Parsing Birdsong with Deep Audio Embeddings

Irina Tolkova[1]*, Brian Chu[1]*, Marcel Hedman[1]*, Stefan Kahl[2] and Holger Klinck[2]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

[2]K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, NY

(*equal contribution)

AI4SG Workshop at IJCAI 2021

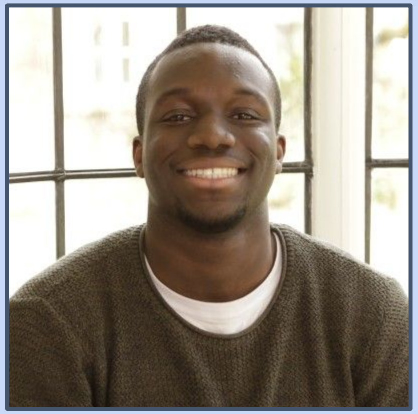
Introduction

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

TheCornellLab
Center for Conservation Bioacoustics

Irina Tolkova

Brian Chu



Marcel Hedman

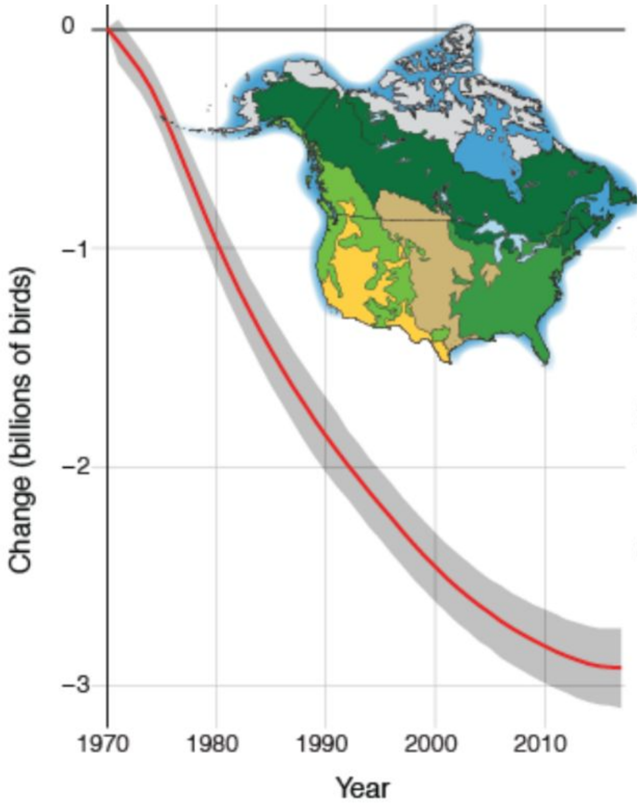
Dr. Stefan Kahl

Prof. Holger Klinck

Thank you to **Prof. Milind Tambe**, **Boriana Gjura**, and **Doria Spiegel** for their support on this project!

Biodiversity Loss

Over the last 50 years, the US has lost an estimated **3 billion birds**.

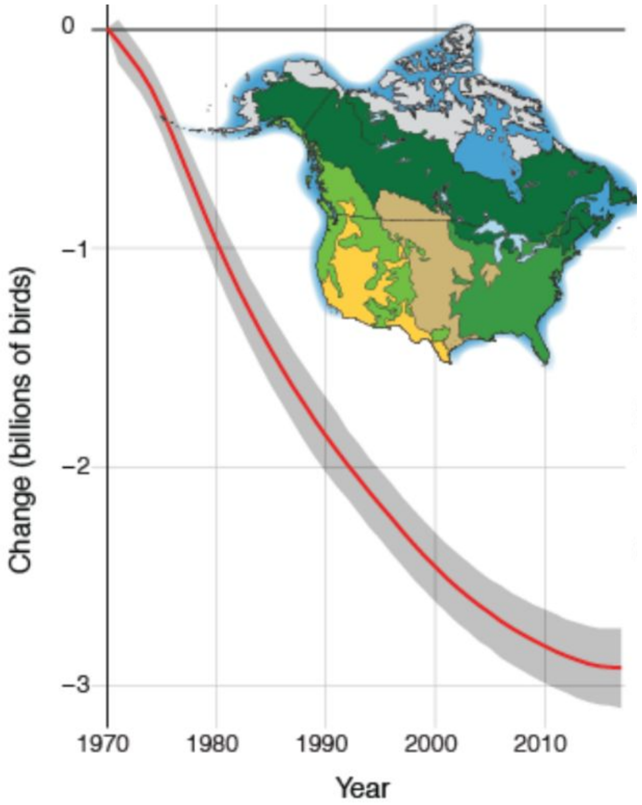


Rosenberg, Kenneth V., et al. "Decline of the North American avifauna." Science (2019).

To prevent further loss, it is necessary to understand **large-scale population dynamics**.

Biodiversity Loss

Over the last 50 years, the US has lost an estimated **3 billion birds**.



Rosenberg, Kenneth V., et al. "Decline of the North American avifauna." Science (2019).

To prevent further loss, it is necessary to understand **large-scale population dynamics**.

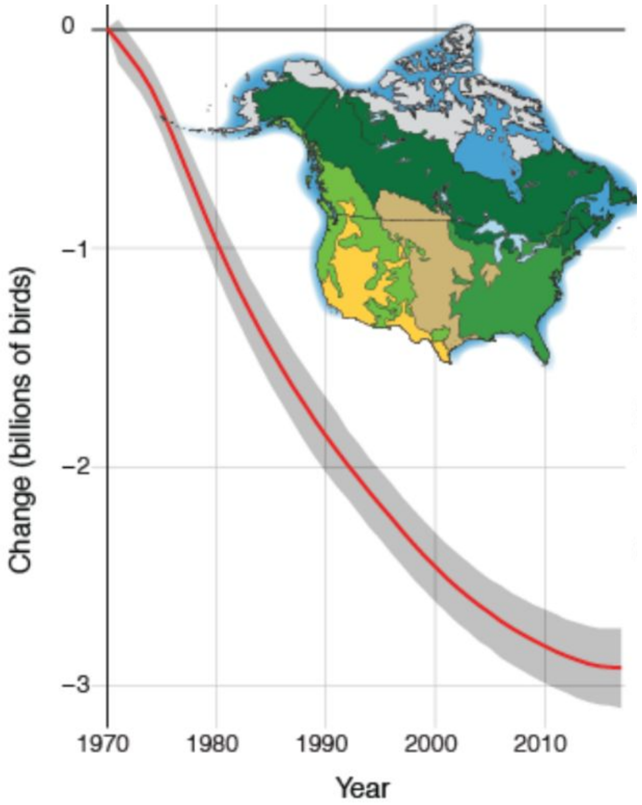
Advances in computing have driven an interest in **automated biodiversity monitoring**, such as through detection and classification of animals within camera traps or drone imagery.

Photo credit: WWF India

Photo credit: ConservationDrones.org

Biodiversity Loss

Over the last 50 years, the US has lost an estimated **3 billion birds**.



Rosenberg, Kenneth V., et al. "Decline of the North American avifauna." Science (2019).

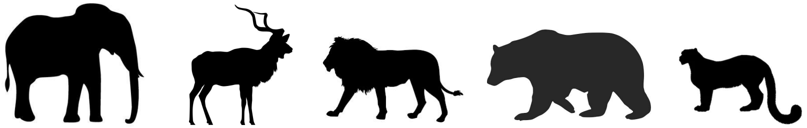
To prevent further loss, it is necessary to understand **large-scale population dynamics**.

Advances in computing have driven an interest in **automated biodiversity monitoring**, such as through detection and classification of animals within camera traps or drone imagery.

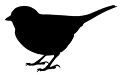
Photo credit: WWF India

Photo credit: ConservationDrones.org

These methods are successful for some taxonomic groups...

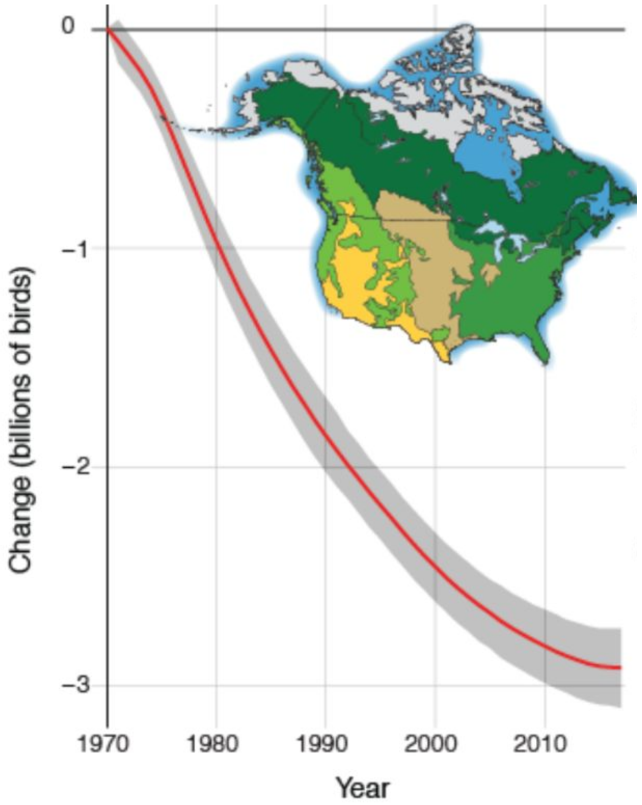


…but challenging to use for **birds**.



Biodiversity Loss

Over the last 50 years, the US has lost an estimated **3 billion birds**.



Rosenberg, Kenneth V., et al. "Decline of the North American avifauna." Science (2019).

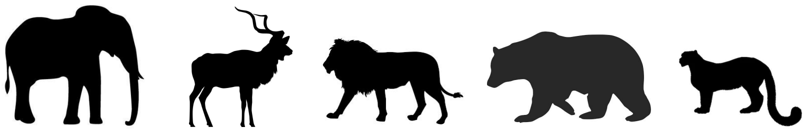
To prevent further loss, it is necessary to understand **large-scale population dynamics**.

Advances in computing have driven an interest in **automated biodiversity monitoring**, such as through detection and classification of animals within camera traps or drone imagery.

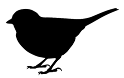
Photo credit: WWF India

Photo credit: ConservationDrones.org

These methods are successful for some taxonomic groups...

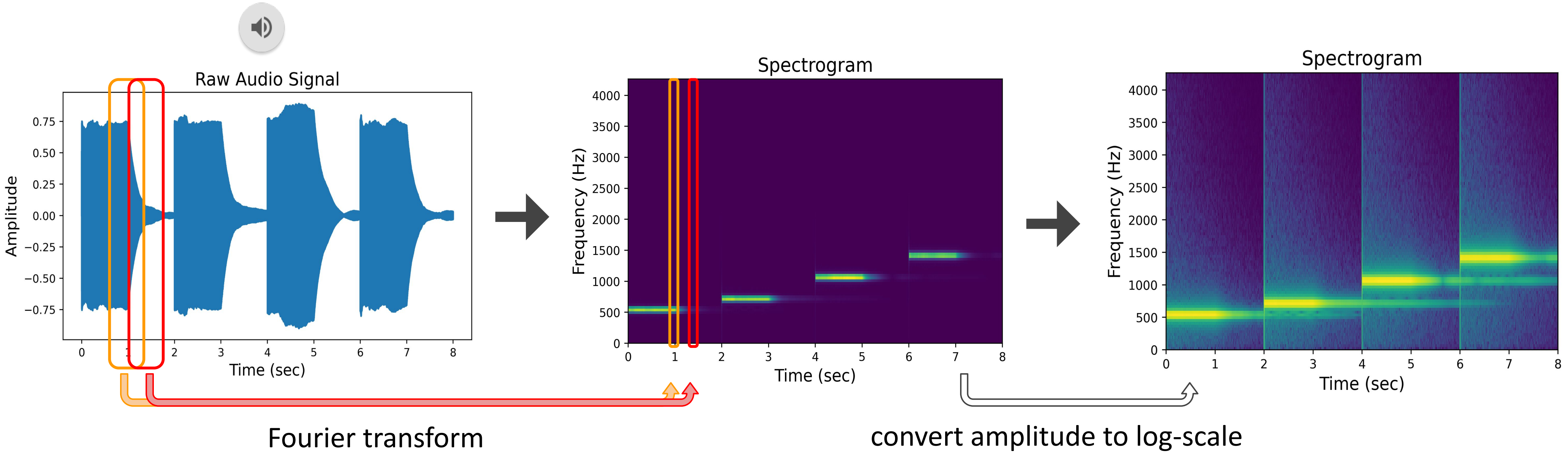


…but challenging to use for **birds**.



Could audio be used for automated monitoring?

How is audio processed?



Note that by transforming audio data to the time-frequency domain, this becomes an **image analysis problem**!

Consequently, detection and classification of birdsong is achieved with **convolutional neural networks** (CNNs).

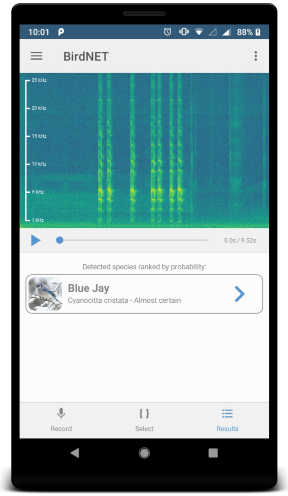
Project Background: BirdNET

BirdNET is a **citizen-science** biodiversity monitoring project developed by the Cornell Lab of Ornithology.

Project Background: BirdNET

BirdNET is a **citizen-science** biodiversity monitoring project developed by the Cornell Lab of Ornithology.

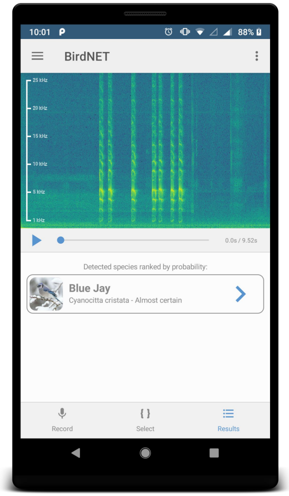
First, a CNN-based classifier is trained to identify and classify birdsong by species.



Project Background: BirdNET

BirdNET is a **citizen-science** biodiversity monitoring project developed by the Cornell Lab of Ornithology.

First, a CNN-based classifier is trained to identify and classify birdsong by species.

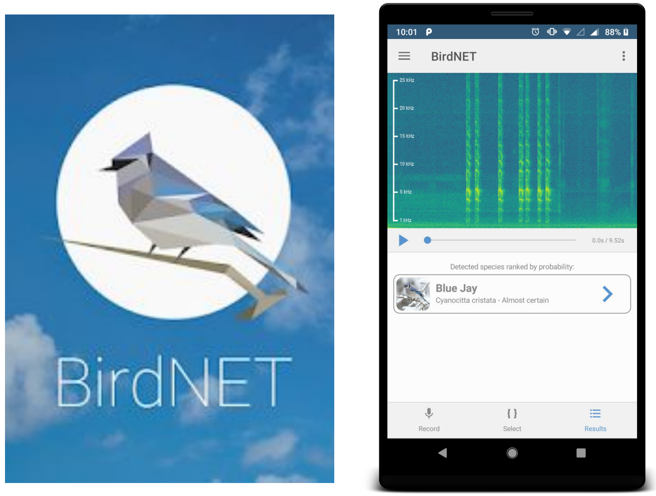


Then, anyone with the BirdNET app can submit recordings of birds and see the classification.

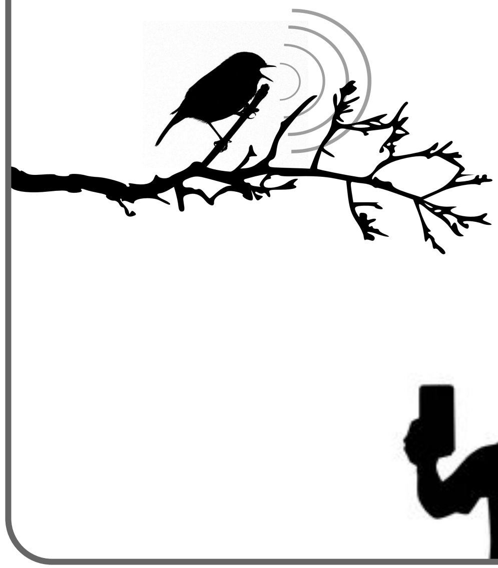
Project Background: BirdNET

BirdNET is a **citizen-science** biodiversity monitoring project developed by the Cornell Lab of Ornithology.

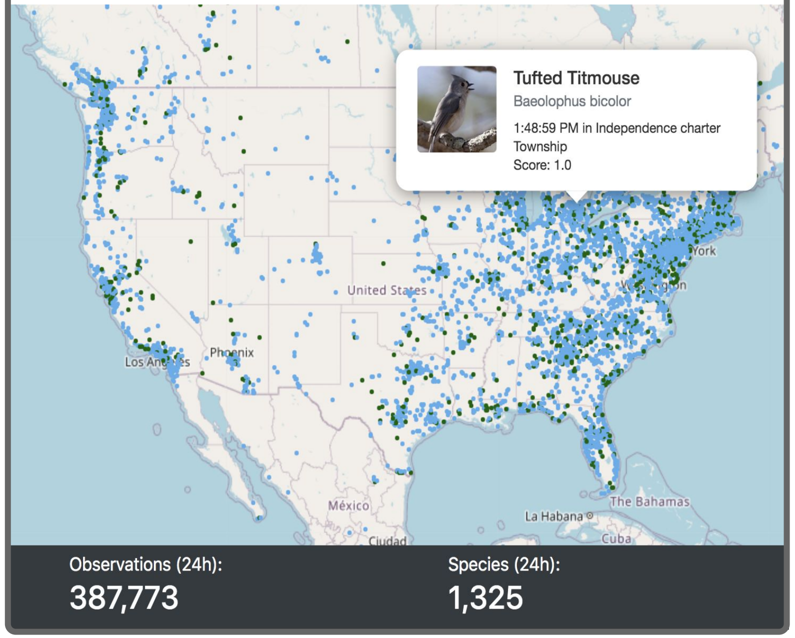
First, a CNN-based classifier is trained to identify and classify birdsong by species.



Then, anyone with the BirdNET app can submit recordings of birds and see the classification.



This data is then sent back for large-scale ecological analysis.



How well does this system perform?

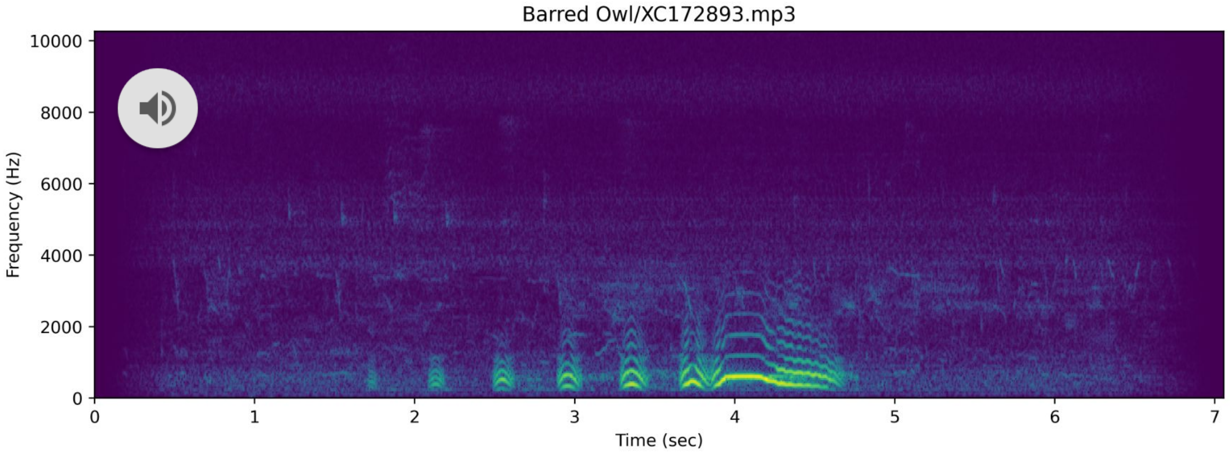
The “Domain Shift”

Evaluations of BirdNET have shown **15-20% false positive rates**.

The "Domain Shift"

Evaluations of BirdNET have shown **15-20% false positive rates**.

The training data from **Xeno Canto** contains high-quality "focal" recordings, often taken with directional microphones.

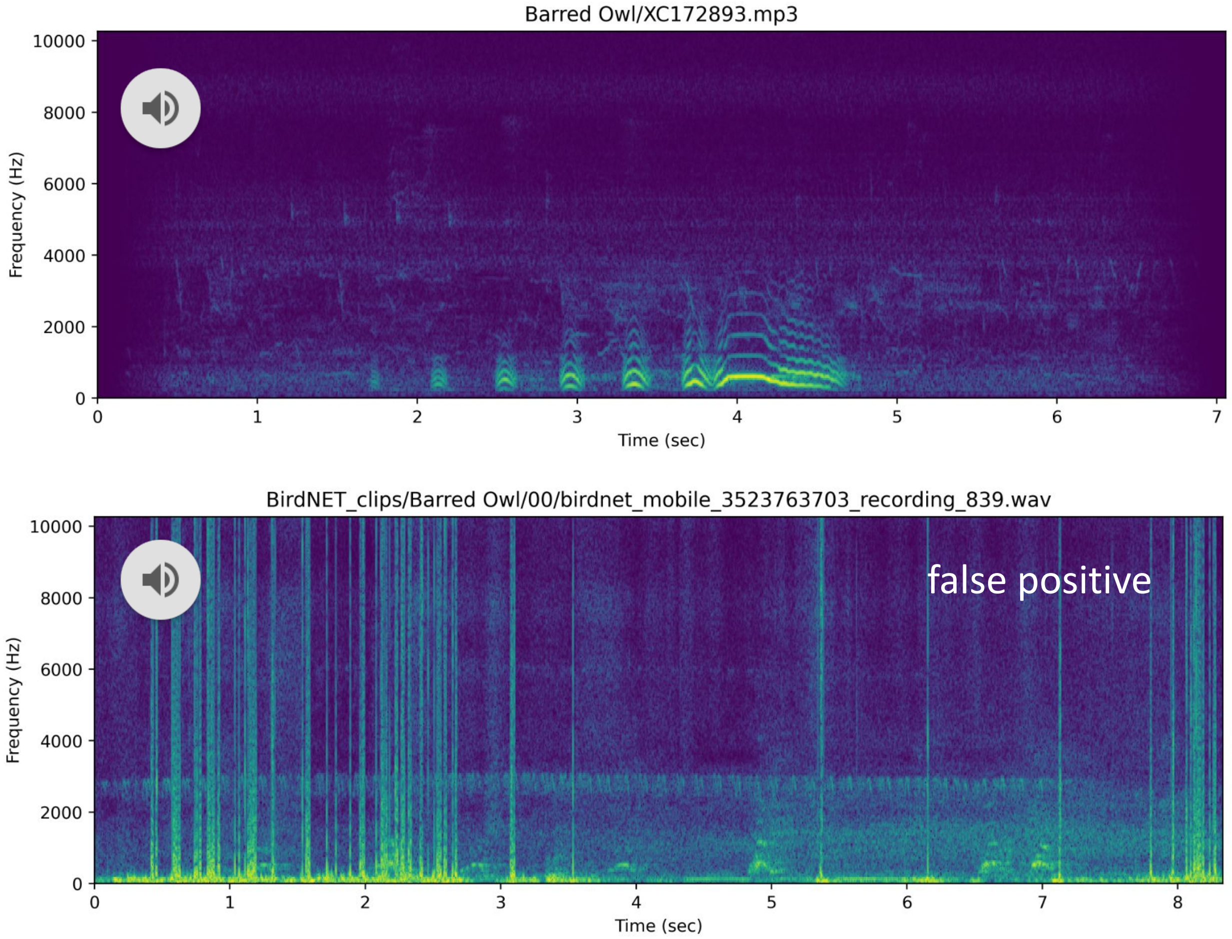


The “Domain Shift”

Evaluations of BirdNET have shown **15-20% false positive rates**.

The training data from **Xeno Canto** contains high-quality “focal” recordings, often taken with directional microphones.

In contrast, **BirdNET submissions** are taken with a cell phone in noisy real-world conditions.

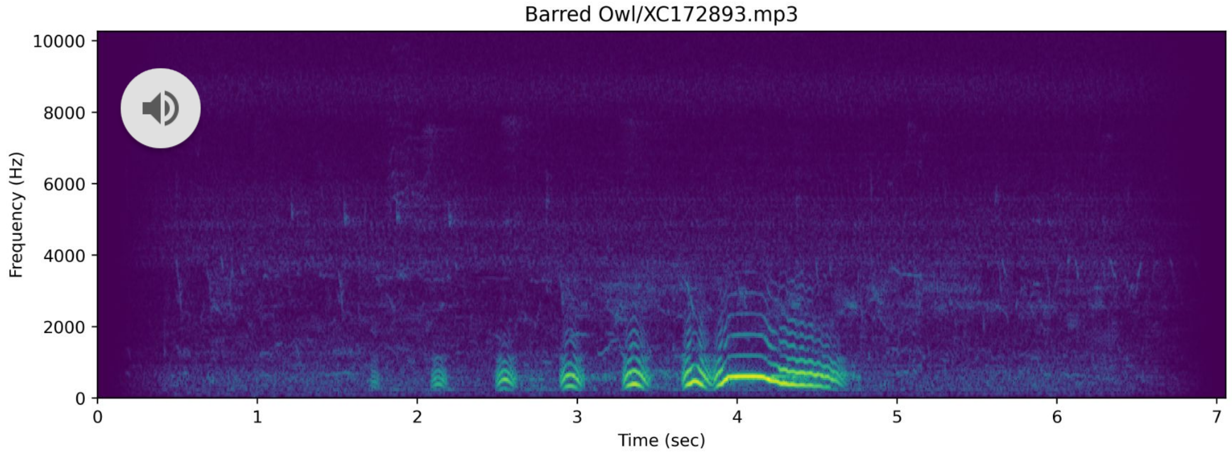


“Domain Shift”

The “Domain Shift”

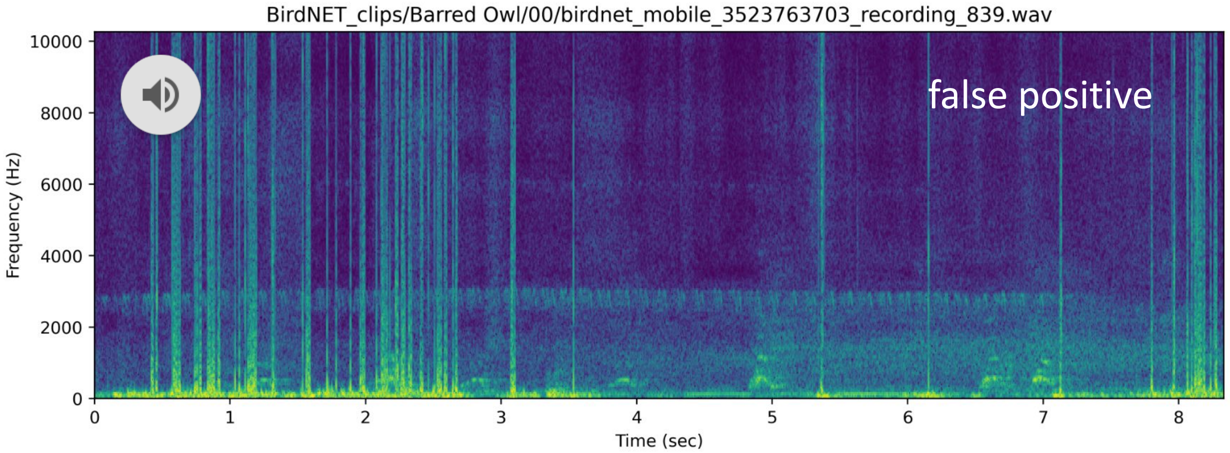
Evaluations of BirdNET have shown **15-20% false positive rates**.

The training data from **Xeno Canto** contains high-quality “focal” recordings, often taken with directional microphones.



“Domain Shift”

In contrast, **BirdNET submissions** are taken with a cell phone in noisy real-world conditions.



Is there a way we can **understand this domain shift and diagnose the misclassifications?**

Approach and Goals

1 **Pre-process** the submitted BirdNET recordings by splitting them into overlapping 1-second segments, and selecting the segments with highest vocal activity.

Approach and Goals

1. **Pre-process** the submitted BirdNET recordings by splitting them into overlapping 1-second segments, and selecting the segments with highest vocal activity.

2. **Calculate embeddings** over these segments with pre-trained networks (VGGish and PANNs), along with a trained convolutional autoencoder.

Approach and Goals

1. **Pre-process** the submitted BirdNET recordings by splitting them into overlapping 1-second segments, and selecting the segments with highest vocal activity.

2. **Calculate embeddings** over these segments with pre-trained networks (VGGish and PANNs), along with a trained convolutional autoencoder.

3. **Cluster the segments** using k-means in the embedded space, and examine inter- and intra-specific structure.

Approach and Goals

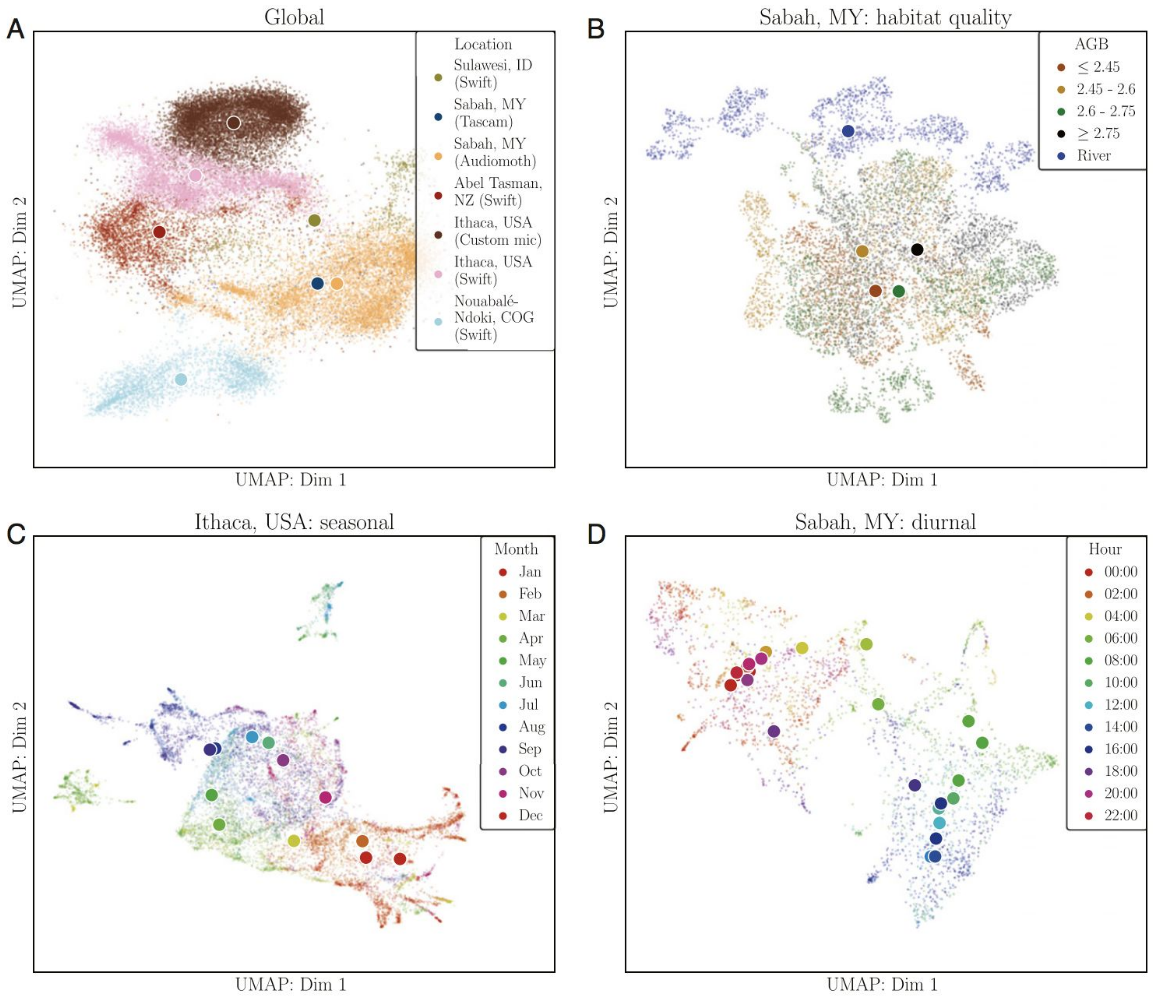
1. **Pre-process** the submitted BirdNET recordings by splitting them into overlapping 1-second segments, and selecting the segments with highest vocal activity.

2. **Calculate embeddings** over these segments with pre-trained networks (VGGish and PANNs), along with a trained convolutional autoencoder.

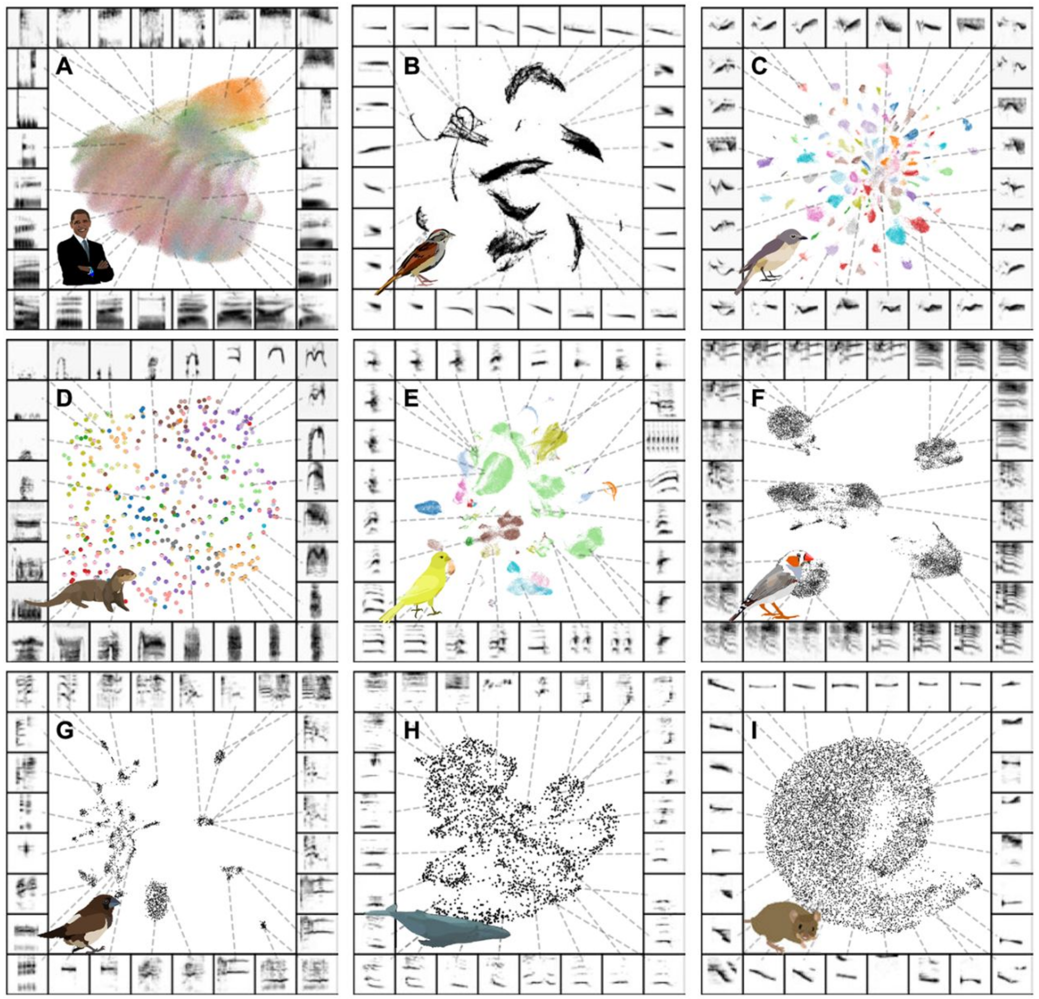
3. **Cluster the segments** using k-means in the embedded space, and examine inter- and intra-specific structure.

4. By **incorporating expert labels** for clusters within intraspecific data, **generate alternative confidence scores** for each BirdNET recording, and **flag false positives**.

Prior Work in Embedding Bio-Acoustic Data



Sethi, Sarab S., et al. "Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set." Proceedings of the National Academy of Sciences 117.29 (2020): 17049-17055



Sainburg, Tim, Marvin Thielk, and Timothy Q. Gentner. "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires." PLoS computational biology 16.10 (2020): e1008228.

VGGish Pre-trained on AudioSet

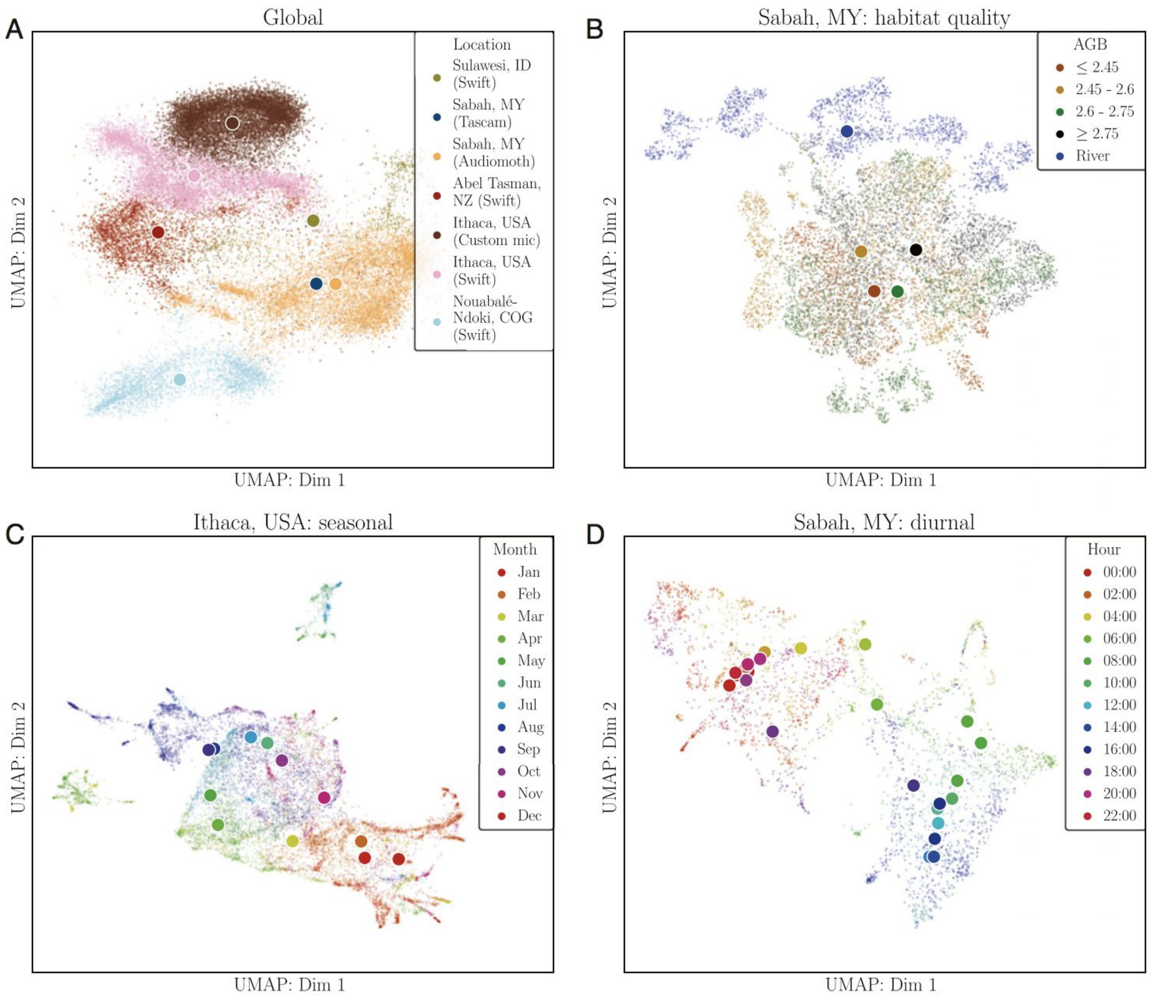
VGGish is a CNN architecture for audio, inspired by the VGG network for image classification.

CNN architectures for large-scale audio classification

S Hershey, S Chaudhuri, DPW Ellis... - ... on acoustics, speech ..., 2017 - ieeexplore.ieee.org

Convolutional Neural Networks (CNNs) have proven very effective in image classification and show promise for audio. We use various CNN architectures to classify the soundtracks of a dataset of 70M training videos (5.24 million hours) with 30,871 video-level labels. We examine fully connected Deep Neural Networks (DNNs), AlexNet [1], VGG [2], Inception [3], and ResNet [4]. We investigate varying the size of both training set and label vocabulary, finding that analogs of the CNNs used in image classification do well on our audio ...

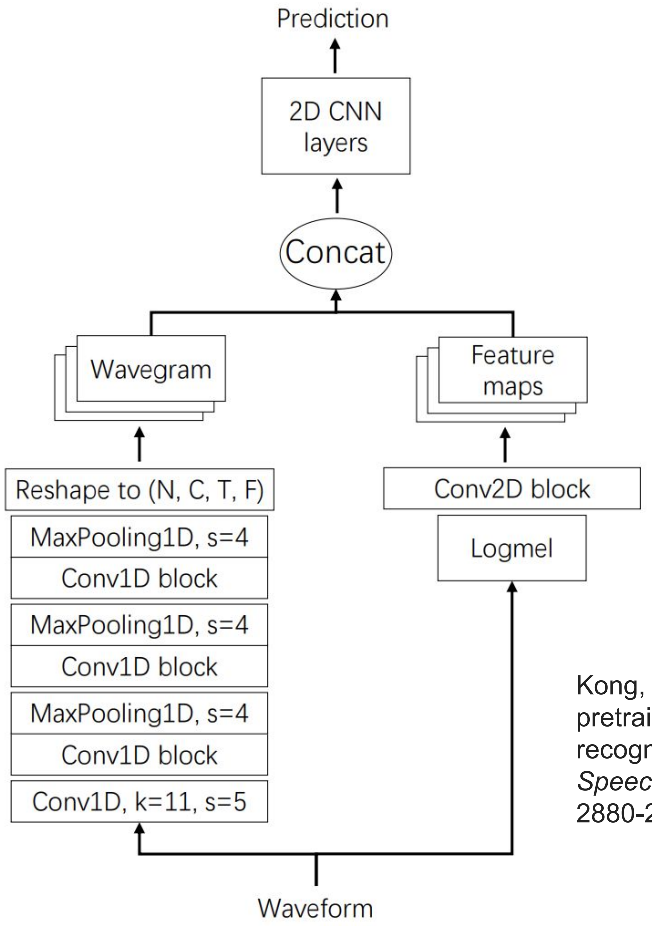
Sethi et al showed very detailed structure of environmental soundscapes when analyzed with VGGish embeddings (right).



Sethi, Sarab S., et al. "Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set." Proceedings of the National Academy of Sciences 117.29 (2020): 17049-17055

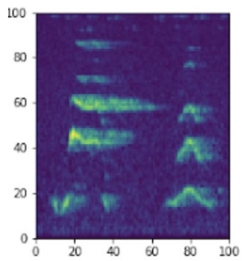
Pre-trained Audio Neural Networks (PANNs)

We looked at a recently-released (2020) set of pre-trained audio networks with high performance on classification tasks and an easy/accessible user interface.



Kong, Qiuqiang, et al. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 2880-2894.

Audio sample of a 1-second window



Pre-trained Network

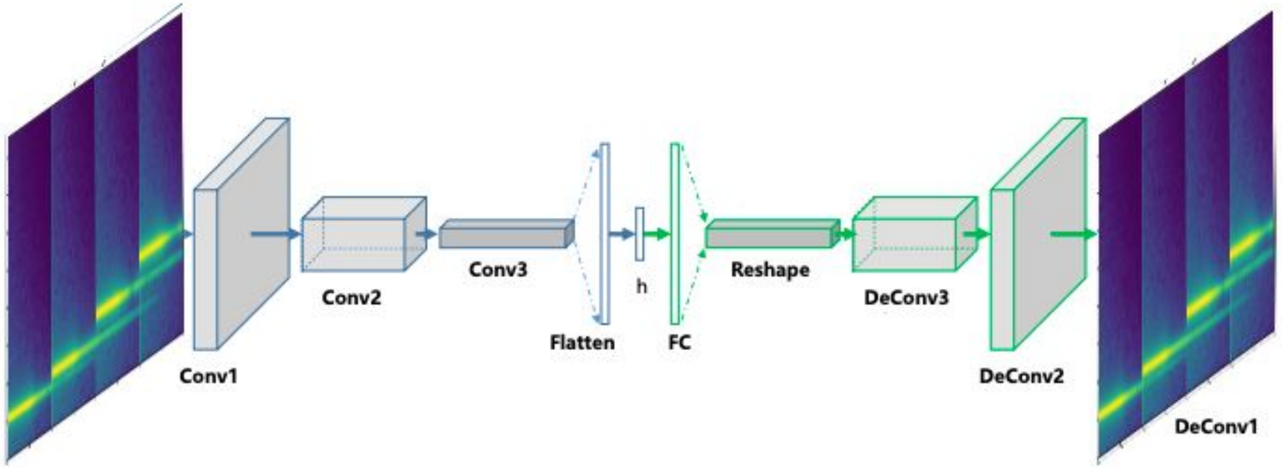
Embedding: vector of size 2048

Use UMAP to reduce to 2 dimensions for visualization

Classification output: probability across 527 categories

Use output as a rough heuristics for bird call likelihood

Convolutional Autoencoder



- An autoencoder aims to optimally reconstruct the data while passing it through an **information bottleneck**.
- The **activations of this "bottleneck" middle layer** can be interpreted as a lower-dimensional embedding.
- We used **4** convolutional layers and an embedding size of **128**.

Autoencoder Parameters:

Layer	Operation	In Size	Out Size	Kernel
1	conv relu	100x100x1 49x49x32	49x49x32 49x49x32	4x4
2	conv relu	49x49x32 23x23x64	23x23x64 23x23x64	4x4
3	conv relu	23x23x64 10x10x128	10x10x128 10x10x128	4x4
4	conv relu flatten	10x10x128 4x4x256 4x4x256	4x4x256 4x4x246 4096	4x4
5	fc	4096	128	
6	fc unflatten	128 4096	4096 1x1x4096	
7	conv relu	1x1x4096 7x7x128	7x7x128 7x7x128	7x7
8	conv relu	7x7x128 20x20x64	20x20x64 20x20x64	8x8
9	conv relu	20x20x64 47x47x32	47x47x32 47x47x32	9x9
10	conv sigmoid	47x47x32 100x100x1	100x100x1 100x100x1	8x8

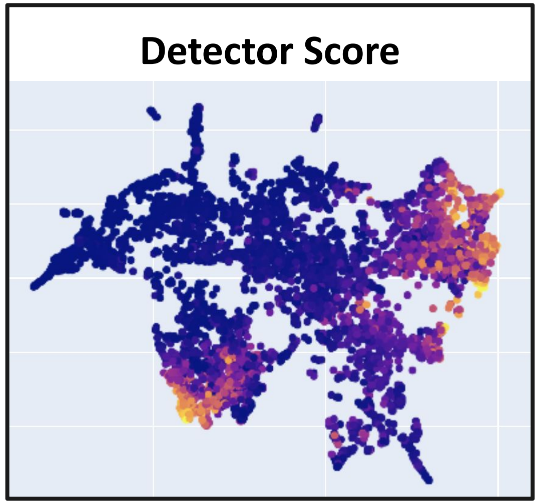
Visualizing Embeddings

We take BirdNET submissions classified as the Tawny Owl, split them into small overlapping windows, and calculate spectrograms.

Visualizing Embeddings

We take BirdNET submissions classified as the Tawny Owl, split them into small overlapping windows, and calculate spectrograms.

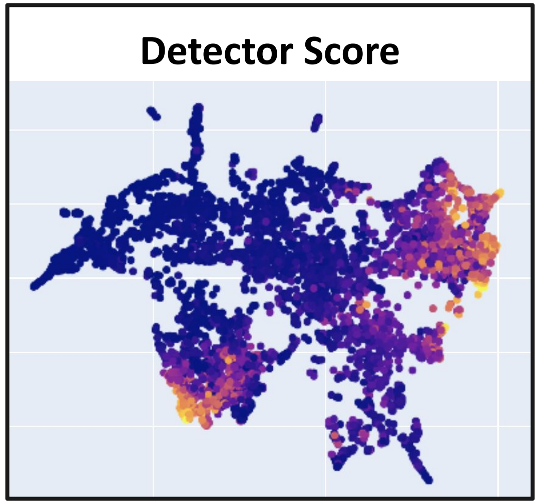
Next, we obtain 2048-D embeddings from the PANNs network, and use UMAP to project them to 2D for visualization.

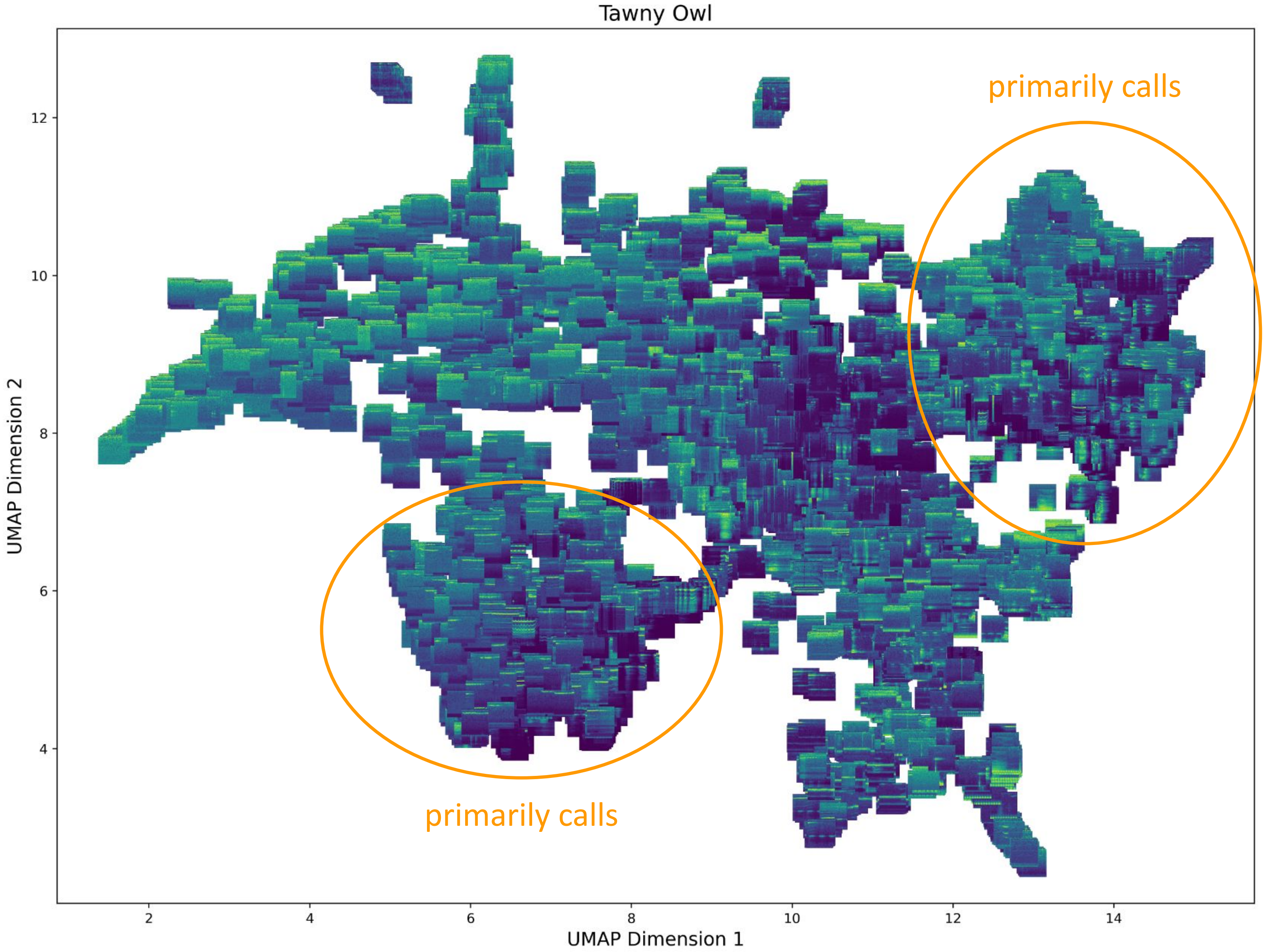


Visualizing Embeddings

We take BirdNET submissions classified as the Tawny Owl, split them into small overlapping windows, and calculate spectrograms.

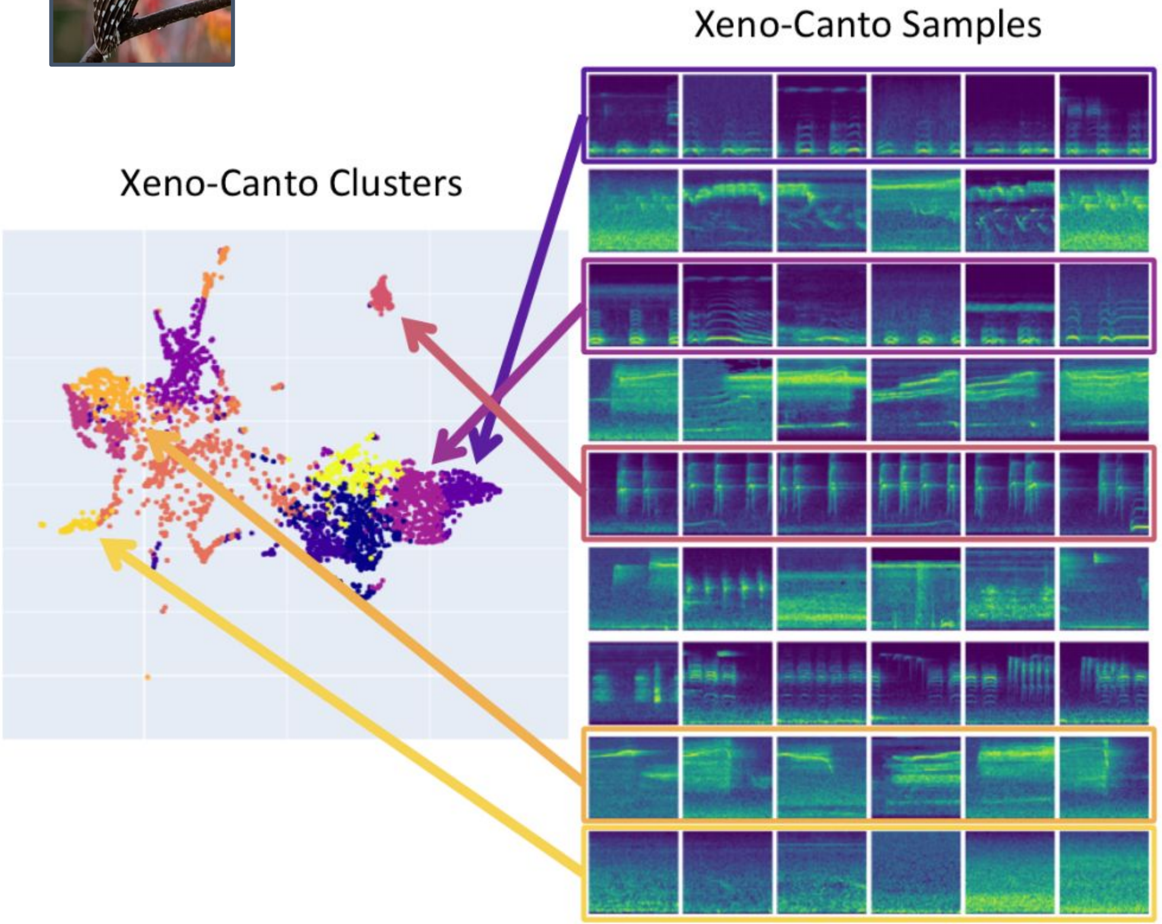
Next, we obtain 2048-D embeddings from the PANNs network, and use UMAP to project them to 2D for visualization.





Clustering and Labeling

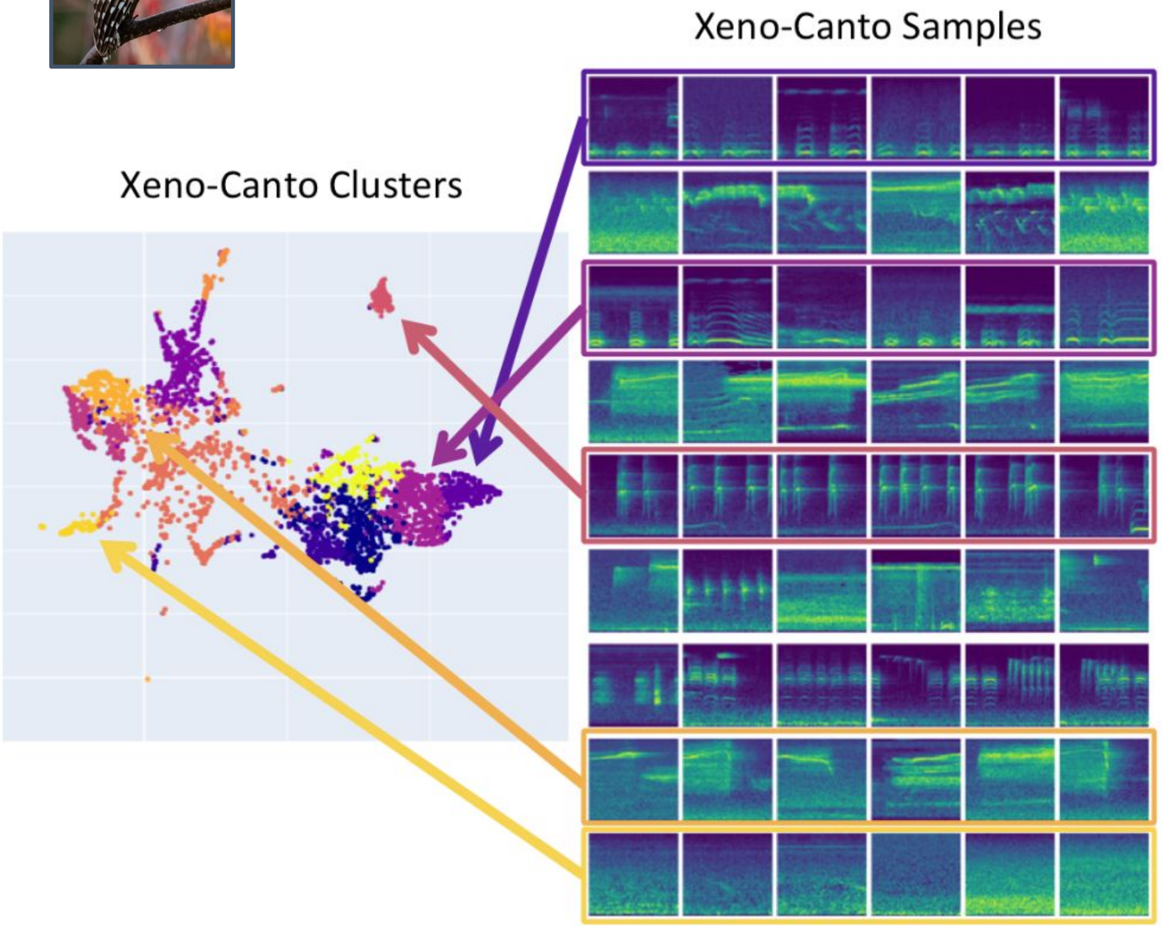
Clusters and Samples for the Barred Owl



- We **apply k-means to obtain 12 clusters** in the high-dimensional embedding space.

Clustering and Labeling

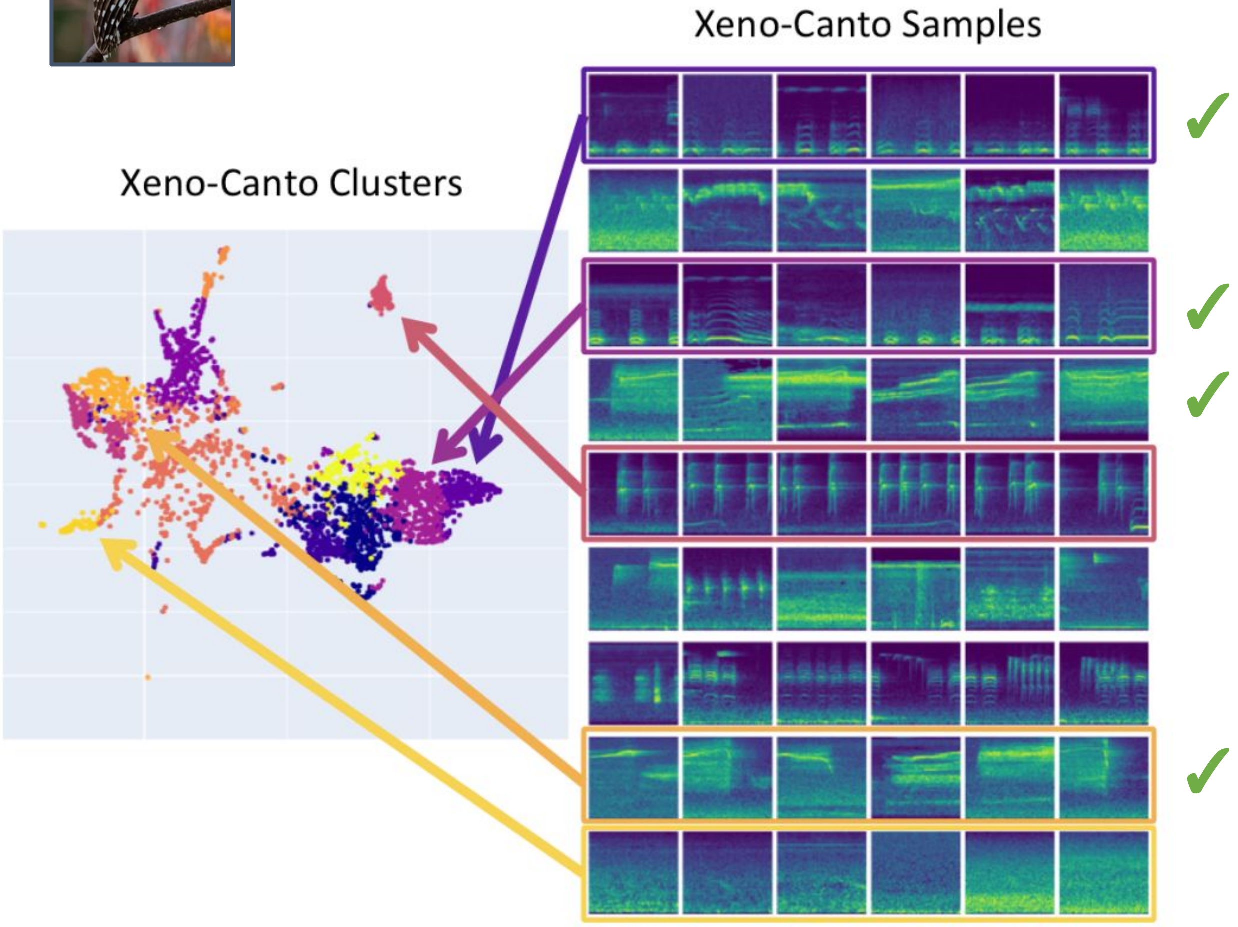
Clusters and Samples for the Barred Owl



- We **apply k-means to obtain 12 clusters** in the high-dimensional embedding space.
- For each cluster:
  - we **present random samples to an expert**
  - **retrieve a binary label** indicating whether a cluster represents calls from this species

Clustering and Labeling

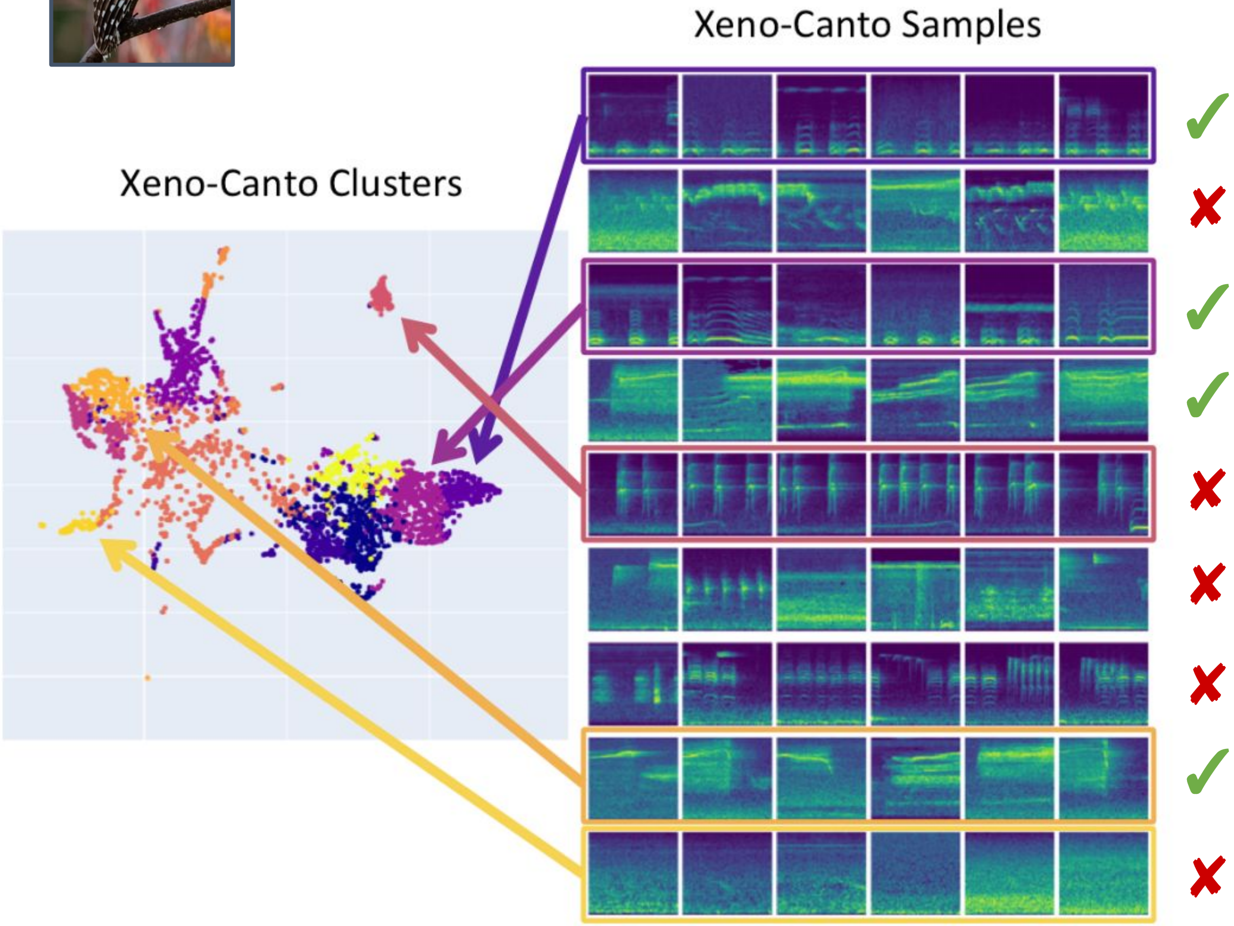
Clusters and Samples for the Barred Owl



- We **apply k-means to obtain 12 clusters** in the high-dimensional embedding space.
- For each cluster:
  - we **present random samples to an expert**
  - **retrieve a binary label** indicating whether a cluster represents calls from this species

Clustering and Labeling

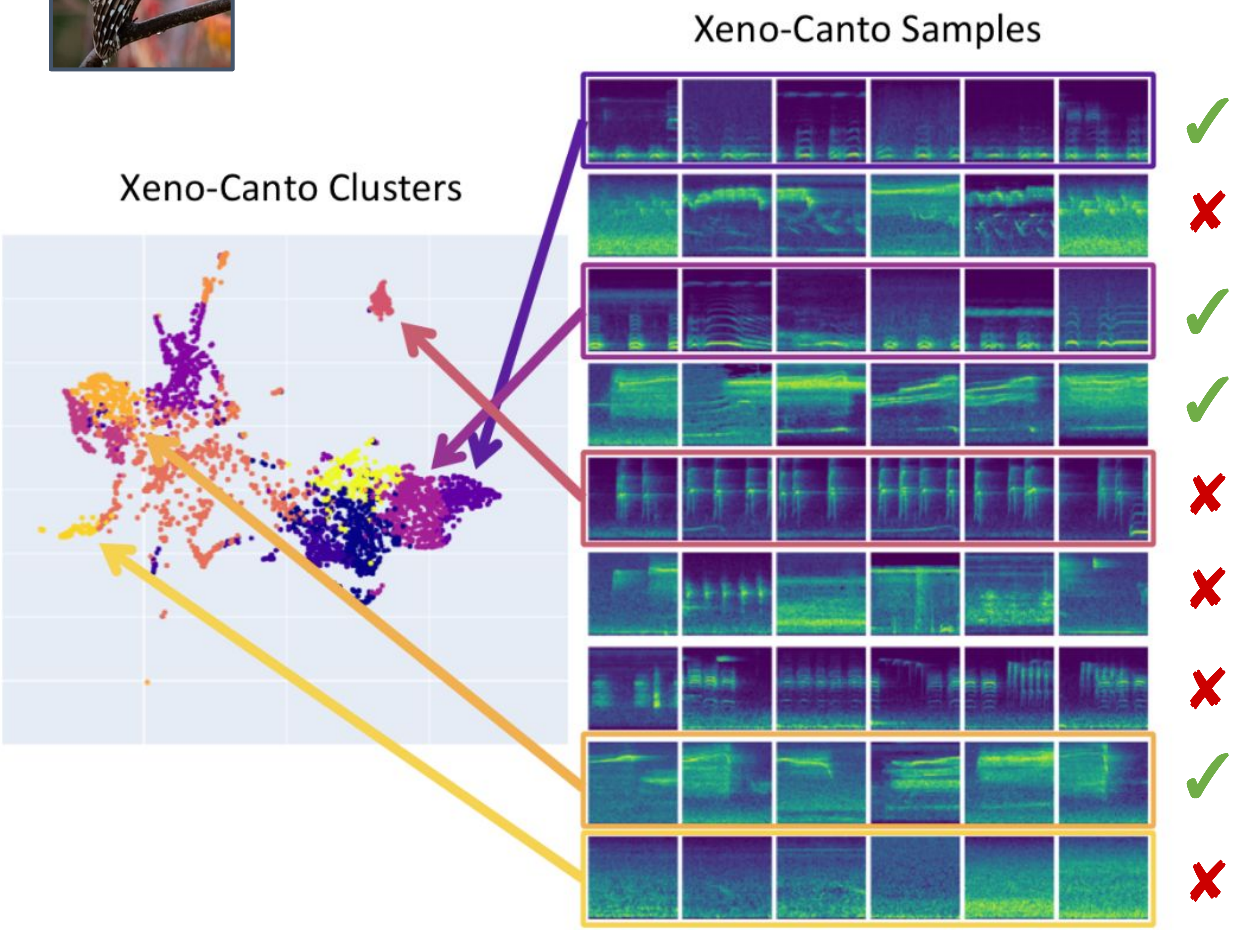
Clusters and Samples for the Barred Owl



- We **apply k-means to obtain 12 clusters** in the high-dimensional embedding space.
- For each cluster:
  - we **present random samples to an expert**
  - **retrieve a binary label** indicating whether a cluster represents calls from this species

Clustering and Labeling

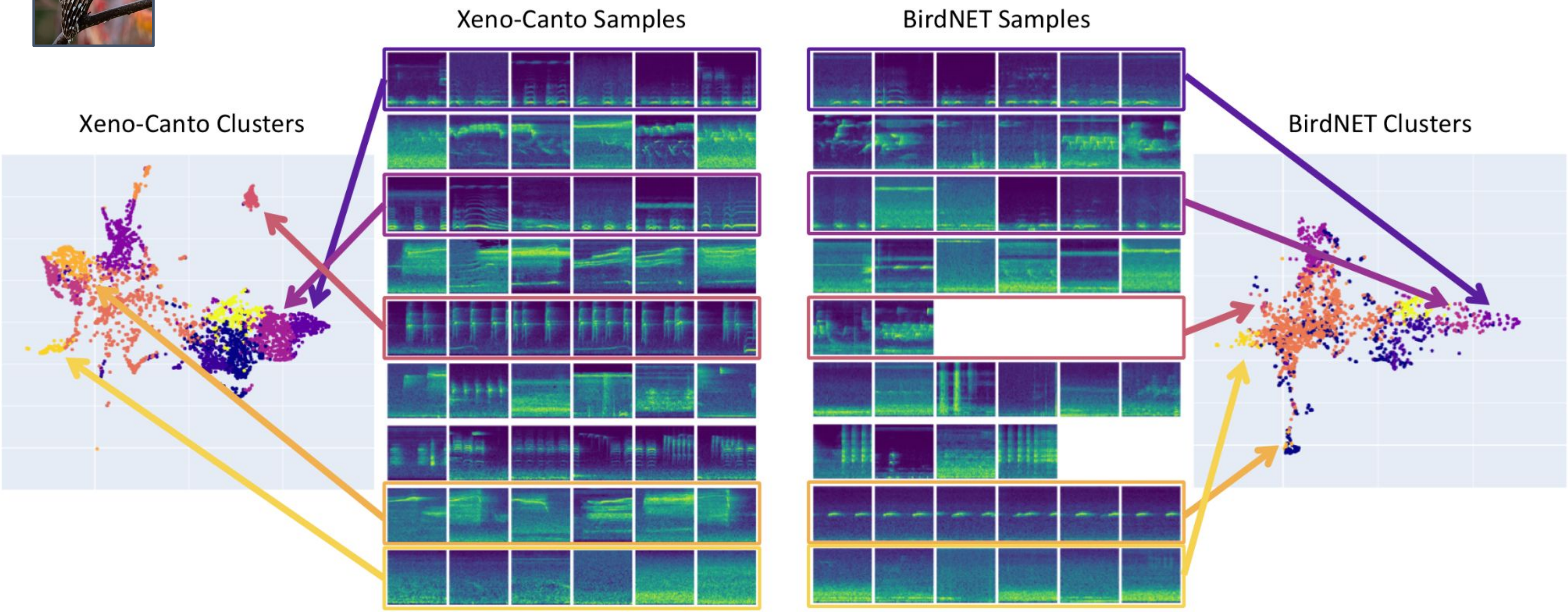
Clusters and Samples for the Barred Owl



- We **apply k-means to obtain 12 clusters** in the high-dimensional embedding space.
- For each cluster:
  - we **present random samples to an expert**
  - **retrieve a binary label** indicating whether a cluster represents calls from this species
- Next, for the BirdNET data:
  - we **assign binary labels to each BirdNET segment** based on its Xeno-Canto neighbors

Clustering and Labeling

Clusters and Samples for the Barred Owl



Results

Qualitative Evaluation:

Embeddings allow **interpretation of the data**, and **visualization of the domain shift** between the Xeno-Canto and BirdNET datasets.

Using audio embeddings also reveals **different call types** within a single species, along with clusters of vocalizations from **other species**. Human-in-the-loop input can help to remove such erroneous samples from analysis.

Results

Qualitative Evaluation:

Embeddings allow **interpretation of the data**, and **visualization of the domain shift** between the Xeno-Canto and BirdNET datasets.

Using audio embeddings also reveals **different call types** within a single species, along with clusters of vocalizations from **other species**. Human-in-the-loop input can help to remove such erroneous samples from analysis.

Quantitative Evaluation:

Finally, we classify a BirdNET recording as a positive if it contains at least two "positive" segments.

We focus on the **precision** of this classification method, and specifically on **improvement on precision** (in bold) over the initial BirdNET classifier.

	Barred Owl			**Common Crane**			**Common Loon**		
Architecture	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
VGG-ish	0.65	0.52 (**+.15**)	0.61	0.47	0.40 (-.03)	0.45	0.58	0.05 (**+.01**)	0.44
W-L-CNN16	0.73	0.74 (**+.37**)	0.43	0.55	0.48 (**+.05**)	0.36	0.95	0.33 (**+.29**)	0.22
Autoencoder	0.57	0.44 (**+.07**)	0.63	0.40	0.39 (-.04)	0.71	0.96	0.50 (**+.46**)	0.11

Quantitatively, the PANNs network (W-L-CNN16) gives the best improvement in precision across species.

Conclusions, Next Steps, and Broader Impact

In summary, we have implemented a pipeline for processing BirdNET submissions, constructing embeddings, and analyzed latent structure and clustering.

We hope that this work can contribute to conservation efforts through passive acoustic monitoring.

Our next steps:

- Expand analysis to a greater number of species (particularly to songbirds)
- Incorporate false positive identification into the BirdNET classifier
- Use contrastive learning to better separate intraspecific call types

Broader Impact and Ethics

- Access to the BirdNET app will be **limited** by smartphone prevalence and public awareness.
- This may cause **geographically variable performance**, which is significant, as over-estimation of species population sizes may inhibit conservation measures for those species.
- Overall, the anticipated benefits to conservation greatly outweigh these risks.

Acknowledgements

We are very grateful to **Daniel Salisbury** for manually labeling the BirdNET data used in this work.

We also thank **Connor Wood** and **Ben Mirin** for feedback and species identification.

Additionally, we would like to thank **Milind Tambe**, **Doria Spiegel**, and **Boriana Gjura** for their support on this project.

HARVARD John A. Paulson School of Engineering and Applied Sciences

TheCornellLab Center for Conservation Bioacoustics

qbio HARVARD QUANTITATIVE BIOLOGY

NSF SIMONS CENTER HARVARD MATHEMATICAL & STATISTICAL ANALYSIS OF BIOLOGY

Thank you all for listening! Questions?