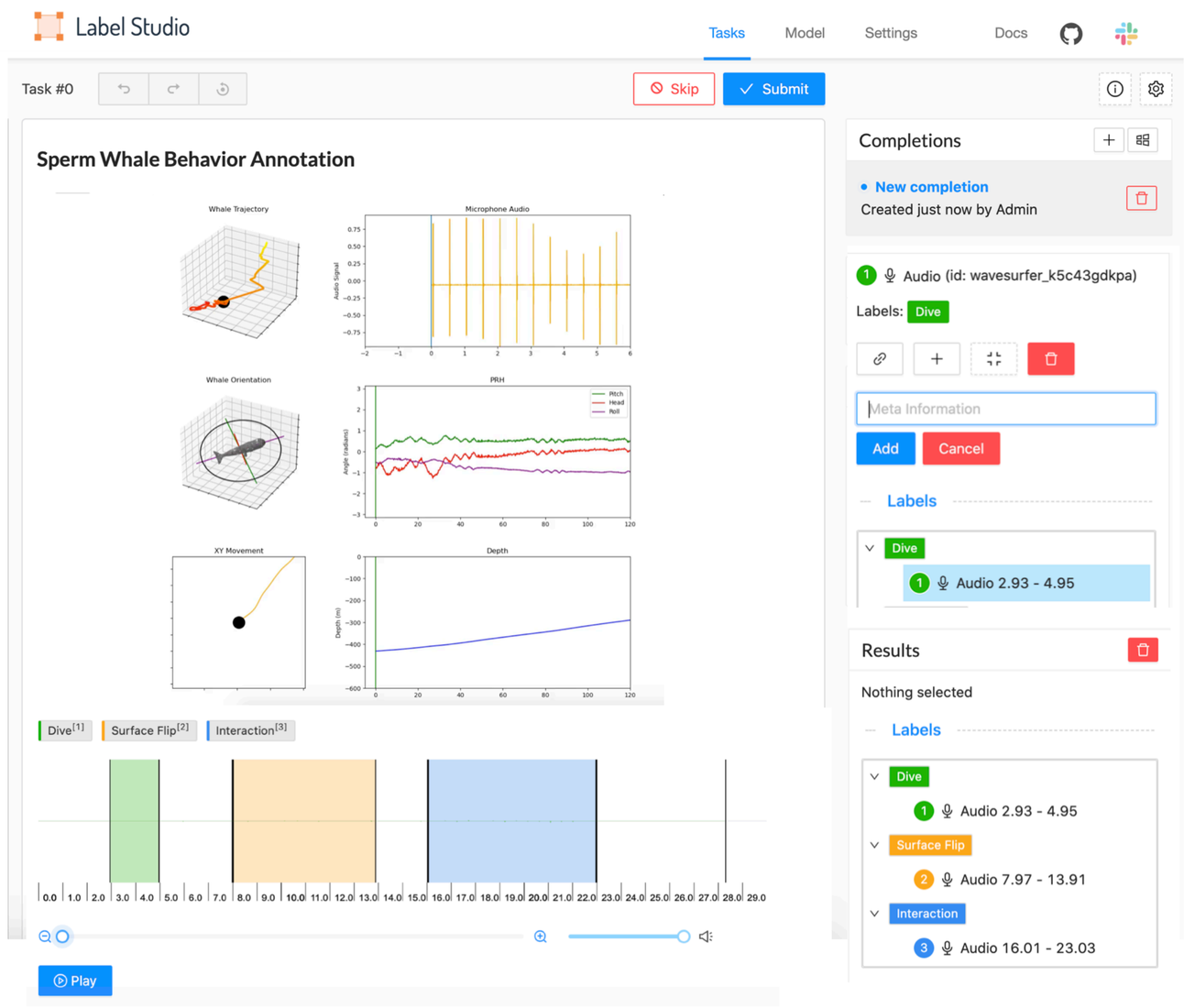


Depth



3D Visualization - Annotation

We want to be able to contextualize the vocalizations with the behavior.



Summary: Automatic Annotation

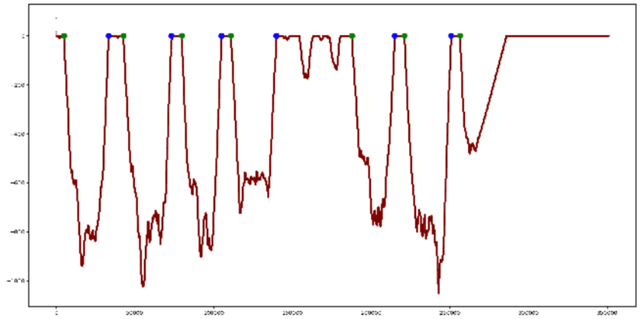
1. Click detection + Source Separation

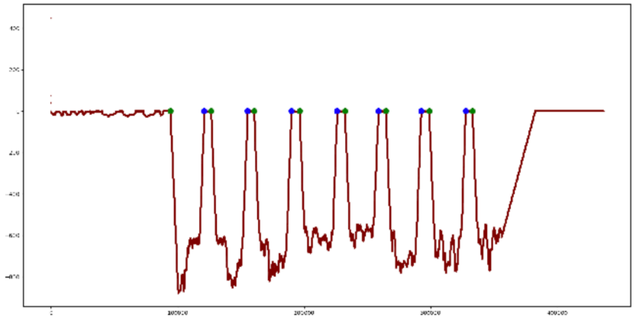
2. Behavior annotation

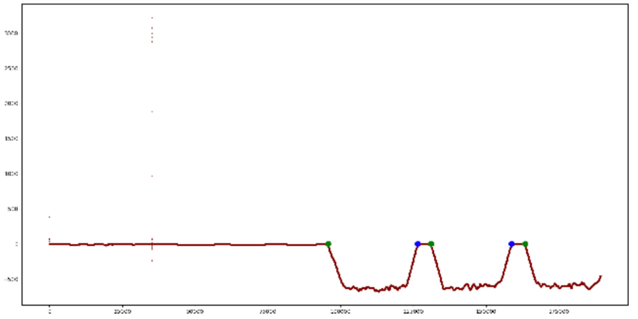
Next : Automatic Behavior Labeling (Increase the context available)

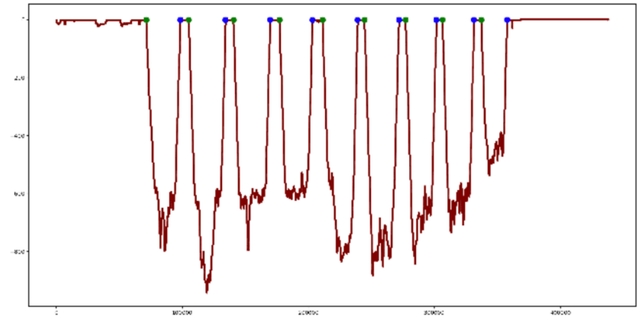
Connecting Sounds to Behavior

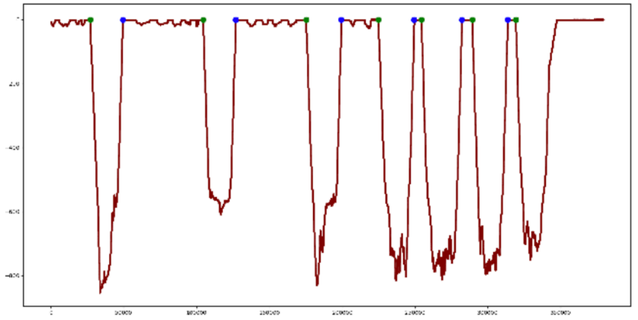
Diving

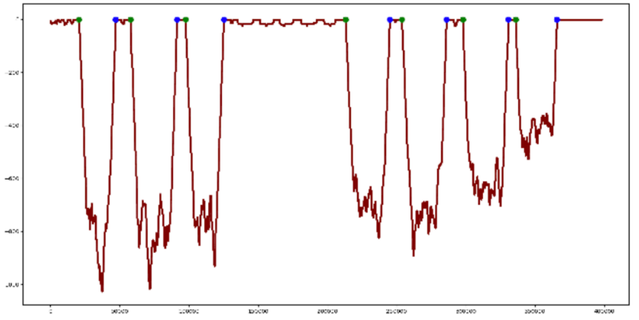


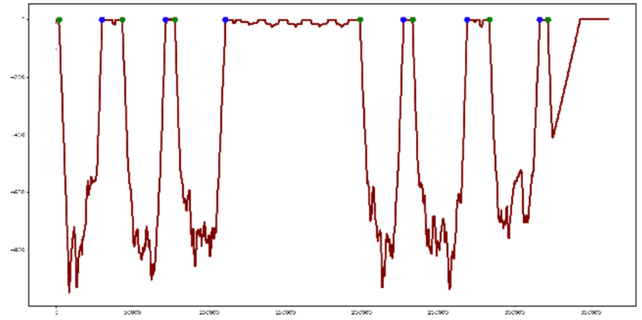


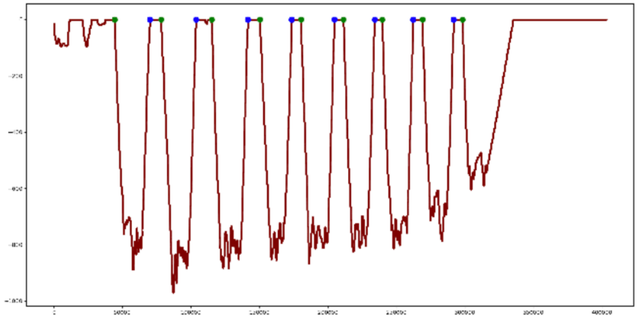


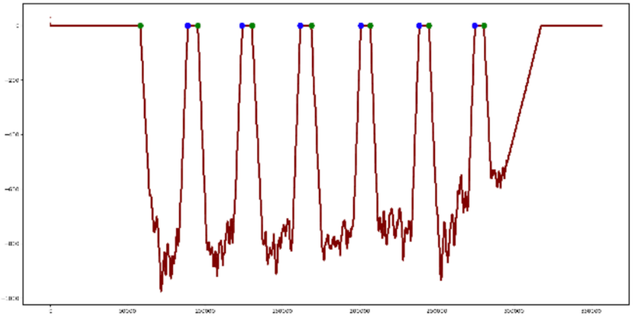




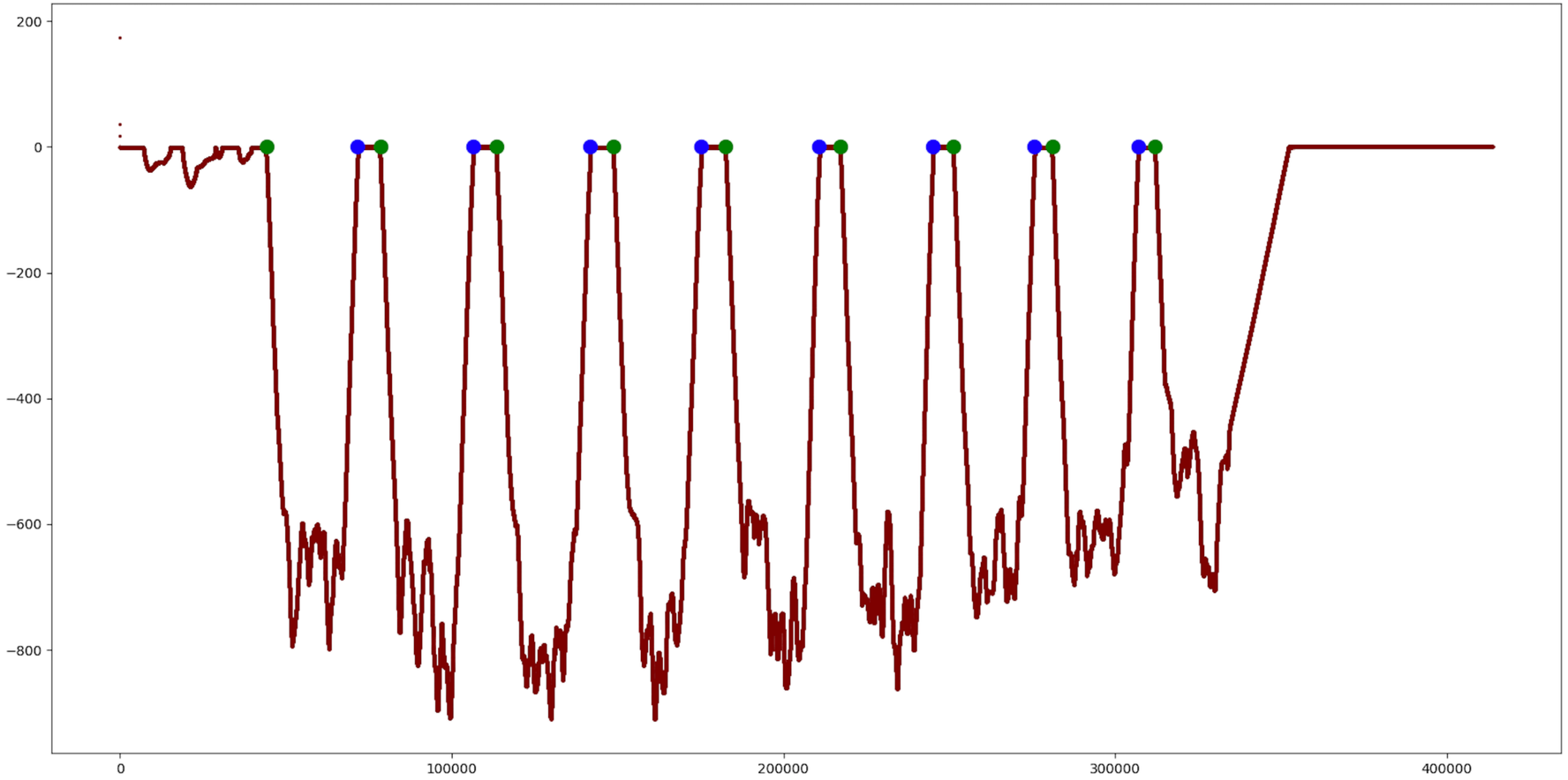


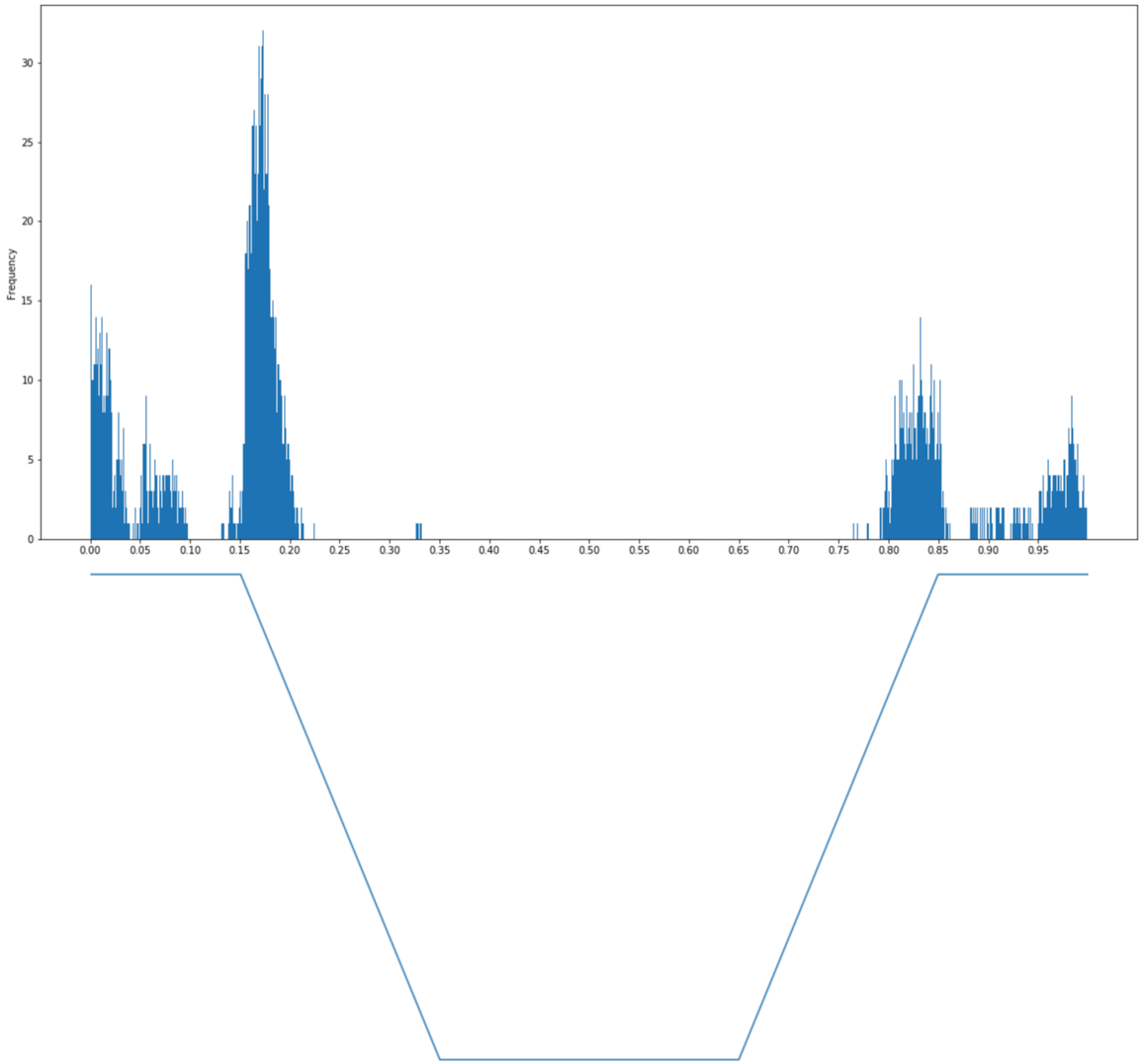






Mean time between dives: 9.95
Mean time of dive: 46.66




Frequency
0
5
10
15
20
25
30
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95

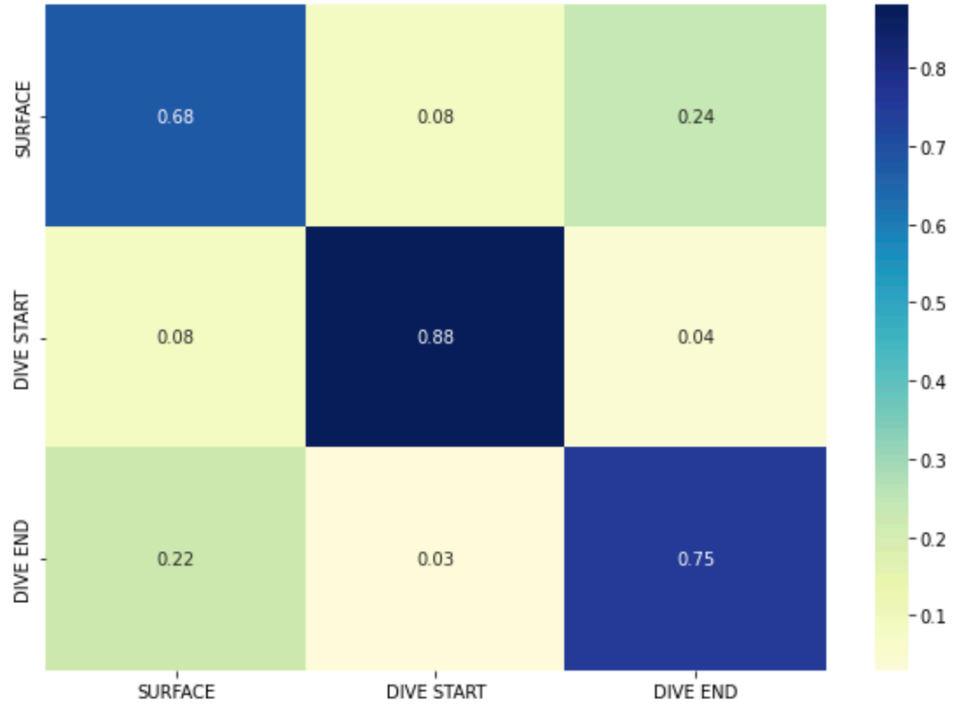
Prediction experiments

Given a conversation can we tell the state :

- diving in,

- returning from a dive,

- socializing on the surface

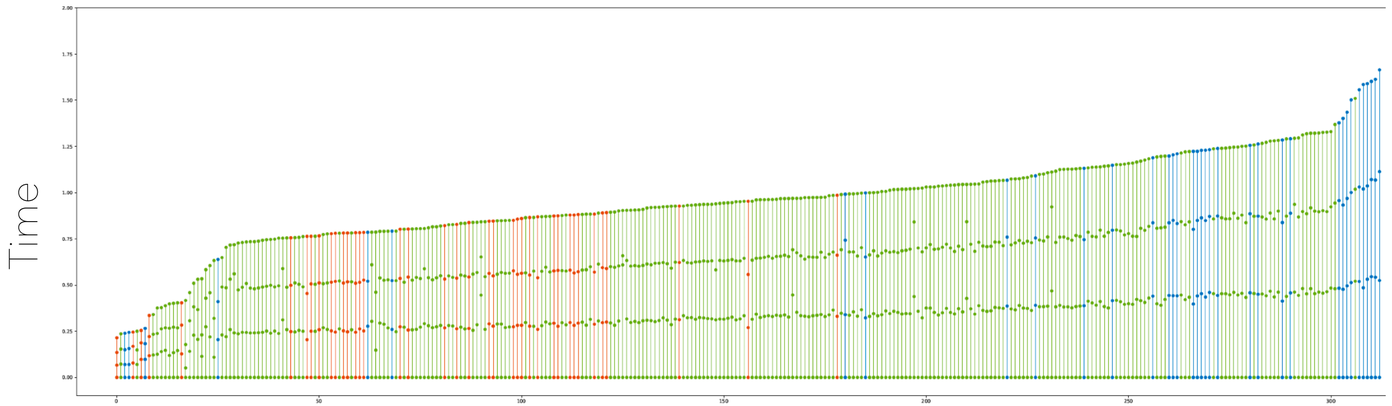


Accuracy: 71%

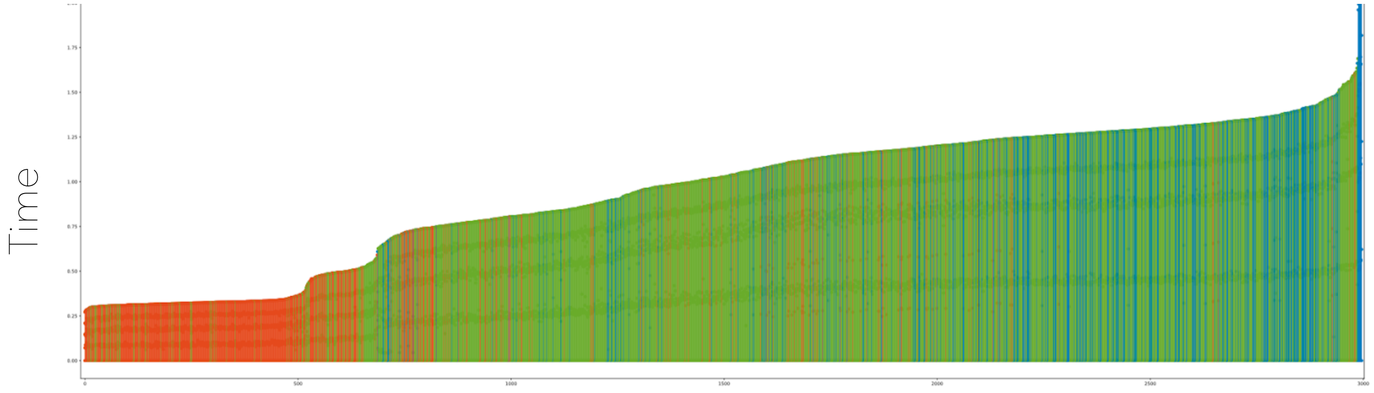
-> Diving ascent

-> codas on the surface

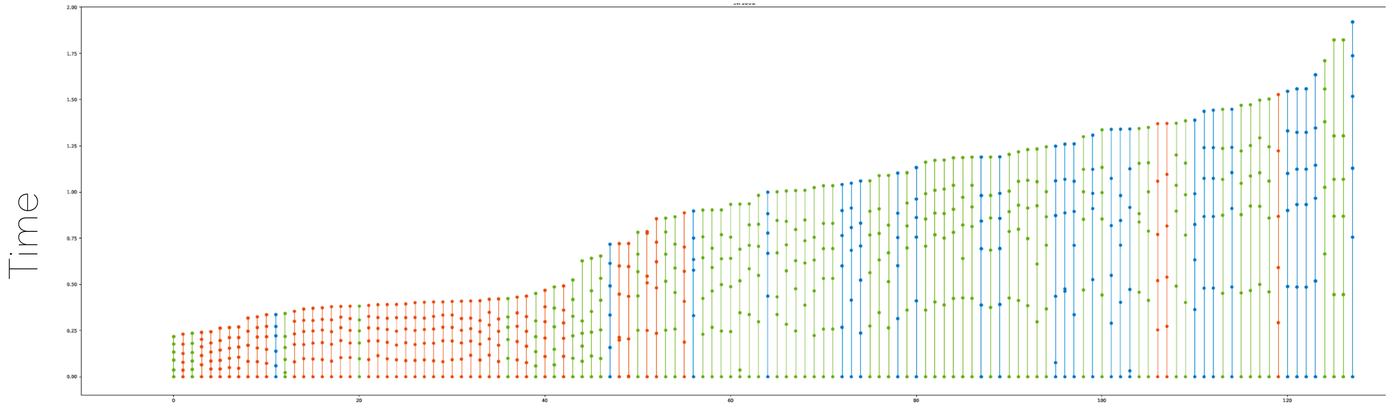
-> Codas on diving descent



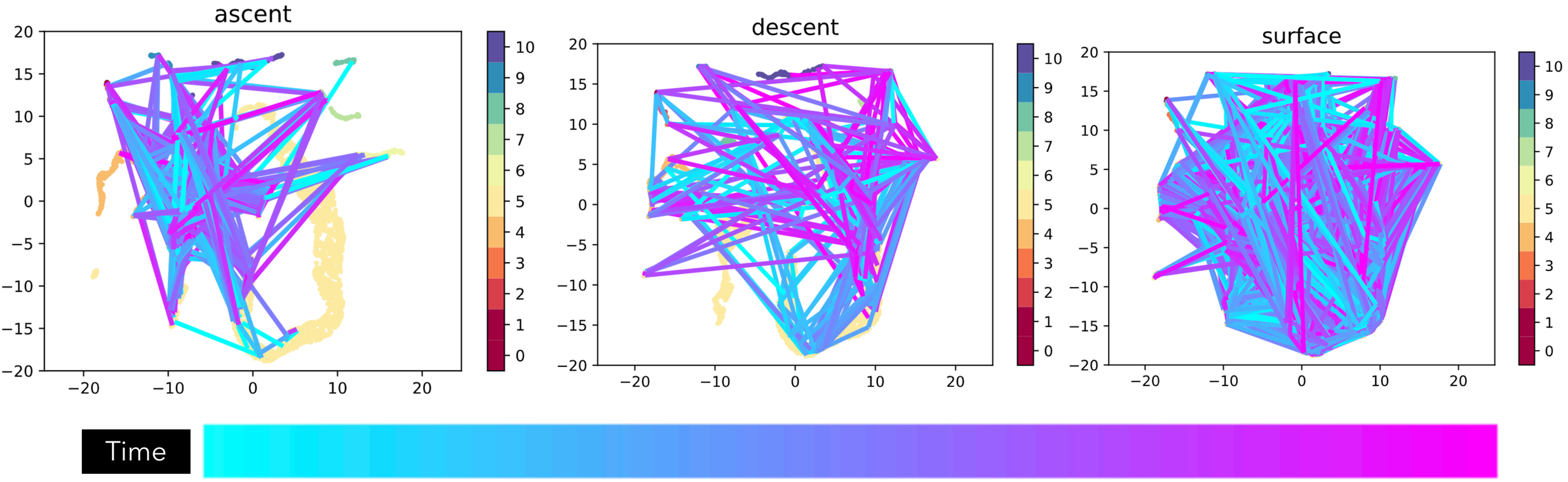
4 click codas



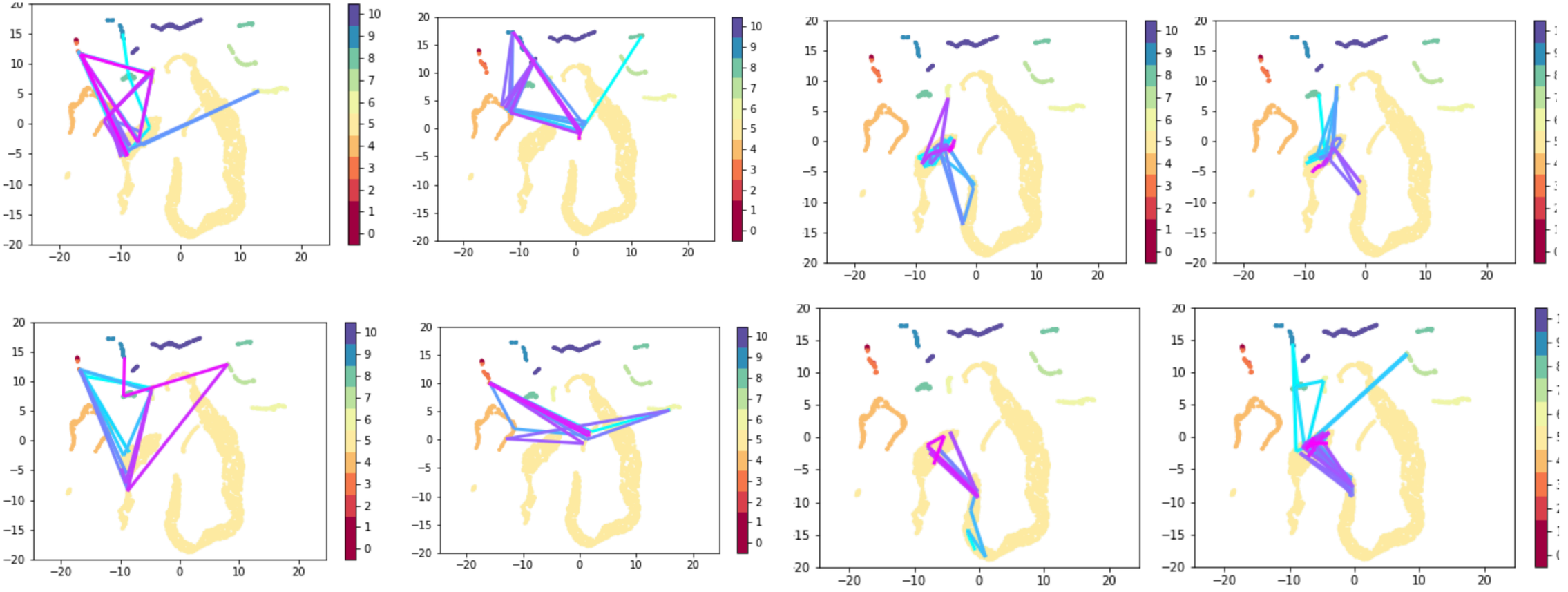
5 click codas



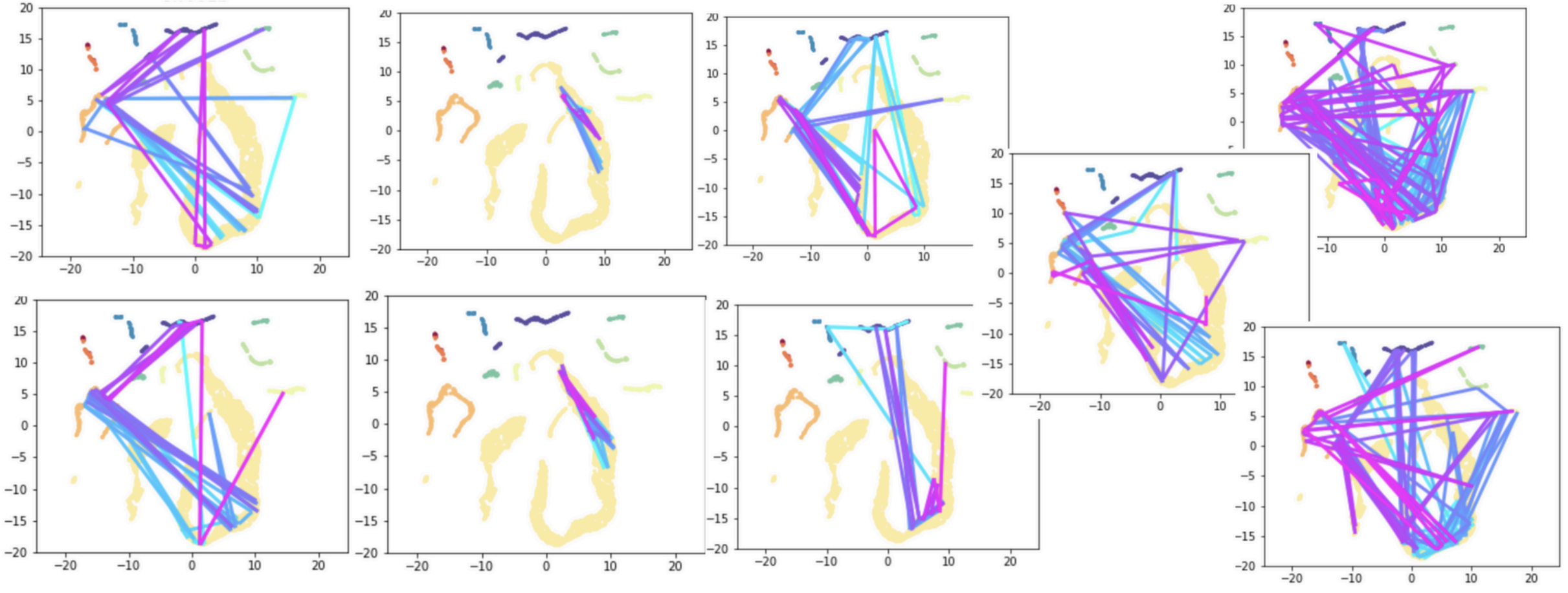
6 click codas


ascent
descent
surface
20
15
10
5
0
−5
−10
−15
−20
−20
−10
0
10
20
0
1
2
3
4
5
6
7
8
9
10
Time

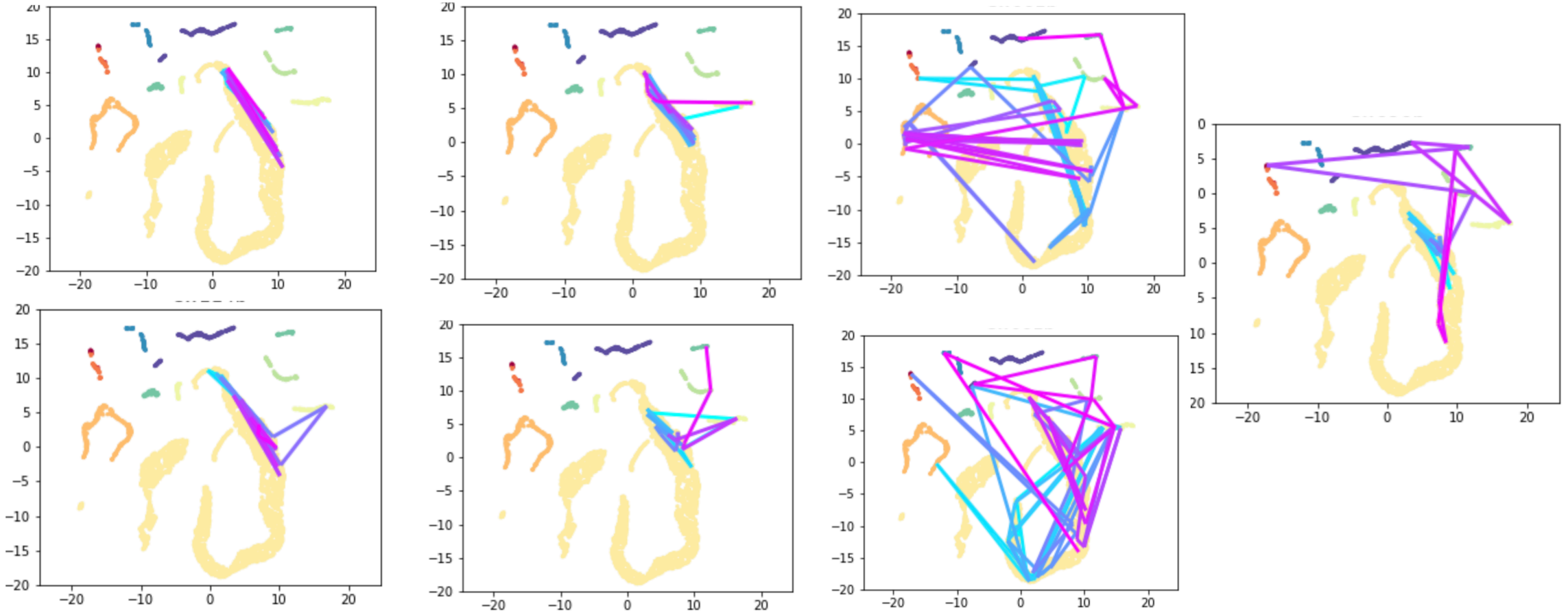
Ascent



Dialogue Surface



Dialogue Descent

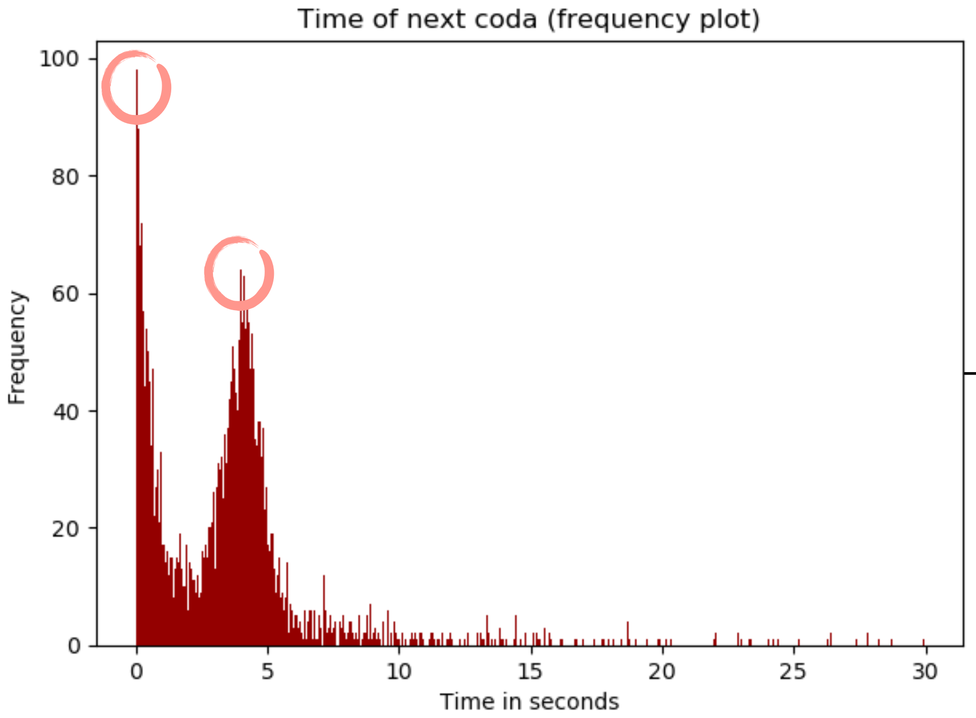


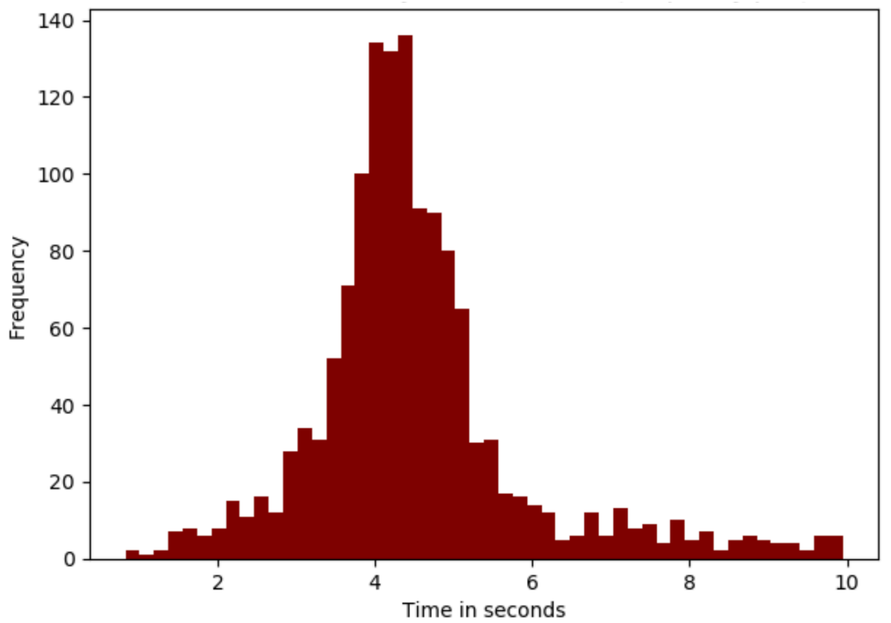
Dialogue and turn-taking

Dialogue

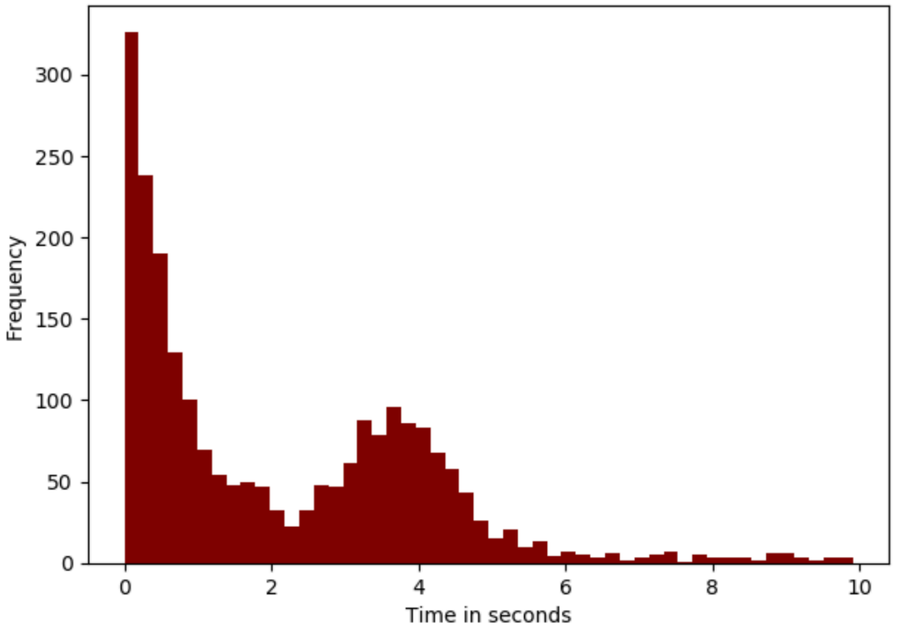
- Same whale diff whale
- At what time is the next coda said
- What is said

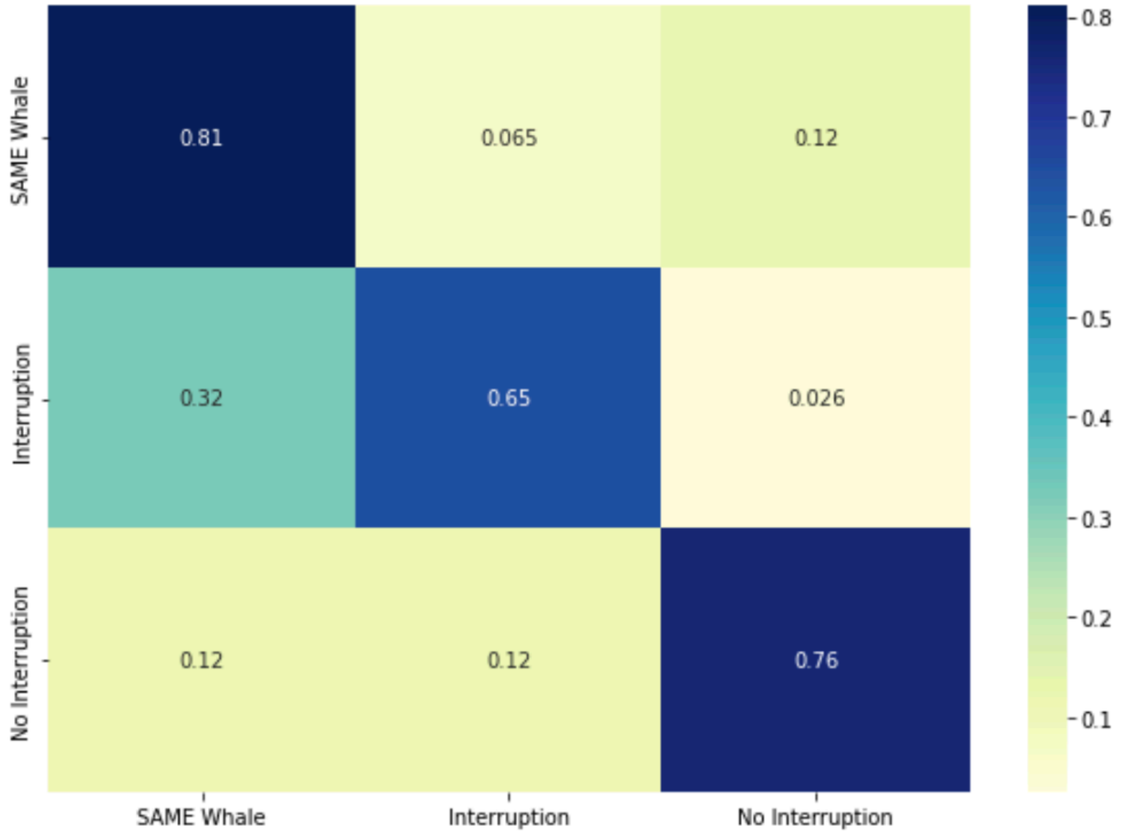
TIME TO NEXT CODA





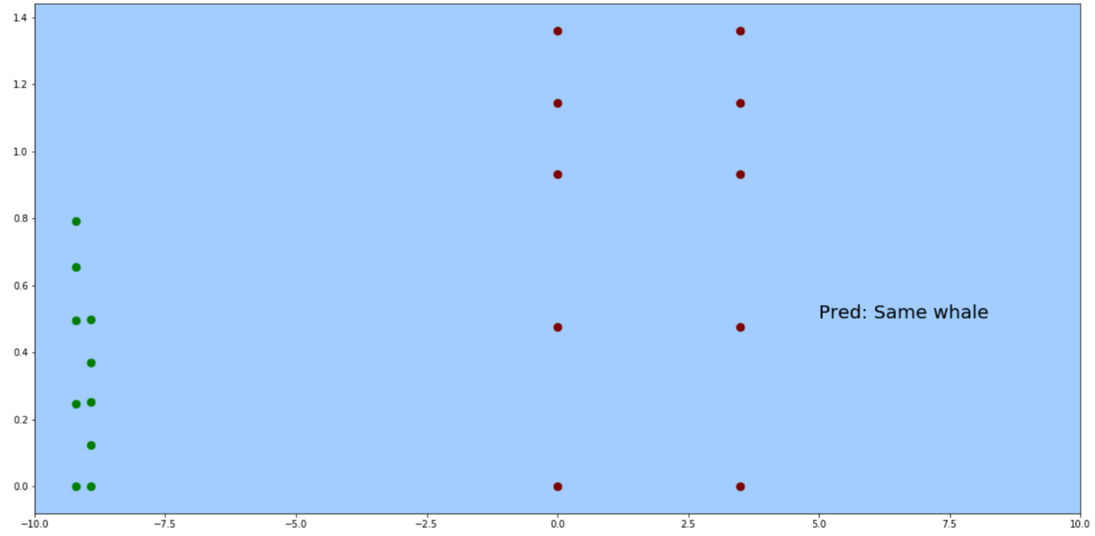
Diff whale

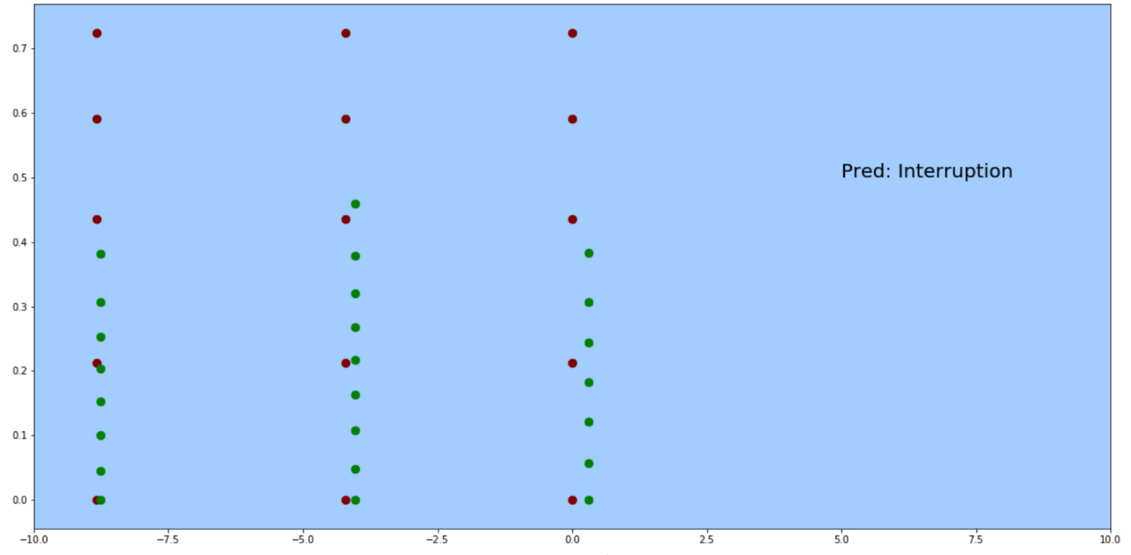


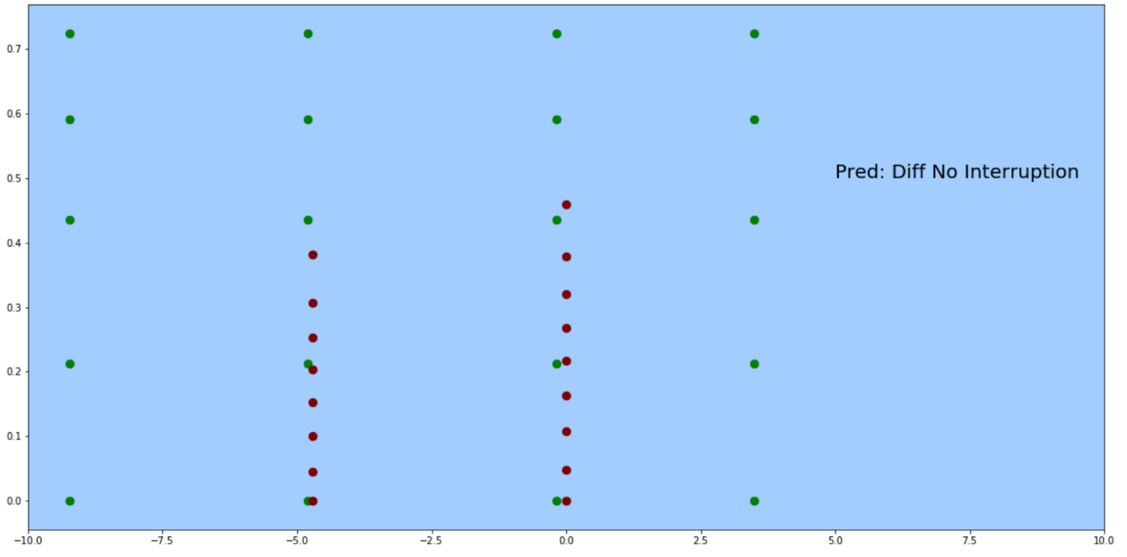


Accuracy: 63.38%

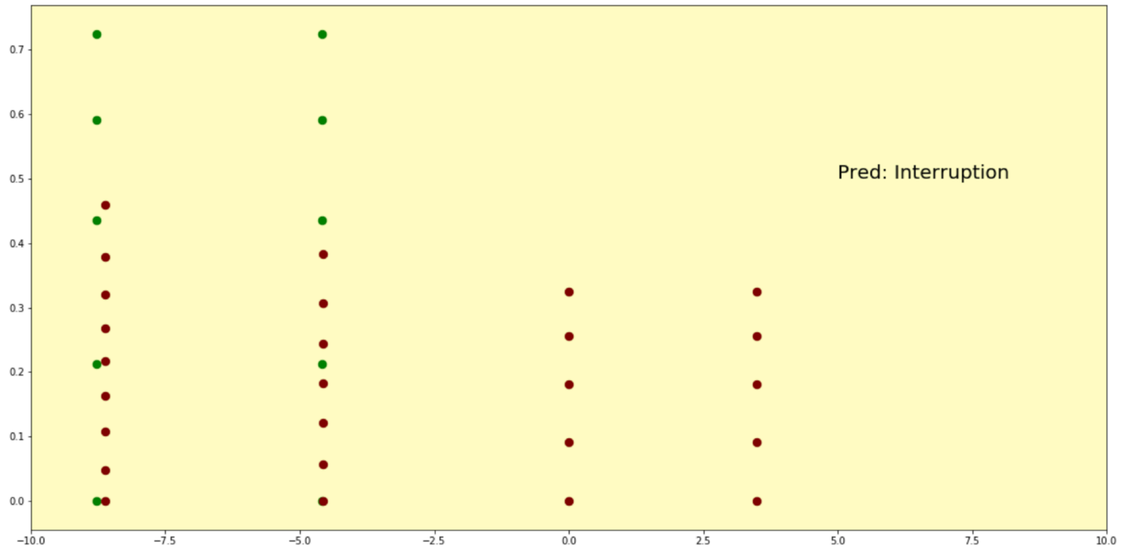
Correct Predictions

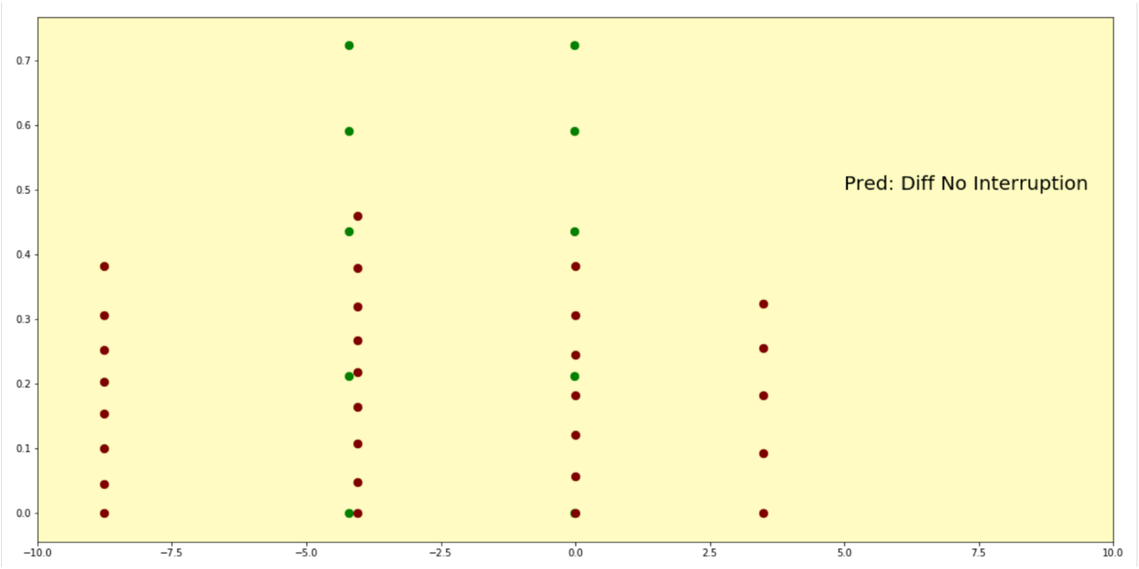


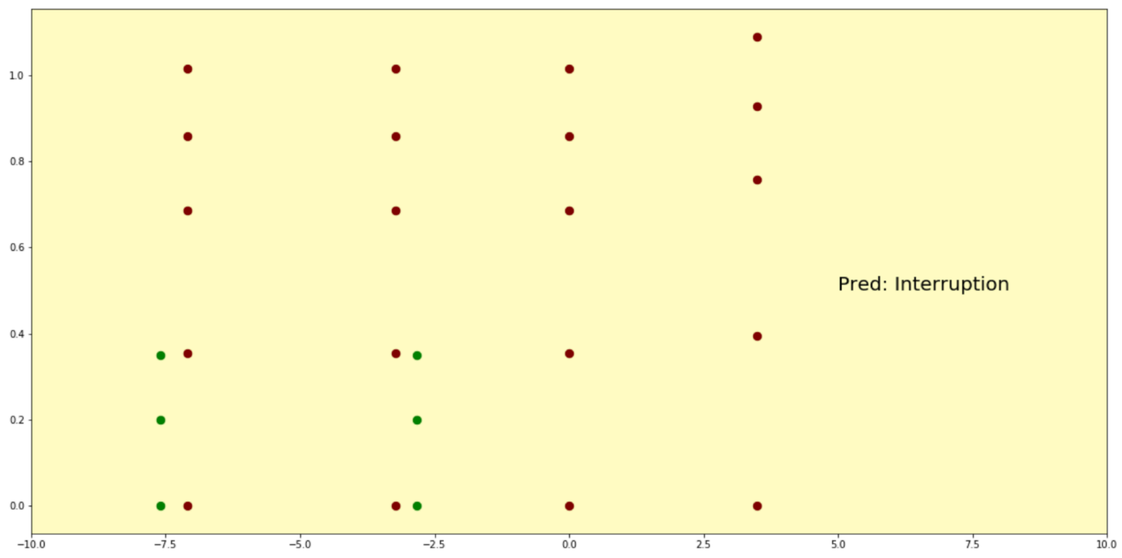




Incorrect Predictions







Summary: Sound <-> Behavior

1. Context Prediction

2. Turn-taking in Dialogue

Next : Increase types of context + controlled experiments

Up Next!

Grammar Induction

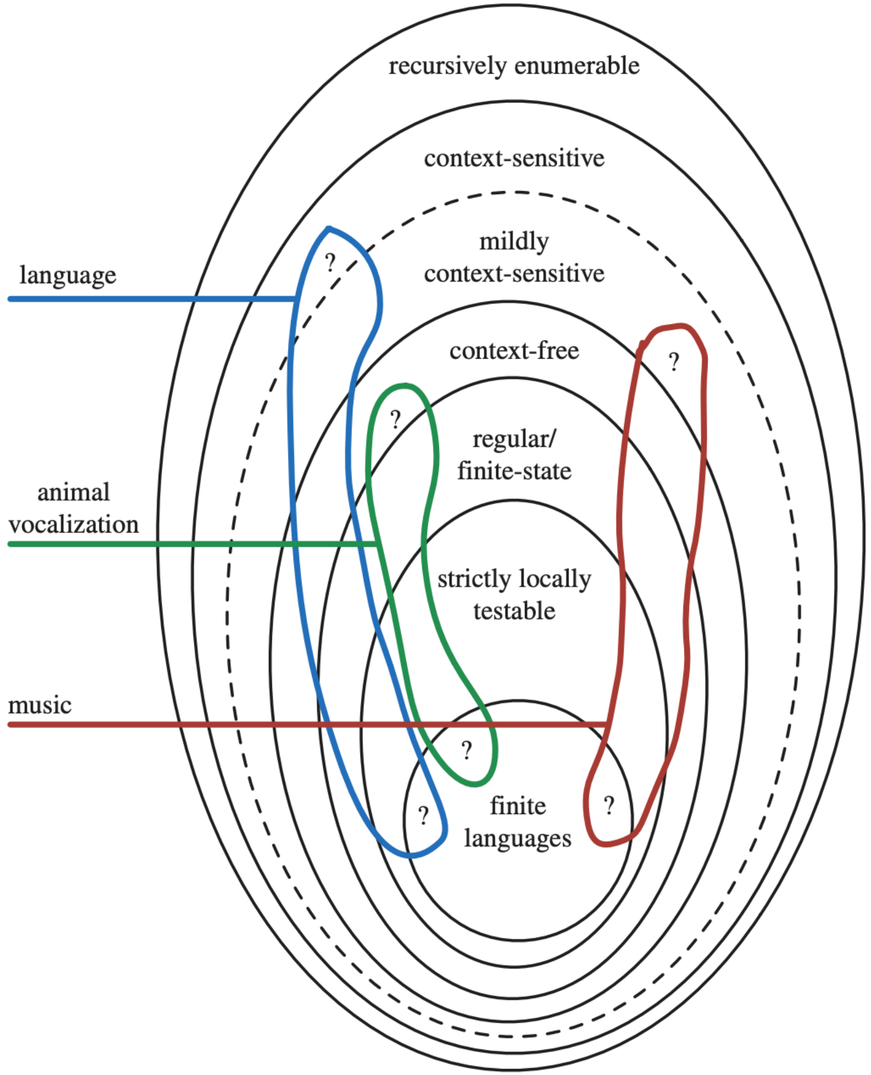
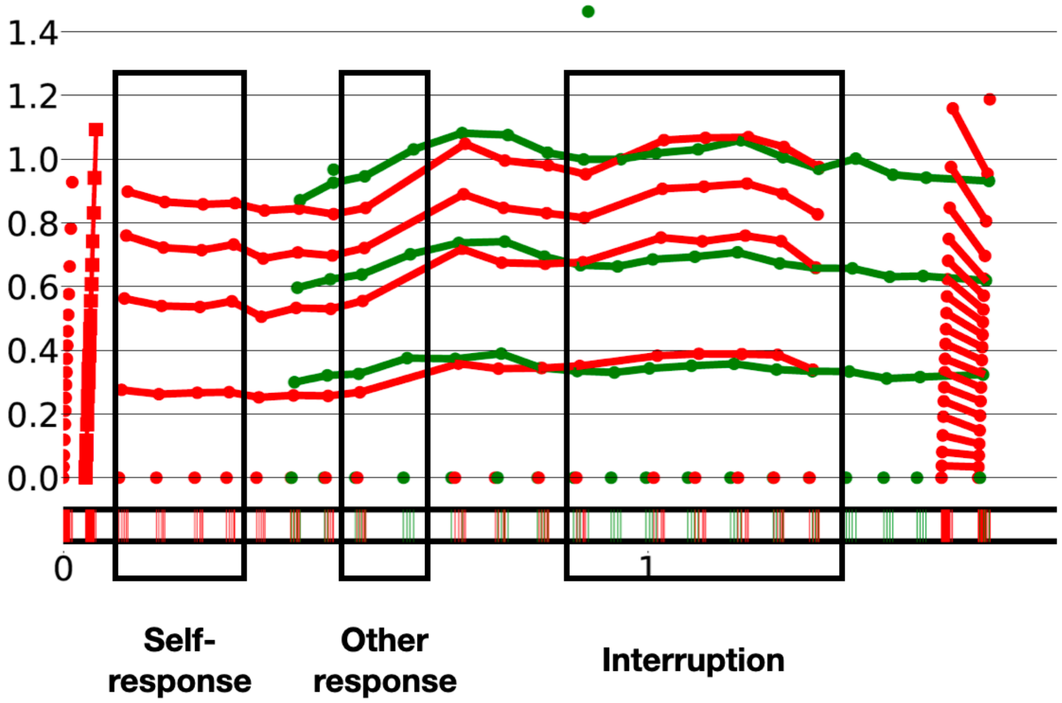
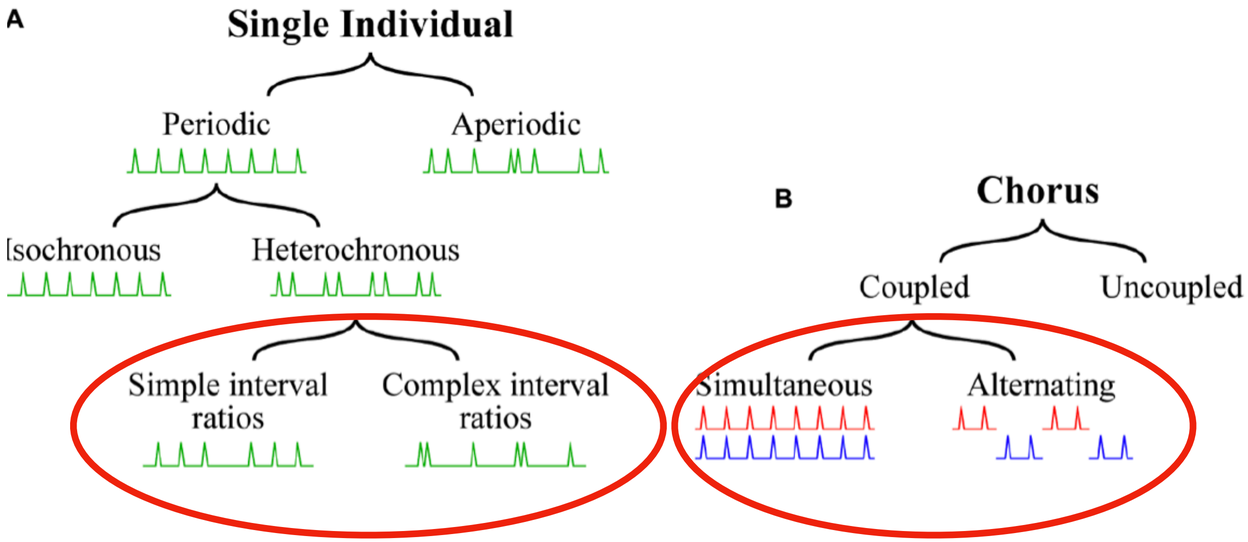


Figure 3. A Venn diagram of the Chomsky hierarchy of formal languages with three extensions annotated with a comparison of the hypothesized classifications of human languages, (human) music and animal vocalization. The areas marked with 'question' signs indicate that further research is required to settle examples for the respective class of complexity in these domains. (Online version in colour.)

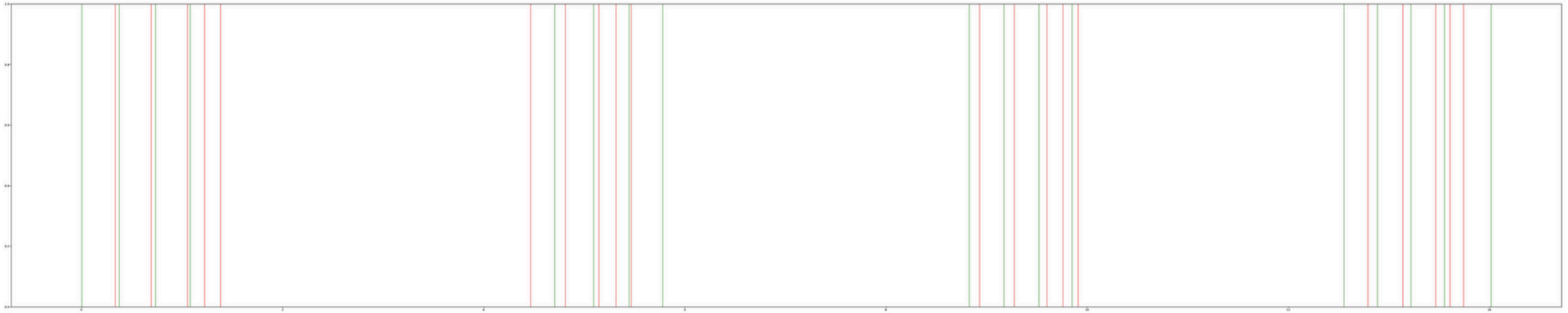
More studies on ties b/w Sound and Behavior

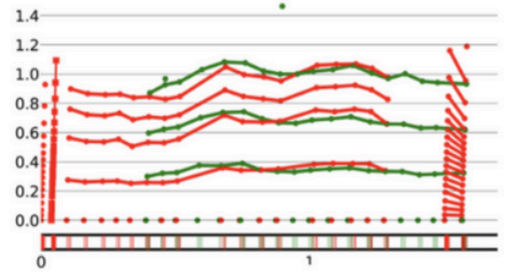
Types of responses

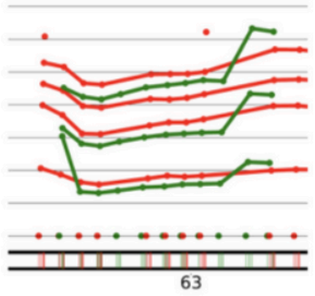


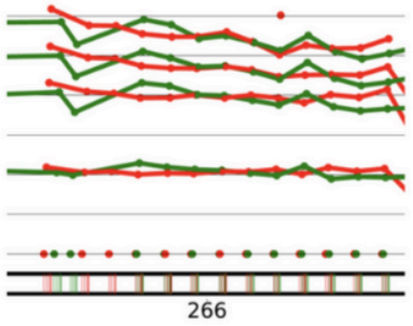


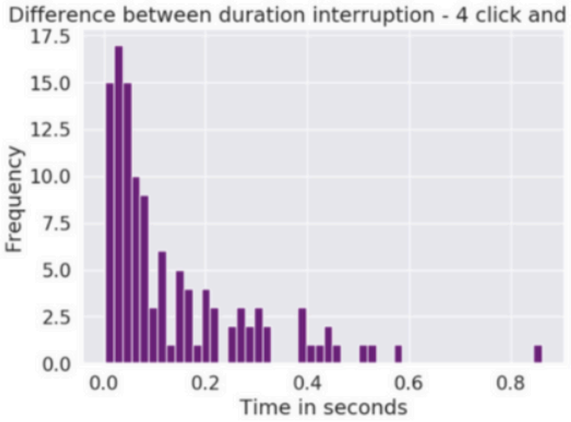
REF: Chorusing, synchrony, and the evolutionary functions of rhythm - Andrea Ravignani , Daniel L. Bowling and W. Tecumseh Fitch

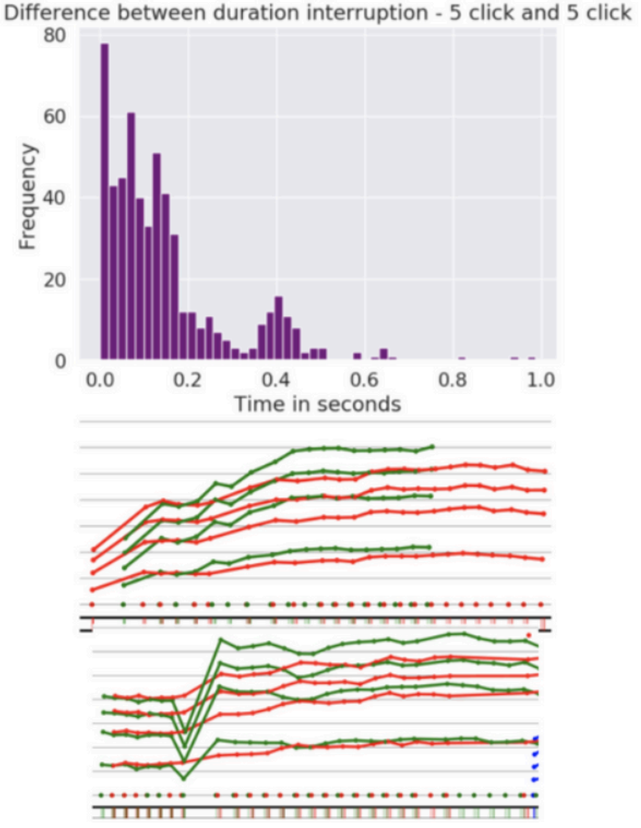












→94% of 4click-5click interruption is from the files recorded on the same day : sw061(98/105)

What can be studied?

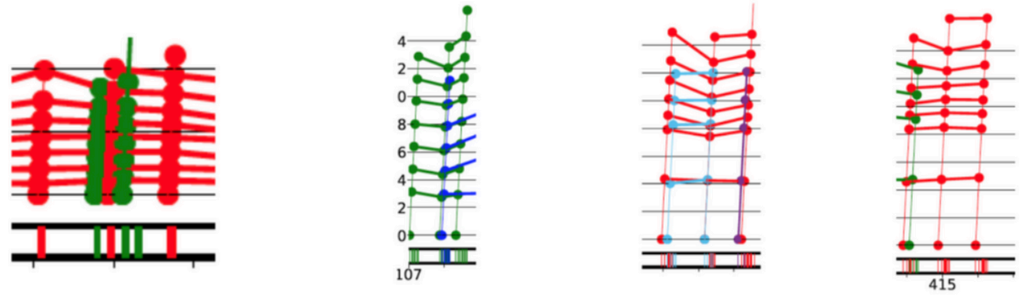
Why do the animals imitate / chorus? - Reasons

What is the protocol of their imitation and chorusing?

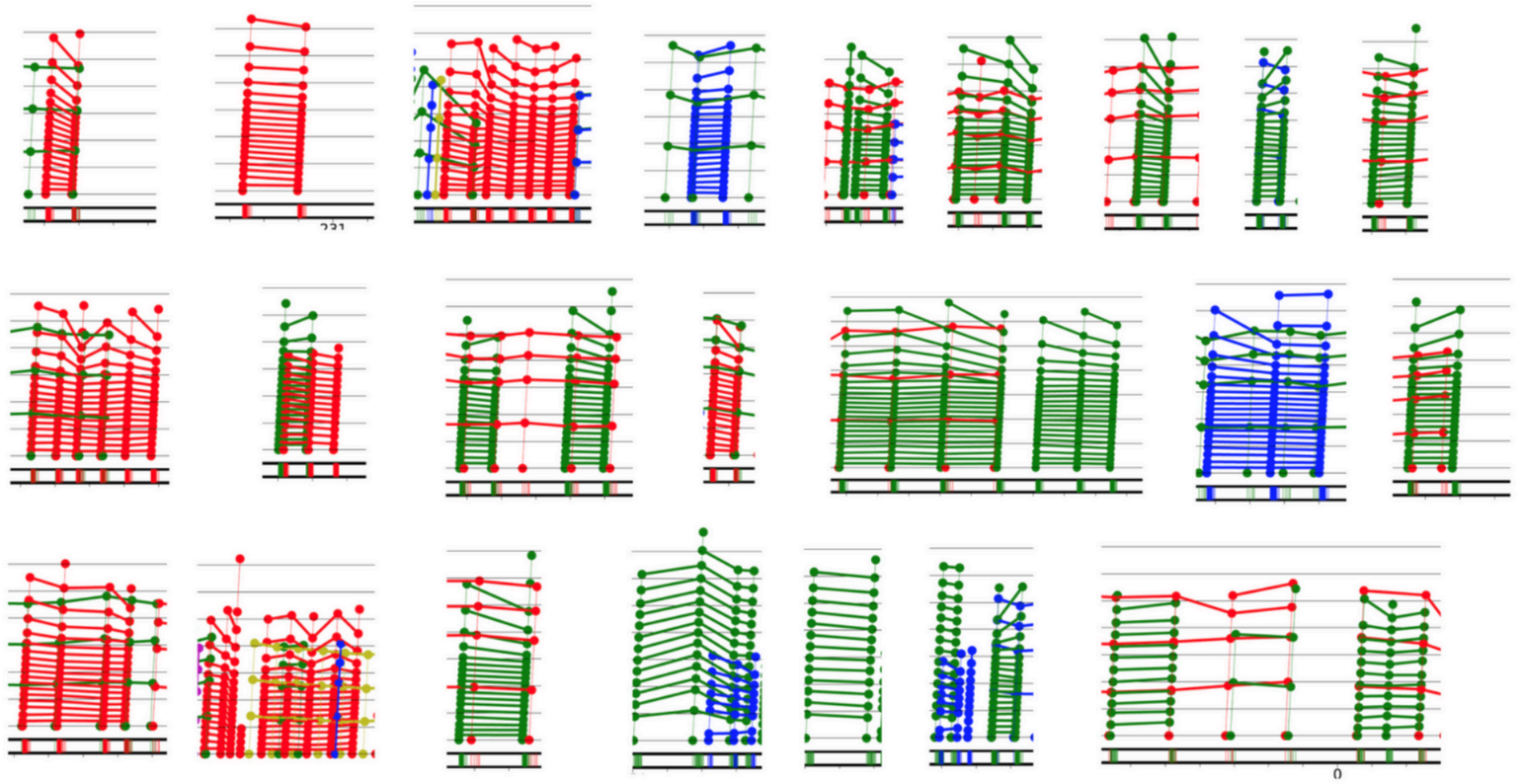
can whales do counting?

Can whales do counting?

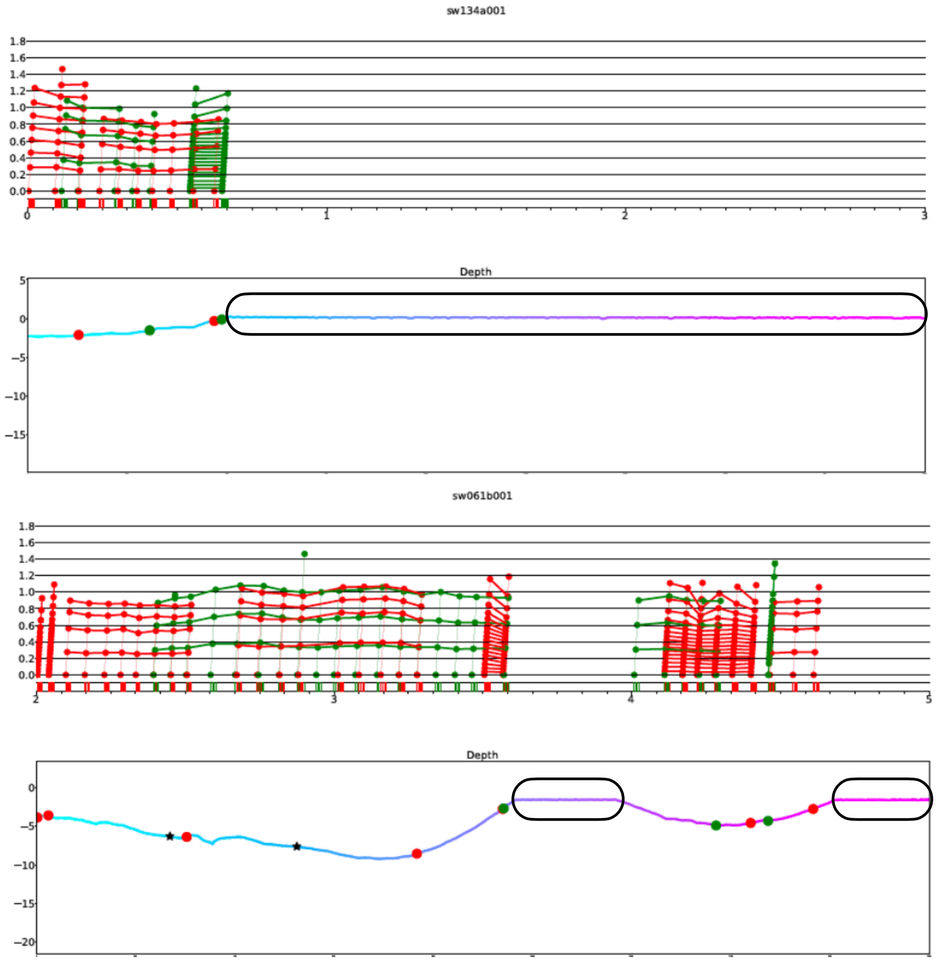
a. Pattern of 7-8-9 → Recurring

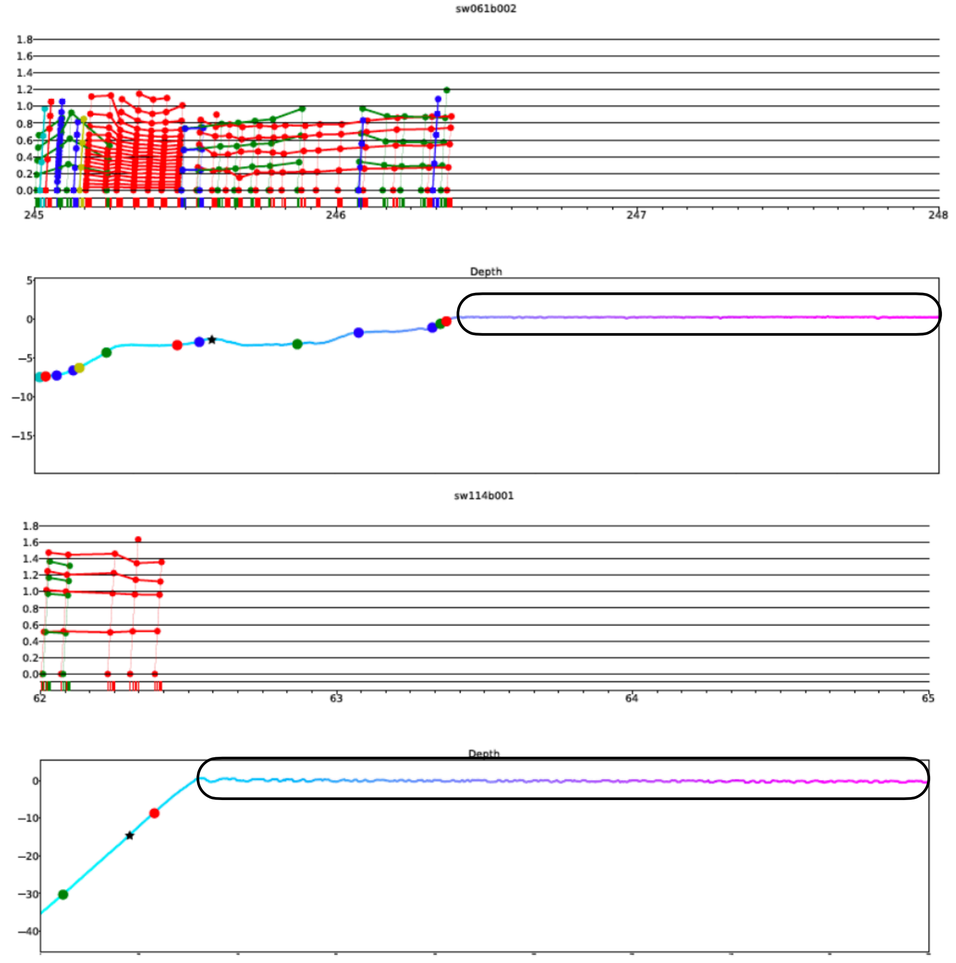


b. Sperm Whales can replicate the number of dots in buzz click codas to the dot to a fairly high degree and only seem to make the extra click in the spaced out end clicks. Furthur, this extra click behaves as the additional click and has an effect on conversation behavior too!

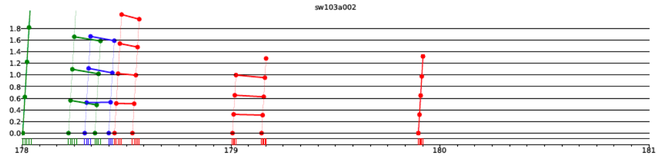


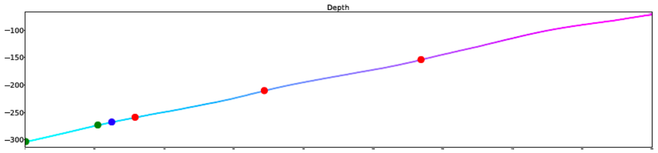
No Conversations at the Surface

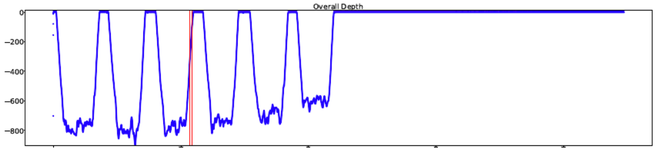


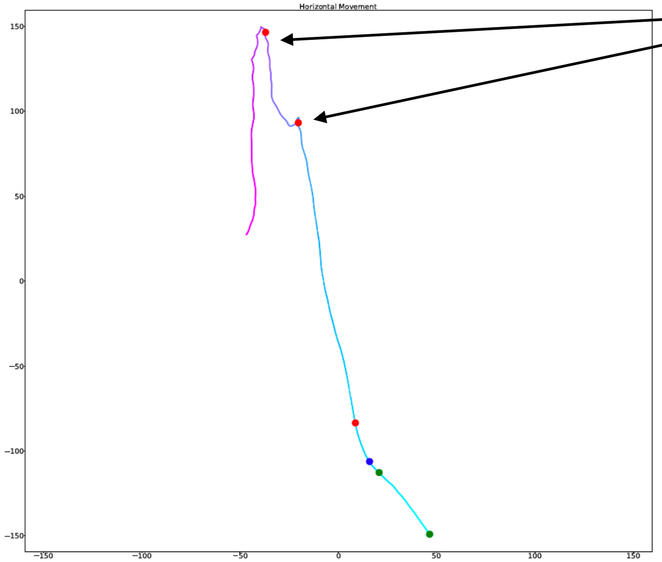


Conversations and Turns

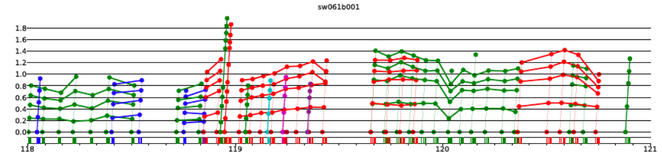


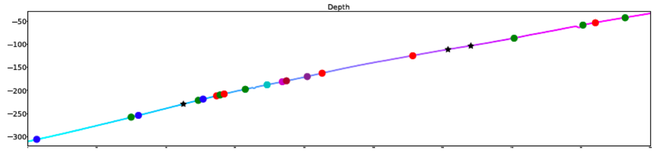


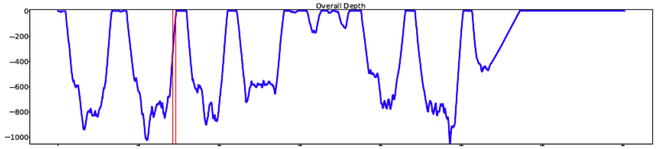


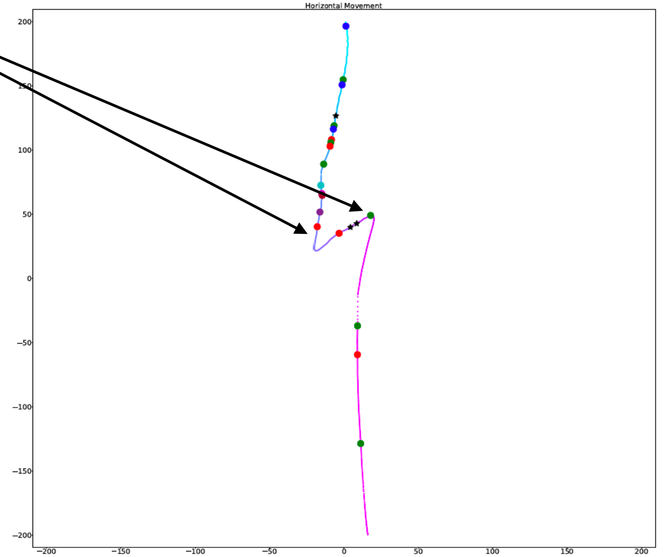


End of speaking -> Turn event









Conclusion

What we have found out so far

- Our visualizations have helped us find patterns of variation within the vocalizations - Imitation of rhythm and interruption (which were earlier treated us repetitions of roughly the same coda by an individual)

- With increase in amount of history as context for prediction the ability of models to predict the next coda improves (Evidence of non Markovian behavior in the vocalizations!)

- "The extra click” is used to initiate change -> Beginning of conversation, end of conversation, switch of turn, sudden turn drop

- We can generate good / highly probable responses to sounds by sperm whales which could help us conduct interventional studies.

Finding Structure in the Sounds

1. Understanding SW Vocalizations : BOW

2. Dialogue: Whale LM (w/ and w/o context) + Protocol

3. What are the smallest units?

4. Deviation from “codas”

    1. Extra click

    2. Patterned interruption

    3. Looking inside a click

Next : Grammar Induction

Automatic Annotation

1. Click detection + Source Separation

2. Behavior annotation

Next : Automatic Behavior Labeling (Increase the context available)

Sound <-> Behavior

1. Context Prediction : Socializing / Diving start / Diving end

2. Predicting turn-taking in dialogue

Next : Increase # of context types + discover reasons for types of chorusing