birdsong in some depth

Feature Space

- **Structure**
  - Simple repeating note, or complex?
  - Long or short phrases?
- **Tonal qualities**
  - Buzzy vs clear whistles, etc.
- **Rhythm**
- **Pitch characteristics**
  - Upswept, downswept, bouncy…
- **Plasticity**
  - How similar are subsequent songs?

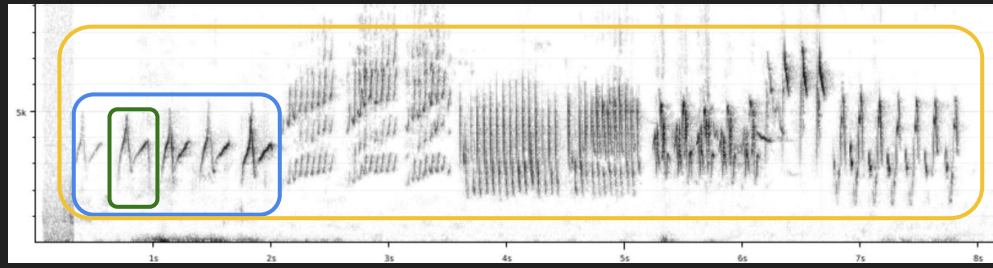
Structure + Plasticity

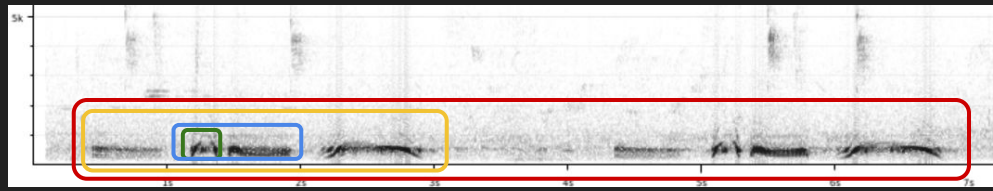
Birdsong is typically divided into units: **notes** (single, separate units of sound), **phrases** (distinct collections of notes), **songs** (a distinct collection of phrases), and **bouts** (session of many vocalizations).

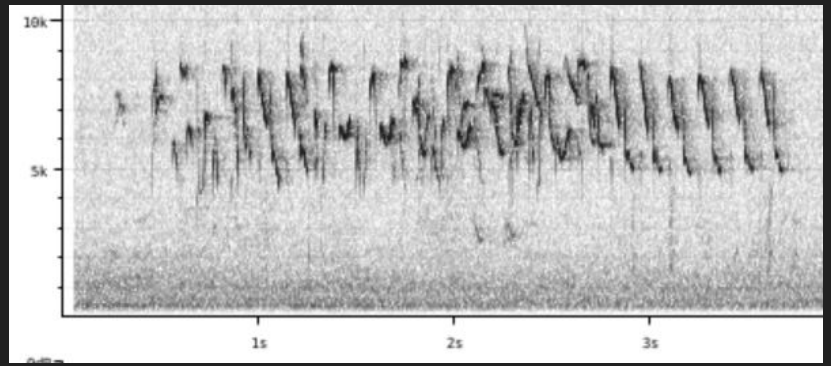
Structures can be simple or complex, at each level!

A phrase might be one note or many.

Plasticity: Subsequent units may be different or repeated.





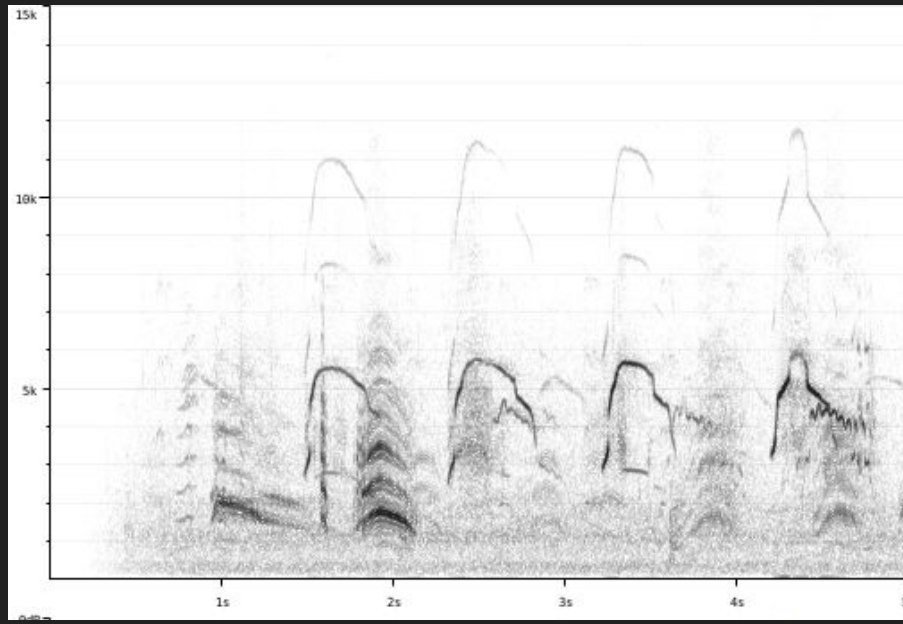


Harmonic Structure

Many sounds have **harmonic** structure:
A 'stack' of the same shape across different frequency bands.

Usually there's a 'fundamental' f,
and then the harmonics are stacked at
f, 2f, 3f, 4f, …

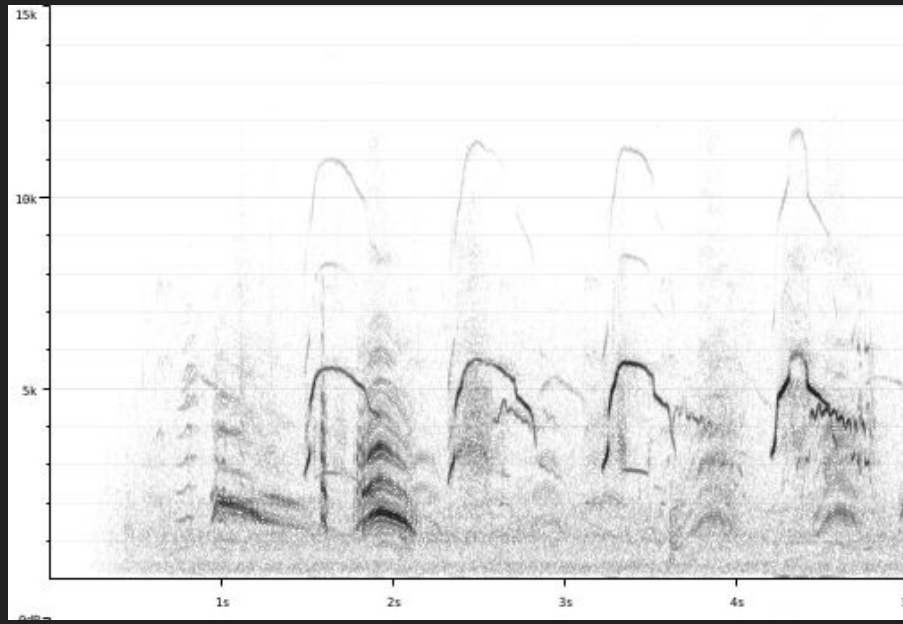
Often the fundamental is the loudest,
but not always!



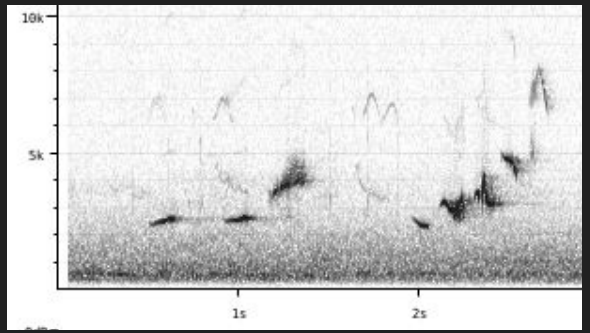
Buzzy notes vs tones

Pure tones have very narrow frequency ranges - they are simple whistles, perhaps with harmonic structure.

Buzzy notes have wider frequency ranges, like the lower-pitched gull in this example.

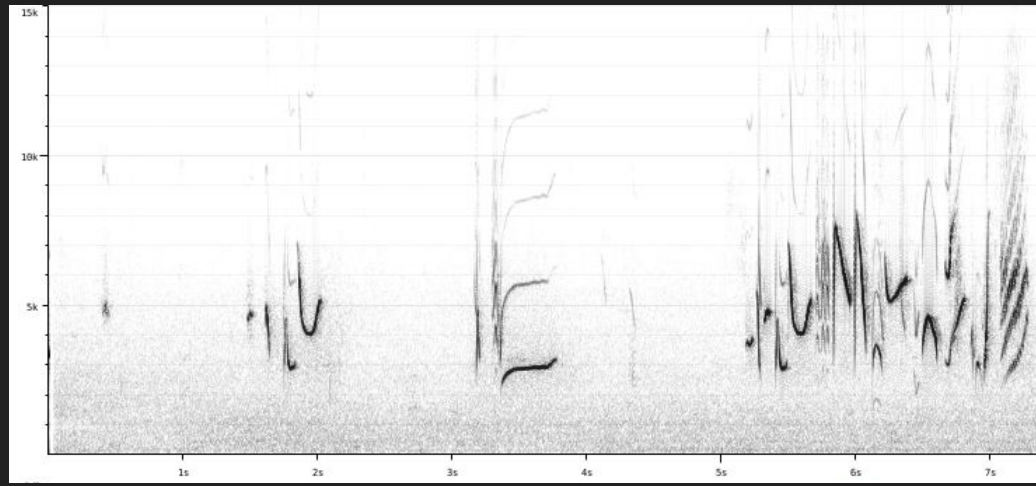


This red-winged starling mixes pure tones and buzzy notes.



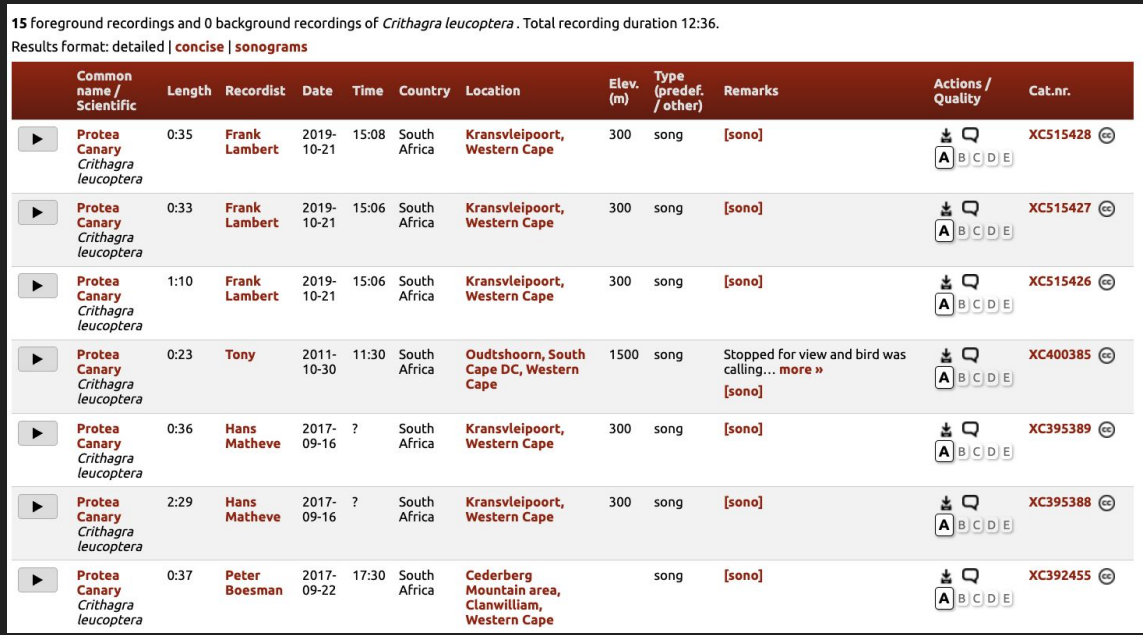
Pitch characteristics

- Overall frequency range utilized by the song (eg, 4-8khz for the fundamental in this Cape Siskin song).
- Each species has some range of frequencies it will vocalize mostly in; some very wide, some narrow.
- Pitch shape: Up-sweep, down-sweep, u-shapes…



How to Learn a Bird's Song

- Read descriptions of the bird on wikipedia and eBird.
- Listen to examples from different days / locations / recordists.
- Look for common features which might help distinguish from other birds.

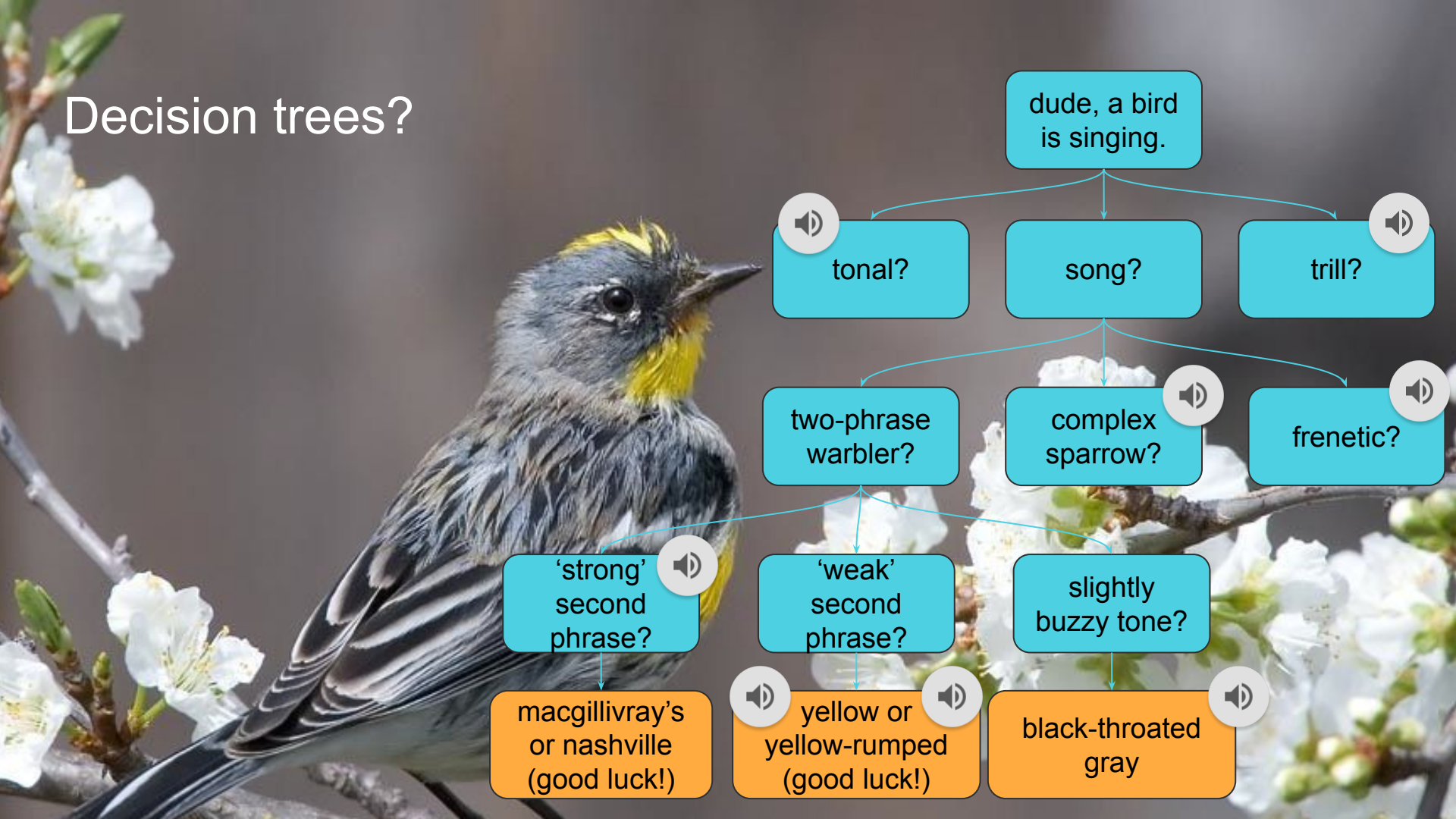


15 foreground recordings and 0 background recordings of *Crithagra leucoptera* . Total recording duration 12:36.

Results format: detailed | concise | sonograms

	Common name / Scientific	Length	Recordist	Date	Time	Country	Location	Elev. (m)	Type (predef. / other)	Remarks	Actions / Quality	Cat.nr.
▶	Protea Canary *Crithagra leucoptera*	0:35	Frank Lambert	2019-10-21	15:08	South Africa	Kransvleipoort, Western Cape	300	song	[sono]	A B C D E	XC515428
▶	Protea Canary *Crithagra leucoptera*	0:33	Frank Lambert	2019-10-21	15:06	South Africa	Kransvleipoort, Western Cape	300	song	[sono]	A B C D E	XC515427
▶	Protea Canary *Crithagra leucoptera*	1:10	Frank Lambert	2019-10-21	15:06	South Africa	Kransvleipoort, Western Cape	300	song	[sono]	A B C D E	XC515426
▶	Protea Canary *Crithagra leucoptera*	0:23	Tony	2011-10-30	11:30	South Africa	Oudtshoorn, South Cape DC, Western Cape	1500	song	Stopped for view and bird was calling… more » [sono]	A B C D E	XC400385
▶	Protea Canary *Crithagra leucoptera*	0:36	Hans Matheve	2017-09-16	?	South Africa	Kransvleipoort, Western Cape	300	song	[sono]	A B C D E	XC395389
▶	Protea Canary *Crithagra leucoptera*	2:29	Hans Matheve	2017-09-16	?	South Africa	Kransvleipoort, Western Cape	300	song	[sono]	A B C D E	XC395388
▶	Protea Canary *Crithagra leucoptera*	0:37	Peter Boesman	2017-09-22	17:30	South Africa	Cederberg Mountain area, Clanwilliam, Western Cape		song	[sono]	A B C D E	XC392455

Decision trees?



Difficulties for humans

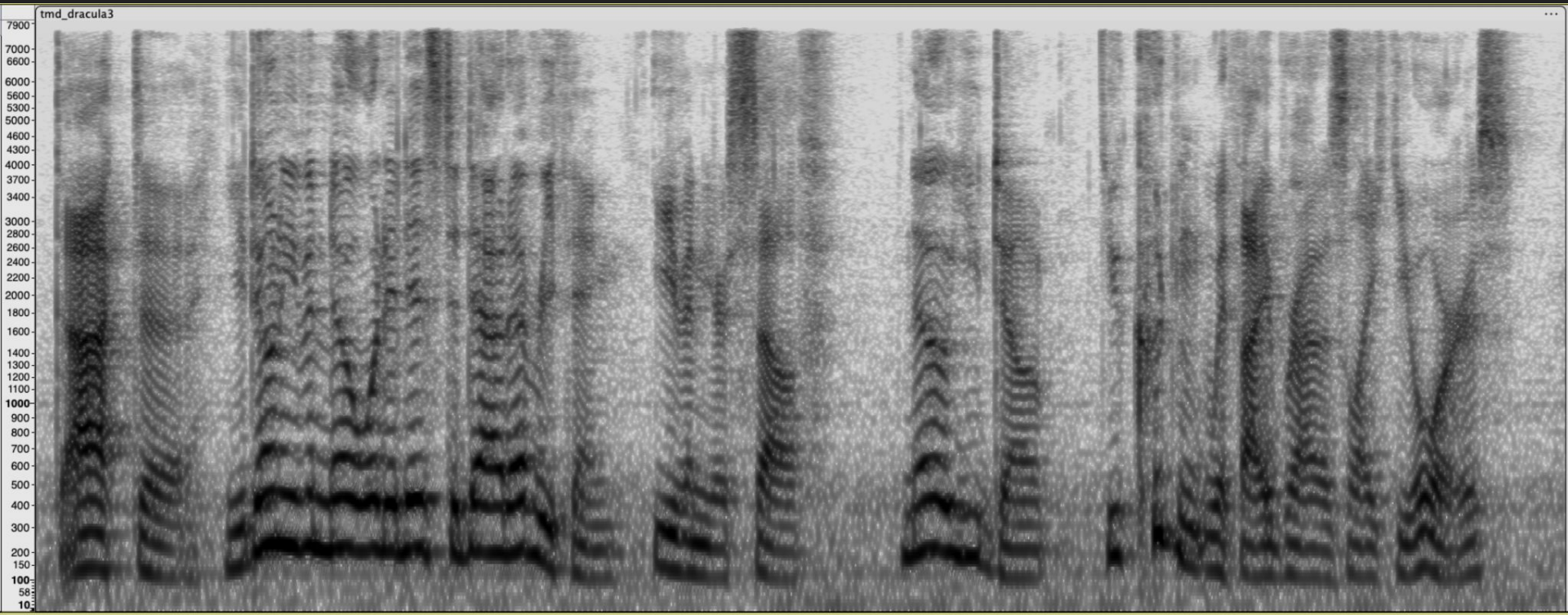
- Relative measures fail in the field, since you can’t directly compare. (eg, ‘junco trills are slower than chipping sparrow’s’)
- ‘Foreign’ sound features, produced by syrinx.
- Time-resolution of the human ear is too low.
- Lifetime of practice ignoring these sounds.

Difficulties for Machines

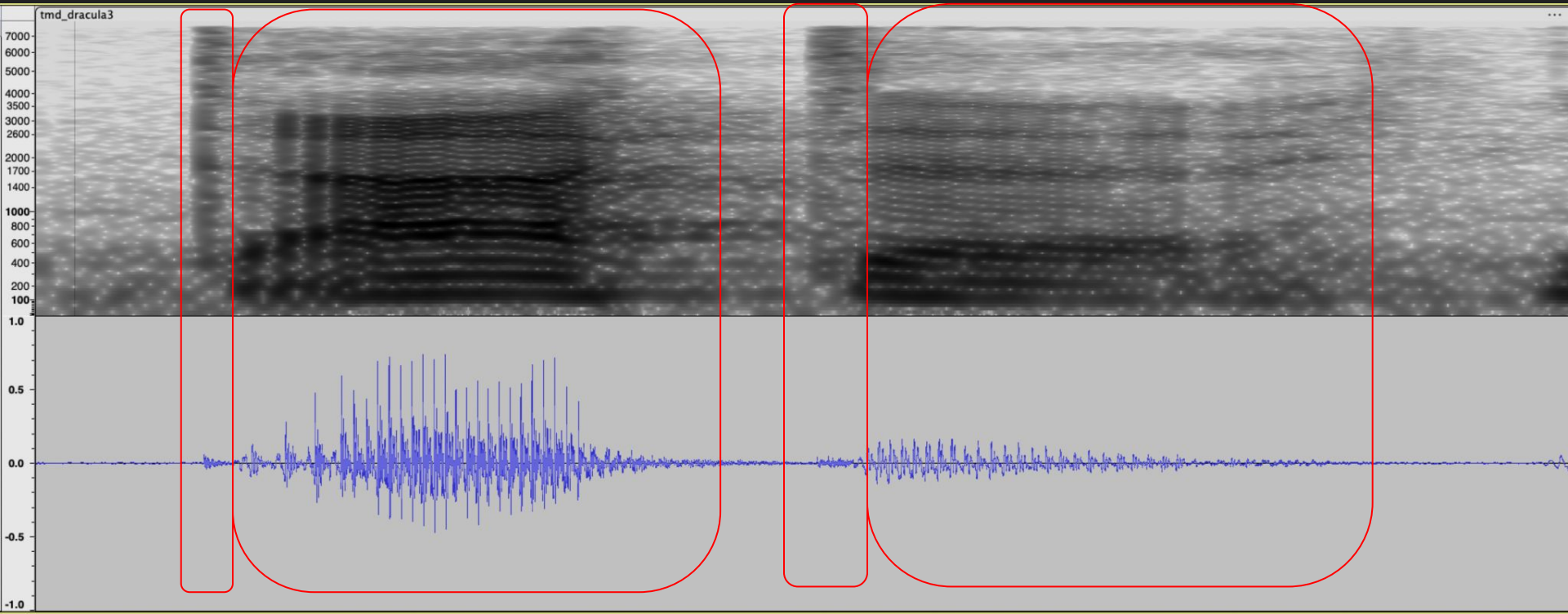
- Hard to get ‘clean’ ground truth:
  Unlike voice applications, there are few-or-no trustworthy studio recordings.
- Lots of variation in the wild!
- Large databases are ‘weakly’ labeled:
  Possibly additional background songs, or other background noises.
- Databases have uneven coverage:
  Many local song variants not represented.
- **?!**

an aside on fourier transforms

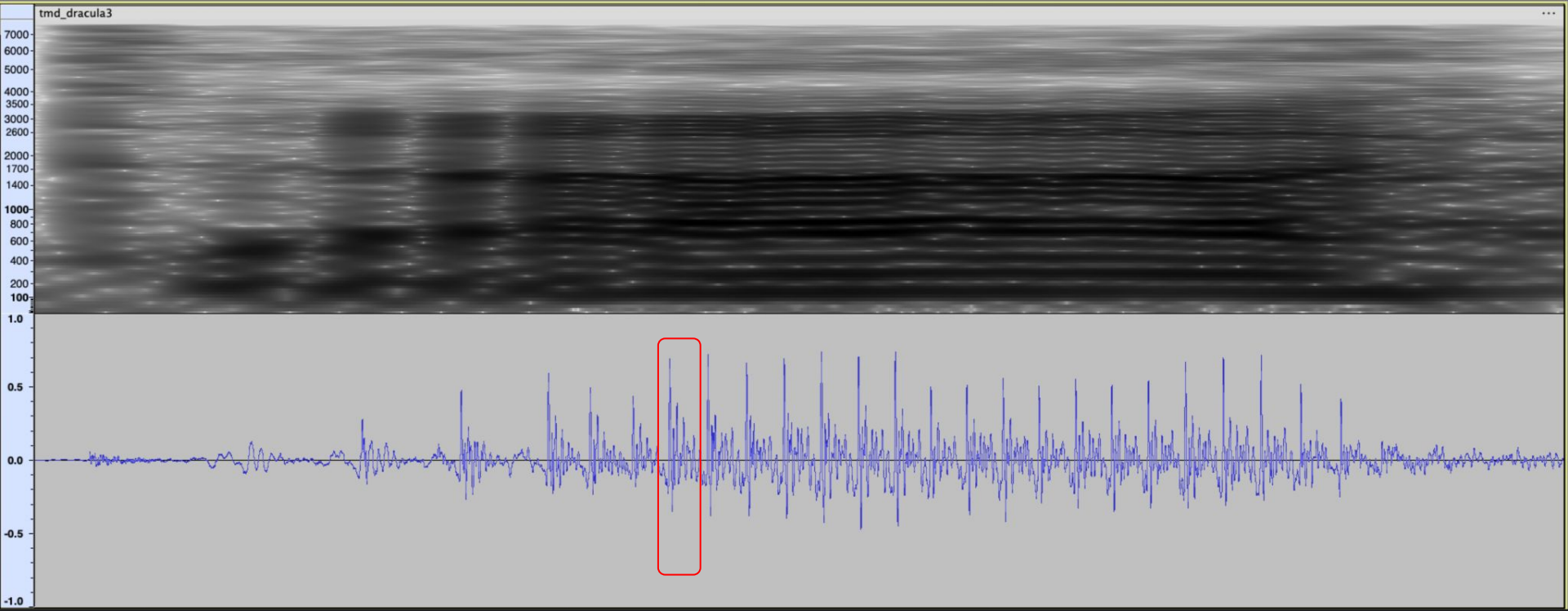
speech spectrogram example



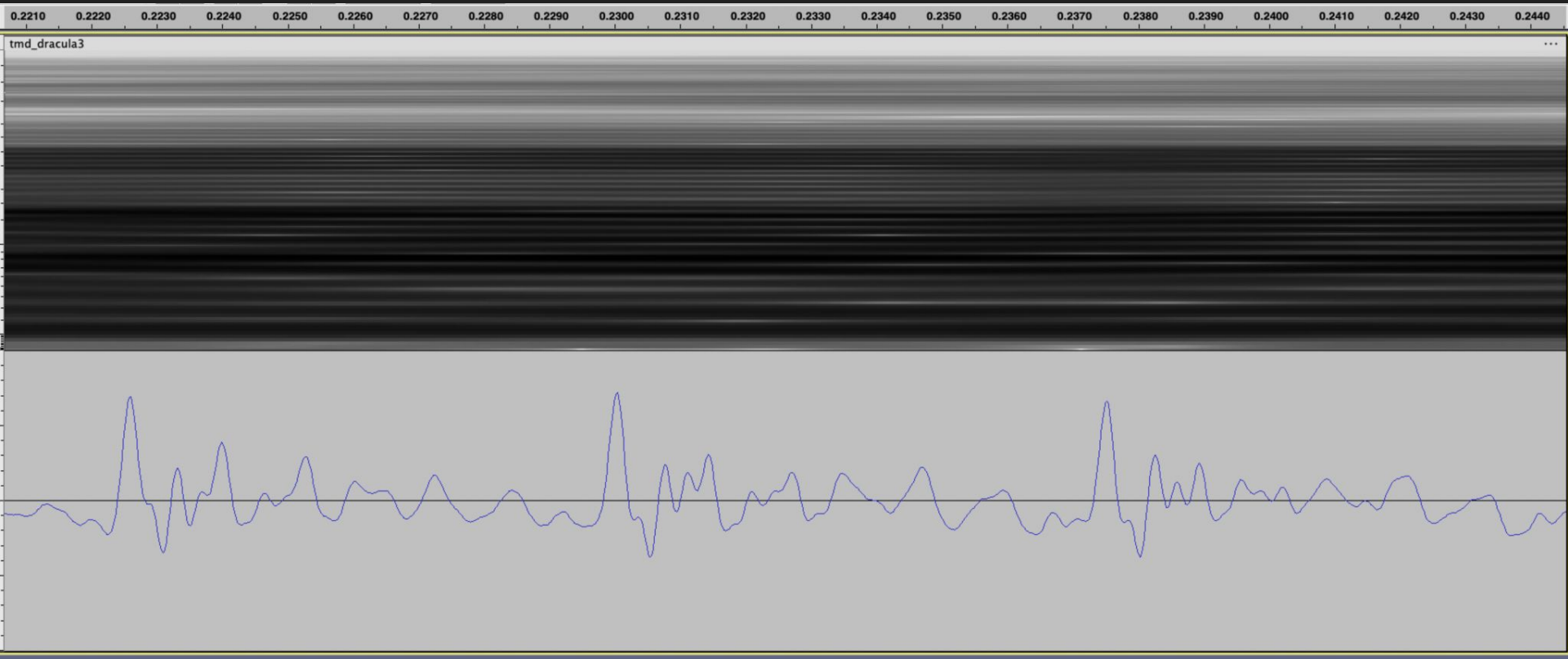
speech spectrogram example: 1s



speech spectrogram example: "dah"



speech spectrogram example: "locally periodic"



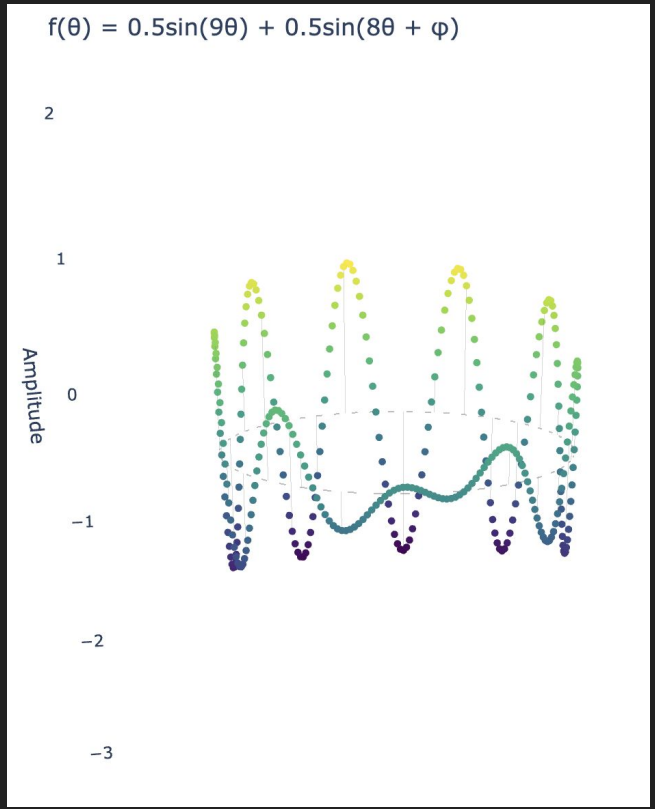
Periodic Functions

Let f: S -> C be a periodic function:
A function on the circle, giving a 'height' at each point on the circle.

The Fourier Transform is a change of basis, using specific **orthonormal basis functions**:

$$x_n = e^{i*n*t} / 2\pi$$

Then $\mathbf{f_n = \int f(t)*x_n(t)}$ gives the n^{th} component.



Discrete Periodic Functions

I find it easier to think about the discrete case!
(but I'm a combinatorialist at heart, ymmv.)

In this case, we have functions Z_N->R.
(If N is big, it starts looking like the continuous case.)
These *functions* can be written as *arrays*!

One basis for these functions is the elementary basis: [0, 0, …, 0, 1, 0, …, 0].
This is how we represent the time domain:
Sample by sample.

The Fourier Transform is a change of basis, using specific **orthonormal basis functions** specifically chosen to measure periodicity.

Discrete Fourier transform

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i 2\pi \frac{k}{N} n} \quad \textbf{(Eq.1)}$$

The raw DFT is complex-valued, so each frequency component has two parts: **magnitude** and **phase**.

When making a visual spectrogram,
we keep only the magnitude, and discard phase.

You can use the Griffin-Lim algorithm to try to reconstruct phase from a spectrogram.
…Or WaveNet!

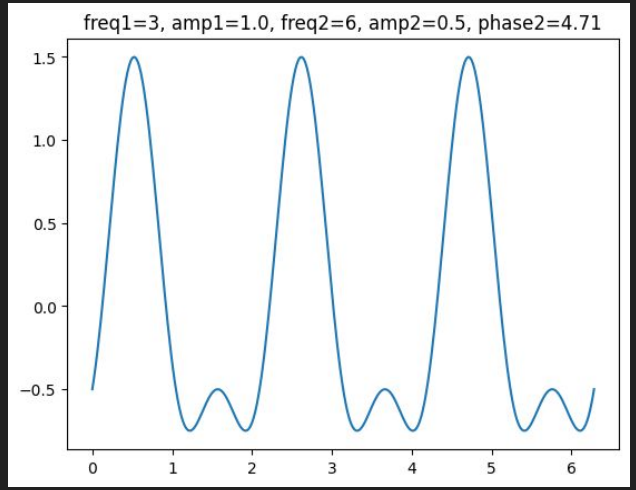
Sinusoidal functions are a basis for periodic functions.

Similar to a Taylor series…

You can express any* periodic function as a **linear combination of sinusoids** (with different frequencies, amplitudes, and phases).

The sinusoidal basis functions are 'spread out' in the time domain, but 'local' in the frequency domain.

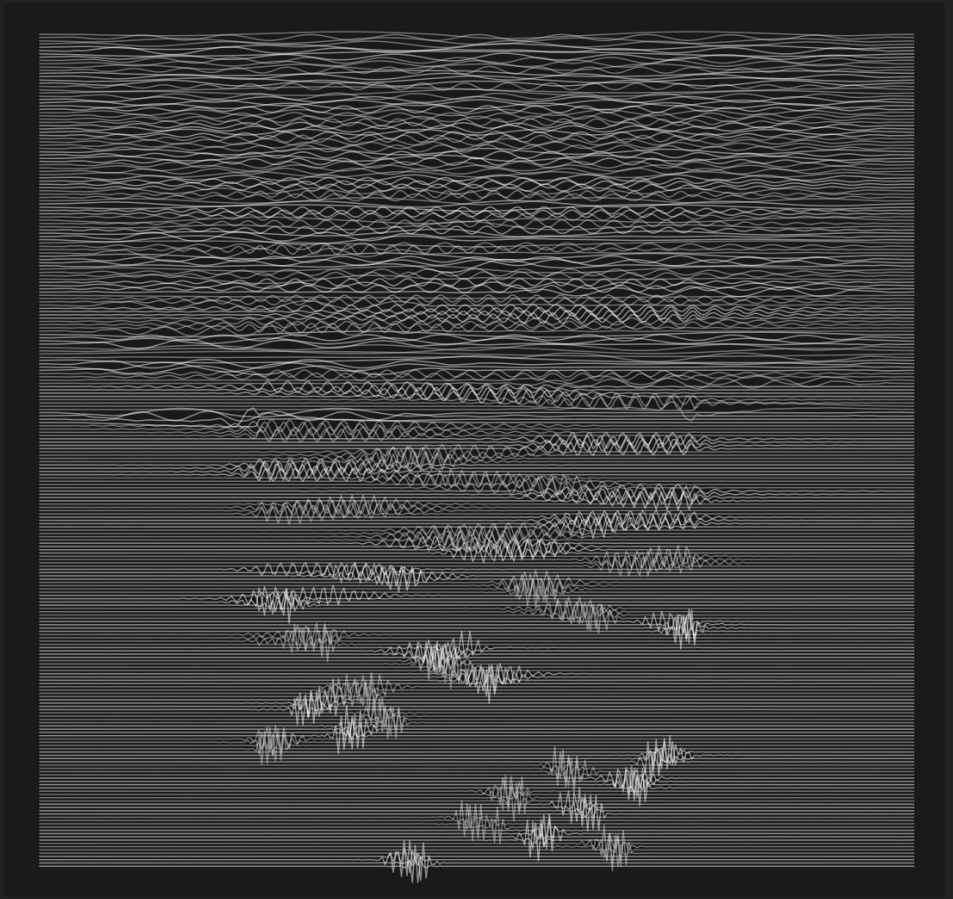
Likewise, the sample basis is 'local' in the time domain, but 'spread out' in the frequency domain.



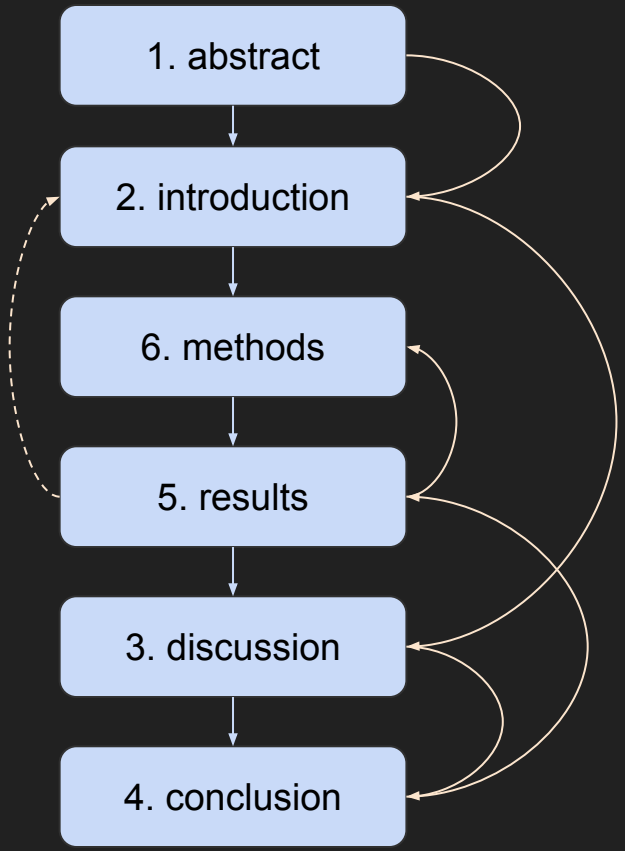
…but the fft isn't the only way to represent audio.

Another major approach is wavelets, which use variable length audio windows to try to localize shorter higher-frequency "chirps".

You can also use a big convolution and try to learn good basis functions directly, like the set pictured at right. These learned functions are very similar to wavelets!



how to read a paper



The **abstract** gives the highest-level view of the paper: Helps decide whether you should read the rest!

The **introduction** is a longer form of the abstract: Why the paper matters and what's new/novel.

Then go to the **discussion and conclusion**, which summarize what we can learn from the experiments.

Then read the **results**: check they match the discussion, and see if there's other things you can observe.

Finally, the **methods** give the gory details of how the experiments were executed.

where we are going

Two Kinds of Problem

- **Broad Monitoring**
  - Biodiversity measures are important, and hard to detect by satellite.
  - Need broad indicators of **species diversity**…
  - So we should develop **multi-species classifiers**.
  - Maximize **precision** then **recall**.

- **Species Monitoring**
  - Identify and track **specific endangered** or **invasive** species.
  - Training data may be very scarce.
  - Often care about fine detail:
    eg, nesting calls vs flight calls, juvenile calls, etc.
  - Maximize **recall** then **precision**.

Problem:
(Pre-Trained) Classification doesn't scale.

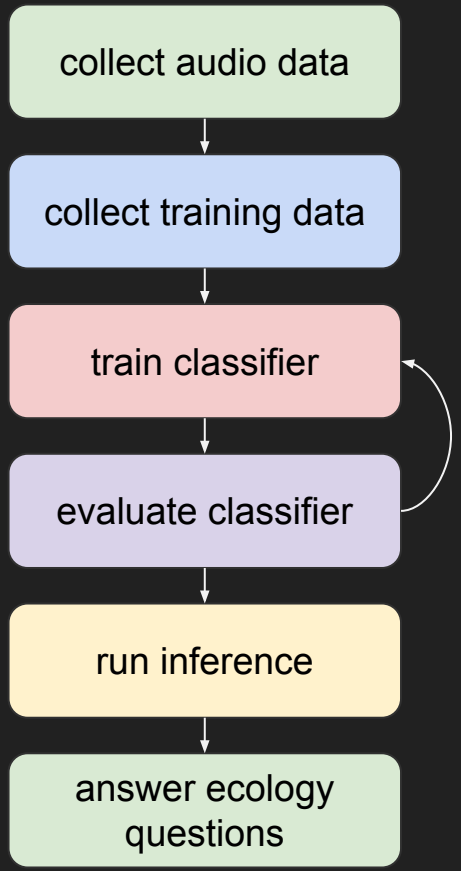
- Need to **re-train per use-case**.
- NorthAm/Euro models fail in **rainforests**.
- **Annotation is too onerous**.
- **Poorer countries** tend to have more species richness and fewer experts!
- Multispecies classifiers **over-index** on high-data species.

- Needs close involvement of **ML experts**.

How can we **build a system**
to efficiently **answer questions**
which have **never been asked before**?

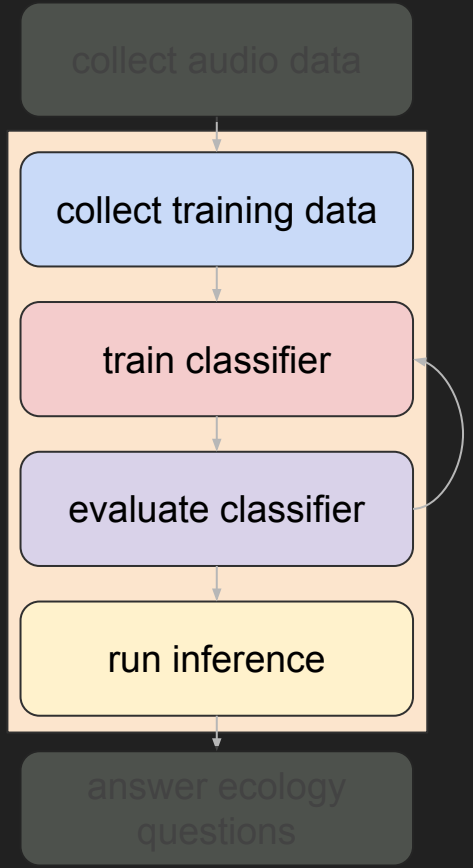
Passive Acoustic Analysis Workflow

- Collection of 1000’s of hours of audio (or more!)
- Limited ecology expert time.
  - Ecologists/conservationists can label some data.
  - Massively underfunded compared to, eg, human health.
- Very limited machine learning expert availability.
  - Possibly a student on staff, possibly an overworked external collaborator.



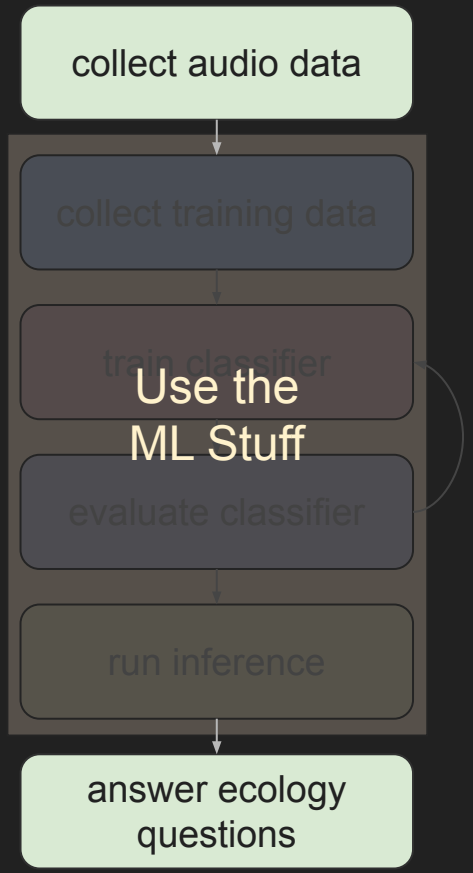
Agile Modeling

- Use a **robust pre-trained model** to **embed** target data.
- Use embeddings to **search** for relevant examples.
- Train a **small classifier** on relevant examples.
- Enable efficient **active learning**:
  Make it easy to **evaluate** and **iterate**.
- Make the system **simple enough for anyone** to use.
  - No waiting to iterate on model quality.
  - Remove need to pass data and models between ecologists and ML experts.

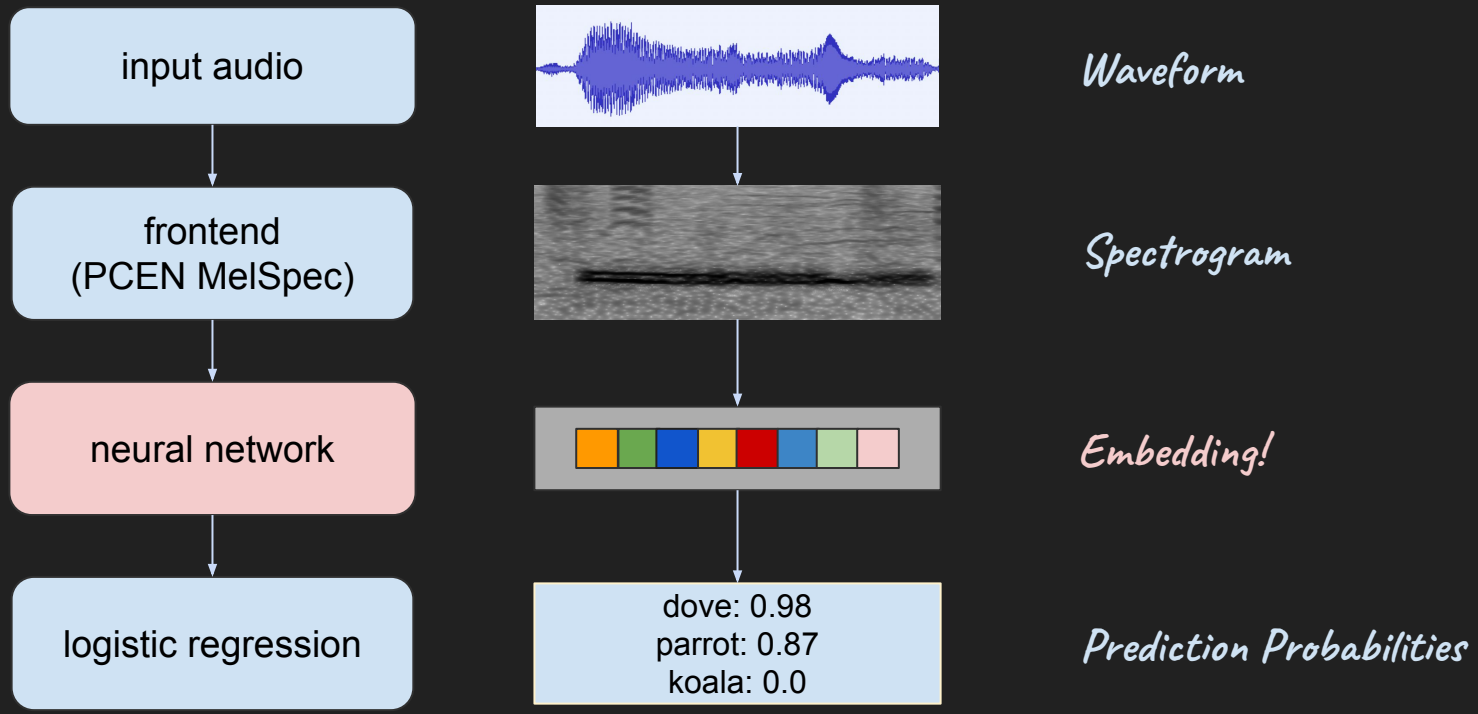


Agile Modeling

- Use a **robust pre-trained model** to embed target data.
- Use embeddings to **search** for relevant examples.
- Train a **small classifier** on relevant examples.
- Make the system **simple enough for anyone** to use.
  - No waiting to iterate on model quality.
    Remove need to pass data and models between ecologists and ML experts.
- If all goes according to plan, allow more focus on '**answering ecology questions**.'



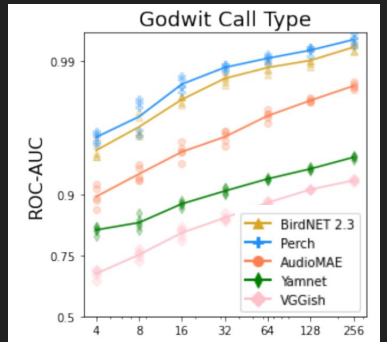
Embeddings: Fingerprints of Sound

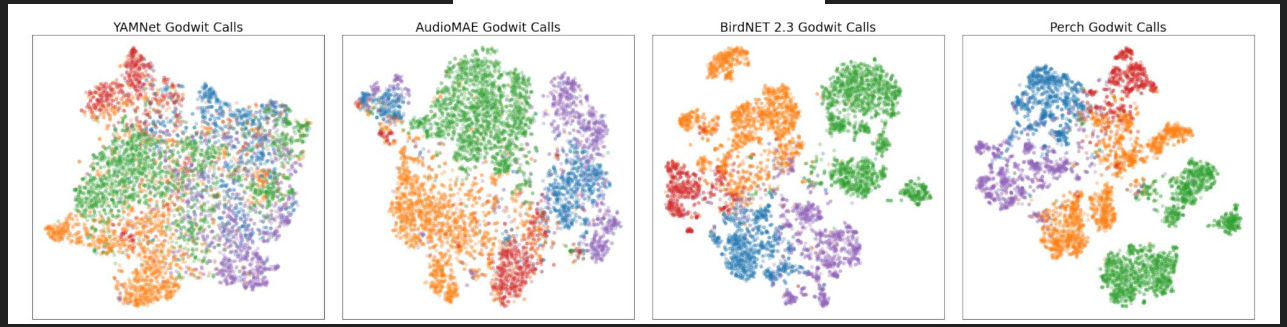


Perch Model

- **EfficientNet B1** architecture, 1280-dim'l embedding.
- Trained on **Xeno-Canto** - ~10k hours of weakly labeled data.
  - That said, the label quality is much higher than typical general audio datasets.
- **Multiclass+Multilabel** problem: Binary Cross-Entropy loss.
- About **15,000 total classes**.
- **Hierarchical labels**: Species, Genus, Family, Order all classified.
- **MixUp augmentation** is the most important, followed by random gain.

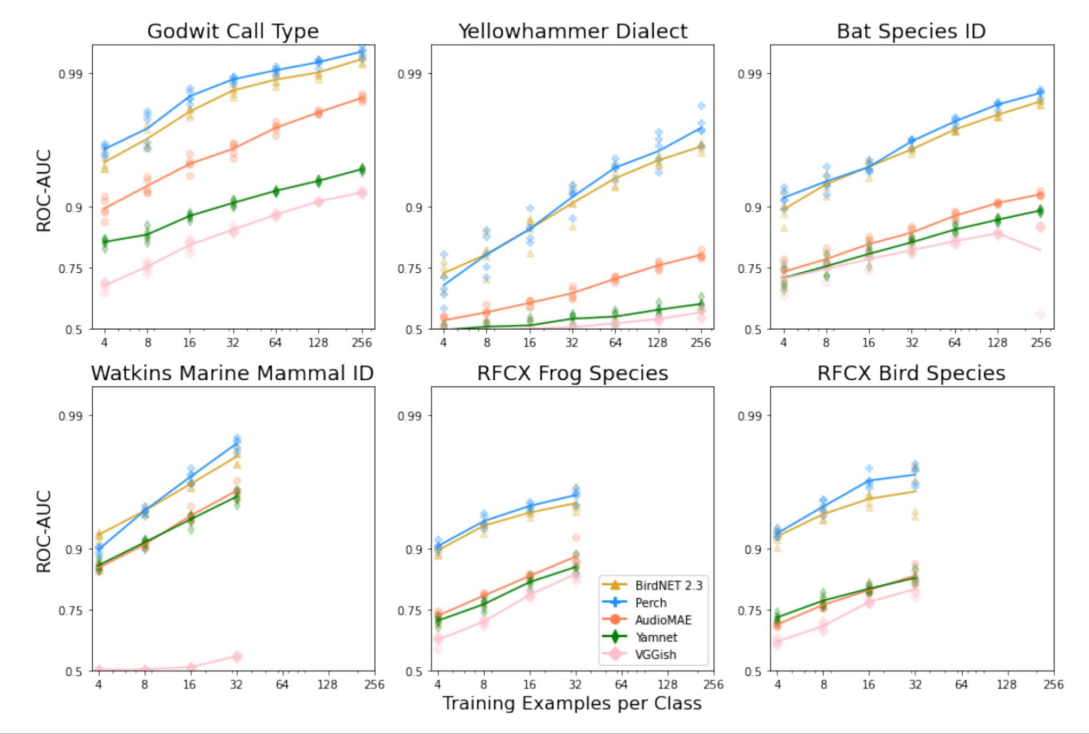
Feature embeddings from global bird classifiers can organize novel data effectively.





[arxiv: ghani+denton +kahl+klinck 2023]

Feature embeddings from global bird classifiers can organize novel data effectively.



[arxiv: ghani+denton +kahl+klinck 2023]