Acoustical Society of America, May 26, 2022

ml-audio-dev-tools

Development tools for deep learning models of acoustical signal processing

Scott H. Hawley, Ph.D.

@drscotthawley

❖ Chem. & Phys. Dept, Belmont U.
❖ Belmont Data Collaborative
❖ Harmonai

Brought to you by ASA Technical Committee on Signal Processing

Thanks: Andrew Parker, Brian McFee, Fabian-Robert Stöter, Jesse Engel, Gene Kogan, Dadabots, David Braun, Zach Evans, scart97, Eric Hallahan

ml audio dev tools

...are *quite* few in number

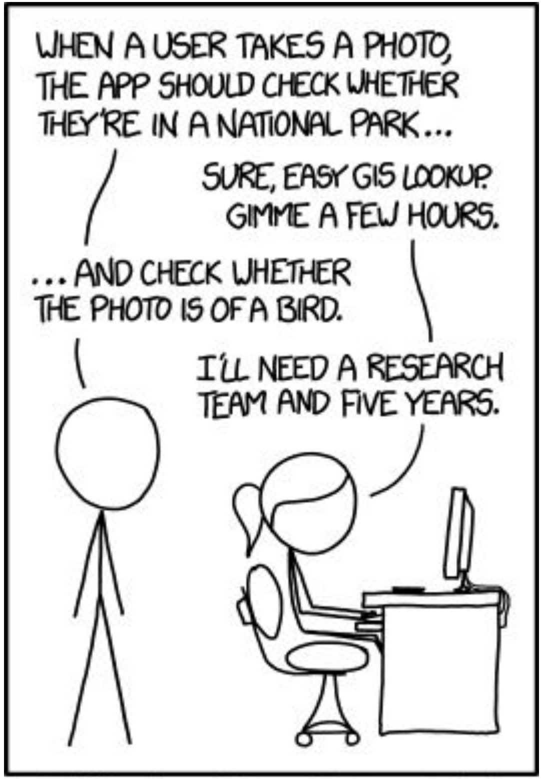
…so we should all build more.

What's too difficult?

Jeremy Howard of fast.ai uses this XKCD cartoon at the start of Lesson 1 in the fast.ai course "Practical Deep Learning for Coders"

...to illustrate the **huge change in the capabilities of Computer Vision systems since the advent of Deep Learning**.

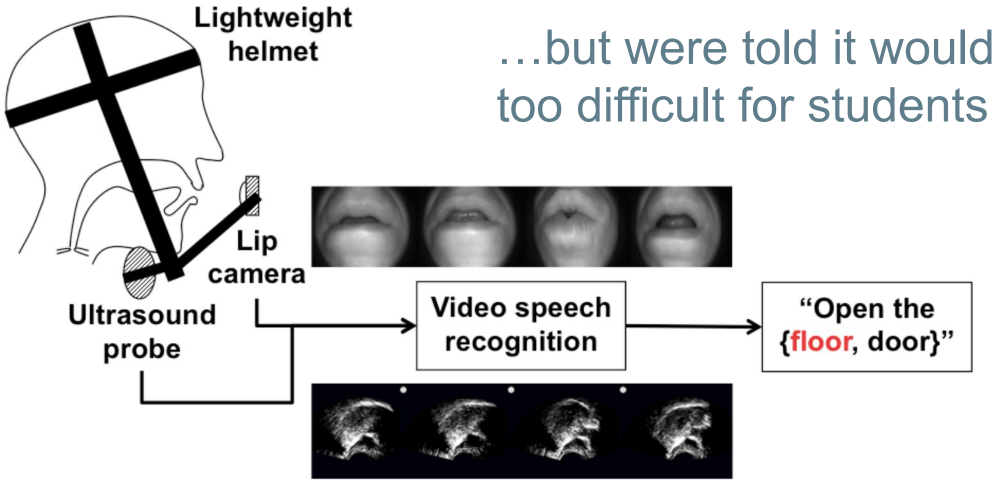
This problem of bird detection is *trivial* nowadays. It is not even worth a homework problem.


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

What’s too difficult?

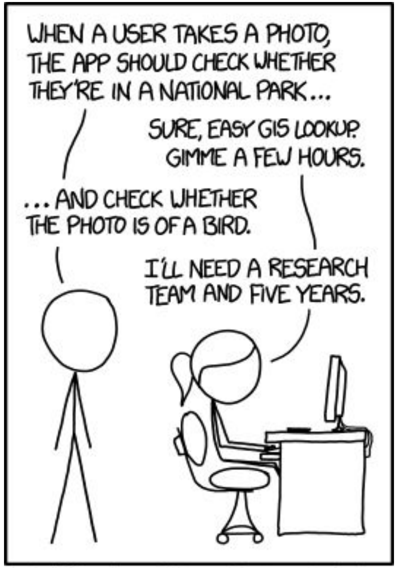
In 2020, Bruce Denby (Inst. Lagevin, Sorbonne) & I wrote to CA TC with a proposal for a 2021 ASA Student Challenge in Speech-To-Text using video imagery ("Silent Speech"):



Video speech recognition

“Open the {floor, door}”

…but were told it would be too difficult for students.

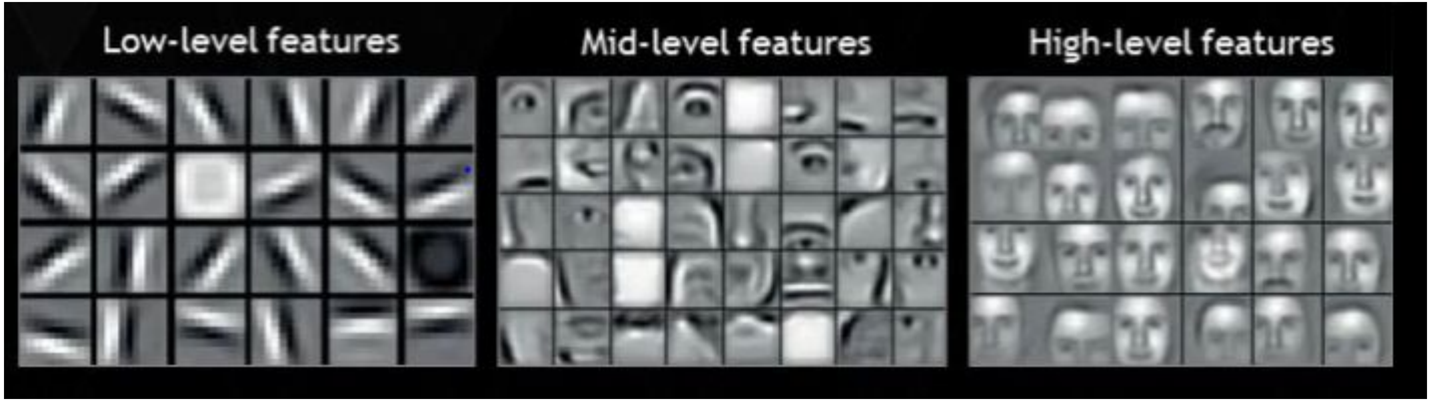

WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

We were *confused*.
With modern DL-CV tools,
This is a *homework problem*,
or at best a minor Kaggle competition.

Some things that were hard become easy with DL

Many human-designed **feature detection** systems can be obviated with Deep Learning systems, which build their own feature detectors as part of the training process.



An example hierarchy of feature detectors for a facial recognition system. (Source: Lee et al, 2009.)

Aside: At SP TC meeting Tuesday night, there was talk about a future session on "Feature Extraction and Dimensionality Reduction" – both of these are built in to DL systems.

Some things that were hard become easy with DL

Various examples, but here's a couple:

Neural networks are great at *pattern matching* and *denoising*:

I wrote my first denoising autoencoder before I knew what a Wiener filter was!

I wrote an object detector for ellipses (Hawley & Morrison JASA 2021 & JASA Express Letters 2022) without bothering to try Hough Transforms (b/c it was complex)

Which means...

- there's a new generation of coder-scientists (🙋‍♂️) who ***hope*** that DL will make up for their *lack of signal processing domain knowledge*. **Not without reason:** DL has shown to *beat* former baselines in many fields for many problems.
- there are veterans in the SP field who are perhaps not up to speed on the rapid pace of advancements in DL for "audio", and may be curious about incorporating DL into their work. ...Good news: This is/(can be) "easy"*!

*compared to hard-core SP/math, if you're already good at coding

More good news: "Everything old is new again". Much of classic SP still finds its way into DL systems & helps drive innovation. (e.g., Vector Quantization!)

Who this talk is for:

- Students and ❤️*Esteemed SP Experts*🙇‍♂️: "Onboarding"
- Experienced DL-Audio researchers: Sharing tips!

Confession/bias of mine:

- When I say "audio" I usually mean "musical" audio:
    - multi-channel
    - high sample rates (44.1kHz+)
    - ...and *not just classification problems*
    - I tend to not even think about *"Speech". No offense*
- For me, DL $\subset$ ML

ml audio dev tools...

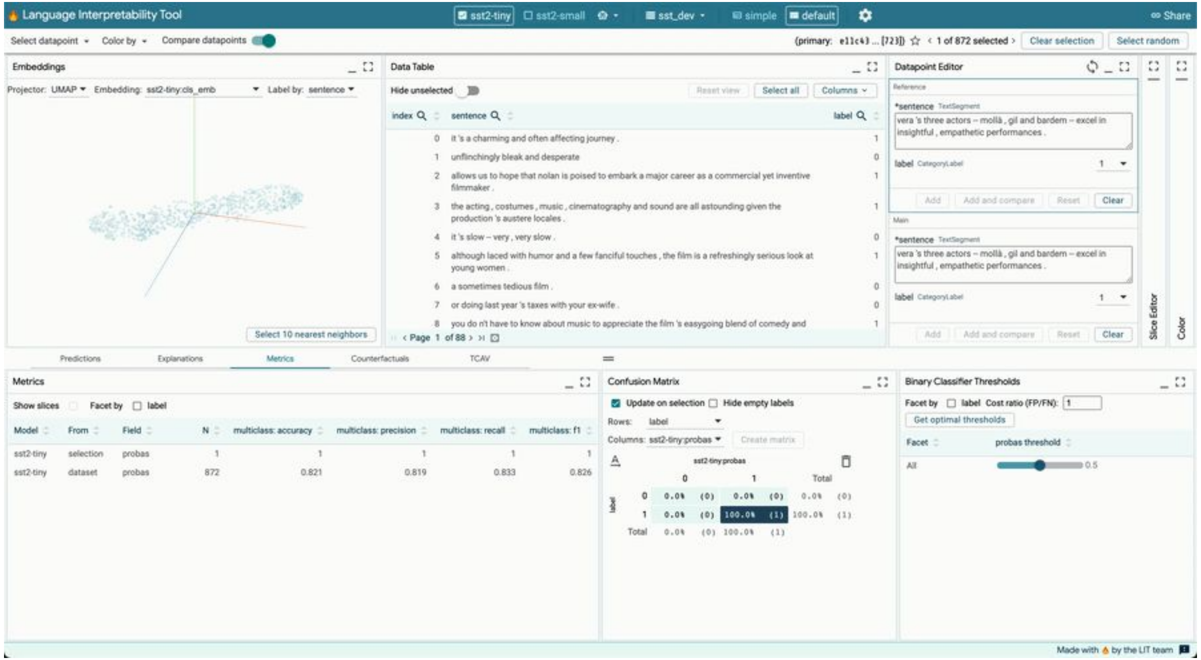
...are few in number

...**Compared** to Tools for DL Models for **Images**, **Text**, ...and even ***Speech***

In these other domains, DL is very mature.*

MANY, many, many tools & demos for viz., analysis, saliency,...

*because money



(primary: e11c43 ... [723]) ☆ ‹ 1 of 872 selected › Clear selection Select random

Projector: UMAP Embedding: sst2-tiny:cls_emb Label by: sentence

index	sentence	label
0	it 's a charming and often affecting journey .	1
1	unflinchingly bleak and desperate	0
2	allows us to hope that nolan is poised to embark a major career as a commercial yet inventive filmmaker .	1
3	the acting , costumes , music , cinematography and sound are all astounding given the production 's austere locales .	1
4	it 's slow -- very , very slow .	0
5	although laced with humor and a few fanciful touches , the film is a refreshingly serious look at young women .	1
6	a sometimes tedious film .	0
7	or doing last year 's taxes with your ex-wife .	0
8	you do n't have to know about music to appreciate the film 's easygoing blend of comedy and	1

‹ Page 1 of 88 ›

Reference

*sentence TextSegment

vera 's three actors -- mollà , gil and bardem -- excel in insightful , empathetic performances .

label CategoryLabel 1

Main

*sentence TextSegment

vera 's three actors -- mollà , gil and bardem -- excel in insightful , empathetic performances .

label CategoryLabel 1

Show slices ☐ Facet by ☐ label

Model	From	Field	N	multiclass: accuracy	multiclass: precision	multiclass: recall	multiclass: f1
sst2-tiny	selection	probas	1	1	1	1	1
sst2-tiny	dataset	probas	872	0.821	0.819	0.833	0.826

☑ Update on selection ☐ Hide empty labels

Rows: label

Columns: sst2-tiny:probas

sst2-tiny:probas

label	0		1		Total	
0	0.0%	(0)	0.0%	(0)	0.0%	(0)
1	0.0%	(0)	100.0%	(1)	100.0%	(1)
Total	0.0%	(0)	100.0%	(1)		

Facet by ☐ label Cost ratio (FP/FN): 1

Facet	probas threshold
All	0.5

There's no audio tool as cool as Google PAIR's Language Interpretability Tool

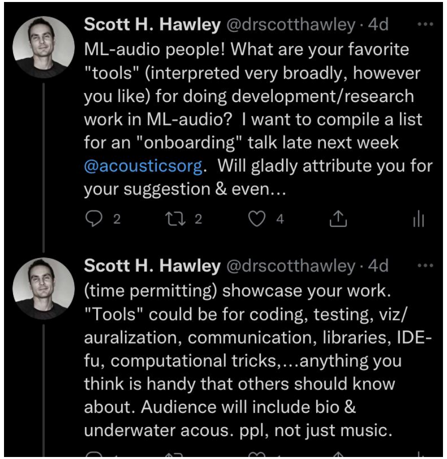
Responses I got...

i.e., tools suggested by DL-audio practitioners...

Thanks: Brian McFee, Fabian-Robert Stöter, Jesse Engel, Gene Kogan, Dadabots, Andrew Parker, David Braun, Zach Evans, scart97, Eric Hallahan, & Christian Steinmetz

...were mostly:

I. Basic ML workflow / 101
II. Normal audio workflow
III. A few extra cool tools



Scott H. Hawley @drscotthawley · 4d

ML-audio people! What are your favorite "tools" (interpreted very broadly, however you like) for doing development/research work in ML-audio? I want to compile a list for an "onboarding" talk late next week @acousticsorg. Will gladly attribute you for your suggestion & even...

Scott H. Hawley @drscotthawley · 4d

(time permitting) showcase your work. "Tools" could be for coding, testing, viz/auralization, communication, libraries, IDE-fu, computational tricks,...anything you think is handy that others should know about. Audience will include bio & underwater acous. ppl, not just music.

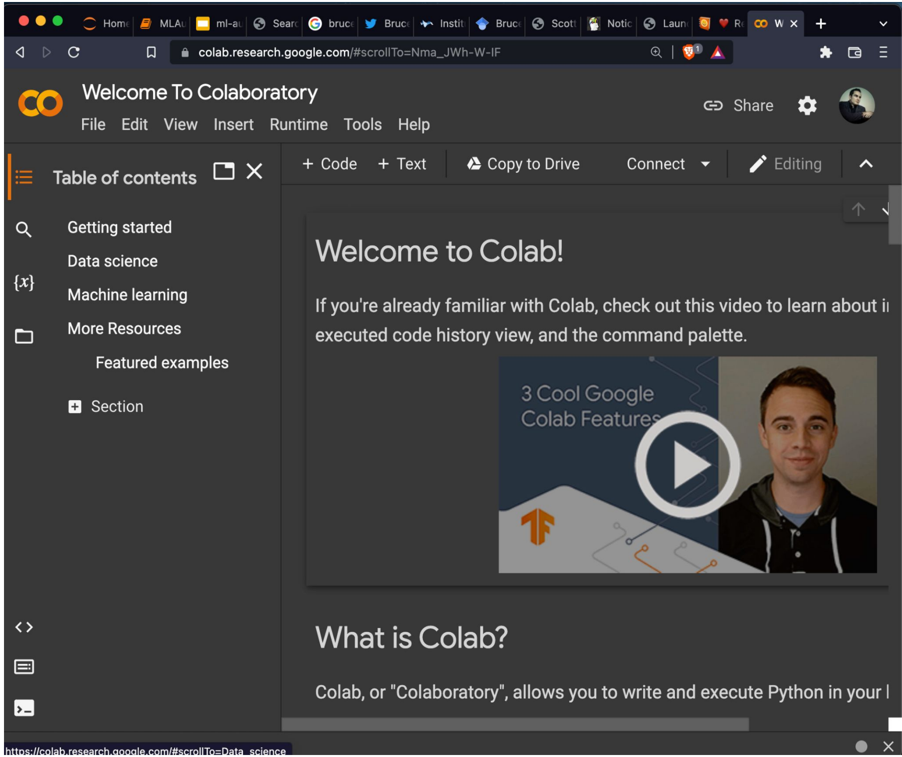
I. Basic ML - GPU Computing

Graphics Processing Units (GPUs) are key: 100x+ faster than CPU

You don't need to buy a GPU. Various cloud-based systems let you do GPU computing for (near) free.

Google Colab, Kaggle, Paperspace Gradient, Amazon Sagemaker,...

These typically make use of Jupyter notebooks.



Welcome to Colab!

If you're already familiar with Colab, check out this video to learn about i executed code history view, and the command palette.

What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your

(For awesome lib-dev via Jupyter notebooks, check out nbdev.)

I. Basic ML - Python Programming

Python is overwhelmingly the most popular & well-supported programming language for ML. (Much more so than MATLAB, C++, Julia, R, JS,...)

There's usually a library/package that does what you want.

Which DL Library? ➡ **PyTorch** ⬅, Tensorflow, JAX,..

Lightning

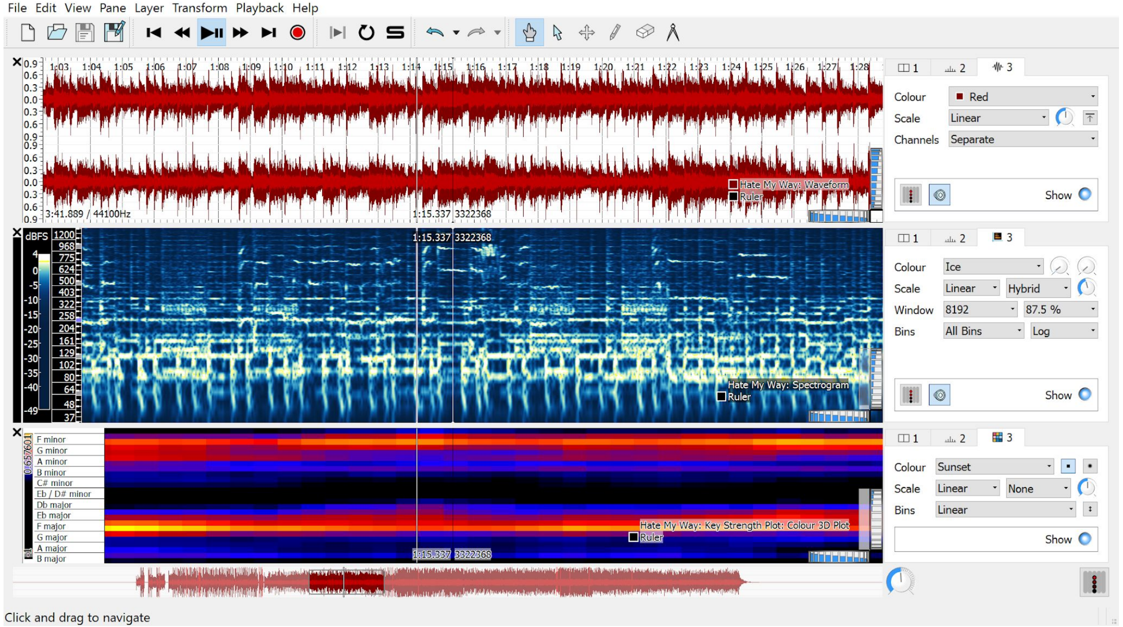
I hate `conda` and only use venv+ `pip`

I. Python Packages for Audio and/or DL

- librosa: "Swiss army knife": (Also by Brian McFee: mir_eval)
- torchaudio: GPU processing
- auraloss: Loss functions (Christian Steinmetz)
- Pedalboard: Data augmentation (Peter Sobot/Spotify)
- DawDreamer: Python DAW (David Braun)
- ONNX: Export to JUCE / C++ / JS

II. Normal Audio Workflow

- Audacity ← We'll come back to this one!
- Sonic Visualizer:
- Reaper
- Logic



Colour Red
Scale Linear
Channels Separate
Show

Colour Ice
Scale Linear Hybrid
Window 8192 87.5 %
Bins All Bins Log
Show

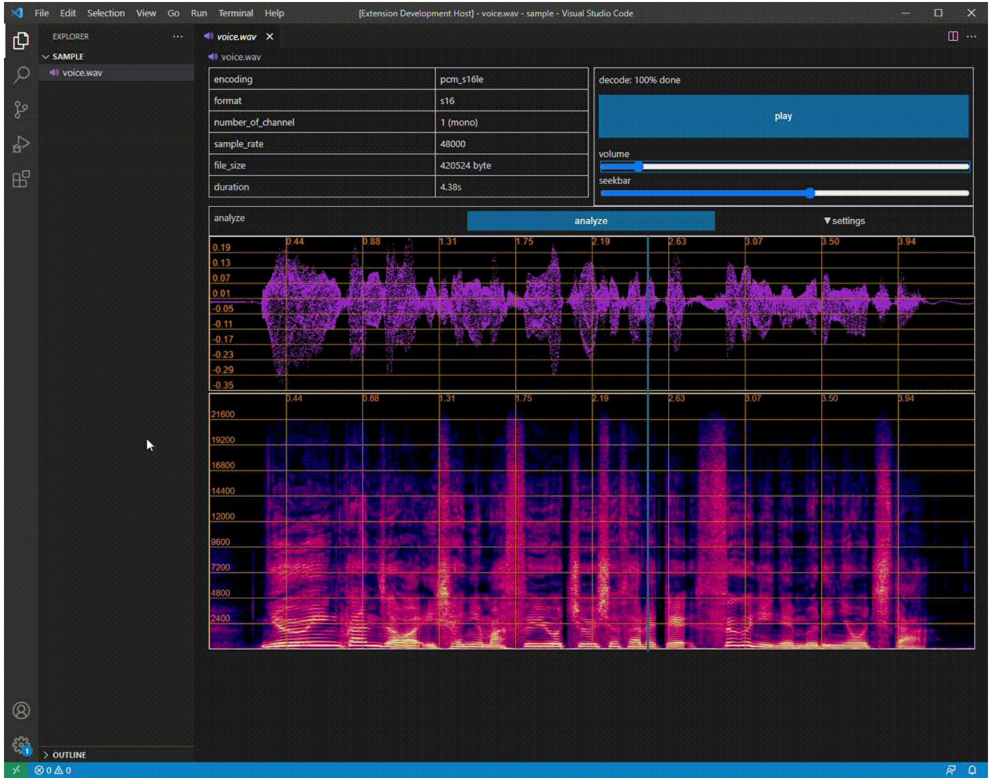
Colour Sunset
Scale Linear None
Bins Linear
Show

Click and drag to navigate

Suggested by Jesse Engel

II. Audio workflow - VSCode audio-preview

(Suggested by
Fabian-Robert Stöter)



[Extension Development Host] - voice.wav - sample - Visual Studio Code

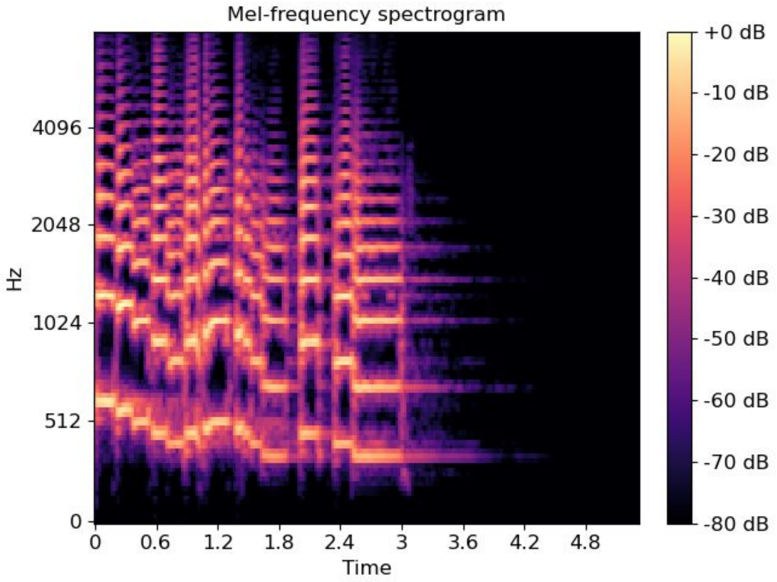
encoding	pcm_s16le
format	s16
number_of_channel	1 (mono)
sample_rate	48000
file_size	420524 byte
duration	4.38s

decode: 100% done

Works in Remote mode!
i.e., play on your laptop
the files on your server.

II.5 Trick: (Mel-)Spectrograms + Image-based DL

- Many posts on "The remarkable effectiveness of Convolutional Neural Networks on (Mel-)Spectrograms"
- Translation equivariance of CNNs fits well with phase (+ pitch) translation invariance of human auditory system.
- Upshot: Just using images of spectrograms with an image-based code can work surprisingly well.
- Makes a great baseline before going for full end-to-end audio DL



Source: LibRosa (McFee et al)

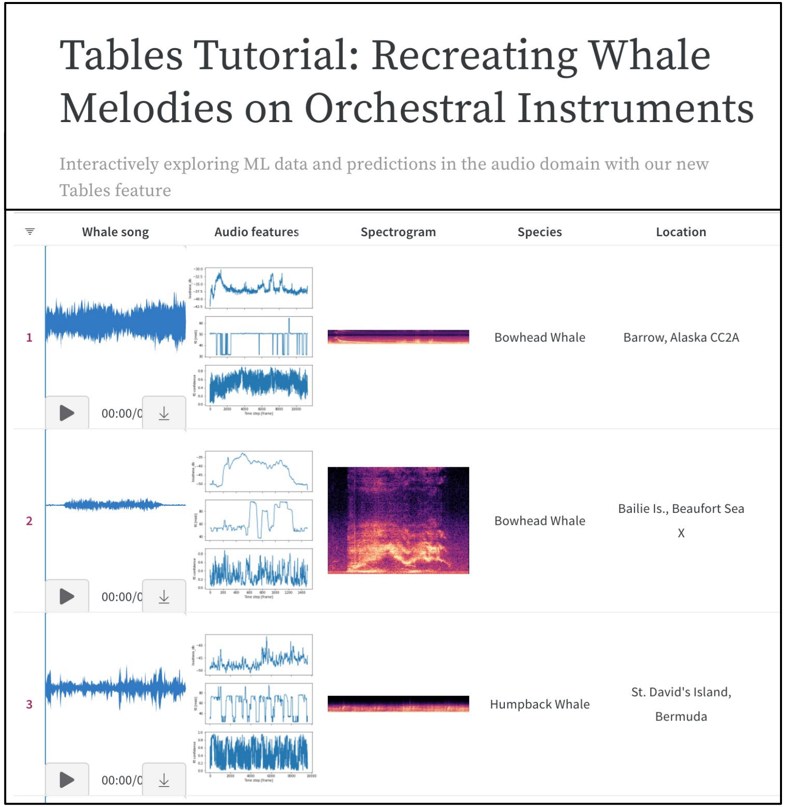
If you're just starting out, probably start with this method

III. Extra Cool - WandB (Audio) Callbacks!

- Weights & Biases ("wandb") is a cloud-based data-logging service you can use for free.

  (Hawley & Morrison JASA-EL 2022 found it "essential" for keeping track of many, many runs.)

- Among the things you can log & playback are audio examples:



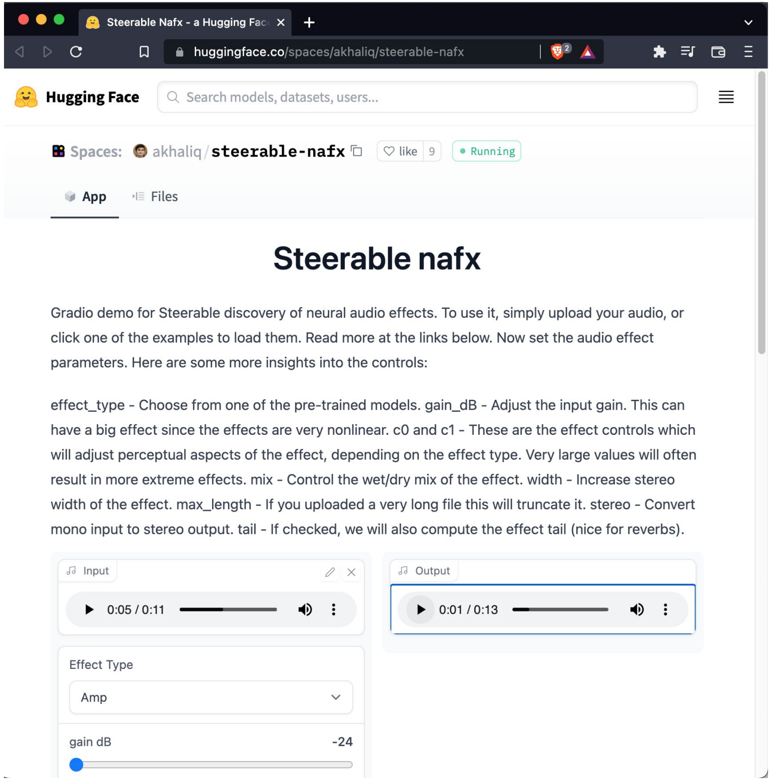
Tables Tutorial: Recreating Whale Melodies on Orchestral Instruments

Interactively exploring ML data and predictions in the audio domain with our new Tables feature

	Whale song	Audio features	Spectrogram	Species	Location
1	00:00/0			Bowhead Whale	Barrow, Alaska CC2A
2	00:00/0			Bowhead Whale	Bailie Is., Beaufort Sea X
3	00:00/0			Humpback Whale	St. David's Island, Bermuda

III. Communication: Demo Hosting via Gradio.app

Demo for Christian Steinmetz & Josh Reiss, NeurIPS 2021


Steerable Nafx - a Hugging Fac
huggingface.co/spaces/akhaliq/steerable-nafx
Hugging Face
Search models, datasets, users...
Spaces: akhaliq/steerable-nafx
like 9
Running
App
Files

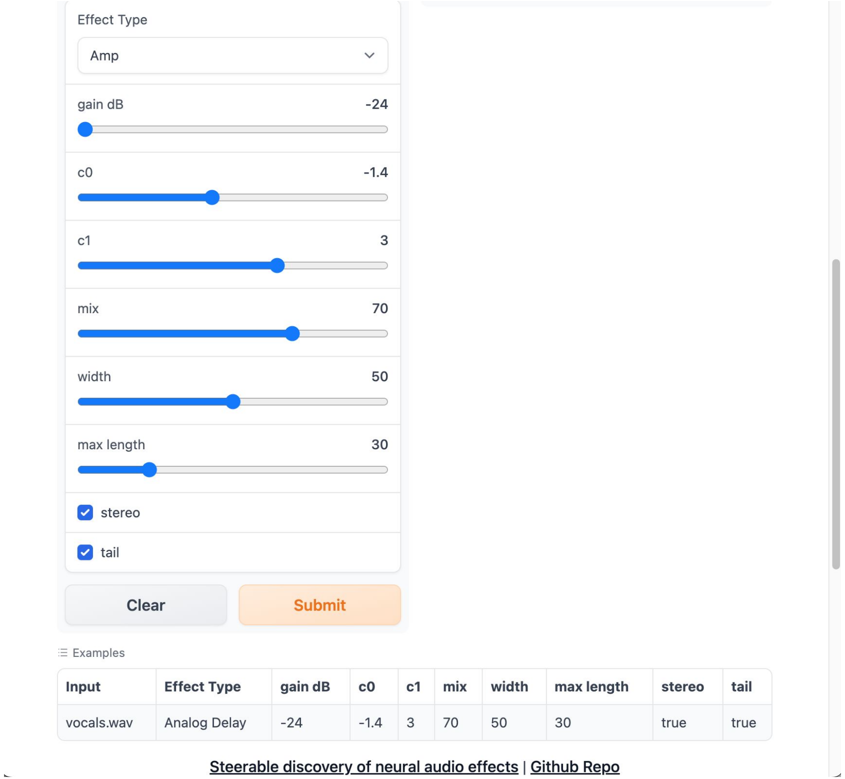
Steerable nafx

Gradio demo for Steerable discovery of neural audio effects. To use it, simply upload your audio, or click one of the examples to load them. Read more at the links below. Now set the audio effect parameters. Here are some more insights into the controls:

effect_type - Choose from one of the pre-trained models. gain_dB - Adjust the input gain. This can have a big effect since the effects are very nonlinear. c0 and c1 - These are the effect controls which will adjust perceptual aspects of the effect, depending on the effect type. Very large values will often result in more extreme effects. mix - Control the wet/dry mix of the effect. width - Increase stereo width of the effect. max_length - If you uploaded a very long file this will truncate it. stereo - Convert mono input to stereo output. tail - If checked, we will also compute the effect tail (nice for reverbs).

Effect Type: Amp

gain dB: -24



Effect Type: Amp

gain dB: -24

c0: -1.4

c1: 3

mix: 70

width: 50

max length: 30

☑ stereo

☑ tail

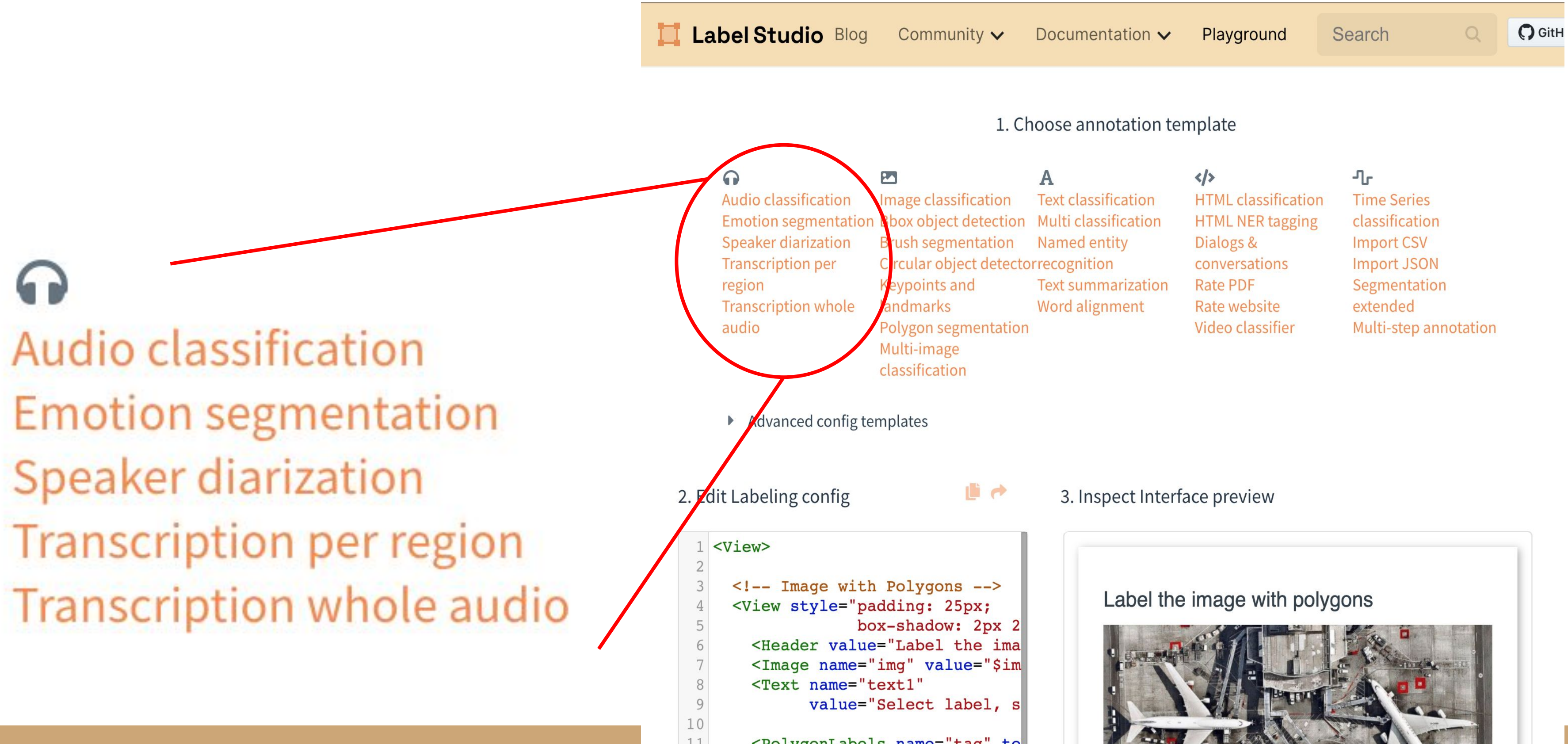
Examples

Input	Effect Type	gain dB	c0	c1	mix	width	max length	stereo	tail
vocals.wav	Analog Delay	-24	-1.4	3	70	50	30	true	true

Steerable discovery of neural audio effects | Github Repo

(Suggested by Fabian-Robert Stöter & scart97)

III. Label Studio



1. Choose annotation template

Audio classification
Emotion segmentation
Speaker diarization
Transcription per region
Transcription whole audio

Image classification
Bbox object detection
Brush segmentation
Circular object detector
Keypoints and landmarks
Polygon segmentation
Multi-image classification

Text classification
Multi classification
Named entity recognition
Text summarization
Word alignment

HTML classification
HTML NER tagging
Dialogs & conversations
Rate PDF
Rate website
Video classifier

Time Series classification
Import CSV
Import JSON
Segmentation extended
Multi-step annotation

Advanced config templates

2. Edit Labeling config

```
<View>

  <!-- Image with Polygons -->
  <View style="padding: 25px;
               box-shadow: 2px 2
    <Header value="Label the ima
    <Image name="img" value="$im
    <Text name="text1"
          value="Select label, s

    <PolygonLabels name="tag" to
```

3. Inspect Interface preview

Label the image with polygons

Audio classification
Emotion segmentation
Speaker diarization
Transcription per region
Transcription whole audio

III. Extra Cool - Audacity DL Models!

- Download models off 🤗

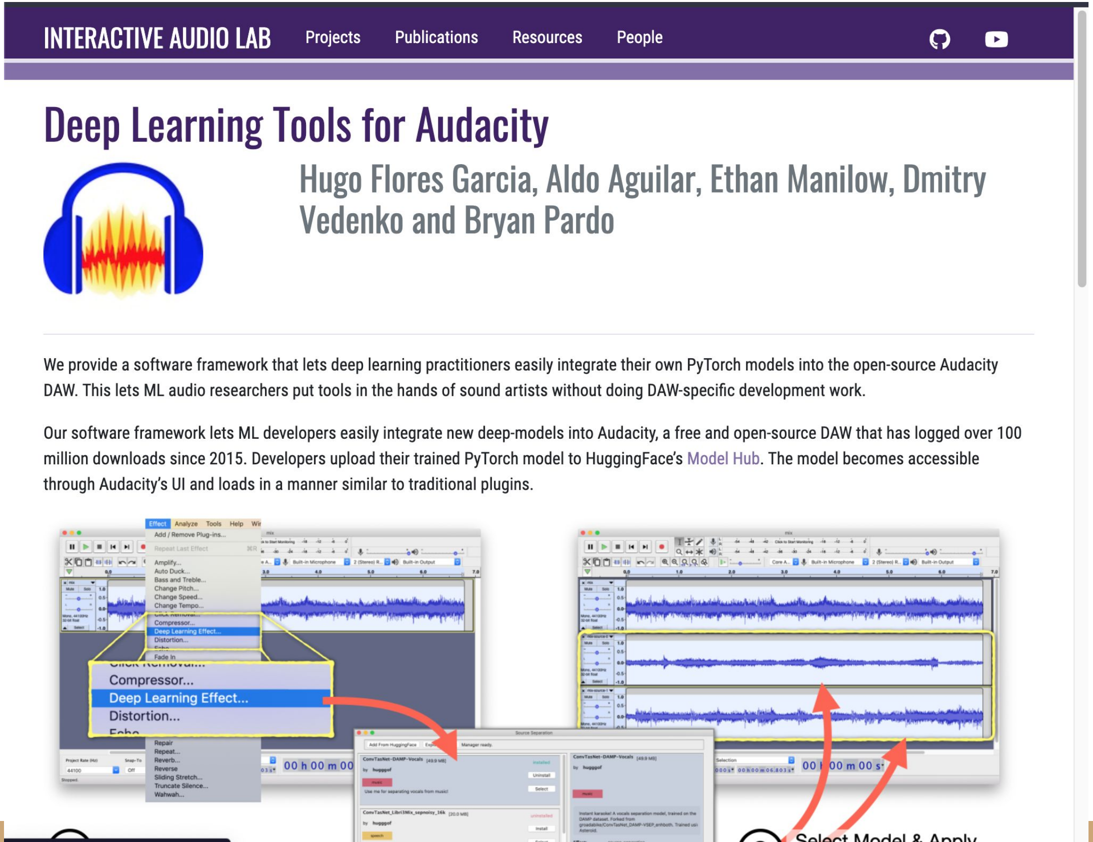
  (& upload your own first!)

- Run model as audio effect

- So far only mono, no knobs

Note: This is a custom Audacity build, download from here

Help them add sliders!



INTERACTIVE AUDIO LAB Projects Publications Resources People

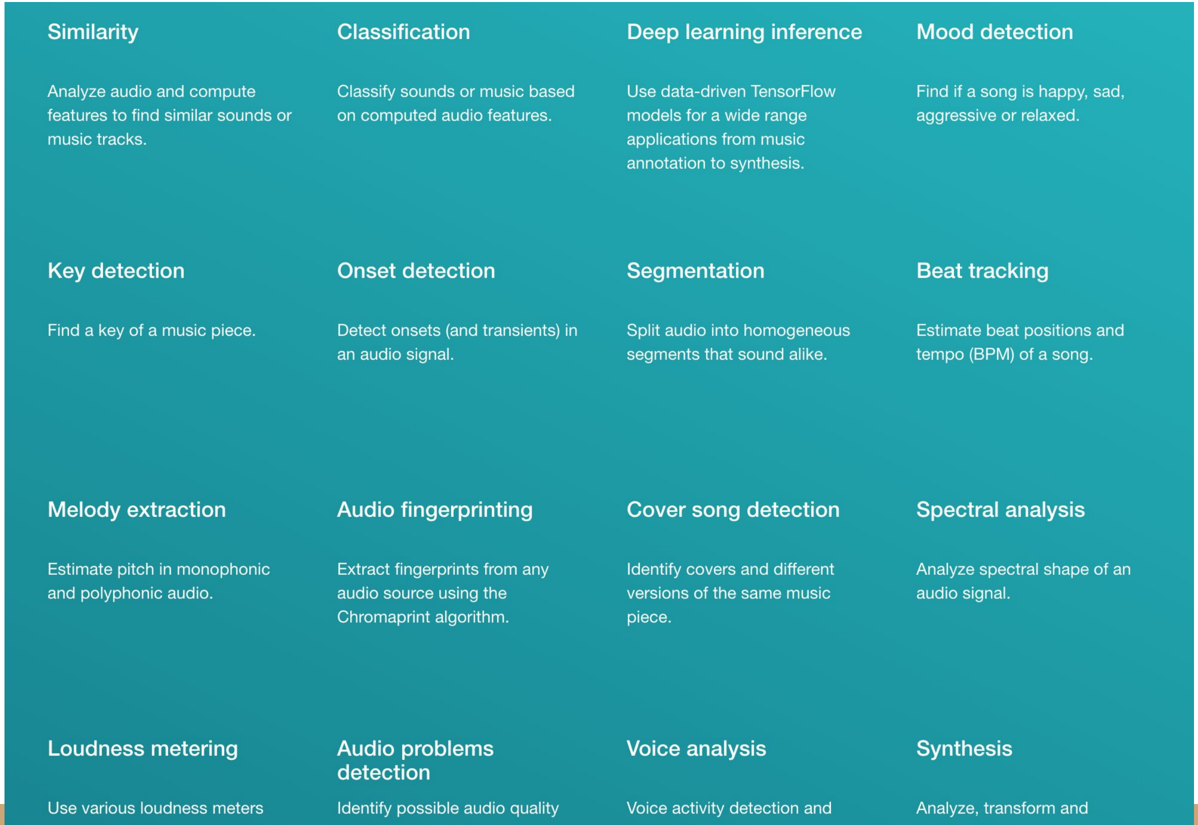
Deep Learning Tools for Audacity

Hugo Flores Garcia, Aldo Aguilar, Ethan Manilow, Dmitry Vedenko and Bryan Pardo

We provide a software framework that lets deep learning practitioners easily integrate their own PyTorch models into the open-source Audacity DAW. This lets ML audio researchers put tools in the hands of sound artists without doing DAW-specific development work.

Our software framework lets ML developers easily integrate new deep-models into Audacity, a free and open-source DAW that has logged over 100 million downloads since 2015. Developers upload their trained PyTorch model to HuggingFace's Model Hub. The model becomes accessible through Audacity's UI and loads in a manner similar to traditional plugins.

III. MPF tools like Essentia(.js)



Similarity

Analyze audio and compute features to find similar sounds or music tracks.

Classification

Classify sounds or music based on computed audio features.

Deep learning inference

Use data-driven TensorFlow models for a wide range applications from music annotation to synthesis.

Mood detection

Find if a song is happy, sad, aggressive or relaxed.

Key detection

Find a key of a music piece.

Onset detection

Detect onsets (and transients) in an audio signal.

Segmentation

Split audio into homogeneous segments that sound alike.

Beat tracking

Estimate beat positions and tempo (BPM) of a song.

Melody extraction

Estimate pitch in monophonic and polyphonic audio.

Audio fingerprinting

Extract fingerprints from any audio source using the Chromaprint algorithm.

Cover song detection

Identify covers and different versions of the same music piece.

Spectral analysis

Analyze spectral shape of an audio signal.

Loudness metering

Use various loudness meters

Audio problems detection

Identify possible audio quality

Voice analysis

Voice activity detection and

Synthesis

Analyze, transform and

III. A Plug for nbdev

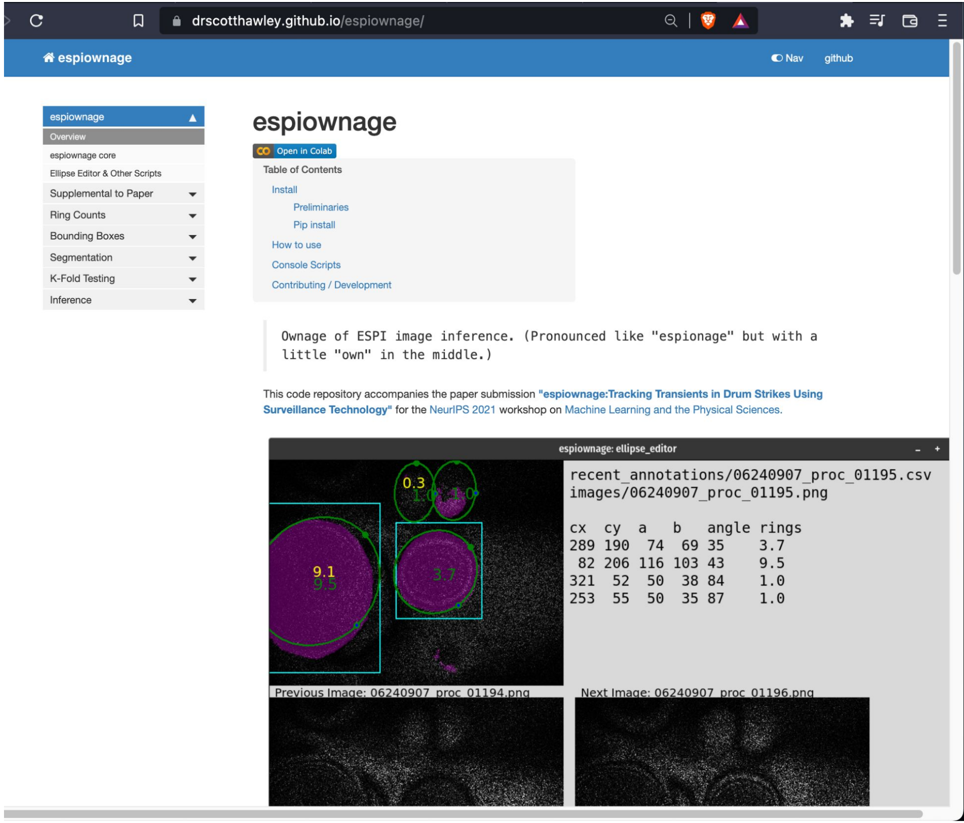
Python library development via Jupyter notebooks

By Jeremy Howard & fast.ai crowd, but doesn't require fastai

is "literate programming" instantiated: code, docs & tests *are one*

Built-in CI via GitHub Actions

Hawley & Morrison (JASA-EL 2022) found it ***essential* for** working efficiently (**staying sane**)



espiownage

Table of Contents

- Install
  - Preliminaries
  - Pip install
- How to use
- Console Scripts
- Contributing / Development

> Ownage of ESPI image inference. (Pronounced like "espionage" but with a little "own" in the middle.)

This code repository accompanies the paper submission **"espiownage:Tracking Transients in Drum Strikes Using Surveillance Technology"** for the NeurIPS 2021 workshop on Machine Learning and the Physical Sciences.

espiownage: ellipse_editor

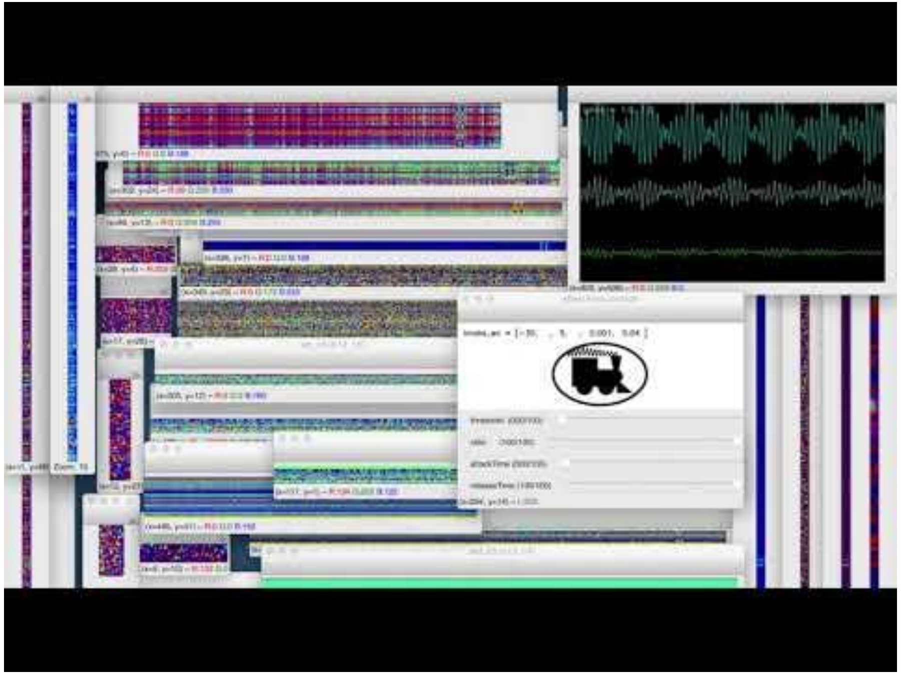
```
recent_annotations/06240907_proc_01195.csv
images/06240907_proc_01195.png

cx  cy  a   b   angle rings
289 190  74  69 35    3.7
 82 206 116 103 43    9.5
321  52  50  38 84    1.0
253  55  50  35 87    1.0
```

Previous Image: 06240907_proc_01194.png

Next Image: 06240907_proc_01196.png

III. WE ALL SHOULD BUILD MORE

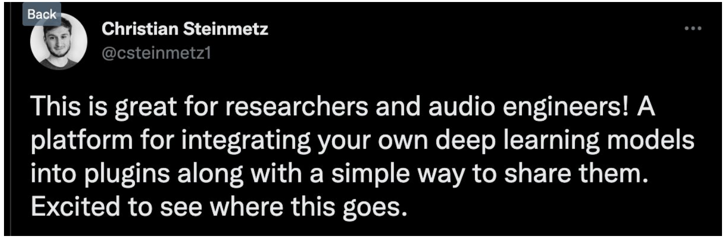


My attempt at interactive viz of PyTorch layer activations: images + oscilloscope.

To some student: *Please* fork this and make it *your own* & make it *good*.

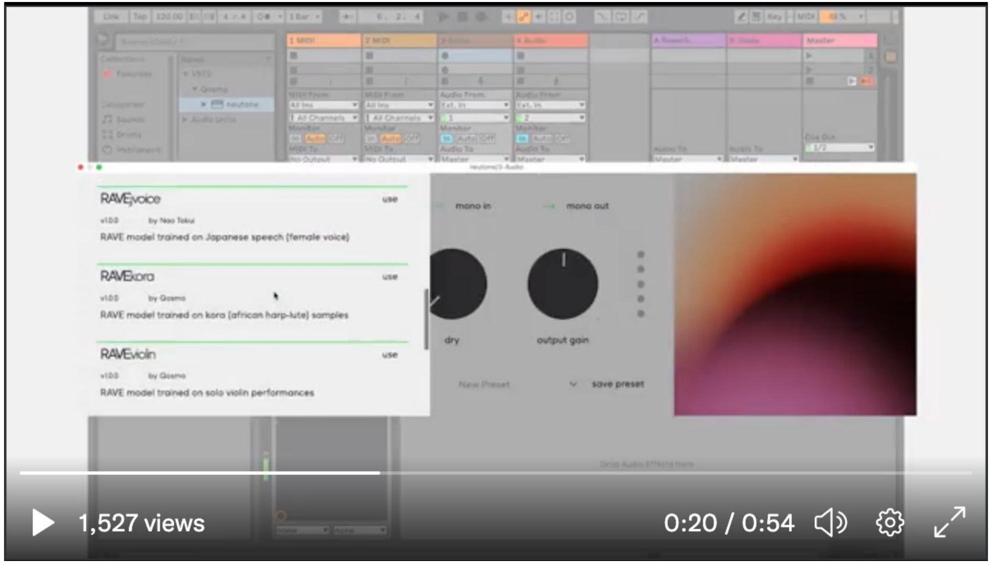
III. Post-Talk: New One! Neutone by Quosmo

"AI audio plugin & community, Bridging the gap between AI research and creativity"



Christian Steinmetz
@csteinmetz1

This is great for researchers and audio engineers! A platform for integrating your own deep learning models into plugins along with a simple way to share them. Excited to see where this goes.



RAVEjvoice
v1.0.0 by Nao Tokui
RAVE model trained on Japanese speech (female voice)

RAVEkora
v1.0.0 by Qosmo
RAVE model trained on kora (african harp-lute) samples

RAVEviolin
v1.0.0 by Qosmo
RAVE model trained on solo violin performances

For audio creators

Neutone makes AI technologies accessible for all to experiment with. You'll find transformative AI audio instruments that will spark endless creative possibilities.

For AI researchers

Neutone is a go-to platform for you to share real-time AI audio processing models with potential users in the audio production community.

Download Plugin

Submit Your Model