Robust Sound Event Detection in Acoustic Sensor Networks

Justin Salamon

Senior Research Scientist
Music & Audio Research Lab / Center for Urban Science & Progress
New York University

Work by MARL, SONYC & BirdVox

 www.justinsalamon.com  @justinsalamon  justin.salamon@gmail.com

With support from:

NYU Music and Audio Research Lab (MARL)

Acoustic Sensor Networks

SONYC

BIRDVOX

Anish Arora
Co-PI

Carlos Bautista
App Developer

Juan Pablo Bello
Lead-PI

Mark Cartwright
Postdoc

Jason Cramer
Masters Student

Yurii Piadyk
PhD Student

Dhrubojyoti Roy
PhD Student

Justin Salamon
Senior Research
Scientist

Ayanna Seals
PhD Student

Mohit Sharma
Asst. Research
Scientist

Harish
Doraiswamy
Res. Asst. Prof.

Graham Dove
Postdoc

R. Luke DuBois
Co-PI

Ben Esner
Senior Personnel

Peter Li
PhD Student

Claudio Silva
Co-PI

Yu Wang
PhD Student

Ho-Hsiang Wu
PhD Student

Yitzchak
Lockerman
Postdoc

Ana Elisa Mendez
PhD Student

Fabio Miranda
PhD Student

Charlie Mydlarz
Senior Research
Scientist

Oded Nov
Co-PI

BirdVox team

Juan Pablo Bello
NYU

Andrew Farnsworth
Cornell

Steve Kelling
Cornell

Vincent Lostanlen
Cornell / NYU

Kendra Oudyk
McGill

Justin Salamon
NYU

Why should we care about noise?

amNEWYORK NEWS POLITICS THINGS TO DO EAT & DRINK ENTERTAINMENT REAL ESTATE TRANSIT OPINION

Washington Heights is among the city's noisiest neighborhoods, according to a recent study of 311 noise complaints. *(Credit: Linda Rosier)*

NEWS

311 noise complaints on the rise in Washington Heights, Inwood

By Lisa L. Colangelo lisa.colangelo@amny.com *January 30, 2018*

New Yorkers are making a lot of noise about noise.

There were 1.6 million noise complaints made to 311 between 2010 and 2015, according to a new report released Monday by State Comptroller Thomas DiNapoli.

"Noise in New York City is a significant quality of life and public health concern," DiNapoli said. "The city ha
noise code and should be commended for taking steps to better enforce local law, but there is more that ci

includ Washington Heights an
(/new-york-kids/sea and C munity Boards 4 and 5

Construction noise complaints build up in the city

BY GLENN BLAIN

NEW YORK DAILY NEWS Thursday, August 31, 2017, 2:20 PM

Time Out New York

NEWS | CITY LIFE

0 ADD COMMENT LOVE IT SAVE I

Yikes! NYC's noisiest neighborhoods are no place for exhausted parents

By **Danielle Valente**
Posted: Wednesday January 31 2018, 12:31pm

The New York Times https://nyti.ms/2vBlsjh

N.Y. / REGION

New York Becomes the City That Never Shuts Up

By WINNIE HU JULY 19, 2017

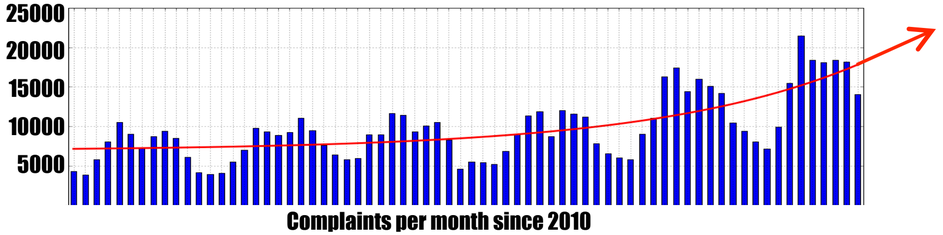
Richard T. McIntosh has never heard such a racket outside his window.

Traffic roars through his neighborhood on the Upper East Side of Manhattan at all hours. The whine of refrigerated grocery trucks by the curb makes things worse. And construction of a new apartment tower across the street forces him to flee his own home. There is the deafening rat-a-tat of jack hammers and the incessant banging and high-pitched wail of construction equipment that echoes in his head.

"I've had two years of absolute violation of my right to peace and quiet," said Mr. McIntosh, a television producer who has lived on the Upper East Side for more than five decades. "I think it's against the Geneva Conventions to have this much noise."

Estimated 9 of 10 adults in NYC exposed to HARMFUL levels of NOISE

Over 3.4 MILLION complaints since 2003 [based on 311 data]



SLEEP LOSS

HEARING LOSS

PRODUCTIVITY

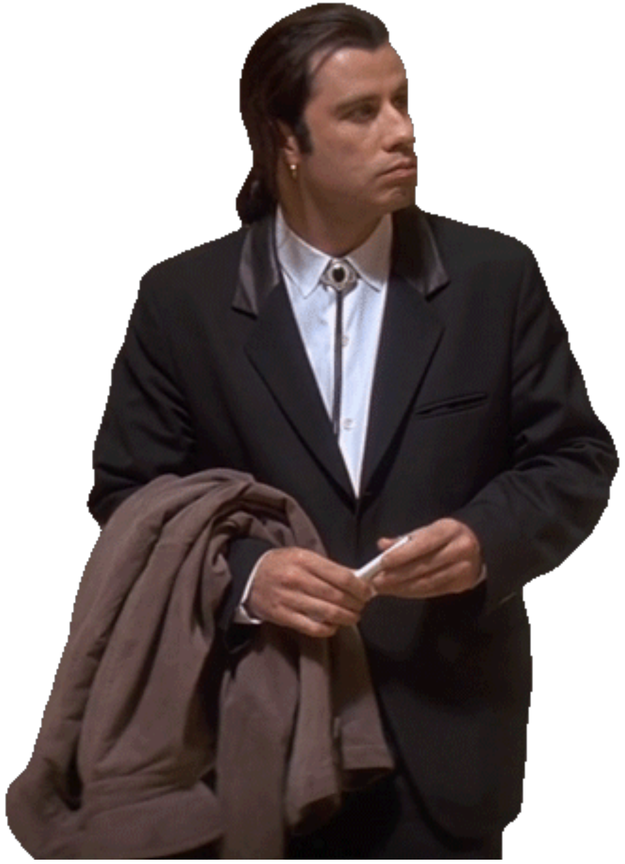
LEARNING IMPAIRMENT

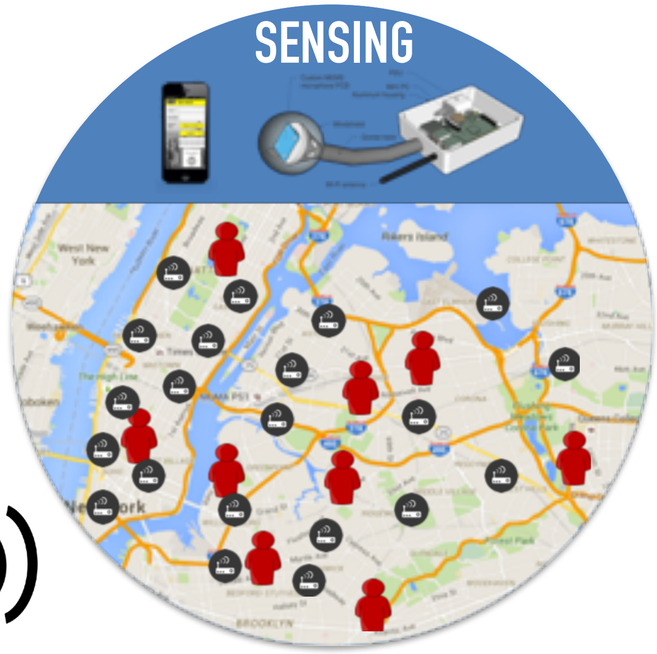
STRESS

HOW DOES THE CITY CURRENTLY TACKLE NOISE?

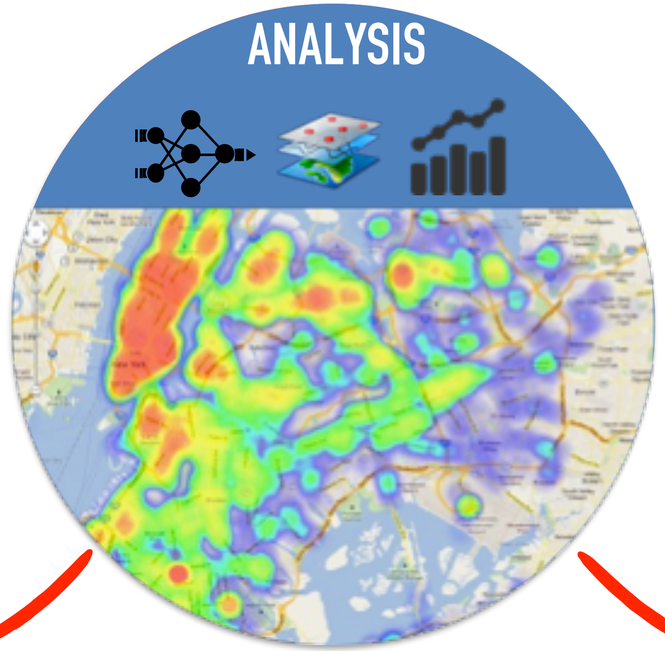
HOW DOES THE CITY CURRENTLY TACKLE NOISE?

HOW DOES THE CITY CURRENTLY TACKLE NOISE?





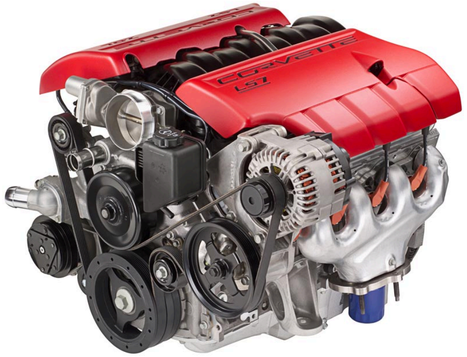
SONYC



Empower Research

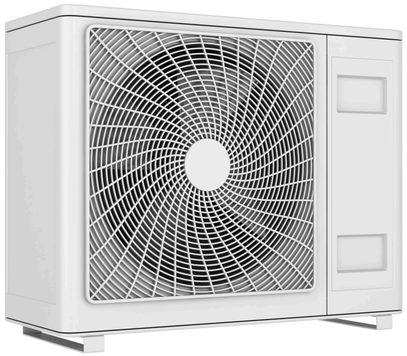
Empower Citizens

SONYC


CORVETTE
LS1


BOSCH


PRD

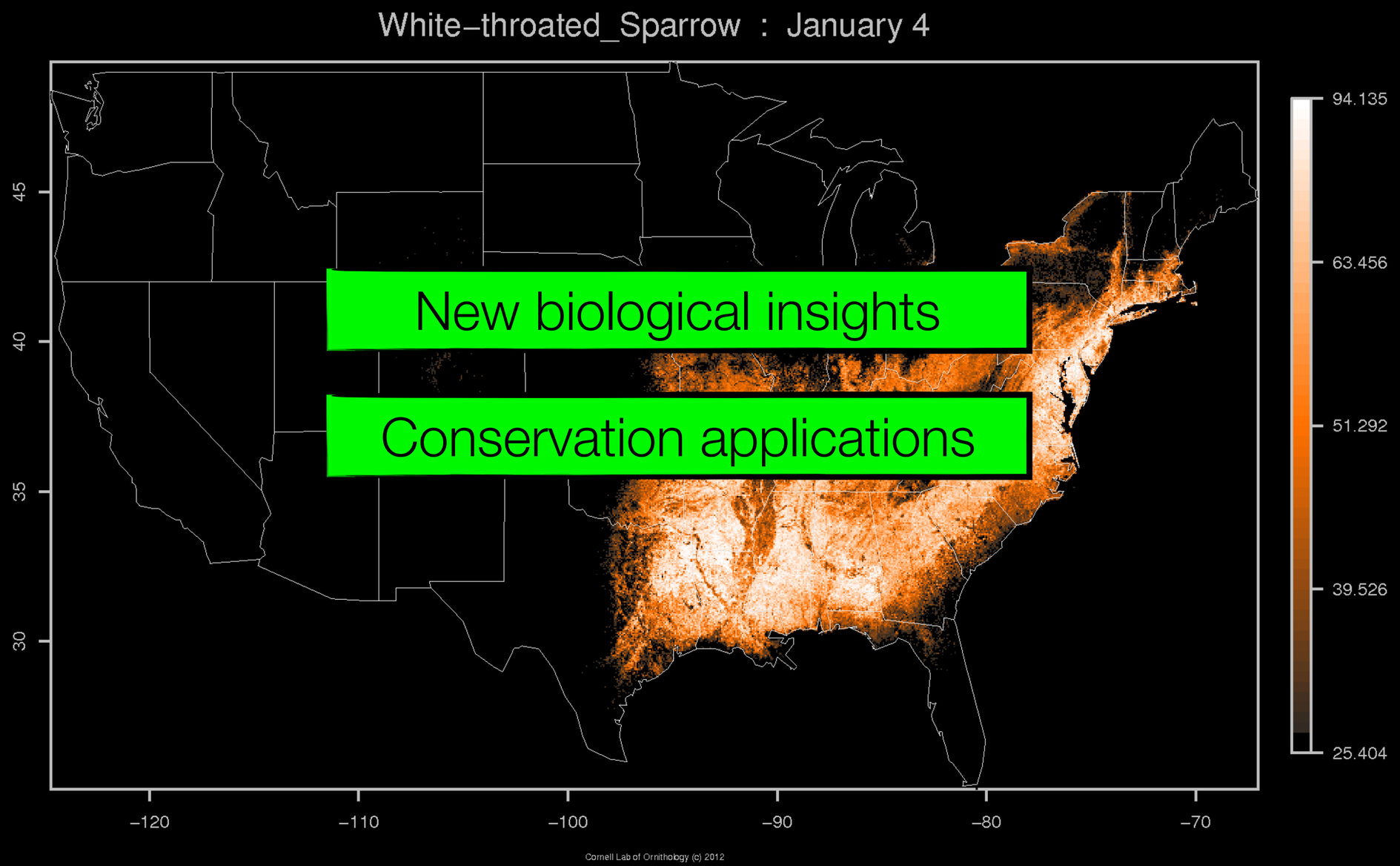



metabo



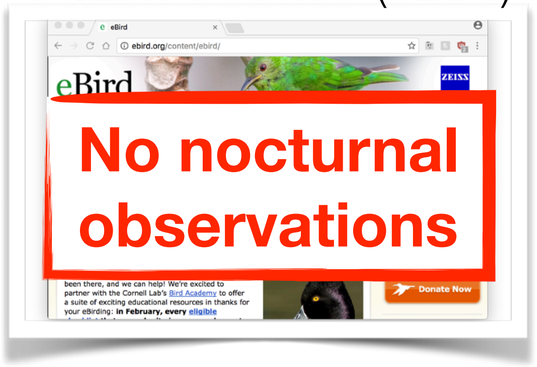



Why do we care about classifying bird vocalizations?



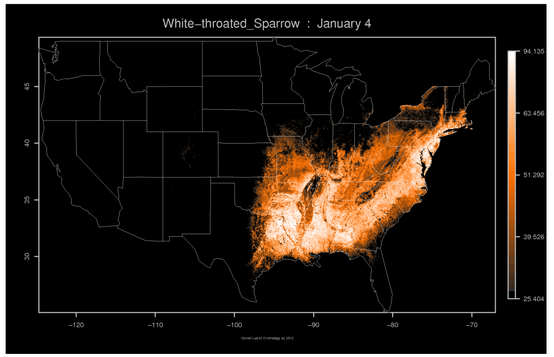
The migration monitoring puzzle

Citizen Science (eBird)

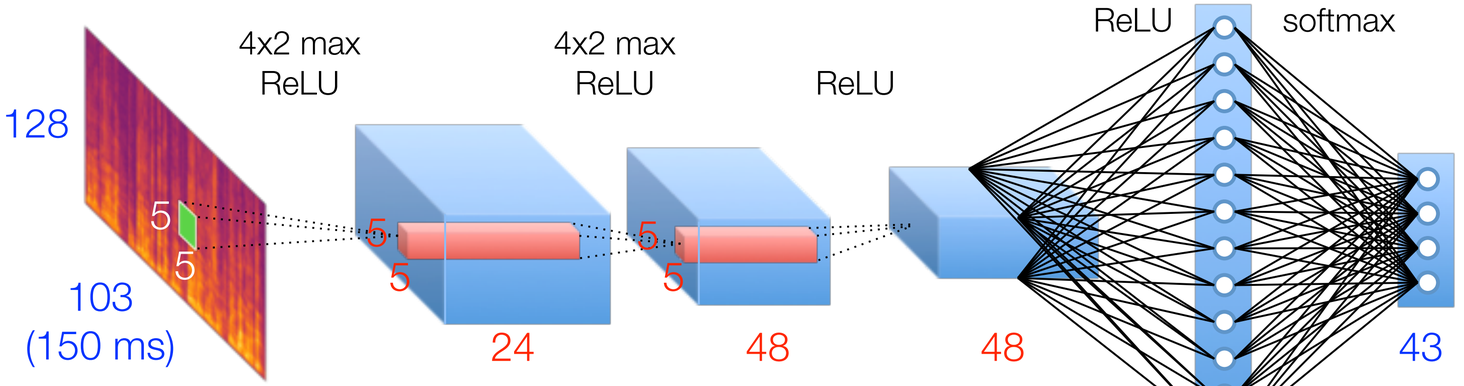


Radar

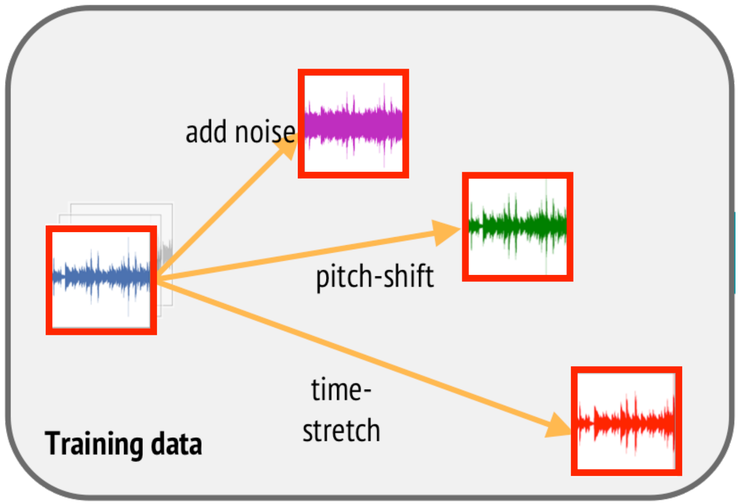
Audio!

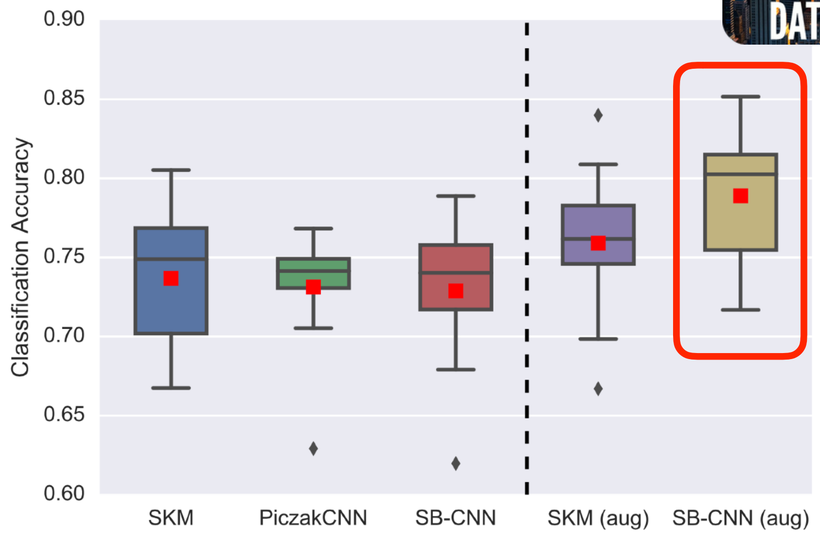


Previously on SONYC…

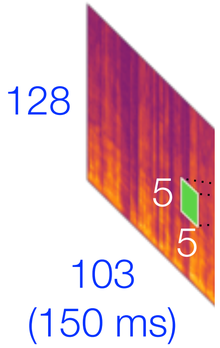


+

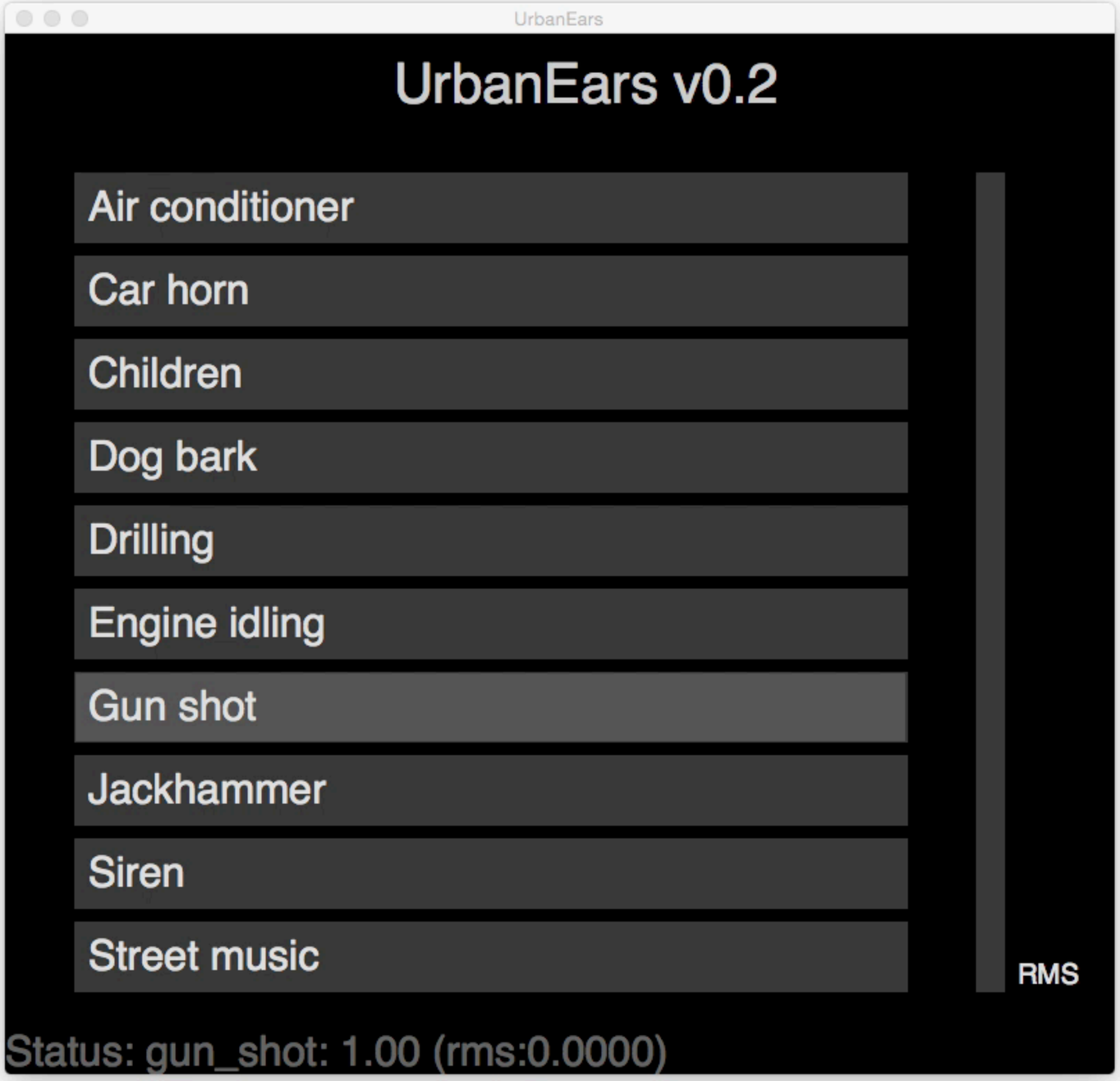




Previc


128
5
5
103
(150 ms)


Training data


UrbanEars
UrbanEars v0.2
Air conditioner
Car horn
Children
Dog bark
Drilling
Engine idling
Gun shot
Jackhammer
Siren
Street music
RMS
Status: gun_shot: 1.00 (rms:0.0000)


SOUND
ASETS

2 Years Later…

- BirdVox: over 6,000 hours of audio
- SONYC: over 26 years of audio
- New data, new challenges:

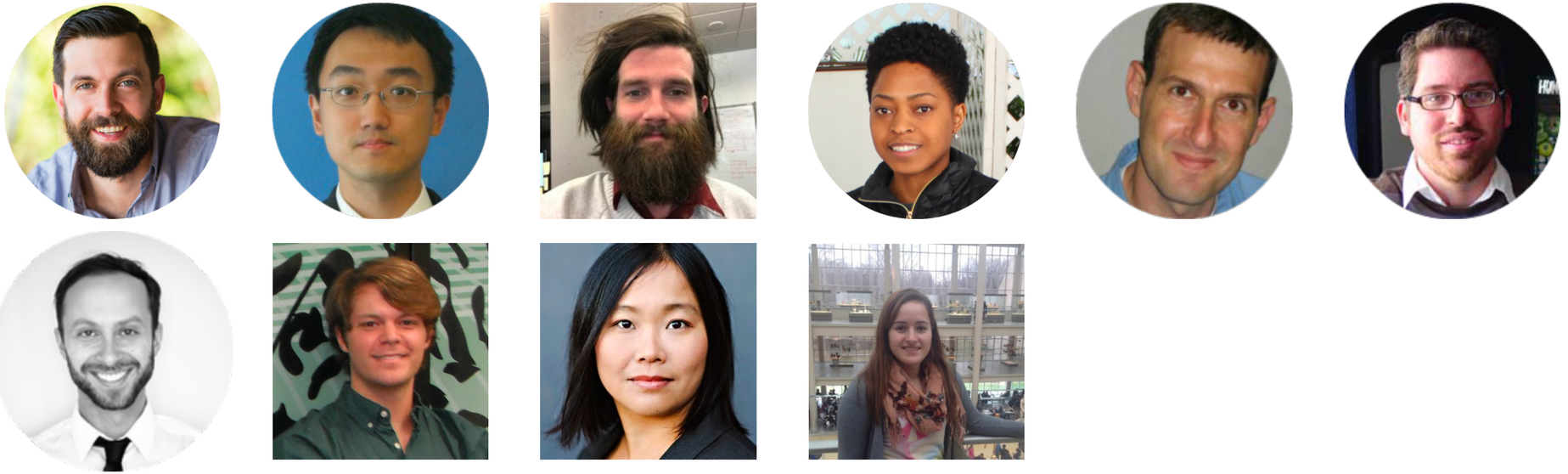
How do we label the data?

How can we best leverage the labeled data?

How can we leverage the unlabeled data?

How do we make our models robust to different sensor locations?

How do we label the data?



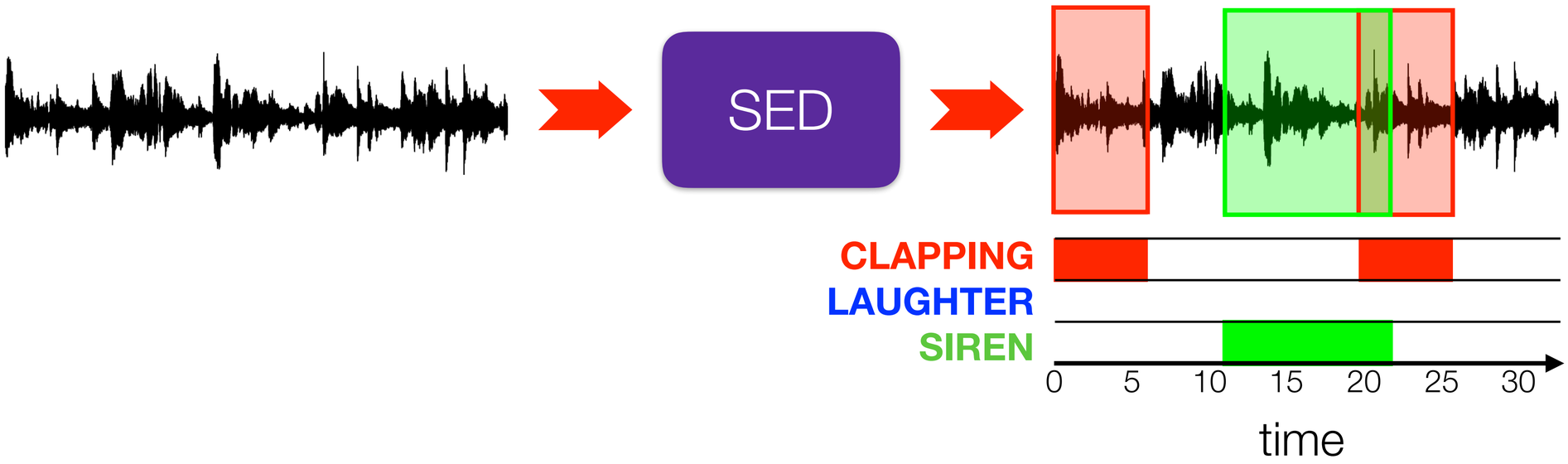
How do we label the data?

Scaper: A Library for Soundscape Synthesis and Augmentation
J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello.
IEEE Workshop on Applications of Sig. Proc. to Audio and Acoustics (WASPAA), New Paltz, NY, USA, Oct. 2017.

Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations
M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J. Bello, and O. Nov.
Proceedings of the ACM on Human-Computer Interaction, 1(2), 2017.

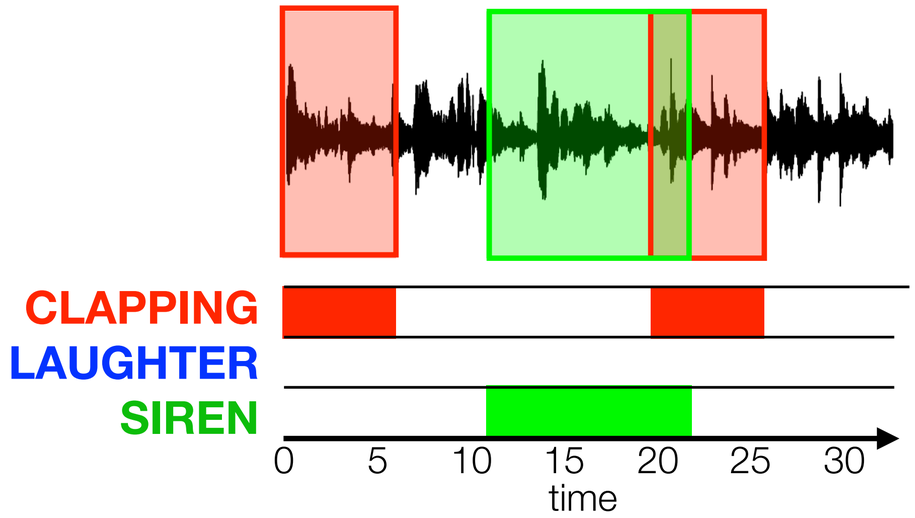
Investigating the Effect of Sound-Event Loudness on Crowdsourced Audio Annotations
M. Cartwright, J. Salamon, A. Seals, O. Nov, and J. P. Bello
In IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, Canada, Apr. 2018.

Sound Event Detection (SED)

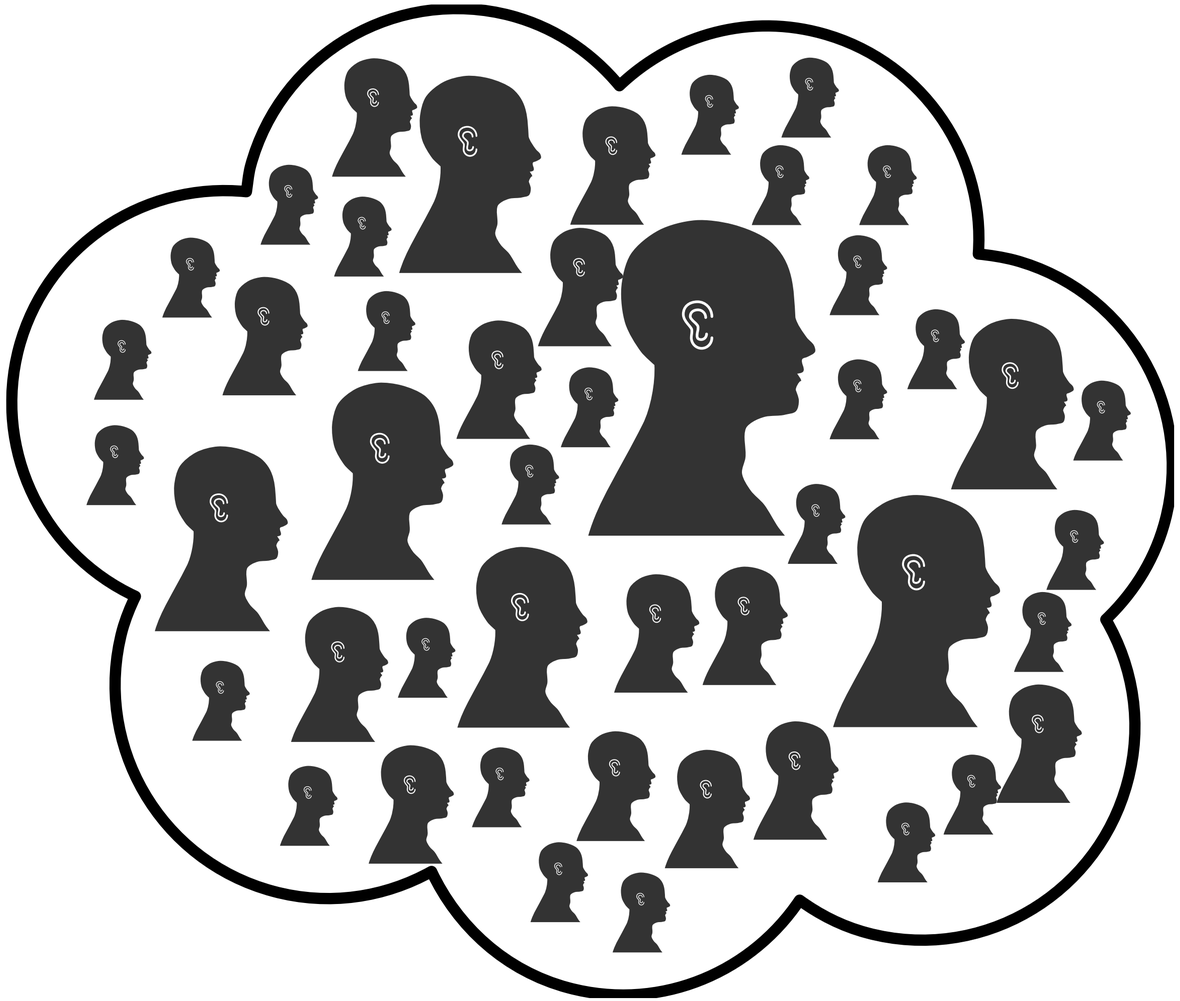


“Strong” sound event labels

- **Strong** labels:
  - Label (sound source)
  - Start time
  - End time



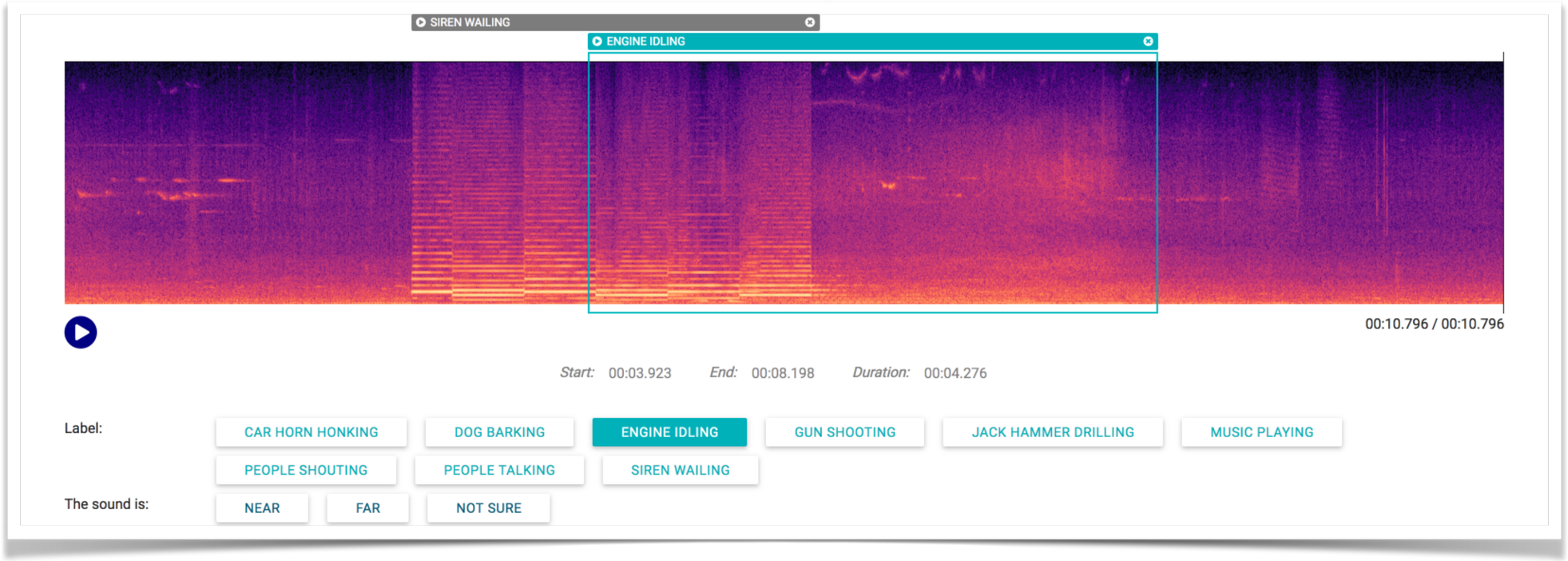
Crowdsourcing



The Audio Annotator

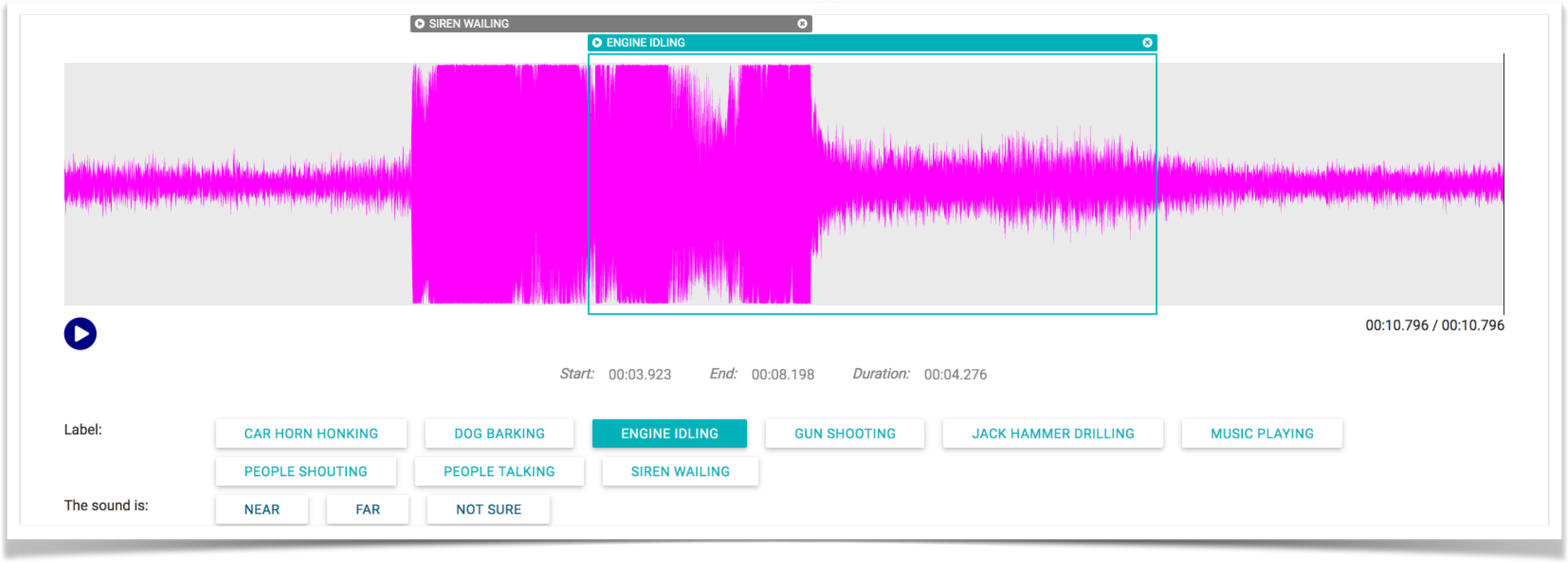
- github.com/CrowdCurio/audio-annotator

Configured with the spectrogram visualization:



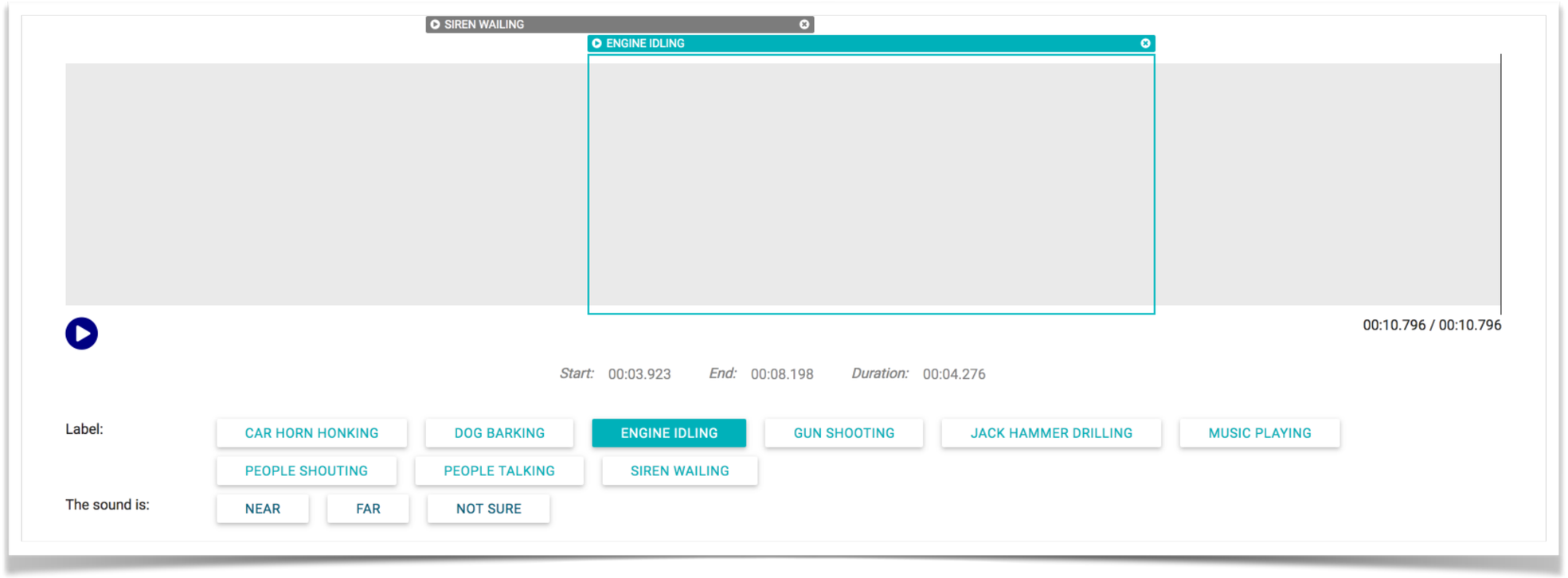
The Audio Annotator

Configured with the waveform visualization:



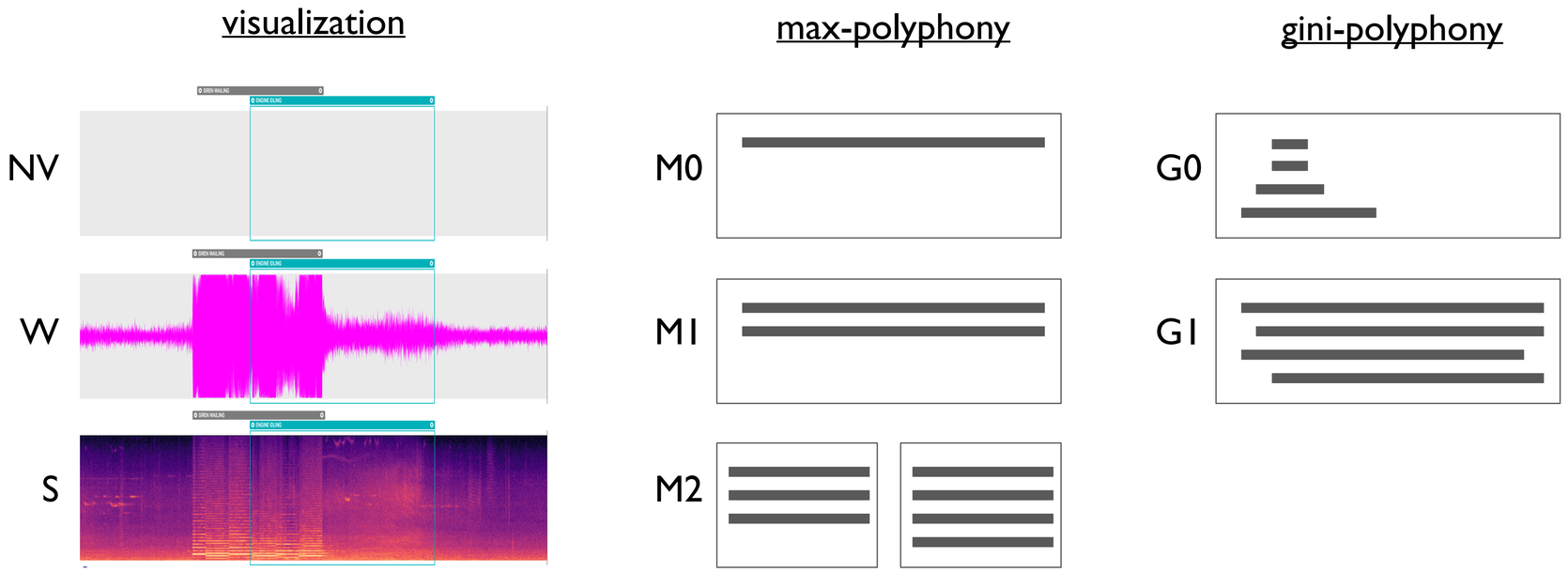
The Audio Annotator

Configured without a visualization:



Experiment

- 3 x 3 x 2 between-subjects factorial design:



- Soundscape examples:
  M0G0 M0G1 M2G0 M2G1

Autio stimuli + labels: Scaper

- Open source python library for soundscape synthesis
- Returns audio + annotation in JAMS format
- github.com/justinsalamon/scaper (pip install scaper)

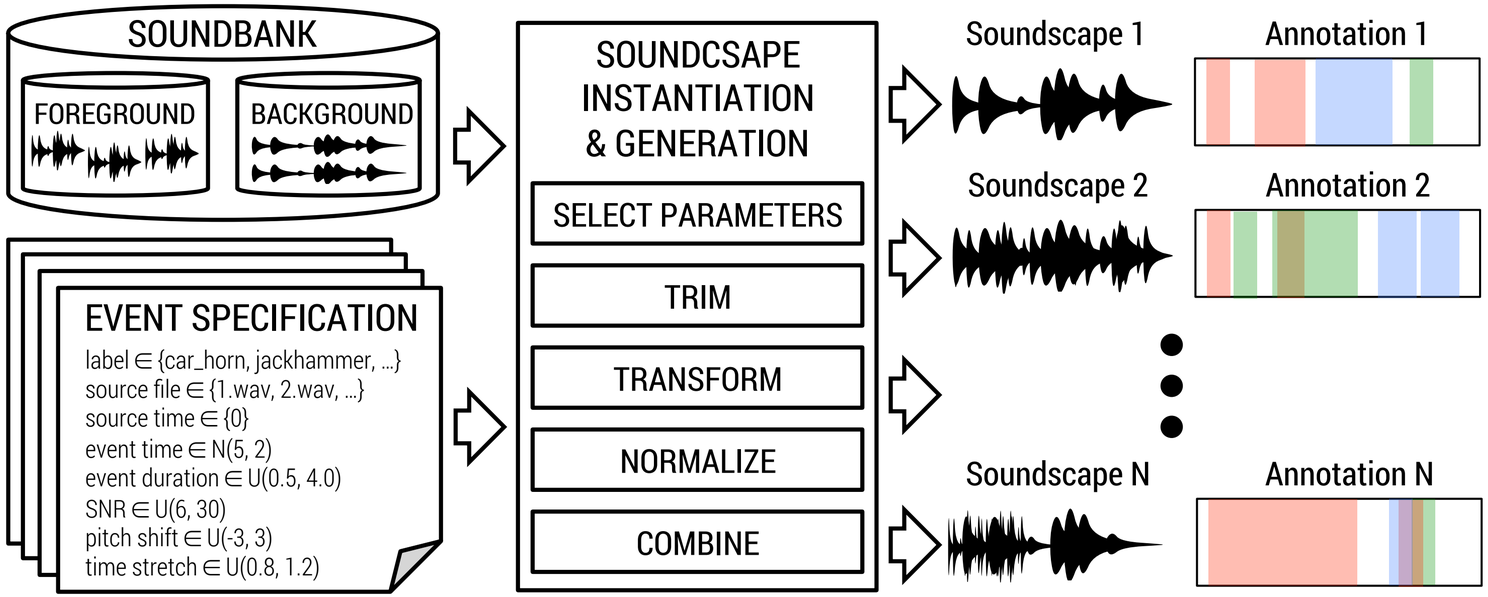
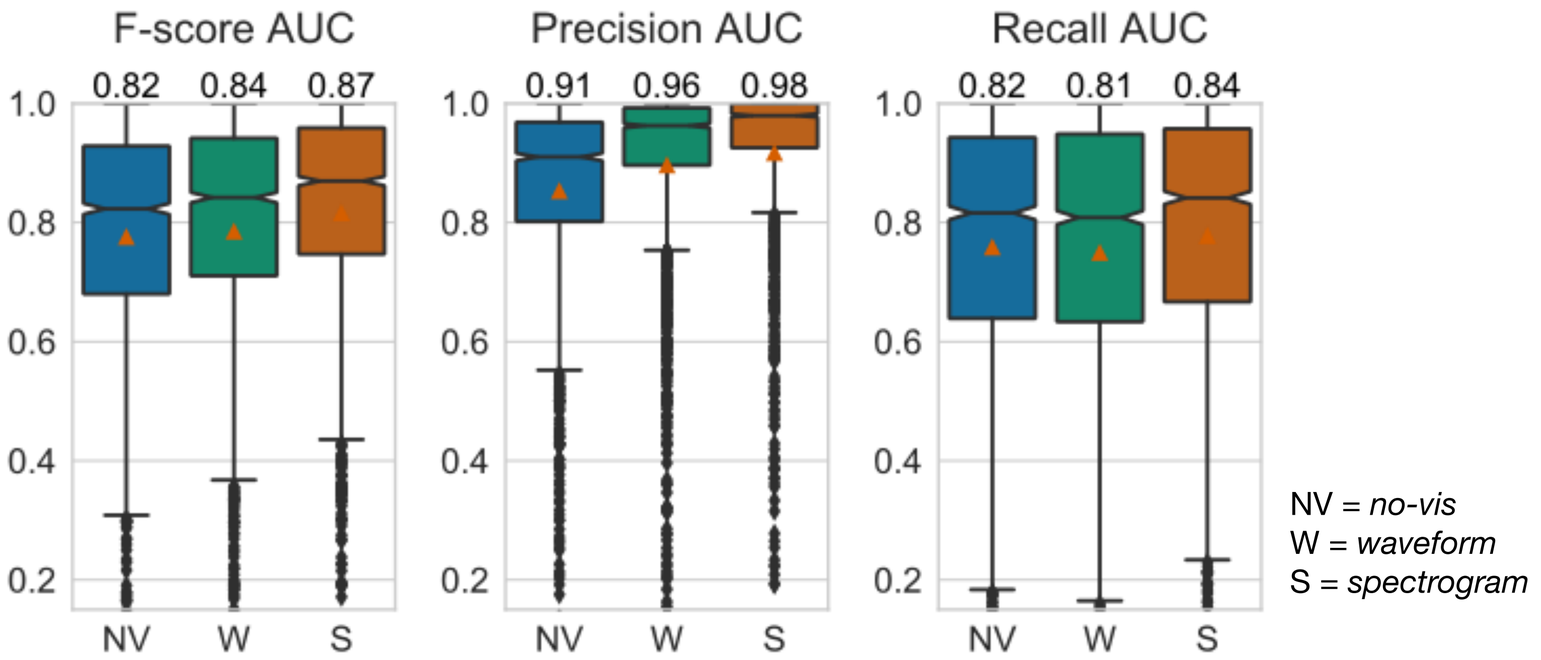


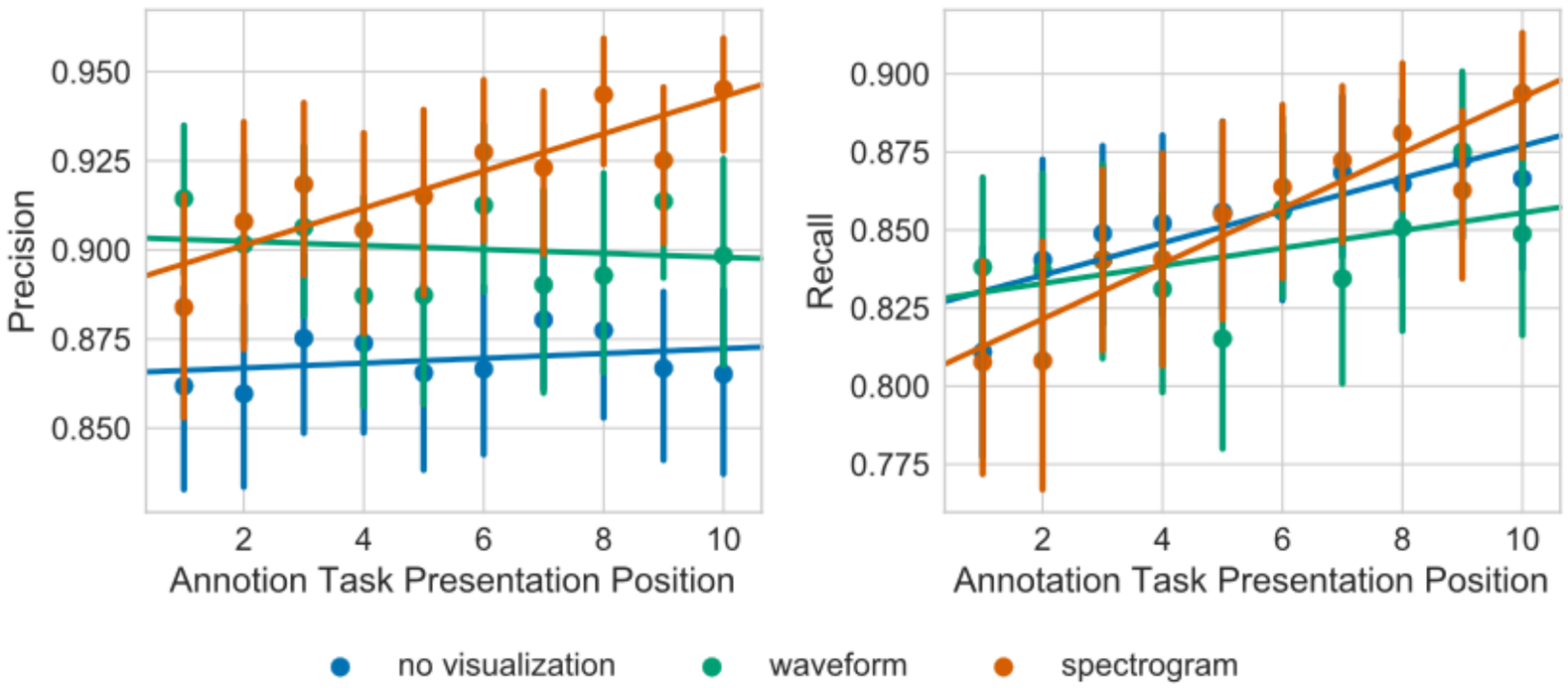
Figure 1: Block diagram of the Scaper synthesis pipeline.

Effect of Visualization on Quality and Speed of Annotations



Spectrogram → higher-quality and faster annotations

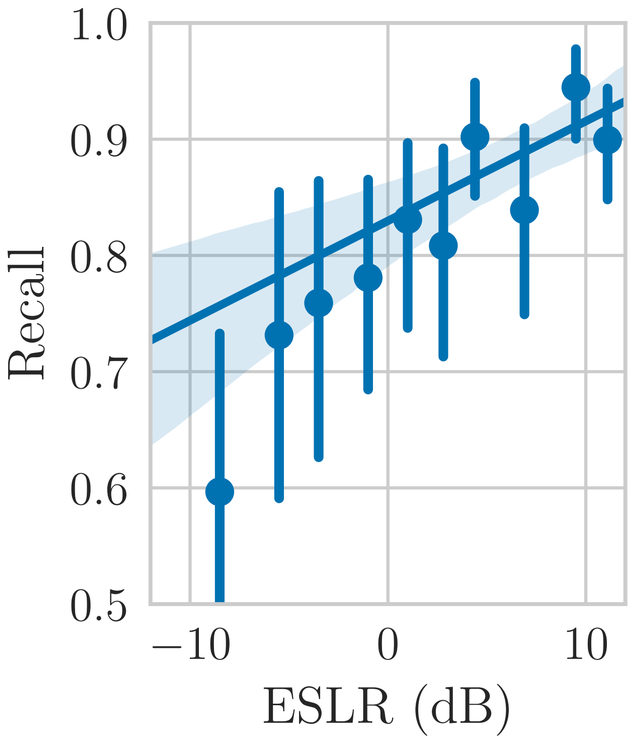
Effect of Visualization on Task Learning

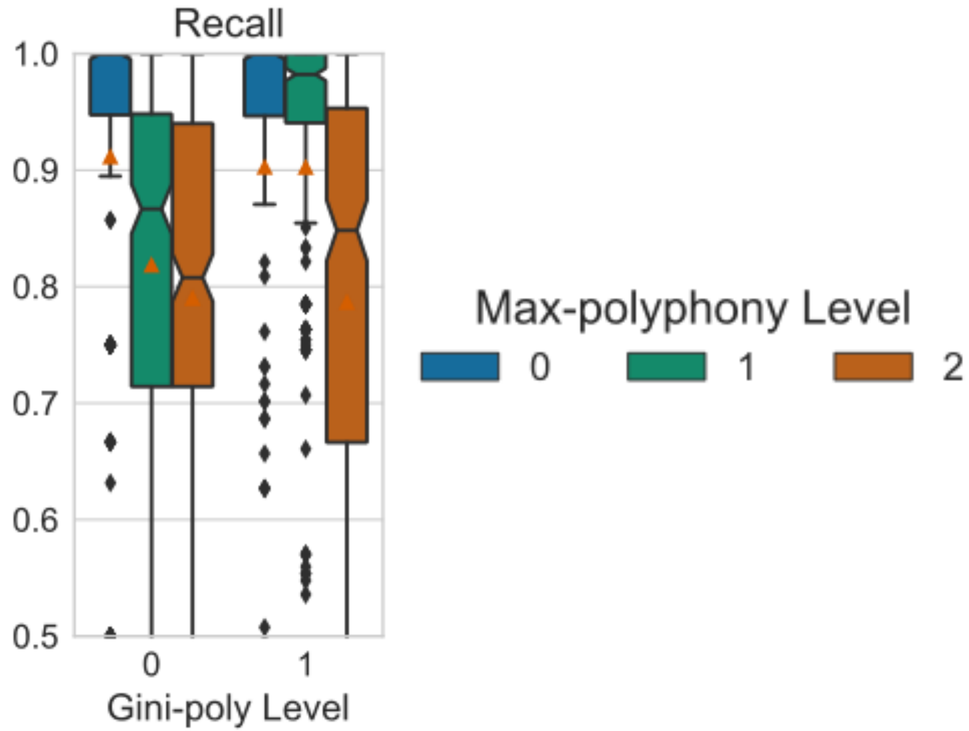


Expect even higher quality annotations after learning period

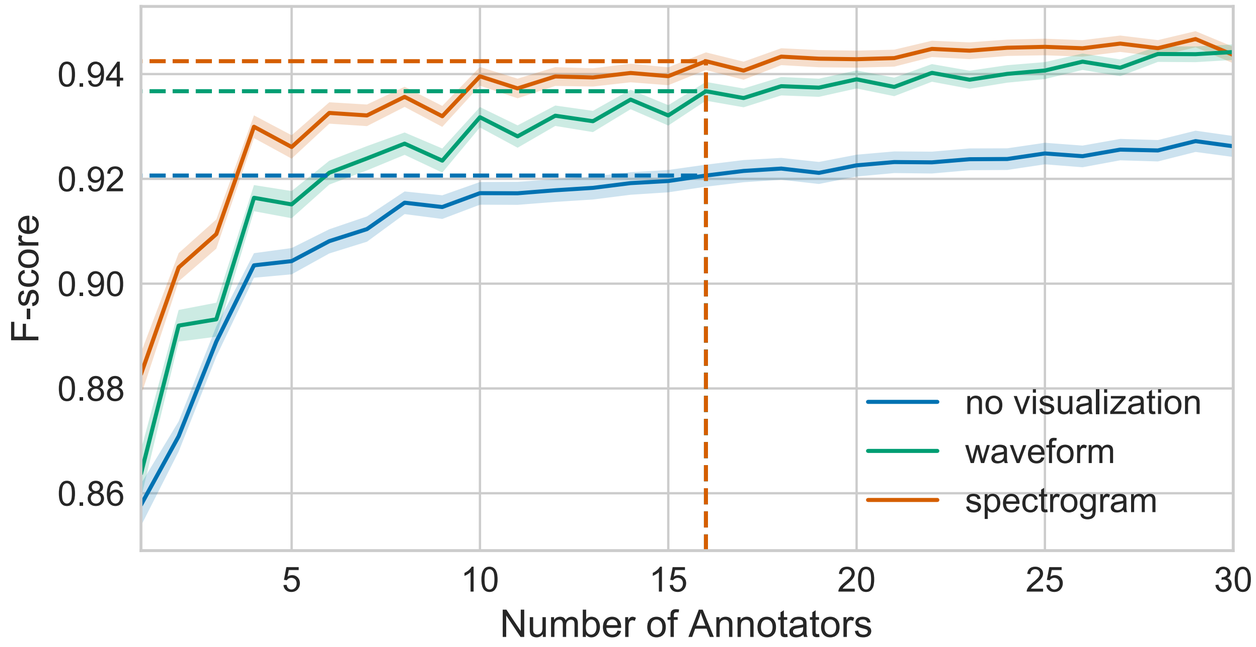
Effect of sound event loudness & scene complexity

Sound event SNR and overlap (polyphony) have direct impact on label recall, precision remains stable



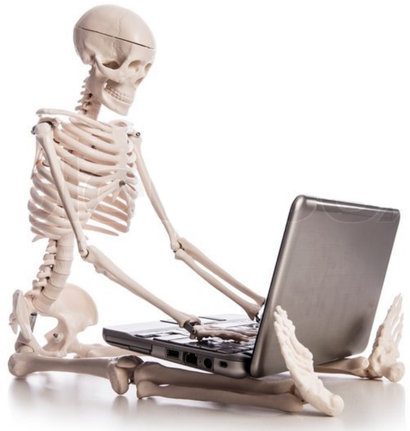


Effect of Number of Annotators on Aggregate Annotation Quality



16 annotators captured 90% of gain in annotation quality, but 5 annotators is reasonable choice with respect to cost/quality trade-off

Annotating strong labels is slow…



How can we best leverage the labeled data?

(can we get away without strong labels?)

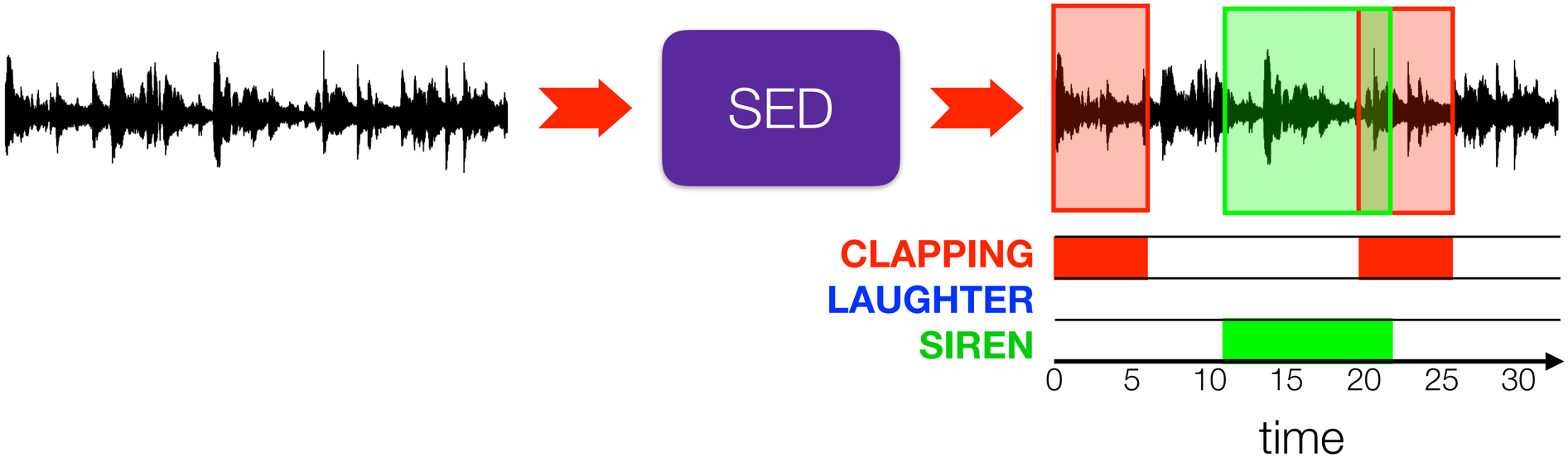


How can we best leverage the labeled data?

(can we get away without strong labels?)

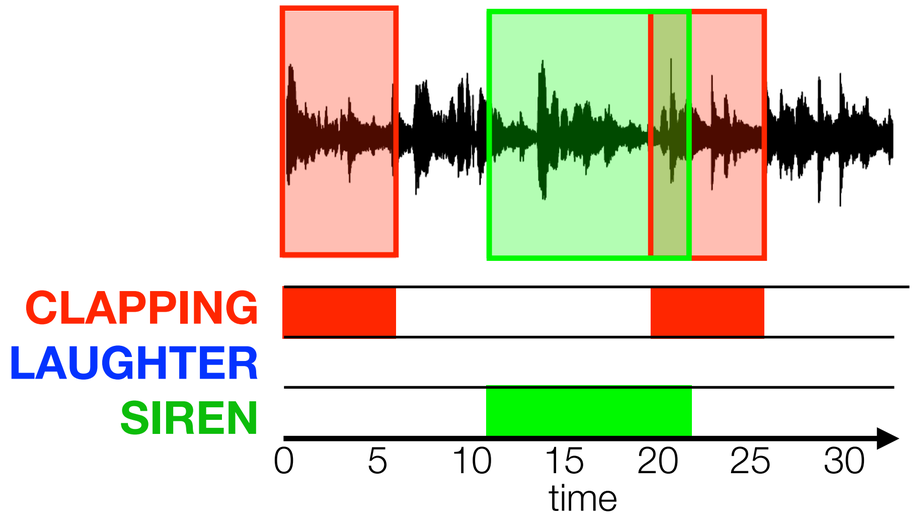
Adaptive pooling operators for weakly labeled sound event detection
B. Mcfee, J. Salamon, and J. P. Bello
IEEE/ACM Transactions on Audio, Speech and Language Processing, 26(11): 2180–2193, Nov. 2018.

Sound Event Detection (SED)

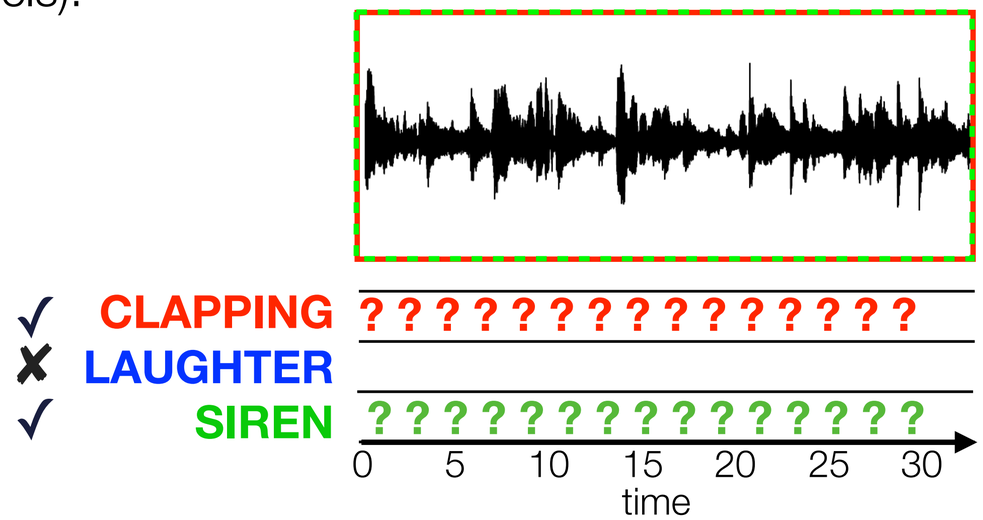


Strong vs. weak labels

- **Strong** labels:
  - Label (sound source)
  - Start time
  - End time



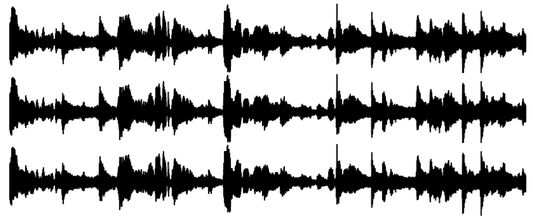
- **Weak** labels (tags, e.g. AudioSet labels):
  - Label (sound source)
  - No timing information
  - **Easier/faster/cheaper to curate!**



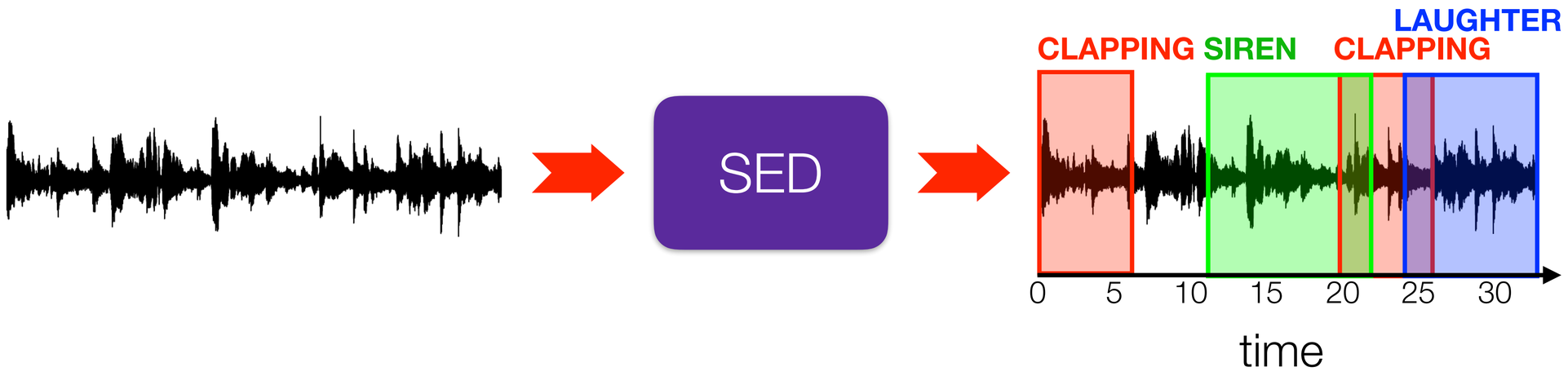
Problem formulation

- Given dataset of tracks with **weak labels**:

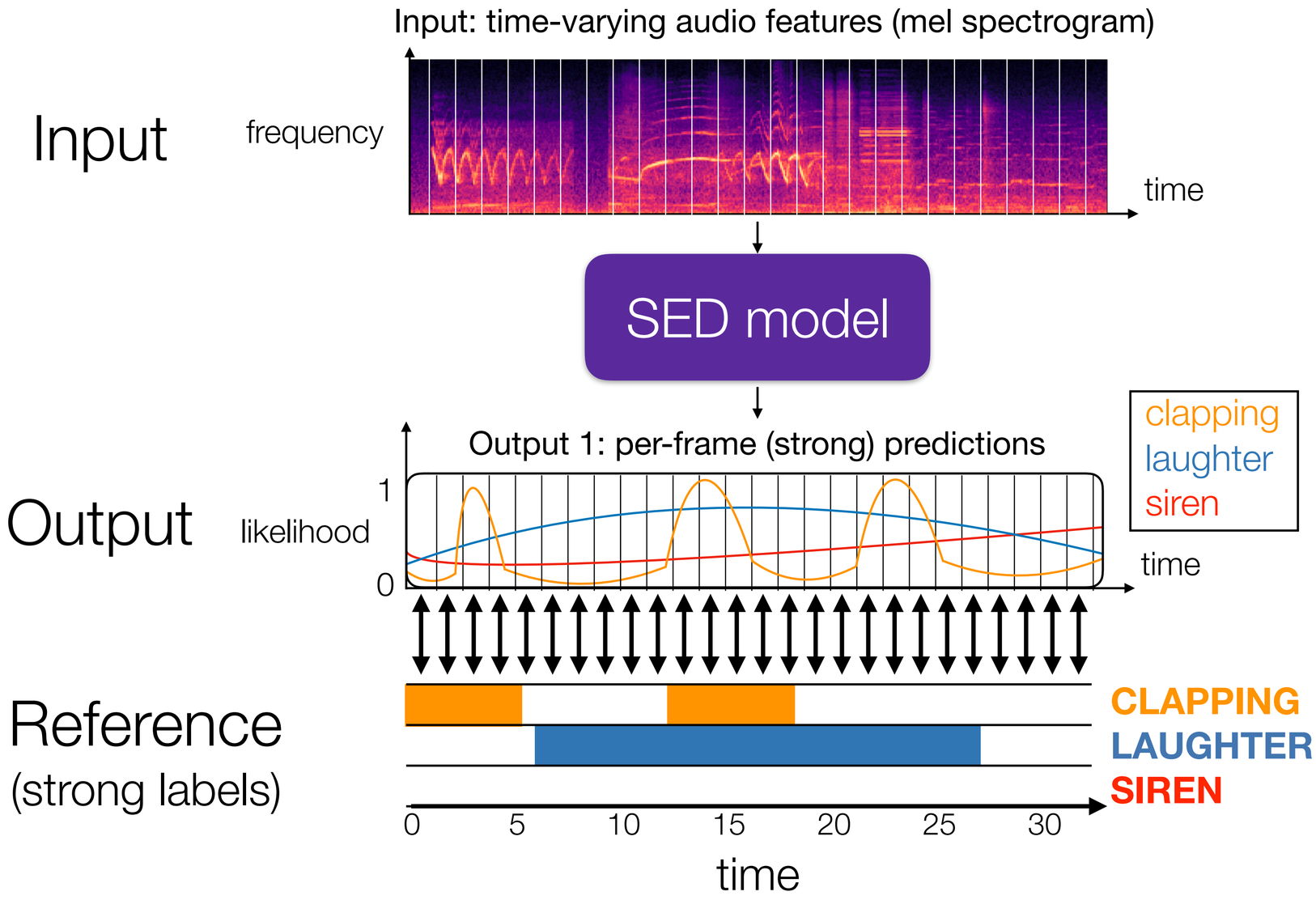
[CLAPPING, SIREN]
[CLAPPING, CAR HONK, SIREN, LAUGHTER]
[LAUGHTER]



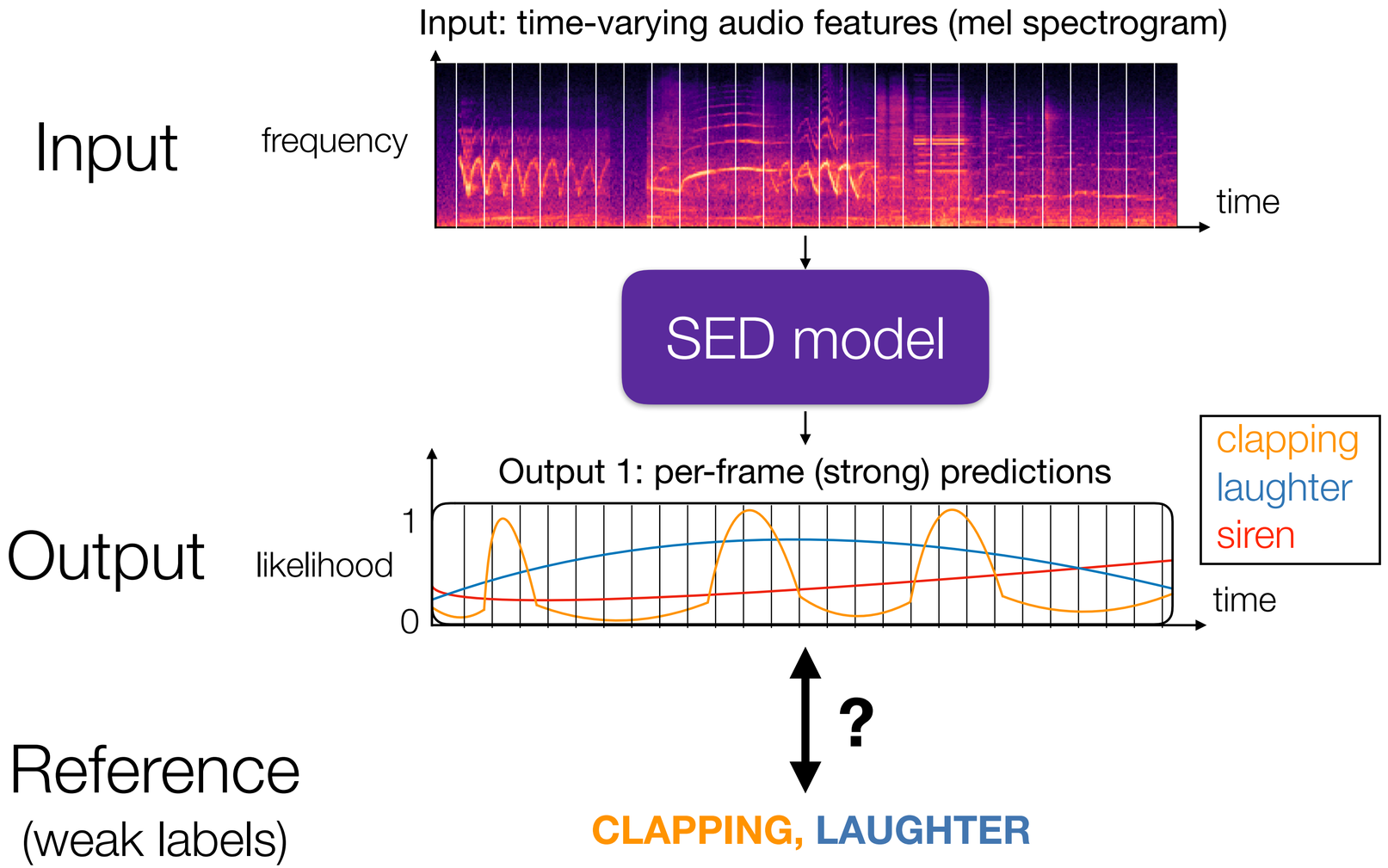
- Train a SED model that outputs **strong labels**:



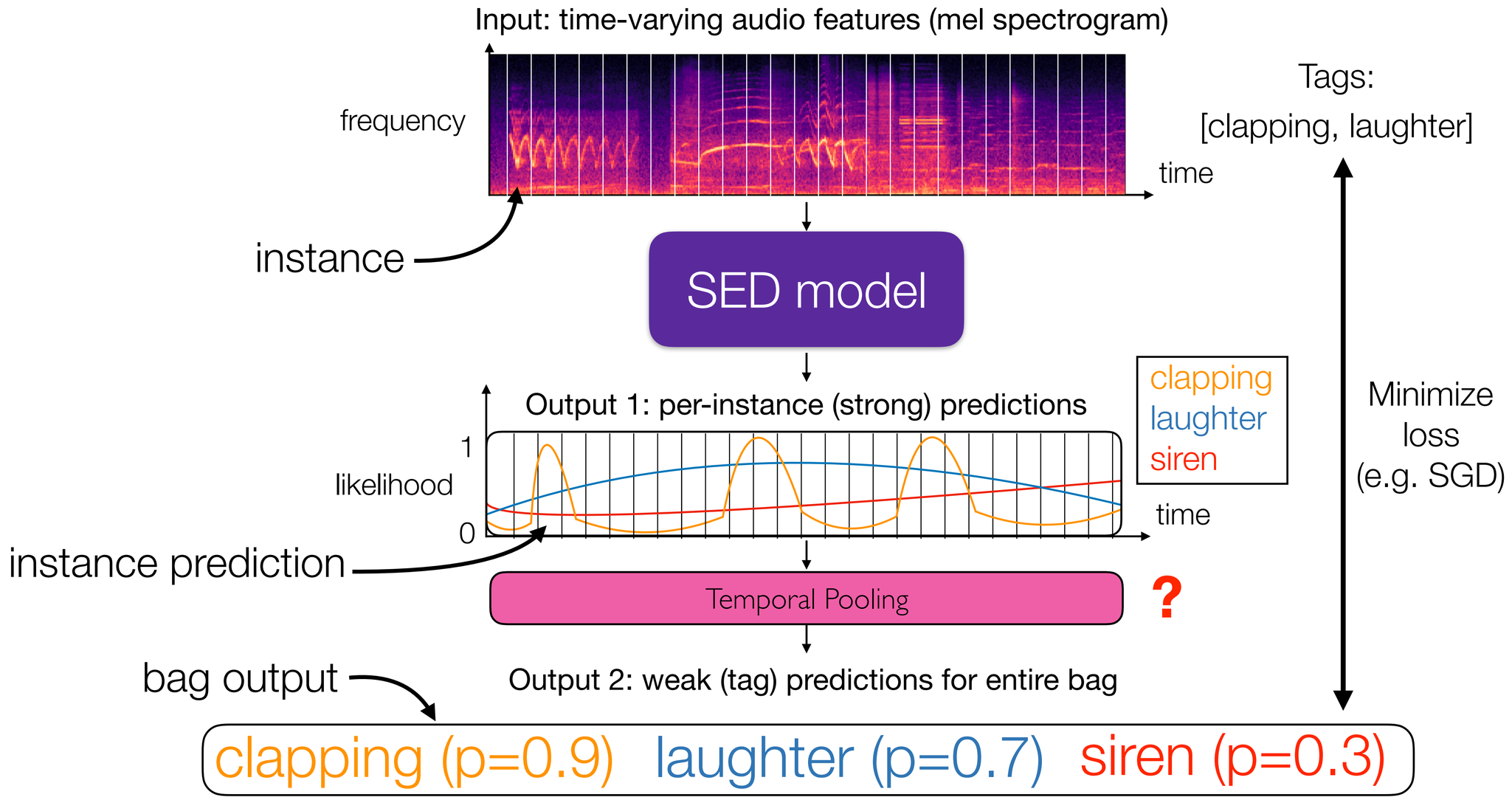
Sound Event Detection: If we had strong labels



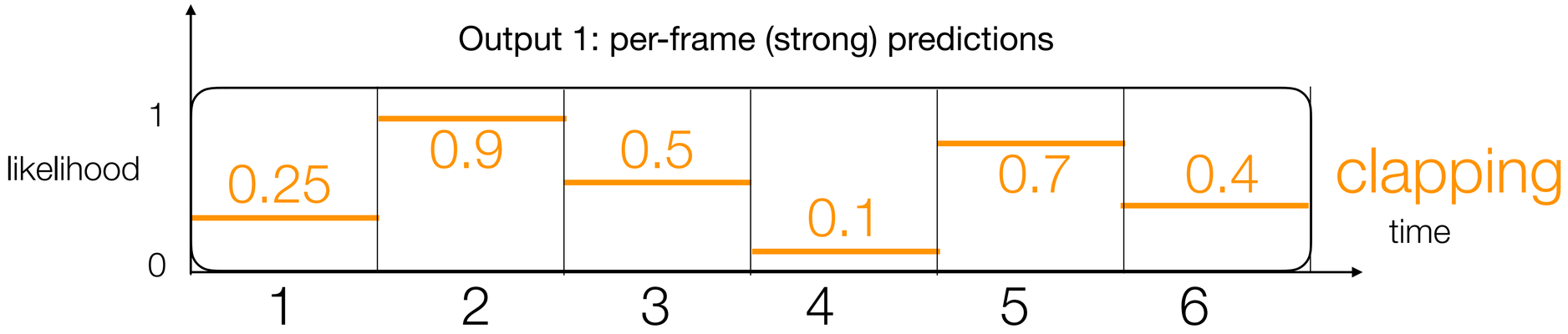
Sound Event Detection: If we had strong labels



Multiple Instance Learning for Sound Event Detection



Training a MIL model for Sound Event Detection

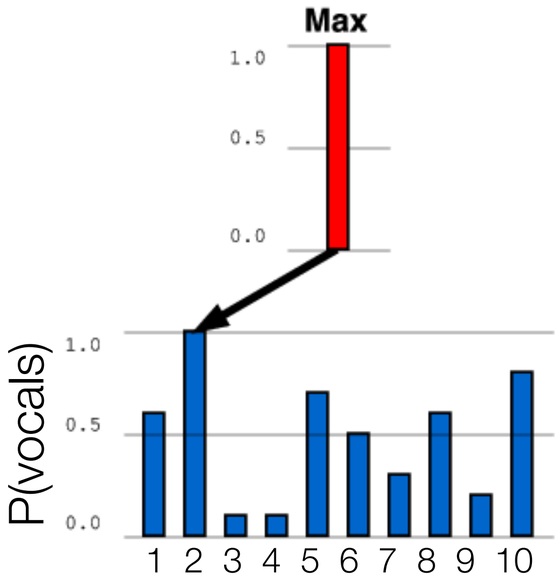


x = [0.25, 0.9, 0.5, 0.1, 0.7, 0.4]

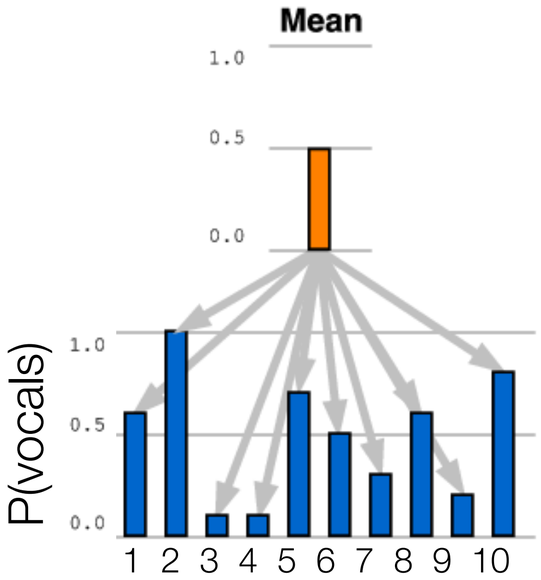
f(x) ?

P(clapping anywhere in track) = 0.9

Temporal pooling

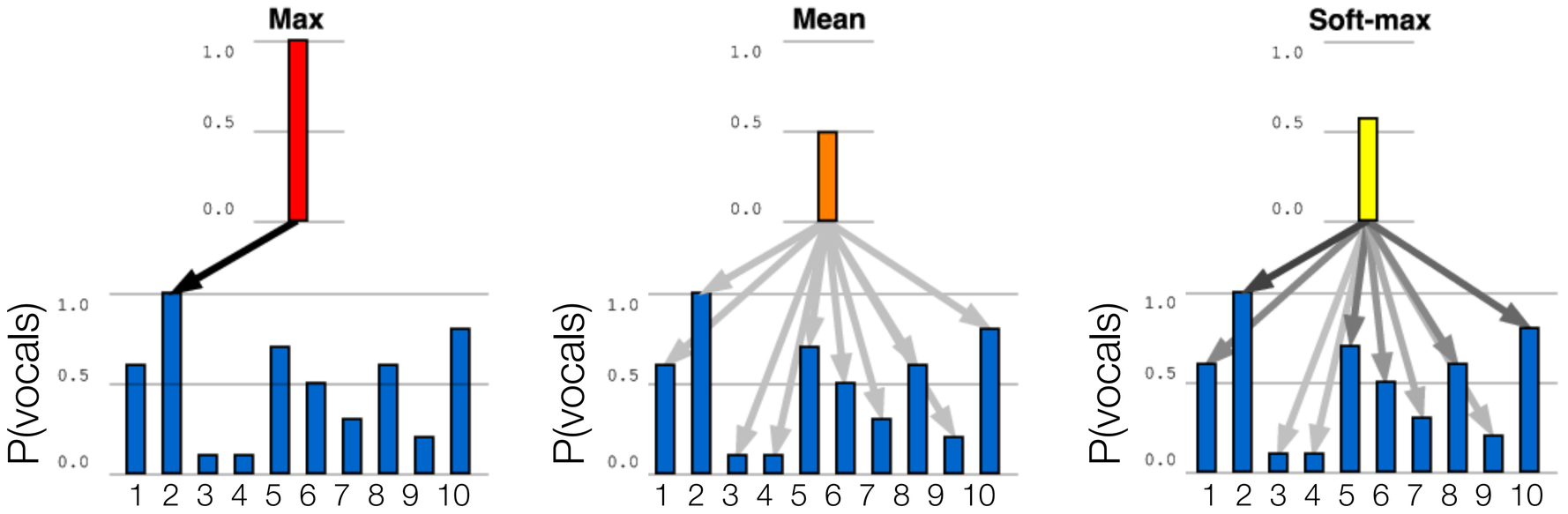


- Brittle
- Gradient descent ignores most predictions
- Learning slow/sensitive to initialization



- Learning is well behaved
- Lose specificity
- Only good events present most of the time

Temporal pooling



[0.2, 0.9, 0.5, 0.3] ➡ [0.18, 0.37, 0.25, 0.20]

0.2*0.18 + 0.9*0.37 + 0.5*0.25 + 0.3*0.20 = 0.55

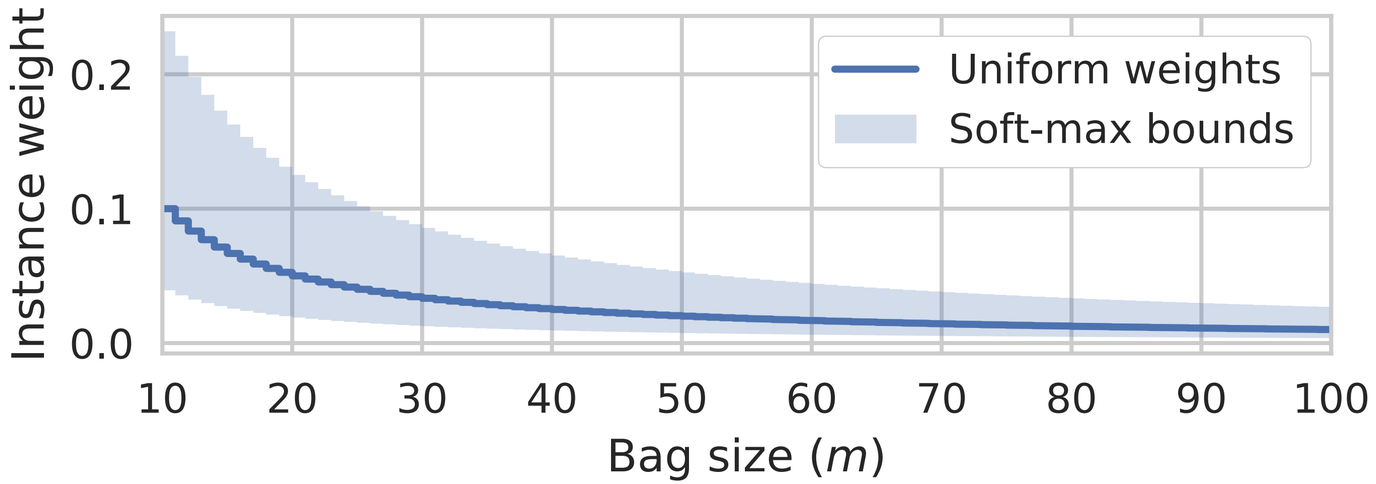
$$\hat{P}_s(Y \mid X) = \sum_{x \in X} \hat{p}(Y \mid x) \left(\frac{\exp \hat{p}(Y \mid x)}{\sum_{z \in X} \exp \hat{p}(Y \mid z)} \right)$$

- Weighted average
- Positive predictions get more weight
- Predicts like max, learns like mean

Softmax: bounded input, bounded output

- Inputs to softmax are probabilities in [0, 1]
- Outputs are positive weights, sum to 1
- Bounded input ⇔ Bounded output
- **Weights converge to uniform for large bags**

$$\frac{\exp \hat{p}(Y \mid x)}{\sum_{z \in X} \exp \hat{p}(Y \mid z)}$$



AutoPool

- Idea: remove input bound!
- Introduce temperature parameter: $\boldsymbol{\alpha}$
- Learn $\boldsymbol{\alpha}$ jointly with model:

$$\hat{P}_\alpha(Y \mid X) = \sum_{x \in X} \hat{p}(Y \mid x) \left(\frac{\exp\left(\alpha \cdot \hat{p}(Y \mid x)\right)}{\sum_{z \in X} \exp\left(\alpha \cdot \hat{p}(Y \mid z)\right)} \right)$$

AutoPool

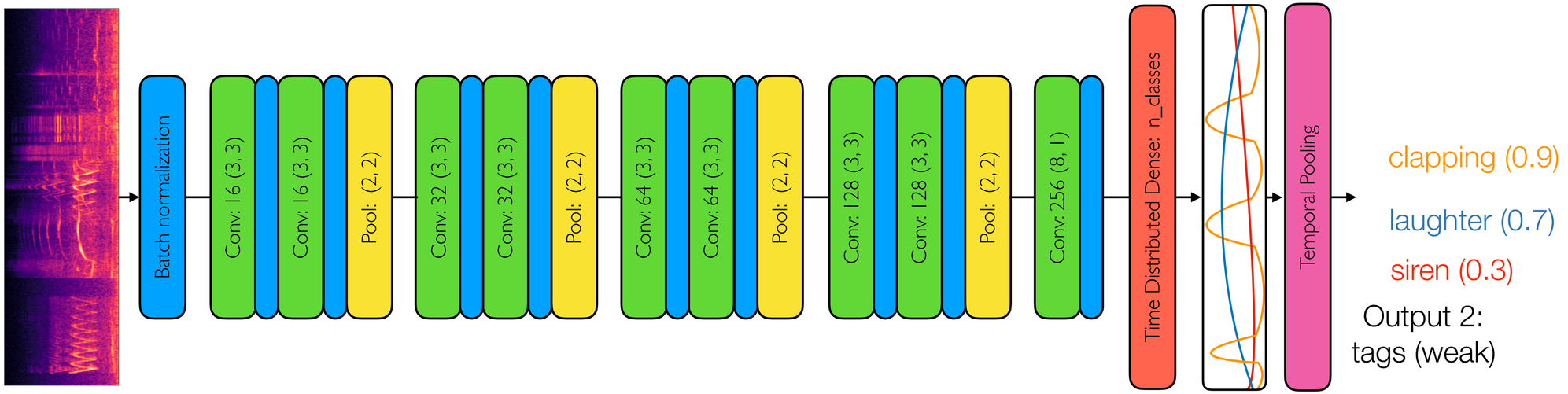
- $\alpha = 0$ ⇒ mean pooling
- $\alpha = 1$ ⇒ softmax pooling
- $\alpha \rightarrow \infty$ ⇒ max pooling
- $\alpha \rightarrow$ -∞ ⇒ min pooling

$$\hat{P}_\alpha(Y \mid X) = \sum_{x \in X} \hat{p}(Y \mid x) \left(\frac{\exp\left(\alpha \cdot \hat{p}(Y \mid x)\right)}{\sum_{z \in X} \exp\left(\alpha \cdot \hat{p}(Y \mid z)\right)} \right)$$

- Multi-label? each class gets its own α
- **Model adapts to temporal extent of each class**
- AutoPool variants: Constrained AutoPool (**CAP**) & Regularized AutoPool (**RAP**)

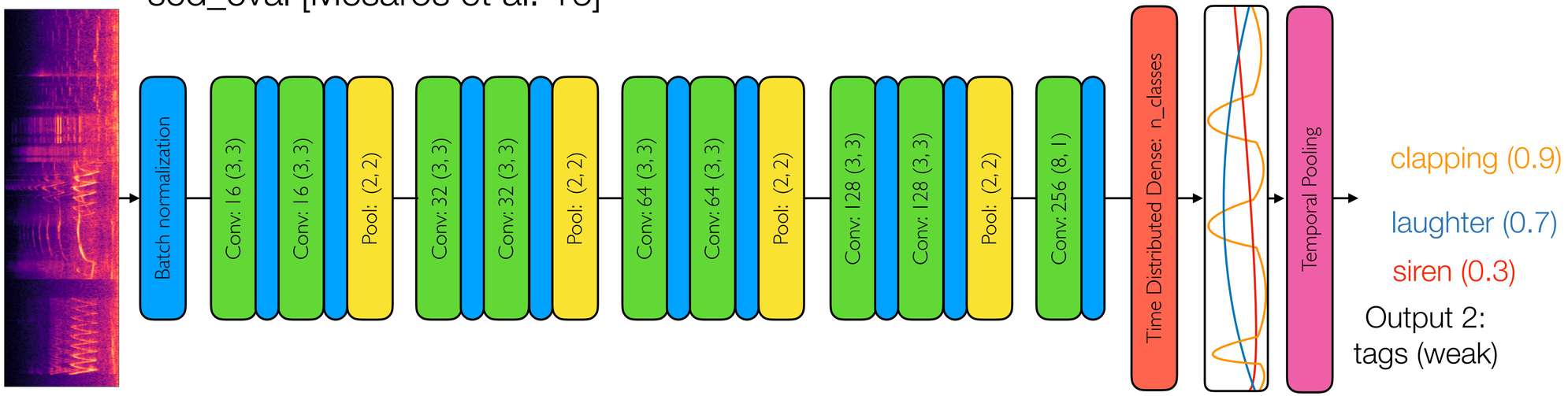
Experimental design

- Keep most of the model architecture **fixed**:
  - Deep convolutional neural network (CNN)
  - Based on audio subnetwork of Look, Listen and Learn (L3) architecture [Arandjelovic & Zisserman'17]
  - Output frame rate: ~3 frames/sec



Experimental design

- Compare **pooling** layers:
  - Max, mean, softmax, AutoPool, CAP, RAP ($\lambda = 10^{-2}, 10^{-3}, 10^{-4}$)
- Compare against **strong training**
- Evaluate strong prediction (instance-level) accuracy: **F1**, P, R (0 = worst, 1 = best)
  - sed_eval [Mesaros et al.'16]



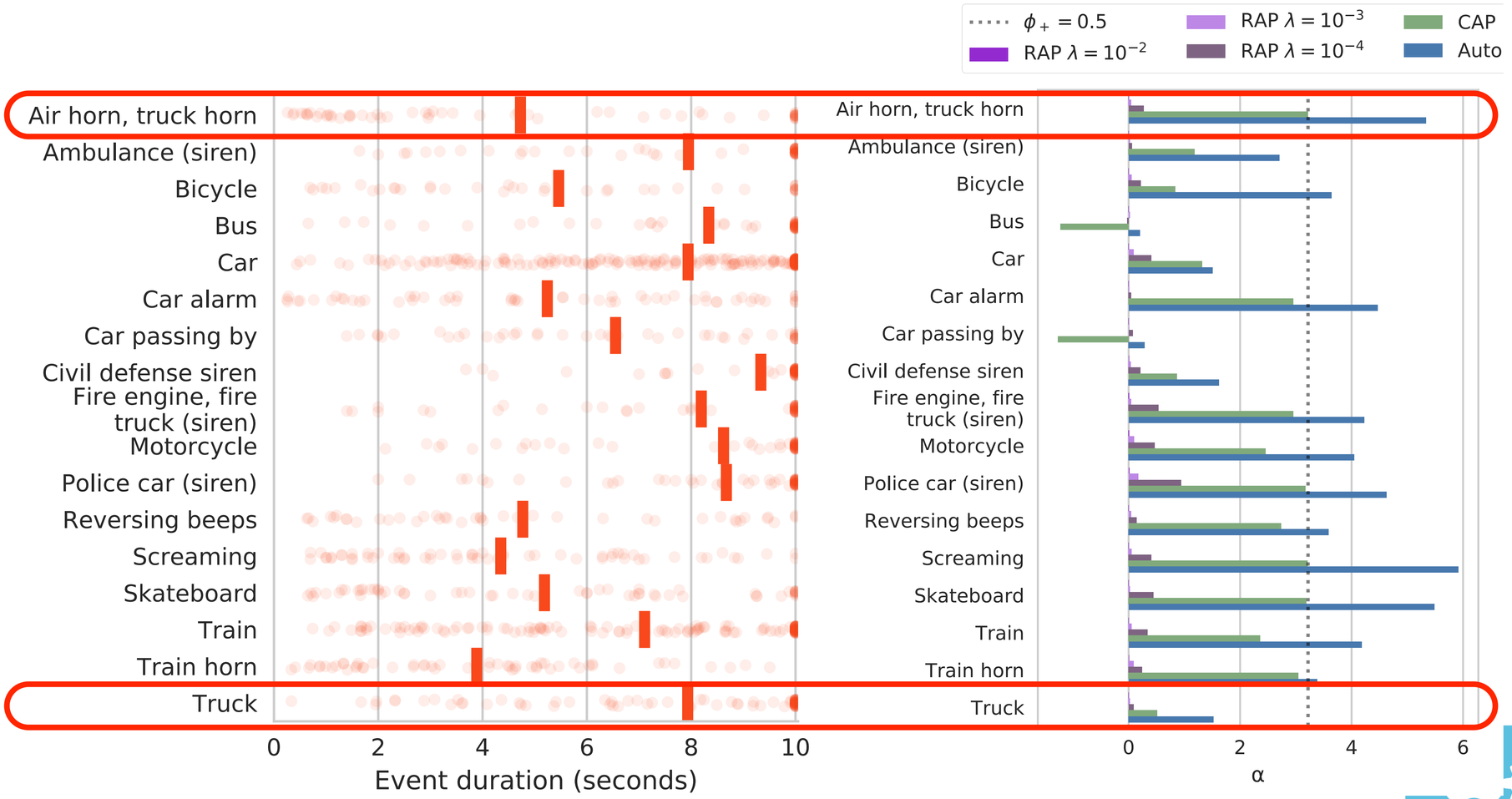
Experimental design: Datasets

- **URBAN-SED** [Salamon et al.'17]
  - 10,000 soundscapes synthesized with Scaper
  - 10 sound classes, **mostly short events (~3 s)**
- **DCASE 2017 (Task 4)** [Mesaros et al.'17]
  - ~50k subset of AudioSet [Gemmeke et al.'17]
  - 17 event classes, **varying event durations (0-10 s)**
- **MedleyDB** [Bittner et al.'15]
  - 122 songs —> 531 with remixing
  - 8 instrument classes, **mostly long events (> 10s)**

Results: URBAN-SED

Model	Strong (time-varying) F_1	P	R	$E_{\downarrow}$
Max	0.463	0.774	0.330	0.695
Mean	0.408	0.280	0.751	2.10
Soft-max	0.492	0.397	0.646	1.22
RAP 10^{-2}	0.419	0.296	0.717	1.88
RAP 10^{-3}	0.529	0.584	0.484	0.731
RAP 10^{-4}	0.526	0.650	0.442	0.681
CAP	0.533	0.622	0.466	0.696
Auto	0.504	0.738	0.382	0.665
Strong	0.551	0.693	0.458	0.642

Results: DCASE 2017



Results: MedleyDB

Instance-level prediction (frames)

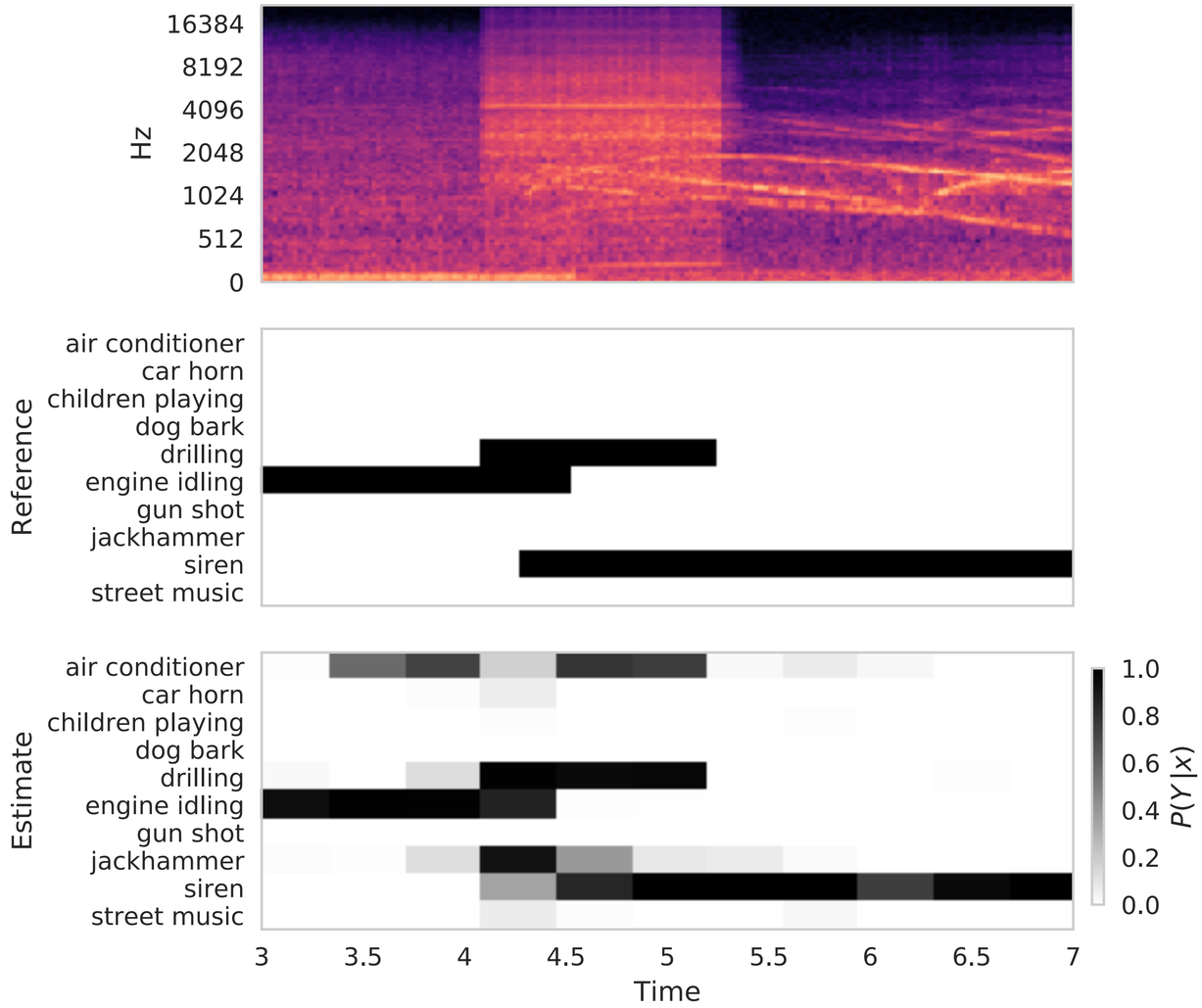
Model	Strong (time-varying) F_1	P	R	$E_{\downarrow}$
Max	0.437	**0.875**	0.292	0.719
Mean	**0.655**	0.594	**0.733**	**0.608**
Soft-max	**0.662**	0.668	**0.658**	**0.524**
RAP 10^{-2}	**0.659**	0.604	**0.727**	**0.593**
RAP 10^{-3}	**0.673**	0.638	**0.714**	**0.545**
RAP 10^{-4}	**0.622**	**0.757**	0.530	**0.540**
CAP	**0.609**	**0.787**	0.498	**0.551**
Auto	0.528	**0.841**	0.386	0.636
Strong	0.675	0.640	0.716	0.540

Results: example

Input

Reference

Model estimate



How can we leverage the unlabeled data?



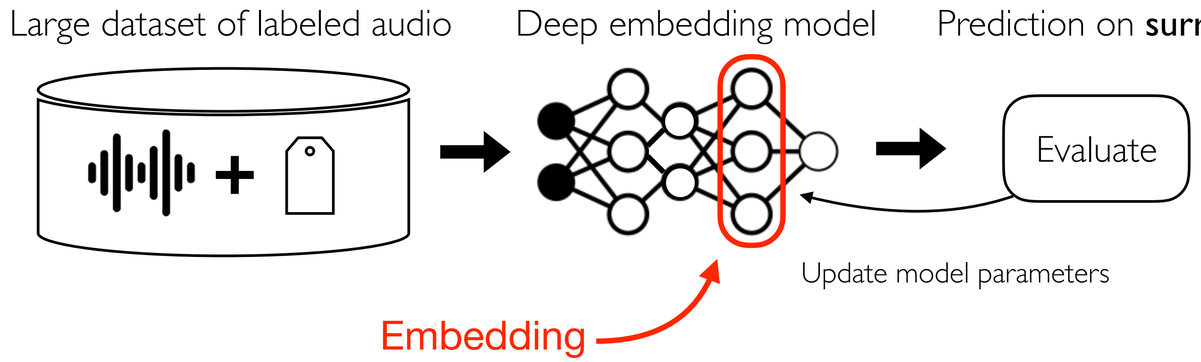

How can we leverage the unlabeled data?

TBD
J. Cramer, H-H. Wang, J. Salamon, J. P. Bello
Coming soon… 2019.

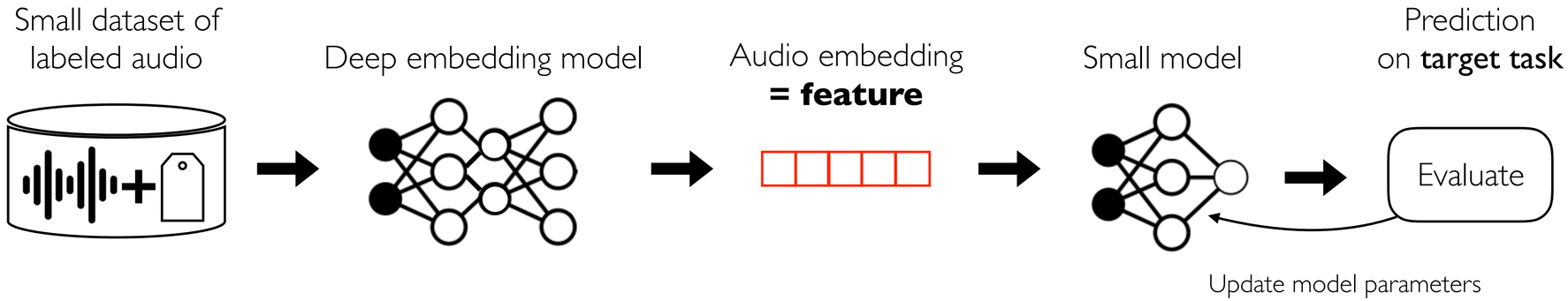
Deep audio embeddings

- The idea:

Step 1: train embedding model on **surrogate task**

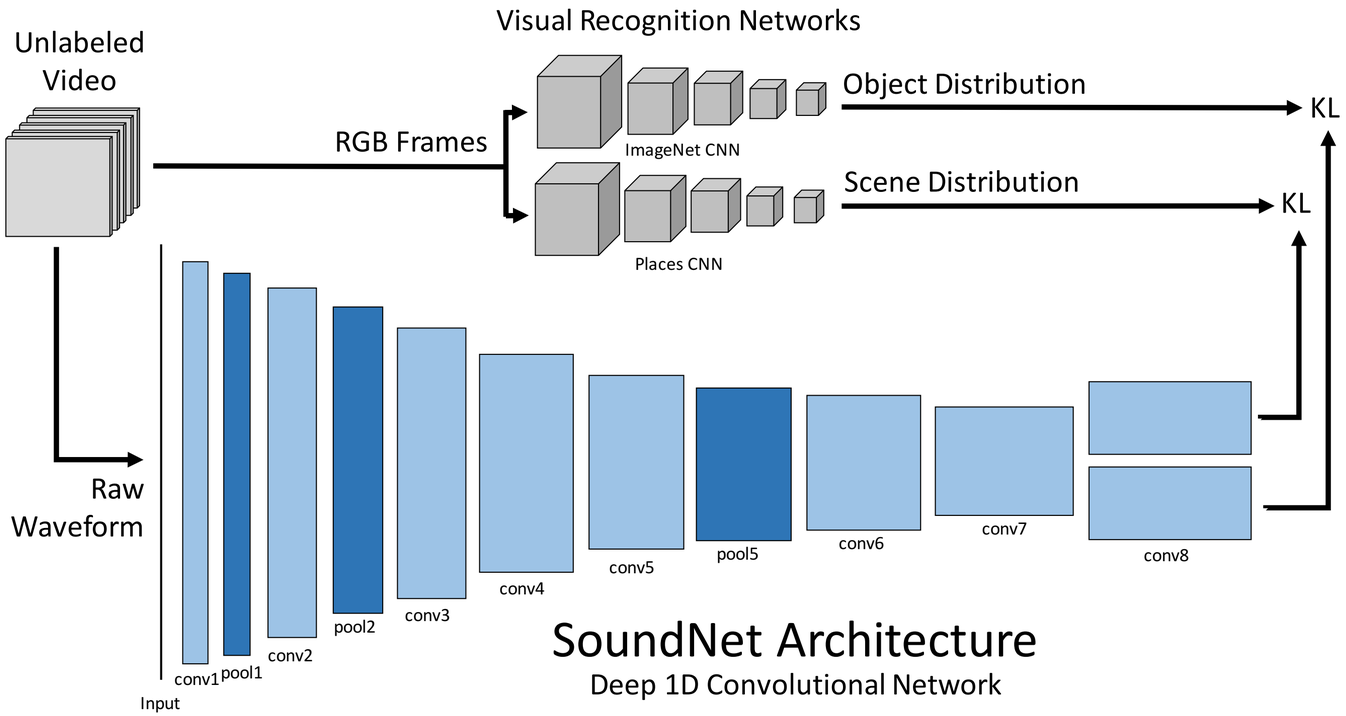


Step 2: use embedding to train "downstream" model on **target task**



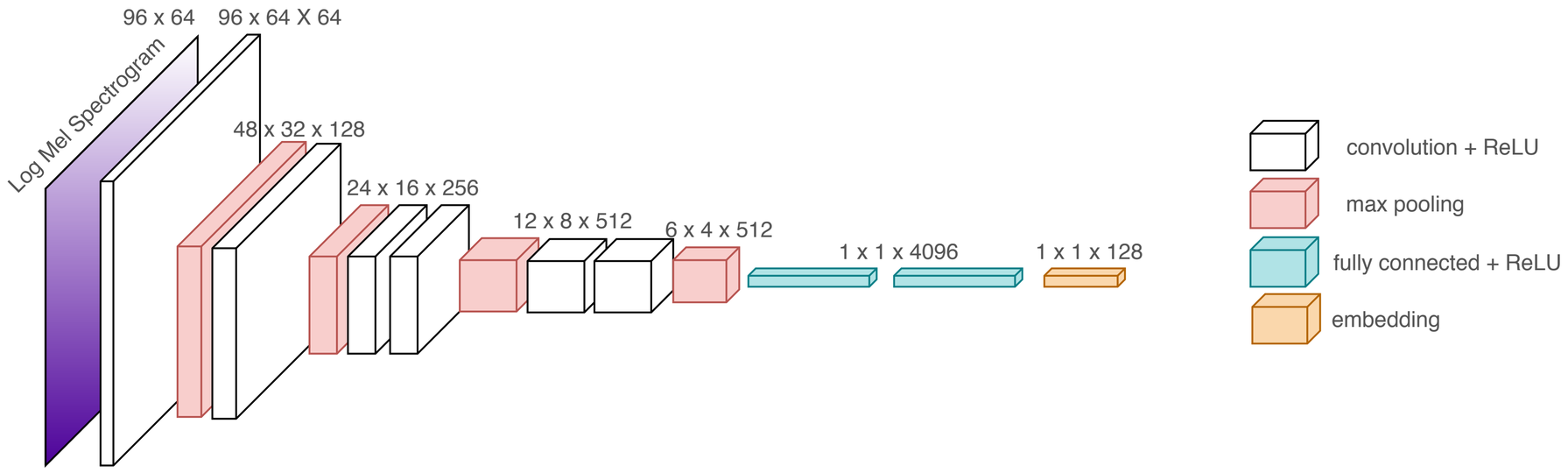
Deep audio embeddings

- SoundNet [Aytar et al.'16]
  - Visual network pre-trained on large image datasets
  - Audio network trained to mimic output of visual network for Flickr videos



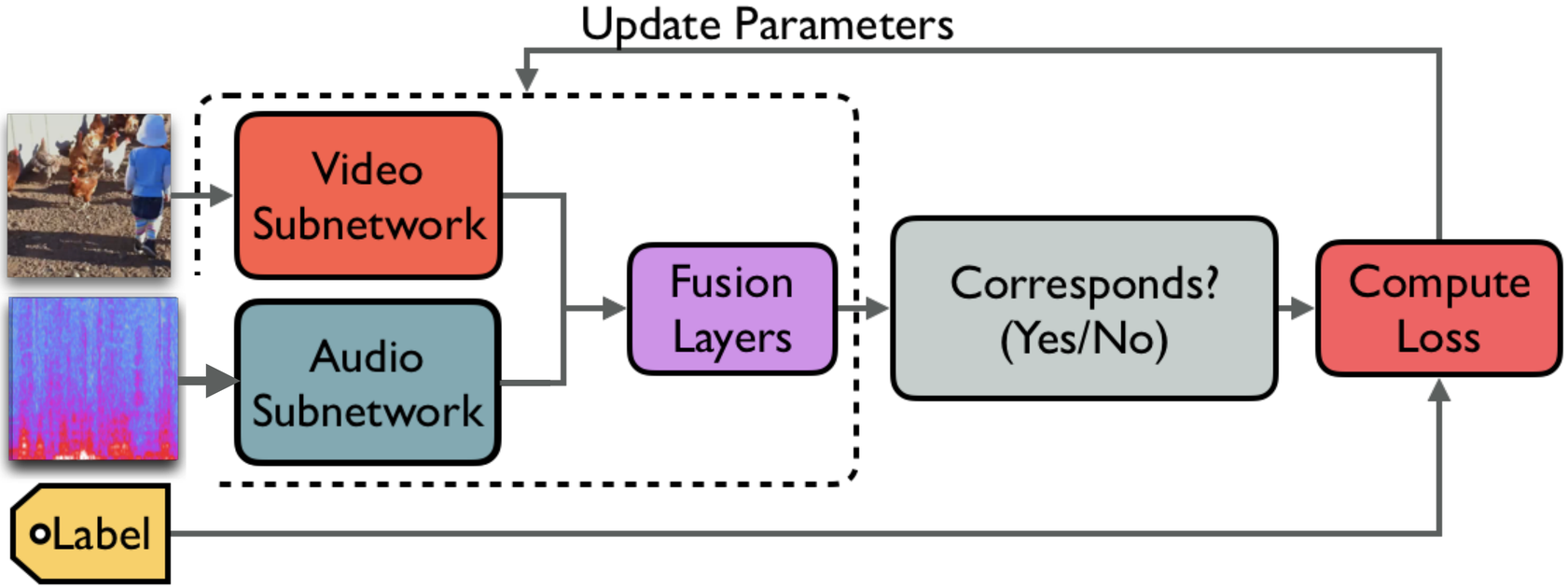
Deep audio embeddings

- VGGish [Hershey et al.'17]
  - Single audio network trained to predict labels on YouTube-8M dataset



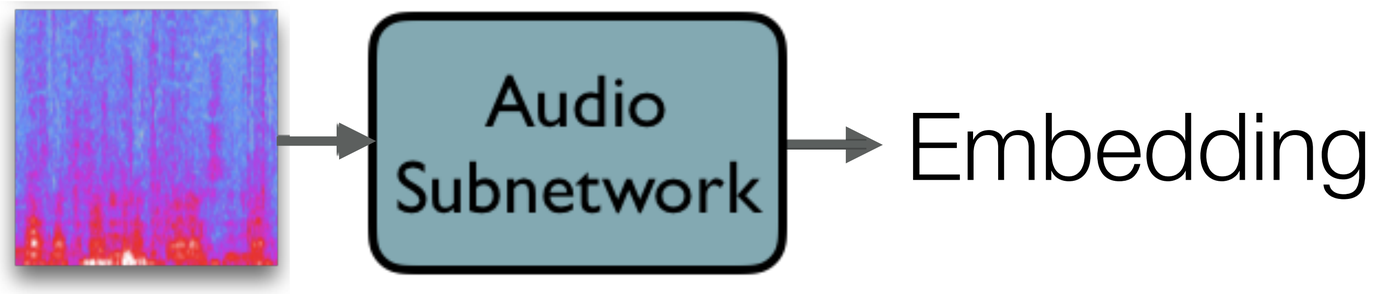
Deep audio embeddings

- Look, Listen, and Learn (**L3**) [Arandjelovic & Zisserman'17]
  - Train model on the task of Audio-Visual Correspondence (AVC)
  - No labels required!



Deep audio embeddings

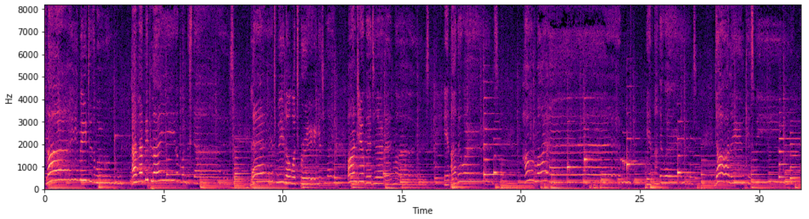
- Look, Listen, and Learn (**L3**) [Arandjelovic & Zisserman'17]
  - Train model on the task of Audio-Visual Correspondence (AVC)
  - No labels required!



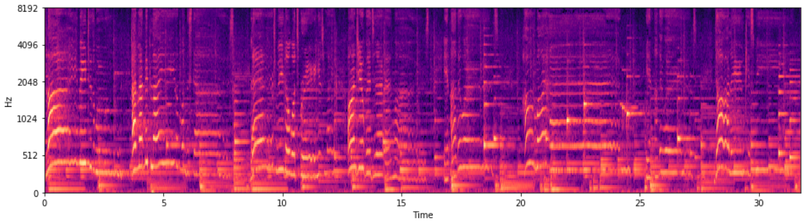
Can we train our own (maybe better) L3 embedding?

- Questions:
  - Does an **audio-informed input representation** give a better embedding?
  - Is it important to use **matched audio domains** between embedding and target?
  - **How much training data** is enough training data for the embedding?
  - Does **data augmentation** still improve performance on target?
  - How does the resulting embedding **compare to state-of-the-art embeddings?**
    - SoundNet [Aytar et al.'16]
    - VGGish [Hershey et al.'17]

Linear



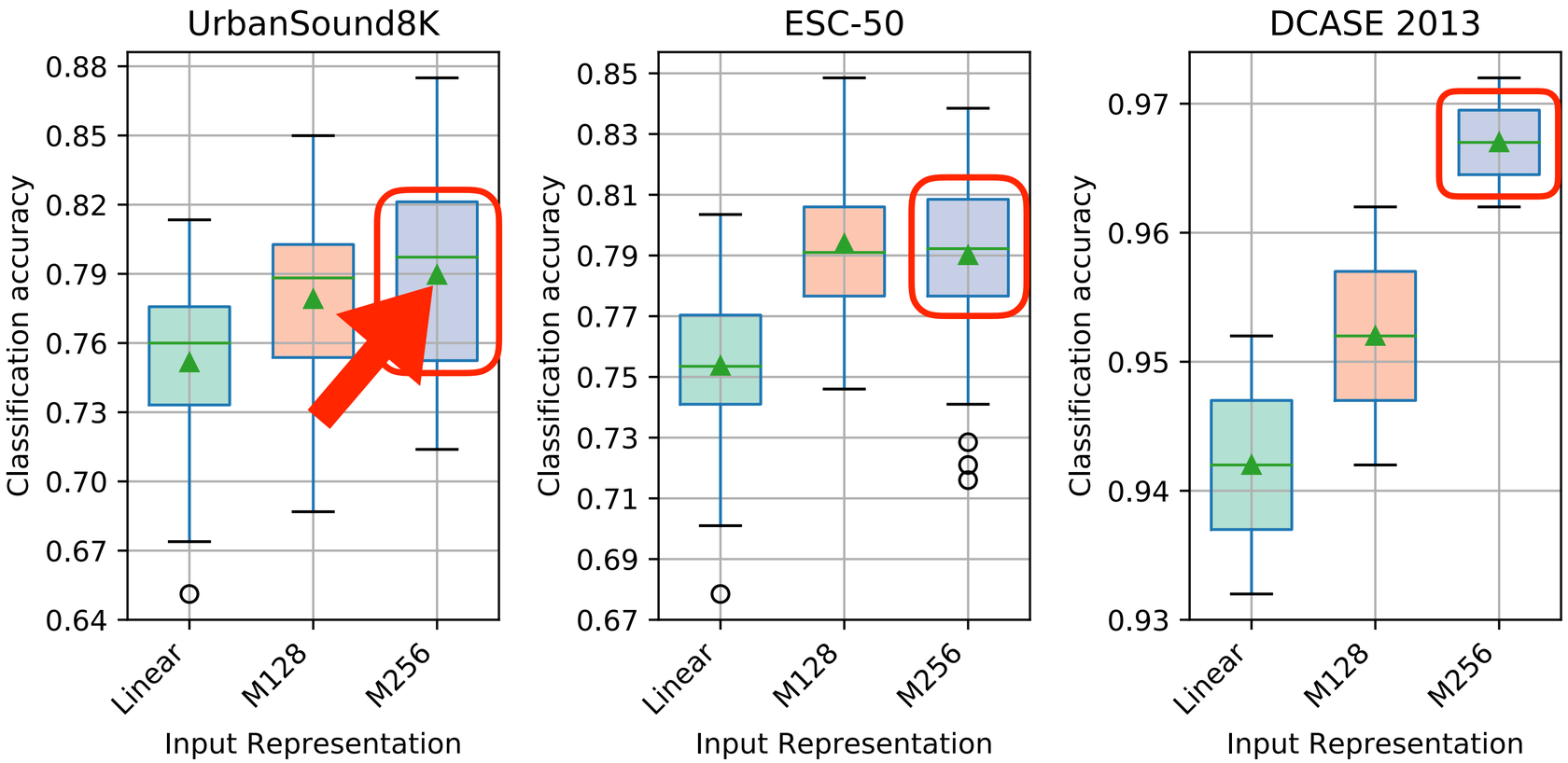
Mel



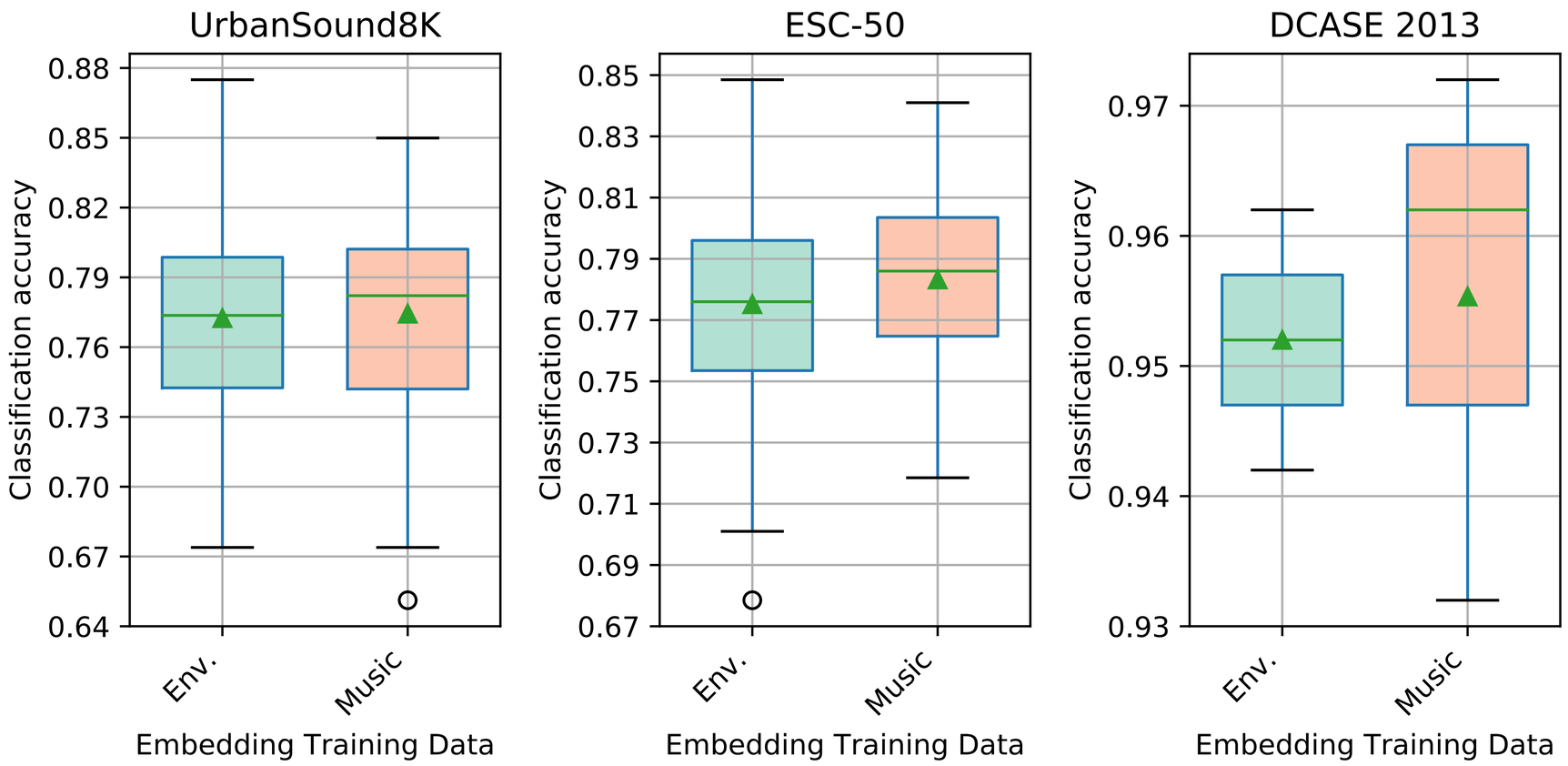
Experimental design (abridged)

- **Step 1: train several variants of L3 embedding**
  - Vary input representation, training data (content & amount)
- **Step 2: evaluate embeddings on target task:**
  - Multi-class sound classification using 2-layer MLP
  - Datasets:
    - UrbanSound8K (**8732** clips, 10 classes) [Salamon et. al'14]
    - ESC-50 (**2000** clips, 40 classes) [Piczak'15]
    - DCASE 2013 (**200** clips, 10 classes) [Stowell et al.'15]
  - Compare to SoundNet and VGGish

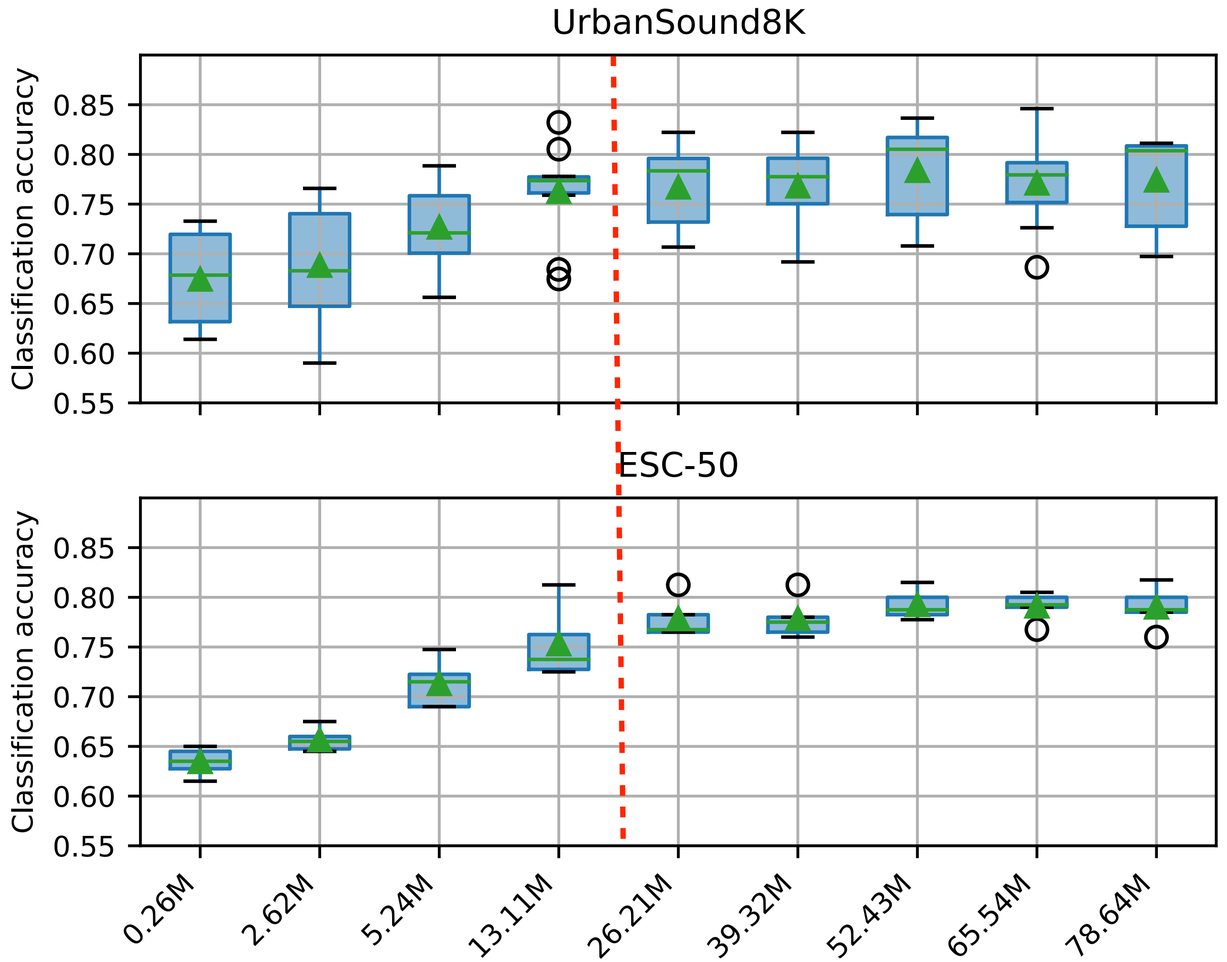
L3: input representation



L3: matched vs mismatched training content

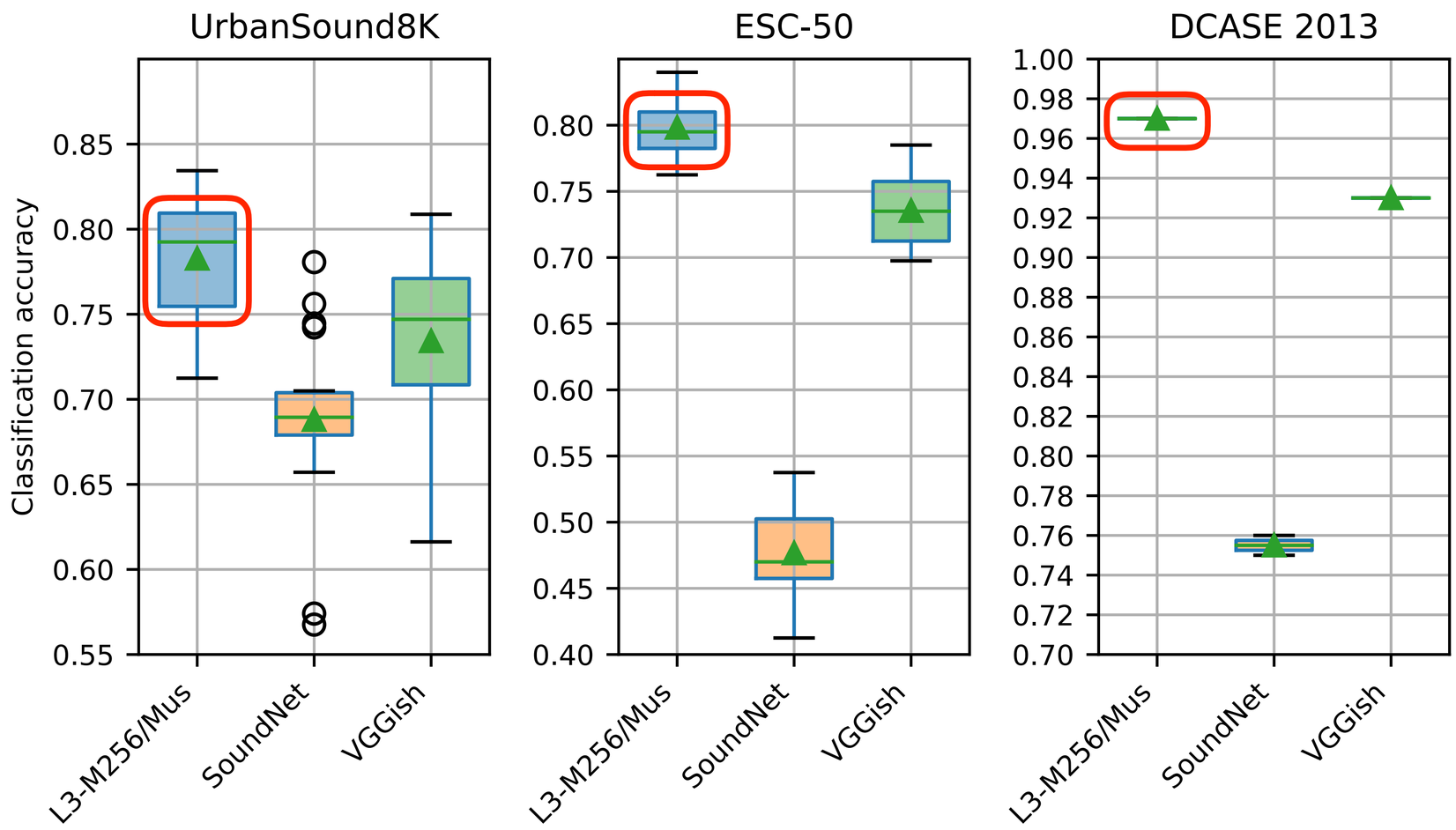


L3: embedding training data vs target performance



Results

- Mel-based L3 vs SoundNet and VGGish:



Try these embeddings out… soon!

github.com/marl/openl3

How do we make our models robust to different sensor locations?

How do we make our models robust to different sensor locations?

…and to changing conditions in each location?

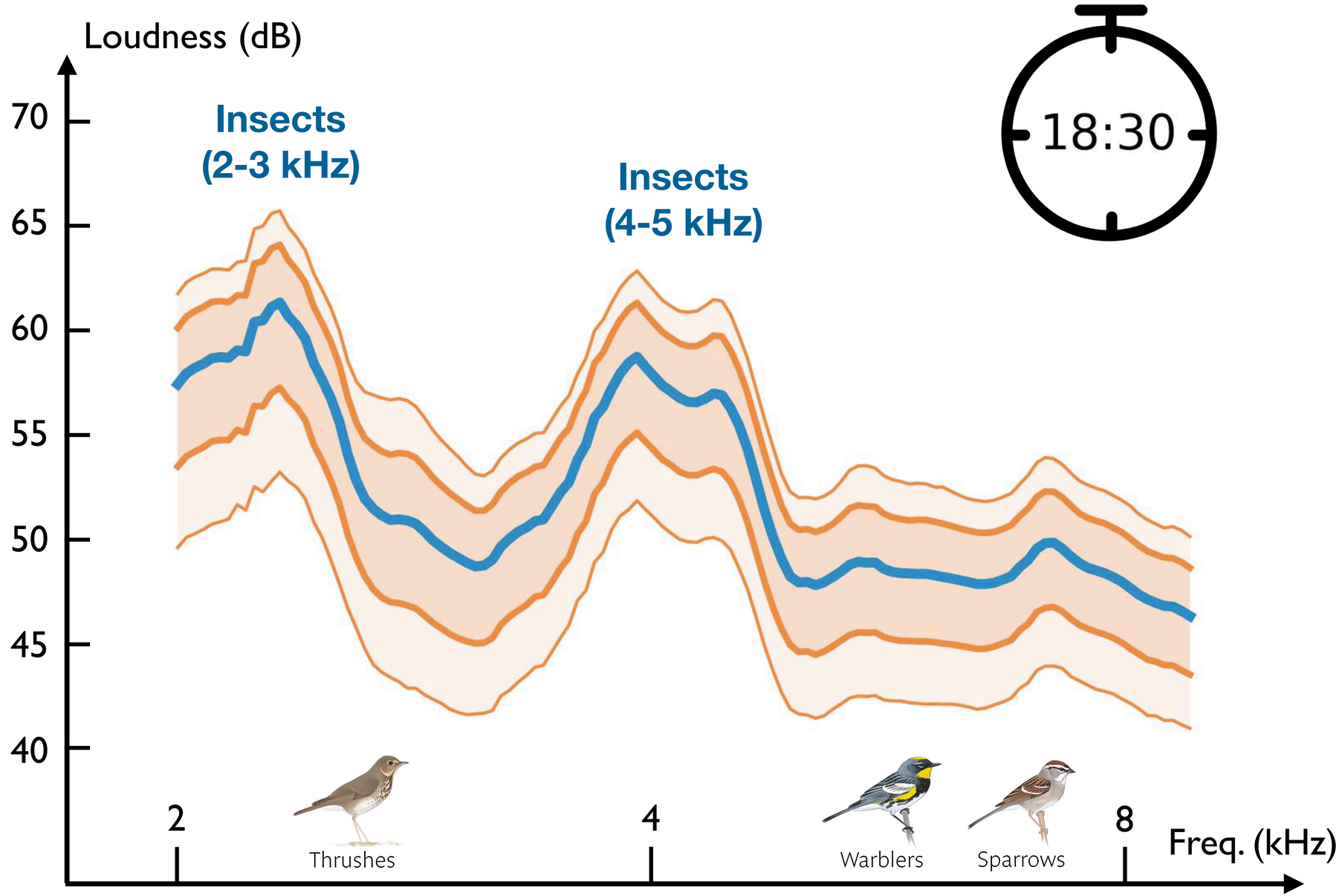
How do we make our models robust to different sensor locations?

…and to changing conditions in each location?

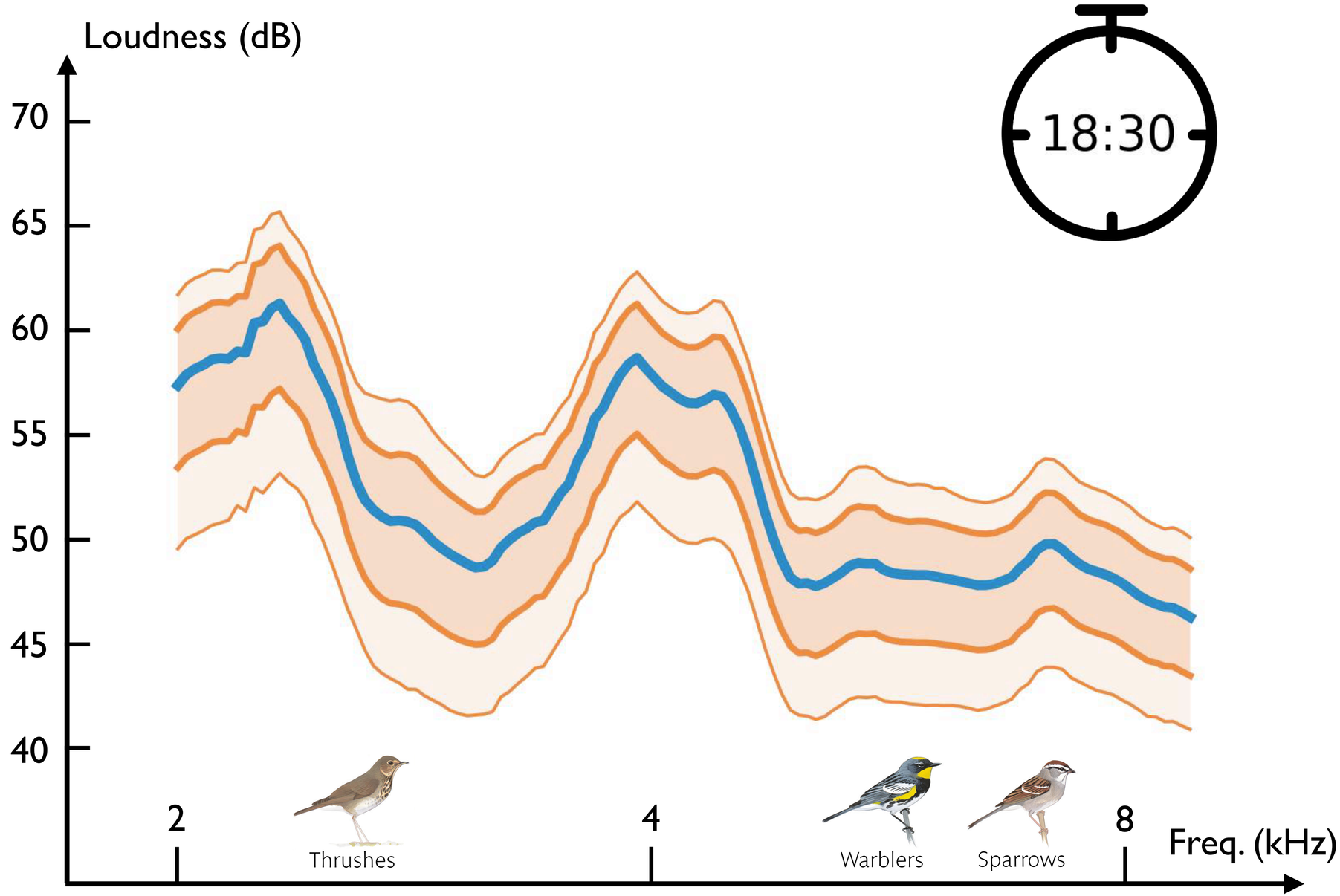
Per-Channel Energy Normalization: Why and how
V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello.
IEEE Signal Processing Letters, in press, 2018.

Robust Sound Event Detection in Bioacoustic Sensor Networks
V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello.
Coming soon… 2019

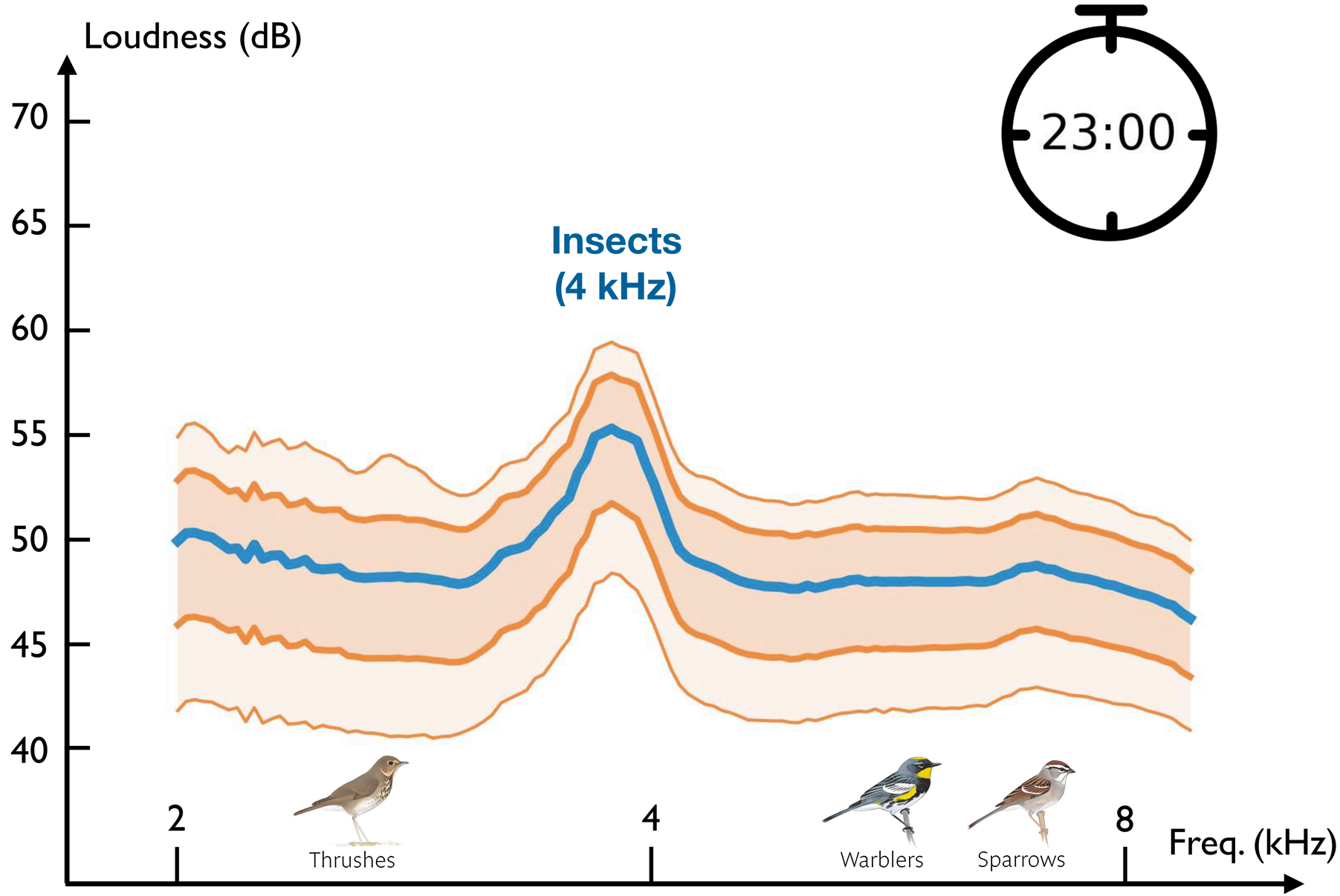
Challenge #1: non-stationary noise



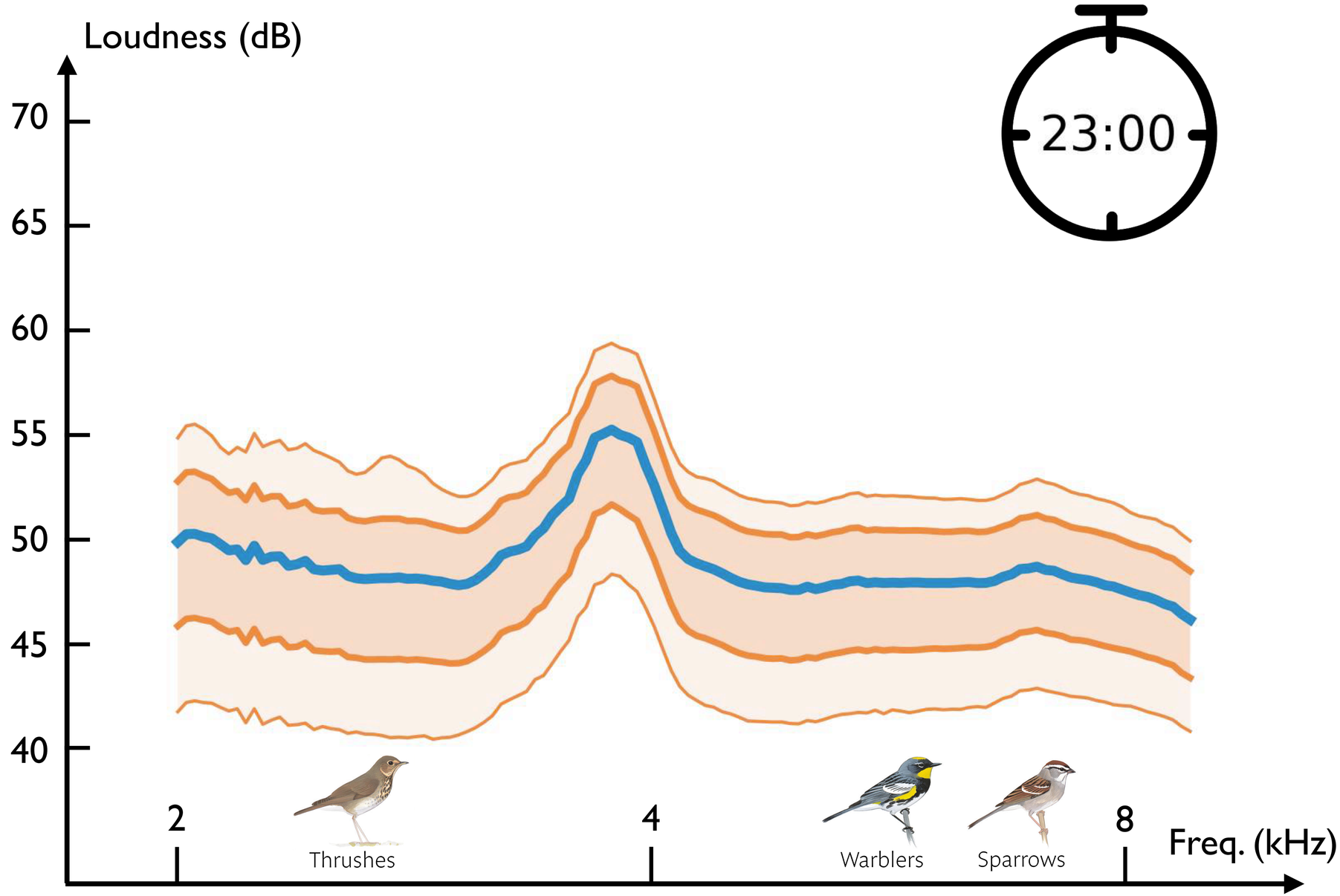
Challenge #1: non-stationary noise



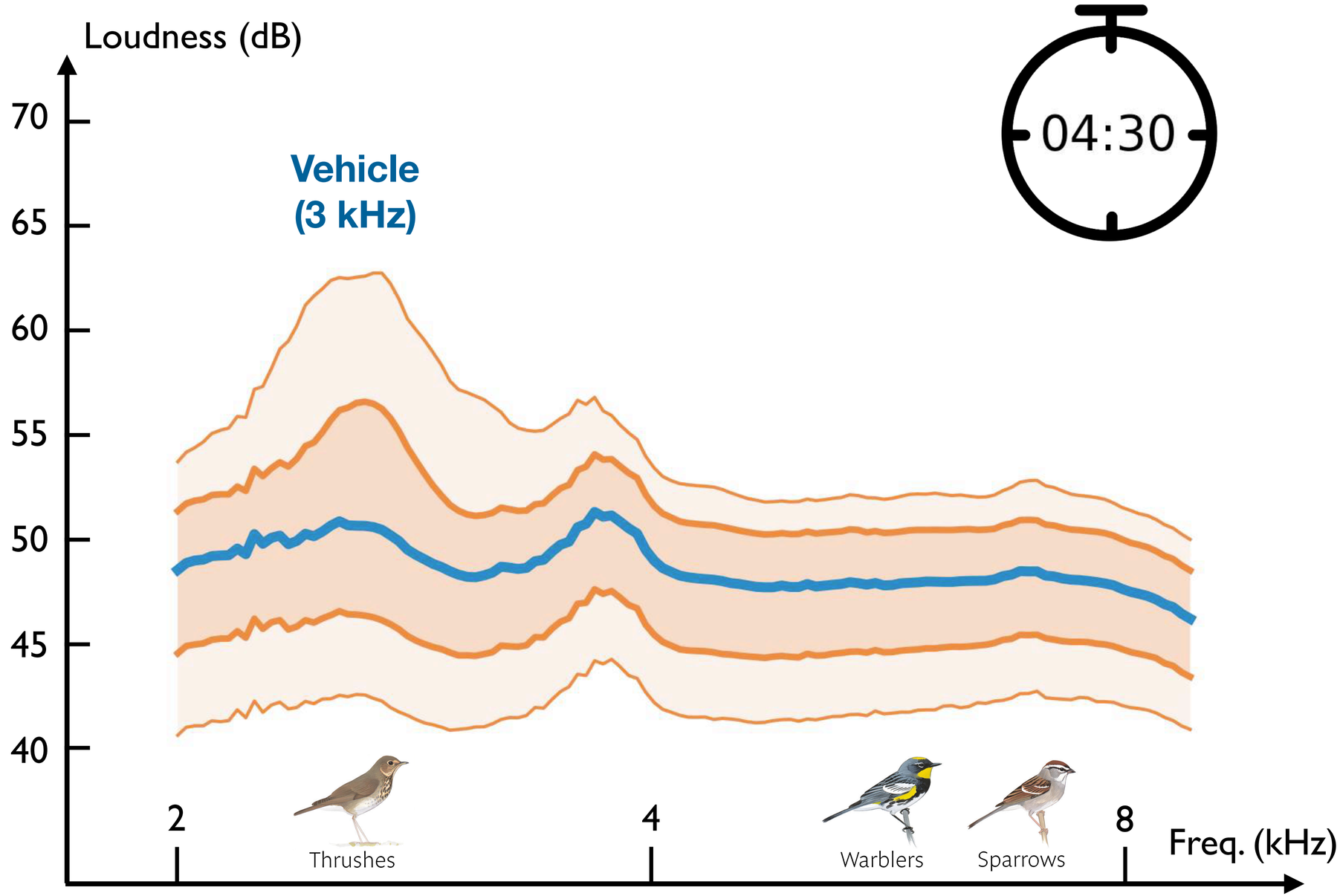
Challenge #1: non-stationary noise



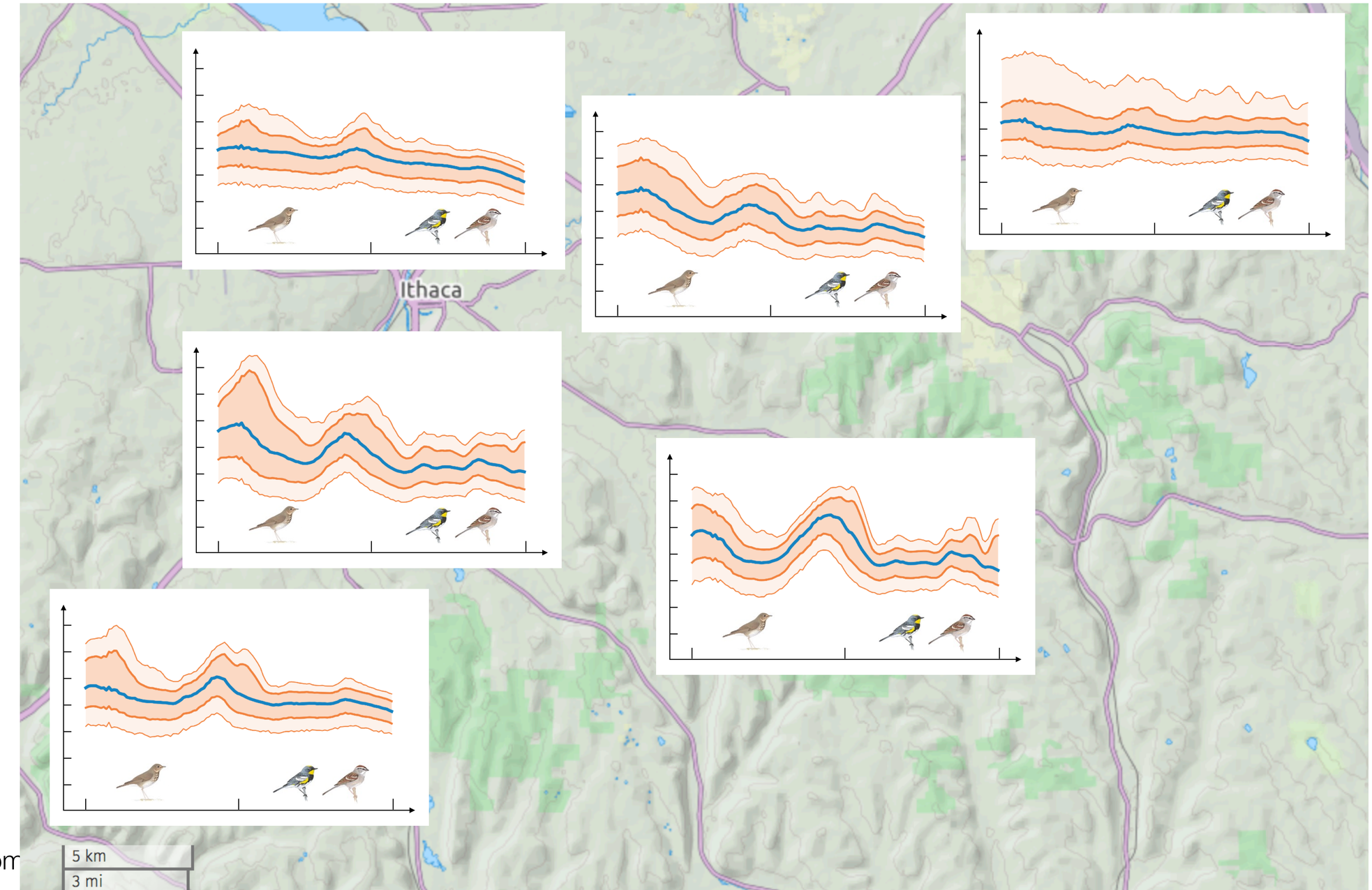
Challenge #1: non-stationary noise



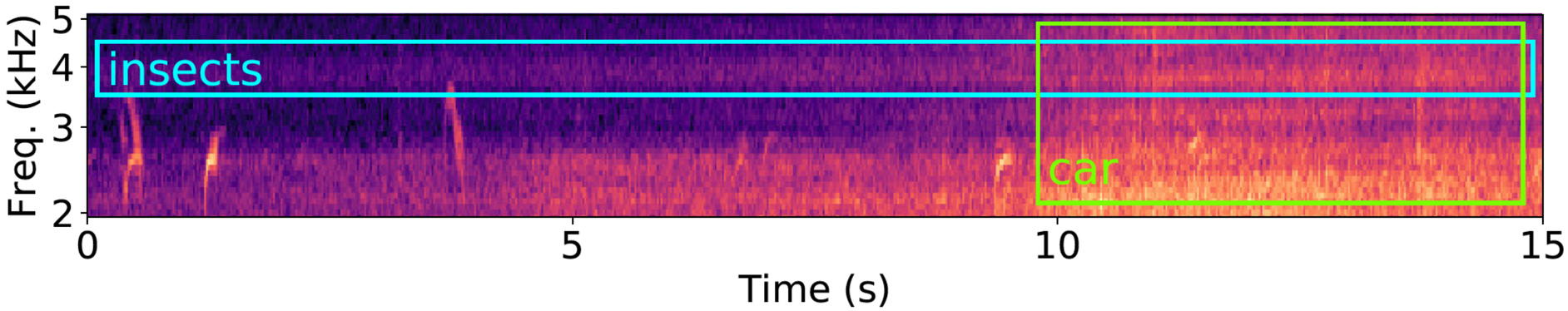
Challenge #1: non-stationary noise



Challenge #2: nonuniform noise



Input representation: log-mel spectrogram?



Per-Channel Energy Normalization (PCEN)

- Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous. **Trainable frontend for robust and far-field keyword spotting**. In IEEE ICASSP, pp. 5670–5674, New Orleans, LA, USA, Mar. 2017.

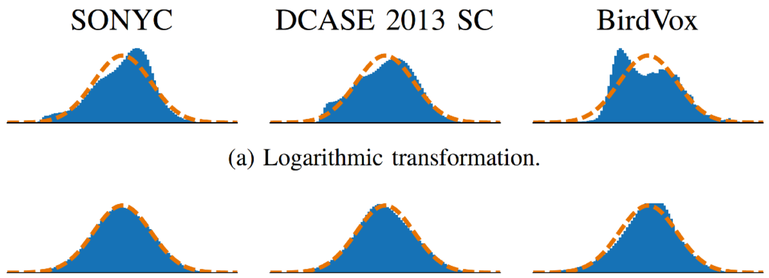
Dynamic range compression

$$\mathbf{PCEN}(t,f) = \left(\frac{\mathbf{E}(t,f)}{\left(\varepsilon + \left(\mathbf{E} \overset{\text{time}}{*} \boldsymbol{\phi}\right)(t,f)\right)^{\alpha}} + \delta \right)^{r} - \delta^{r}$$

Adaptive gain control

Per-Channel Energy Normalization (PCEN)

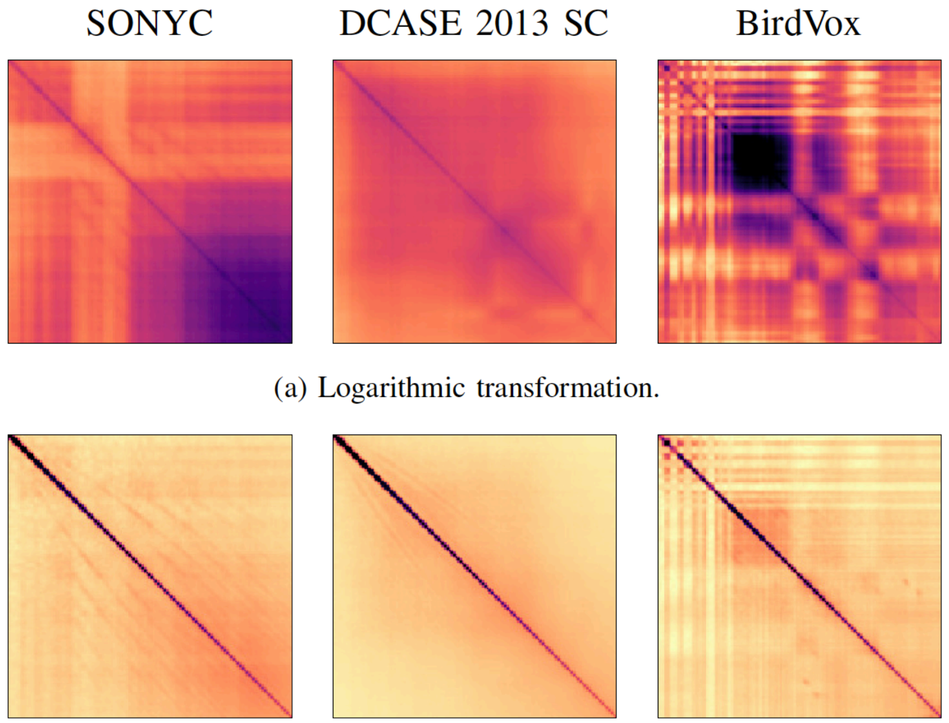
- V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello. **Per-Channel Energy Normalization: Why and how.** IEEE Signal Processing Letters, in press, 2018.



(a) Logarithmic transformation.

(b) Per-channel energy normalization (PCEN).

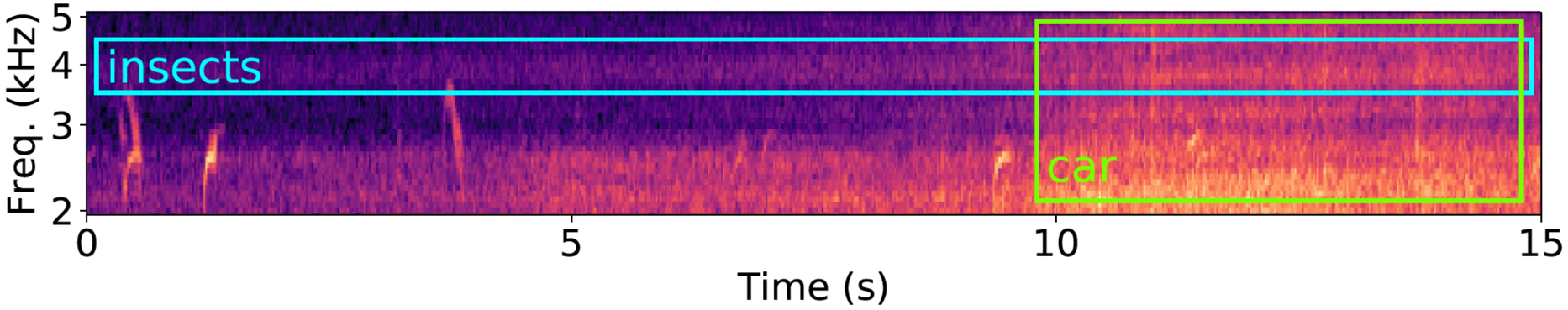
	indoor	outdoor
ε	10^{-6}	10^{-6}
α	0.98	0.80
δ	10	2
r	0.5	0.25



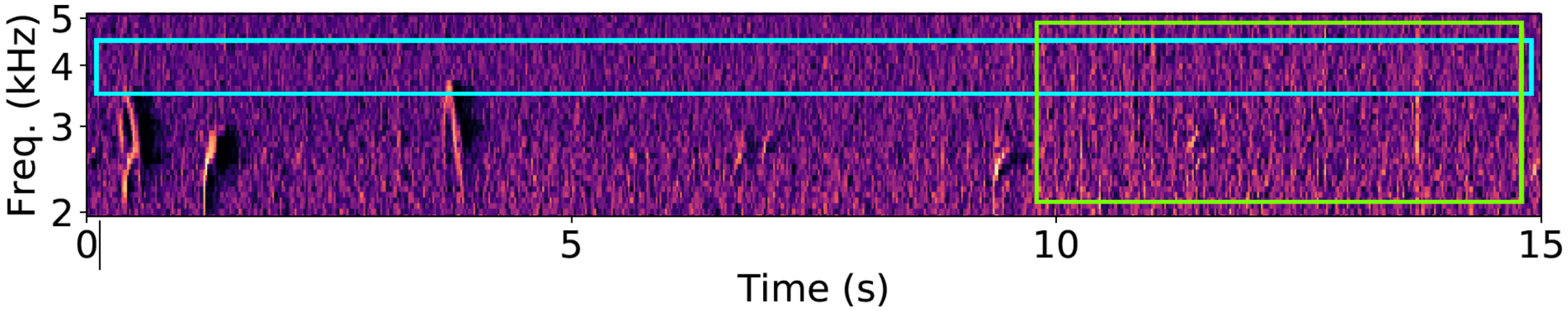
(a) Logarithmic transformation.

(b) Per-channel energy normalization (PCEN).

Per-Channel Energy Normalization (PCEN)

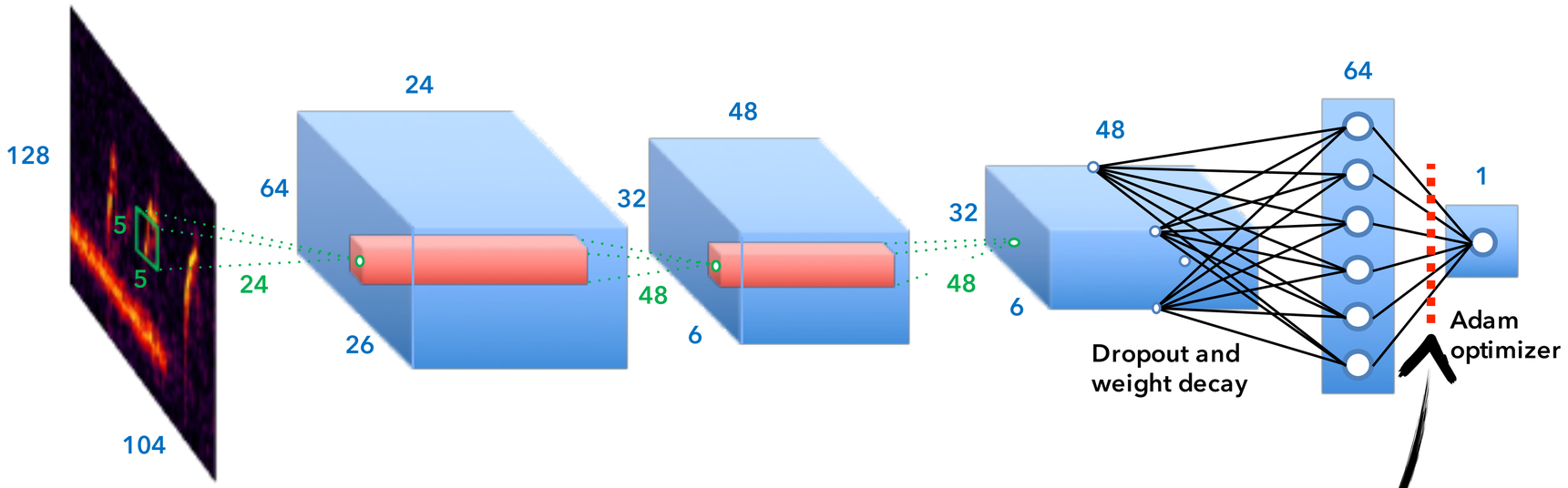


(a) Logarithmic transformation.

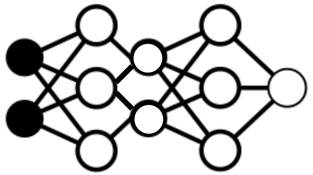


(b) Per-channel energy normalization (PCEN).

Context-Adaptive Neural Networks



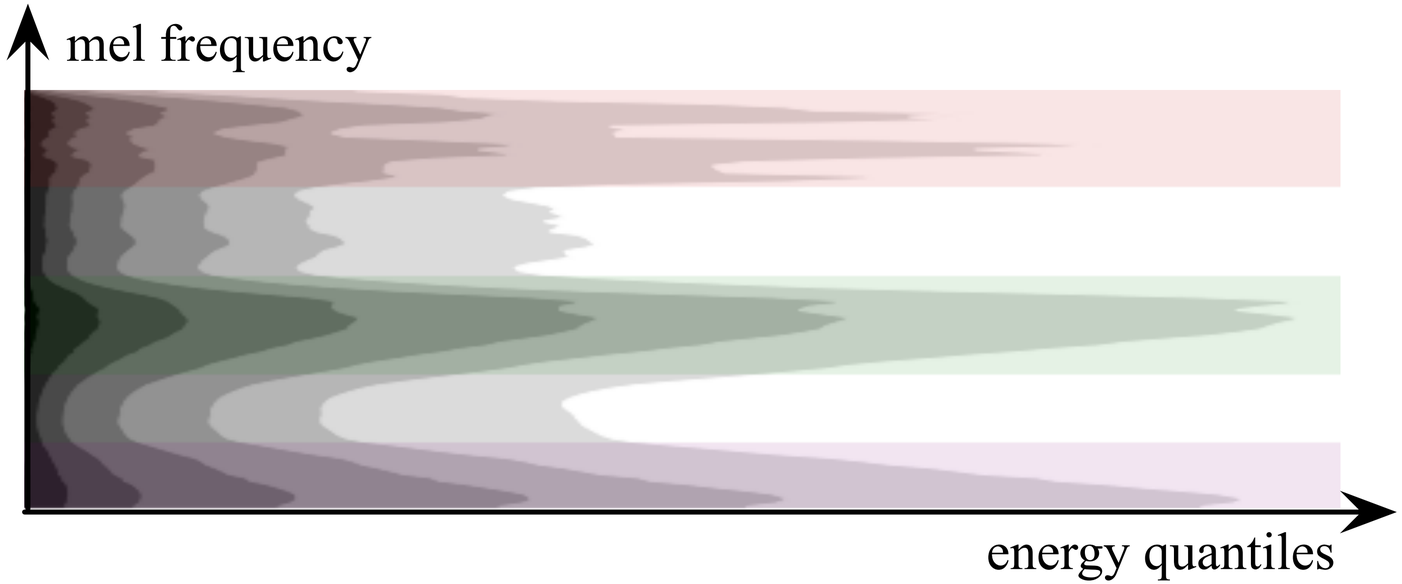
Representation of background (???)



Background (auxiliary) Features

Summarize background at scale T = 30 minutes: energy quantiles (median, quartiles, centiles, etc.)

Auxiliary features describe noise only, not the class of interest.



Context-Adaptive Neural Networks

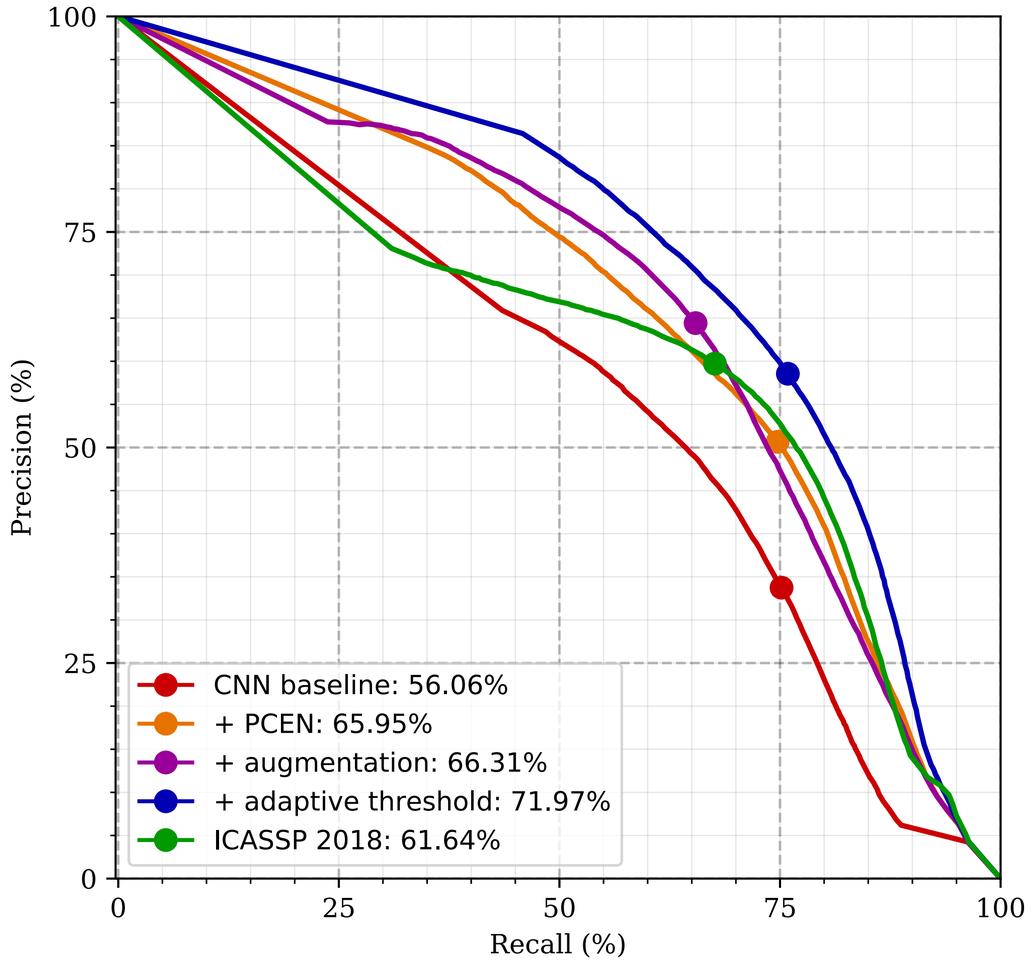
static sigmoid layer: $$\boldsymbol{y}(t) = \sigma\Big(\sum_k \big\langle \boldsymbol{w}_k \,|\, \boldsymbol{x}_k(t) \big\rangle + b \Big)$$

dynamic filter network: $$\boldsymbol{y}(t) = \sigma\Big(\sum_k \big\langle \boldsymbol{w}_k(t) \,|\, \boldsymbol{x}_k(t) \big\rangle + b \Big)$$

mixture of experts: $$\boldsymbol{y}(t) = \sigma\Big(\sum_k \alpha_k(t) \big\langle \boldsymbol{w}_k \,|\, \boldsymbol{x}_k(t) \big\rangle + b \Big)$$

adaptive threshold: $$\boldsymbol{y}(t) = \sigma\Big(\sum_k \big\langle \boldsymbol{w}_k \,|\, \boldsymbol{x}_k(t) \big\rangle + b(t) \Big)$$

Context-Adaptive Neural Networks



- Results (AUPRC):
- **Spectral flux: 15%**
- **CNN: 56%**
- **CNN + aug: 62%**
- **CNN + PCEN + aug: 66%**
- **CNN + PCEN + CA + aug: 72%**

Robust Sound Event Detection in Acoustic Sensor Networks

- How do we label the data?
  - Crowdsourcing: github.com/CrowdCurio/audio-annotator
  - Synthesis: github.com/justinsalamon/scaper
- How can we best leverage the labeled data?
  - Multiple instance learning + AutoPool: github.com/marl/autopool
- How can we leverage the unlabeled data?
  - Deep audio embeddings: github.com/marl/openl3
- How do we make our models robust to different & changing sensor locations?
  - PCEN: github.com/librosa/librosa
  - Context-adaptive networks: github.com/BirdVox/bv_context_adaptation

Thanks!