more signal processing basics

waveforms

Audio is a measurement of changes in air pressure over time.

It is usually stored as **an array**, with some **sample rate**. The sample rate is the number of samples measured per second of audio.

We'll often work with 32kHz sample rate: 32,000 samples per second.

Thus, 5s of audio at 32kHz is an array with shape (160_000,)!

To emphasize:
Audio is an array, and the transformations of audio into a spectrogram are matrix operations.

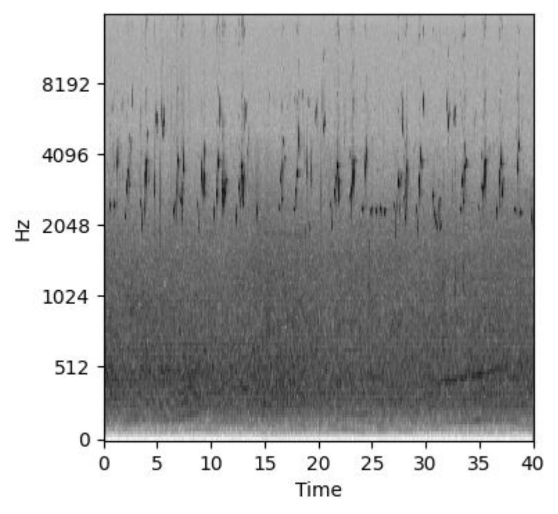
- The Hann window is pointwise multiplication.
- The FFT can be written as an invertible matrix.
- The melspec is a matrix.

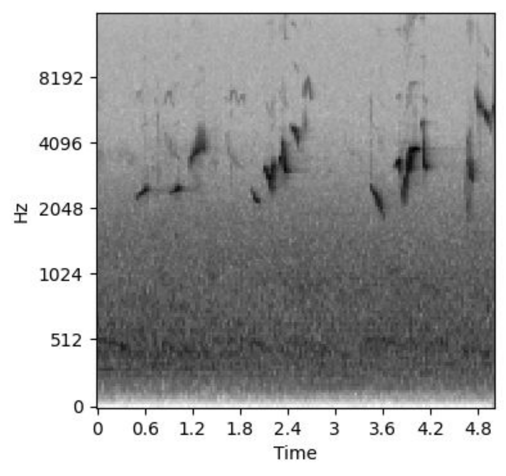
audio windows

Plotting lots of audio in a single spectrogram will 'squish' the time detail.

You can use **slices** to select a smaller audio window:

```
arr[offset*sr:
    (offset+window_size)*sr]
```





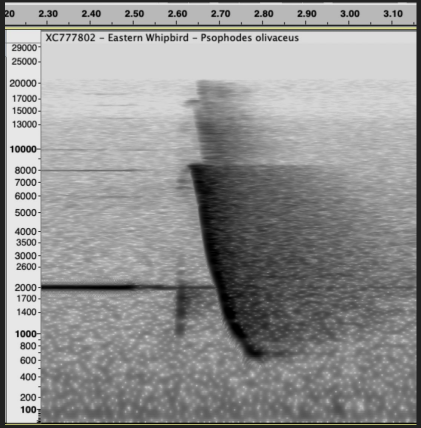
spectrogram frequency scale

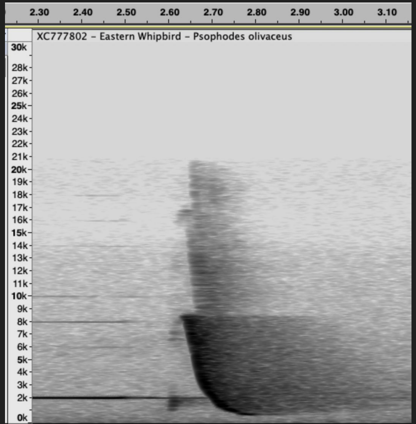
The y-scaling of the spectrogram changes its appearance.

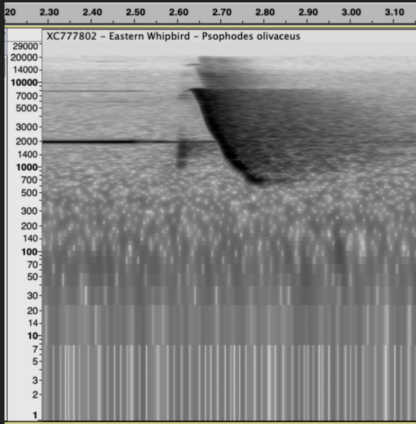
Linear scale tends to 'squish' low-frequency features.

Log scaling gets weird at low frequencies (but can be cut off).

Mel-scale is similar to log-scale, but handles low frequencies better.







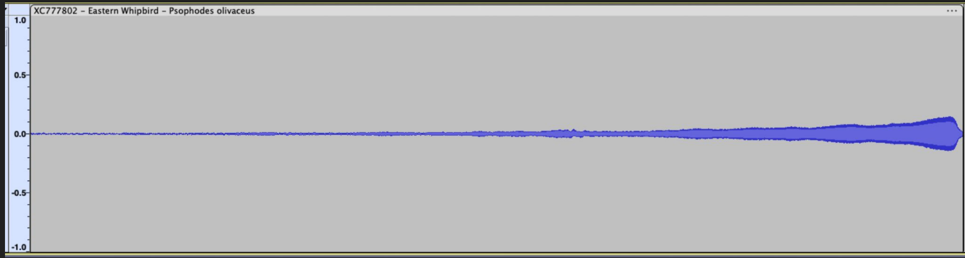
fourier transform

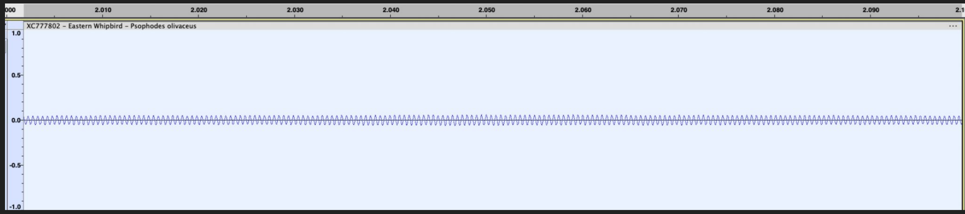
The Fourier transform converts the waveform ("time-domain") to frequencies ("frequency domain").

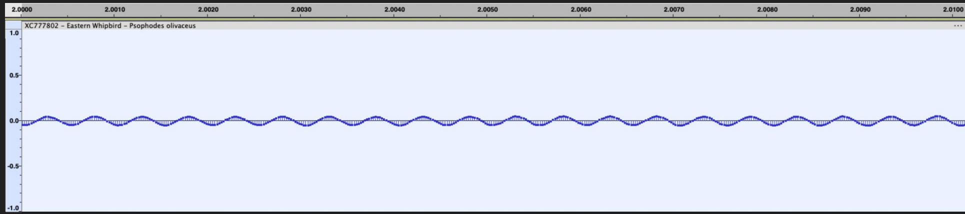
In brief: For each frequency, we measure **similarity of the signal** to a **pure tone of that frequency**.

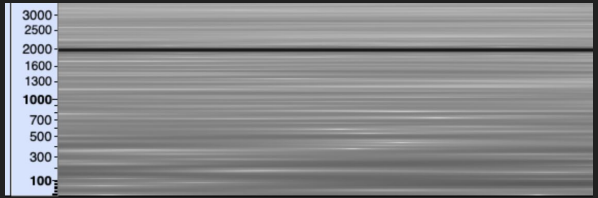
The collection of all of these measurements is the Fourier transform.

There are many mathematical tricks for computing this quickly!
"Fast Fourier transform."









nyquist frequency

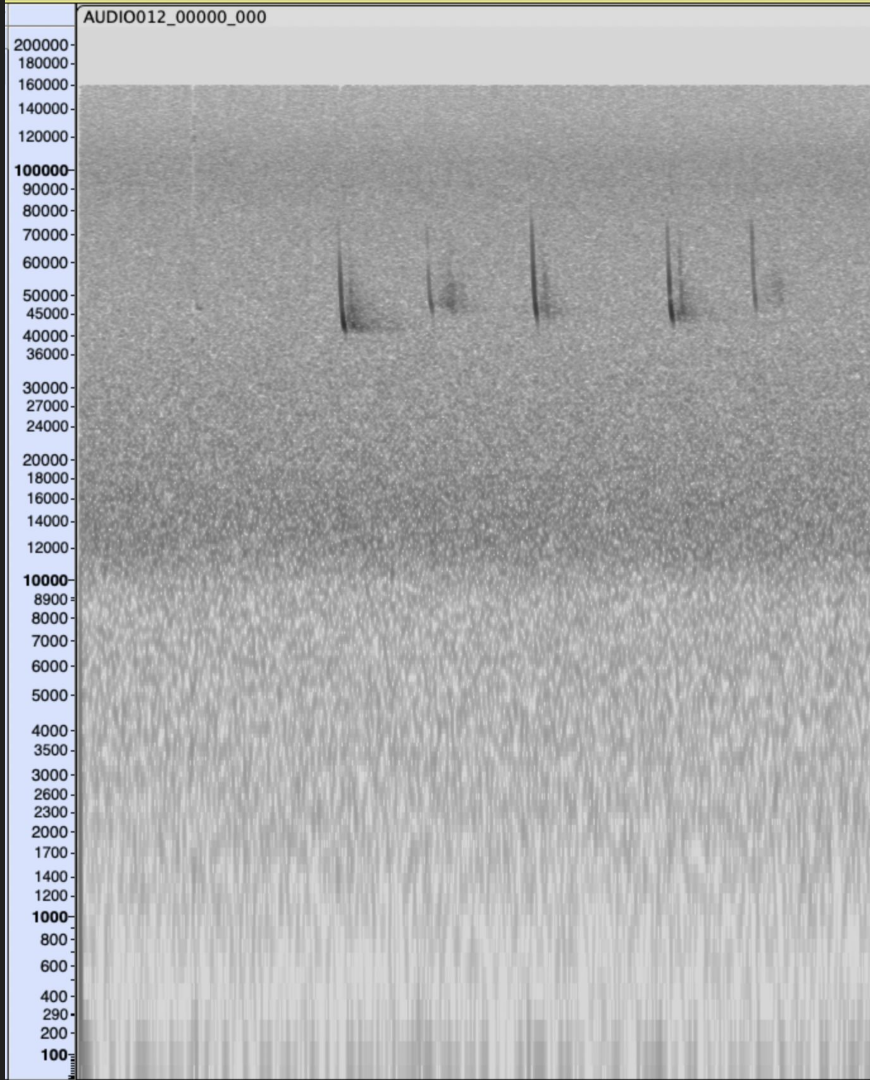
The **highest frequency** we can measure in an audio signal is **half the sample rate** (Nyquist theorem).

Higher sample rates produce lots of data.

Bats calls range from 14kHz to >100kHz. Often recorded at >=192kHz sample rate.

Speech is often recorded at 8kHz or 16kHz.

High-pitched bird vocalizations may be 13-14kHz in frequency.

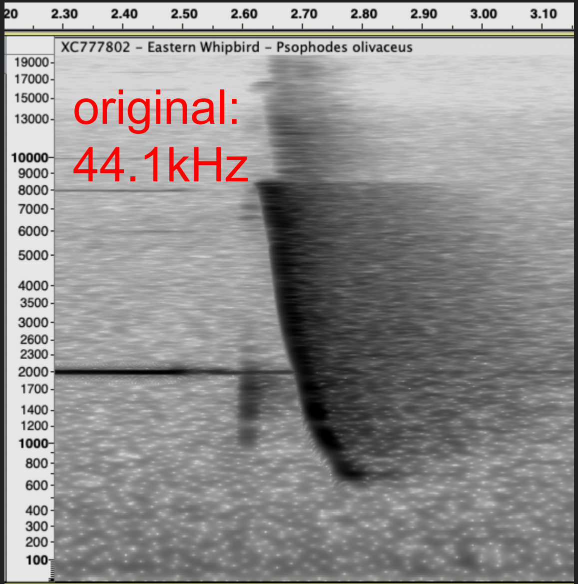


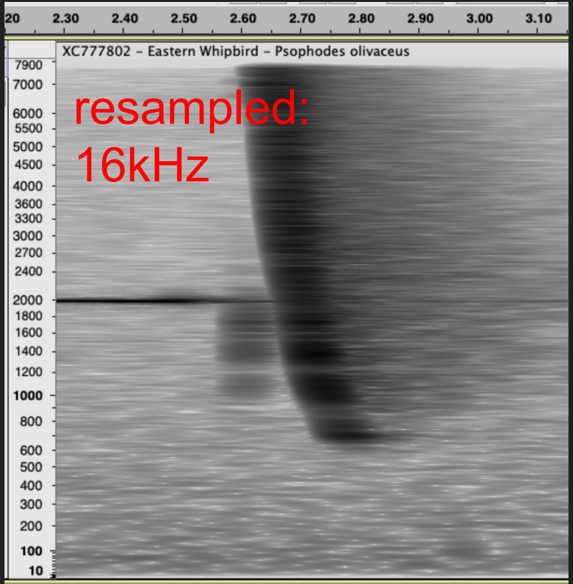
resampling

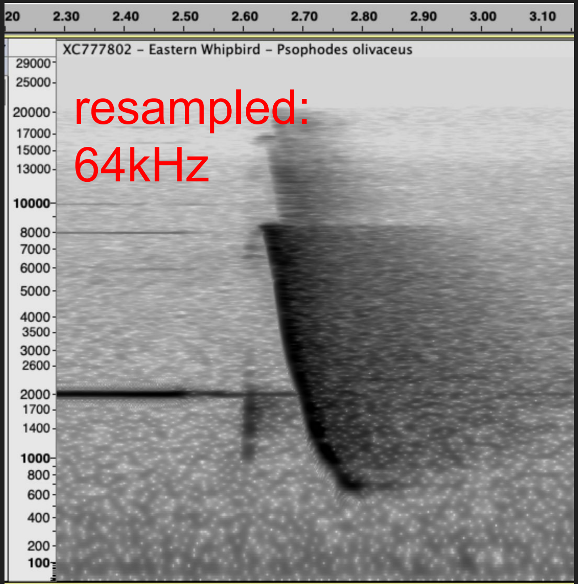
Good algorithms exist for changing the sample rate of audio.

Lowering the sample rate cuts off high-frequency signals.

Raising the sample rate gives a dead zone (no signal) at the top of the spectrogram.



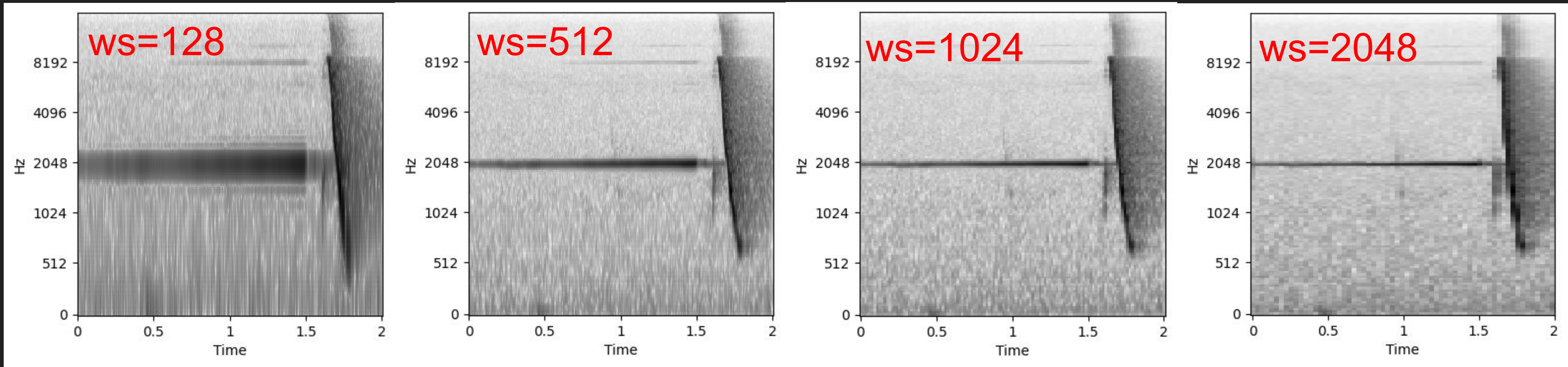




window size

Set `hop_length=window_size//2`.

Then play with the `window_size`…



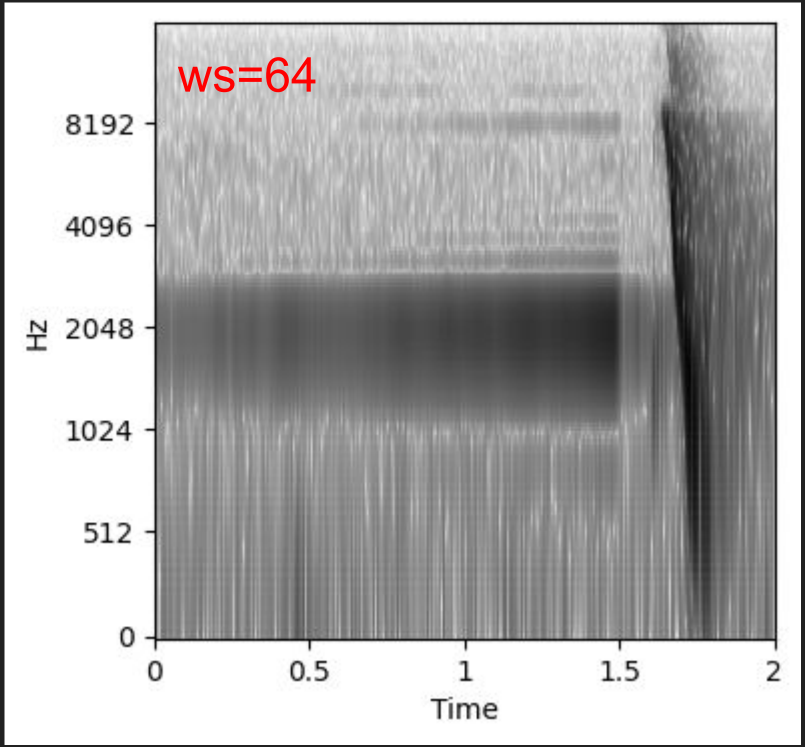
rayleigh frequency

The **lowest frequency** the fft can measure is f_{min} = 1/T, where T is the window length in seconds (Rayleigh frequency).

Example:

64 samples / 32,000 Hz = 1/500 s.

Then f_{min} = 500Hz.



Important note for machine learning…

We typically feed spectrograms to a computer vision model for learning.

We need to use the same spectrogram parameters for inference as were used when training the model!

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?

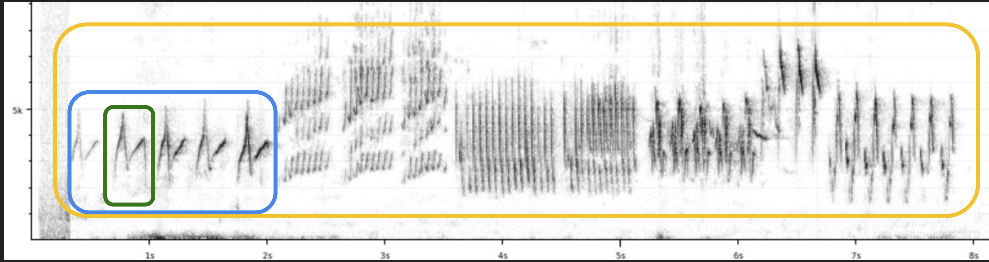
Structure + Plasticity

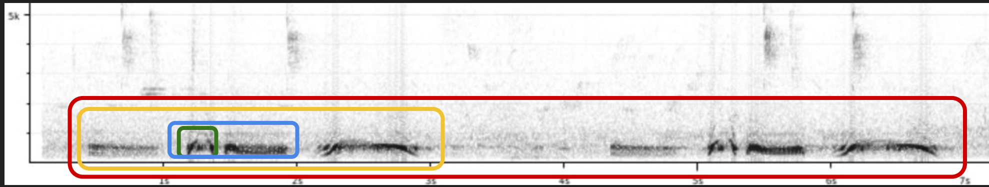
Birdsong is typically divided into units: **notes** (single, separate units of sound), **phrases** (distinct collections of notes), **songs** (a distinct collection of phrases), and **bouts** (session of many vocalizations).

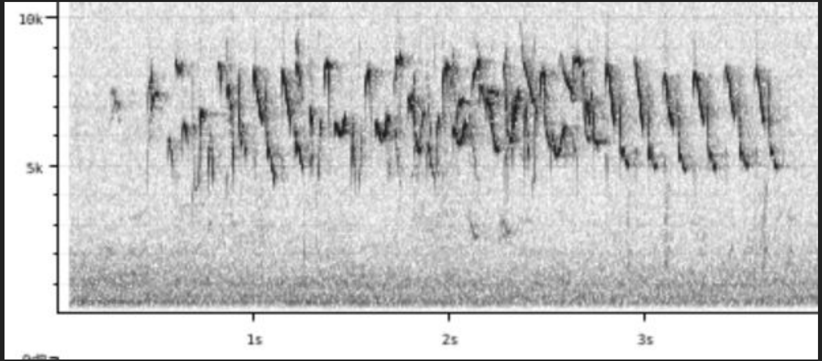
Structures can be simple or complex, at each level!

A phrase might be one note or many.

Plasticity: Subsequent units may be different or repeated.







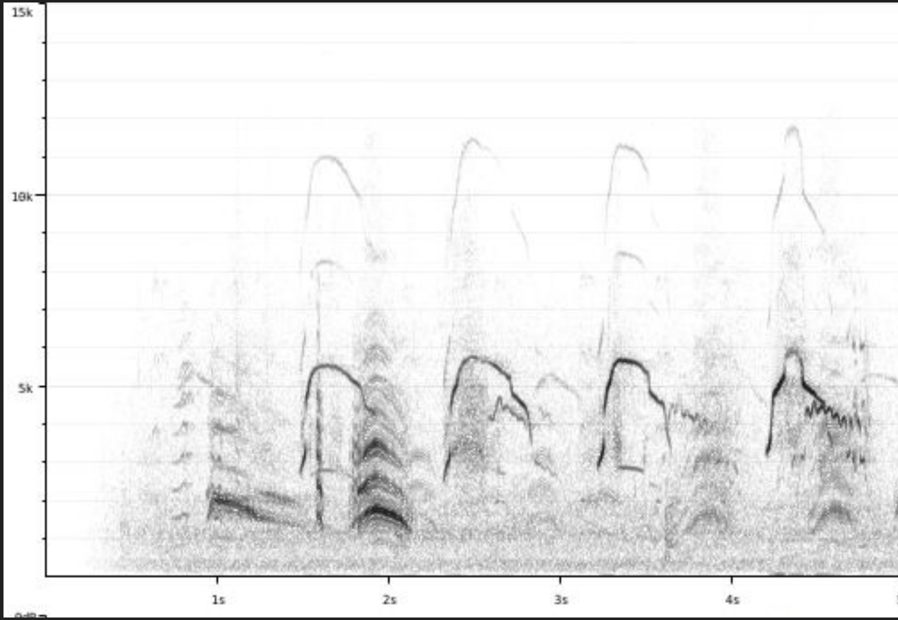
Harmonic Structure

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, …

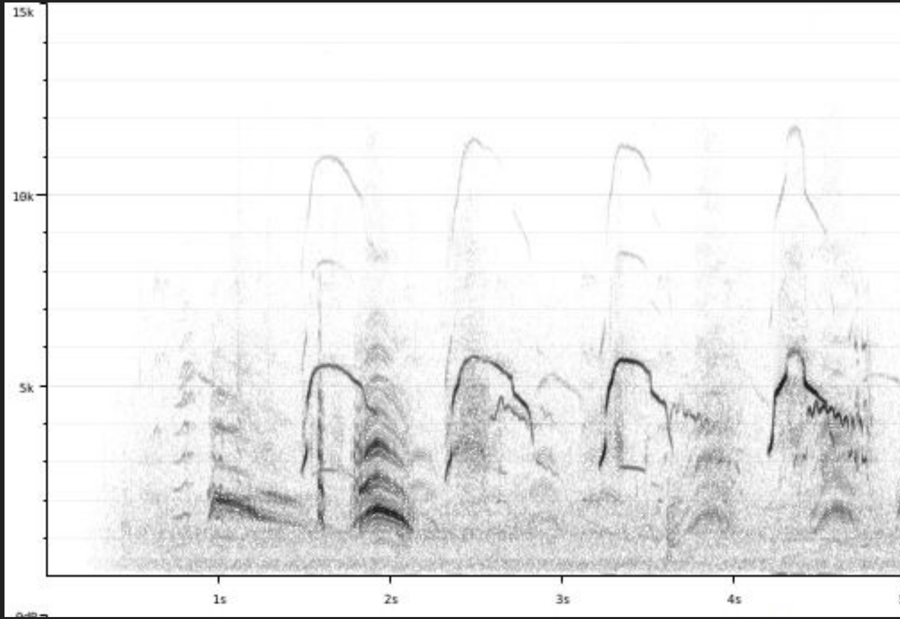
Often the fundamental is the loudest,
but not always!



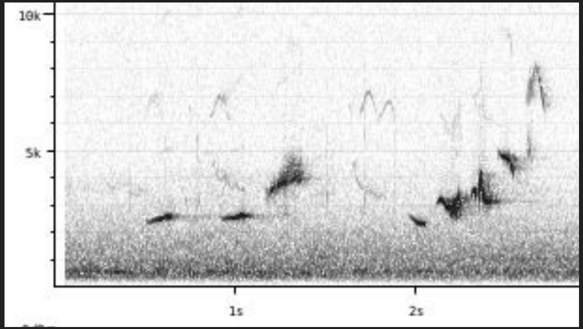
Buzzy notes vs tones

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.

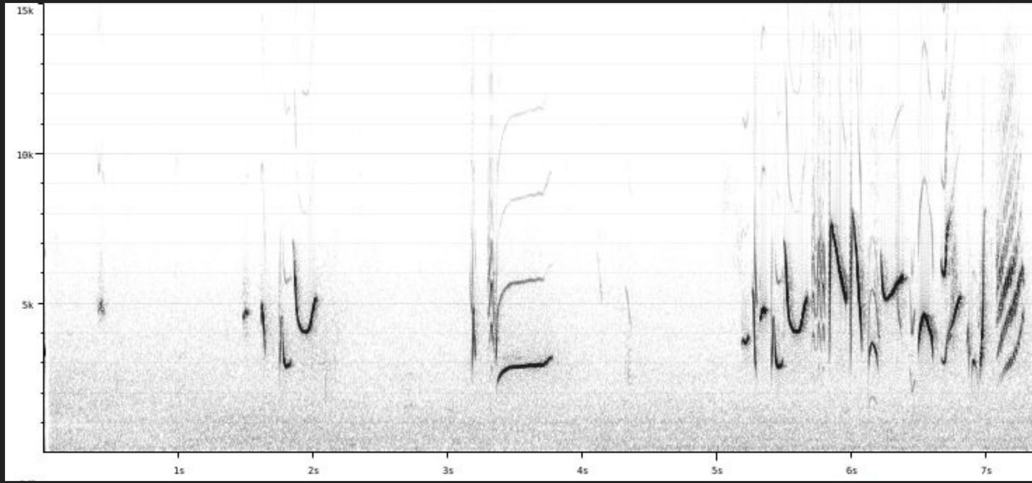


Pitch characteristics

- 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

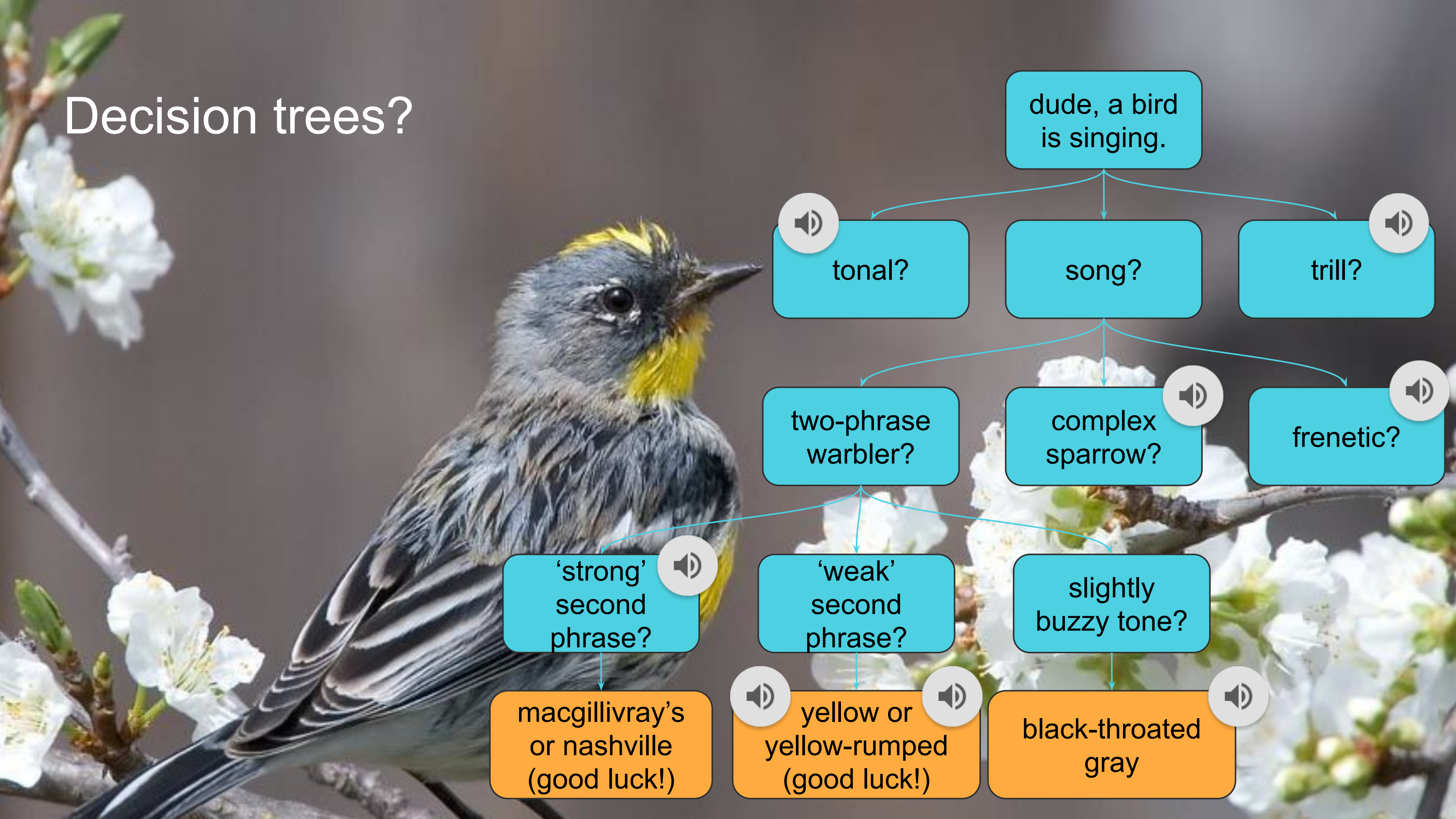
- Read descriptions of the bird on wikipedia and eBird.
- Listen to examples from different days / locations / recordists.
- Look for common features which might help distinguish from other birds.

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.
- ?!