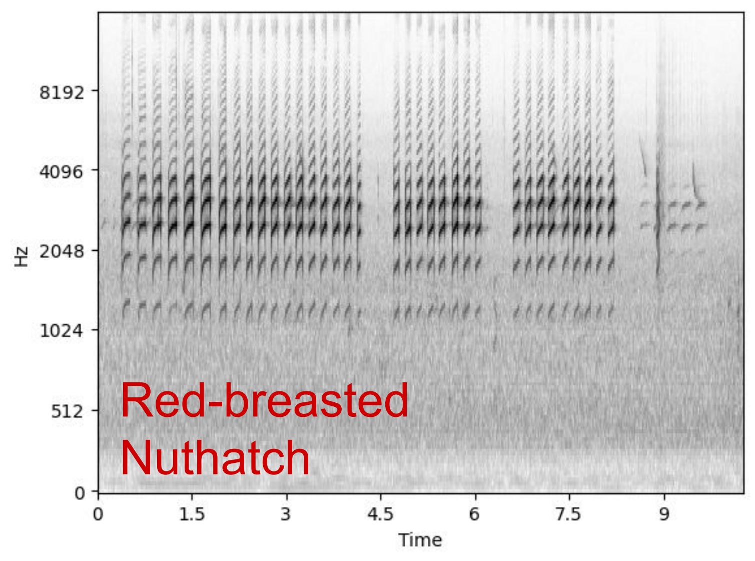
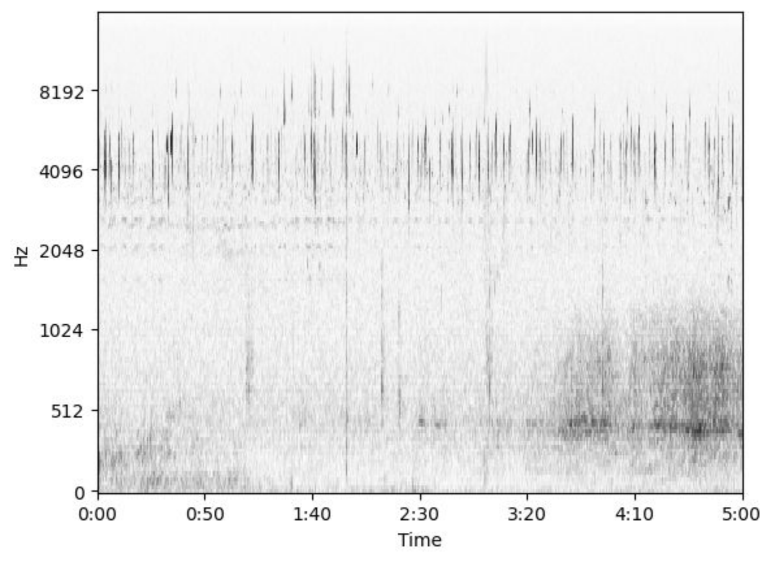
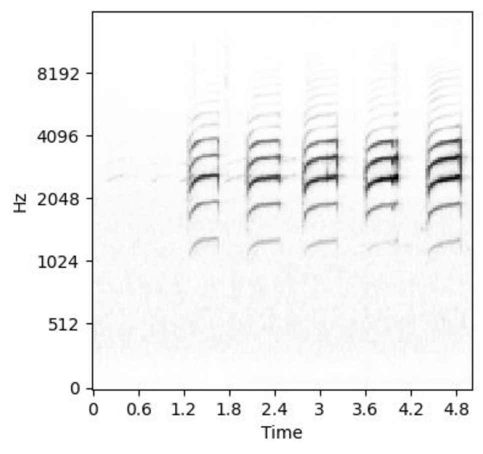
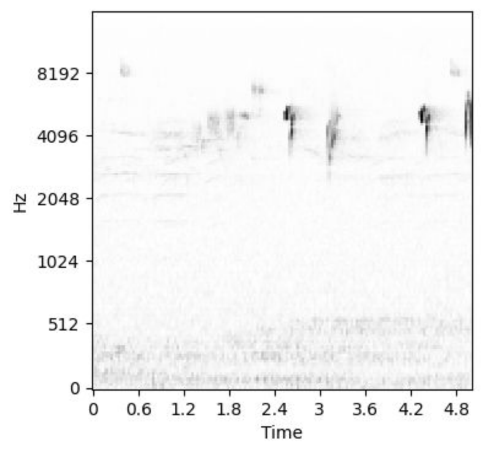
bioacoustics datasets

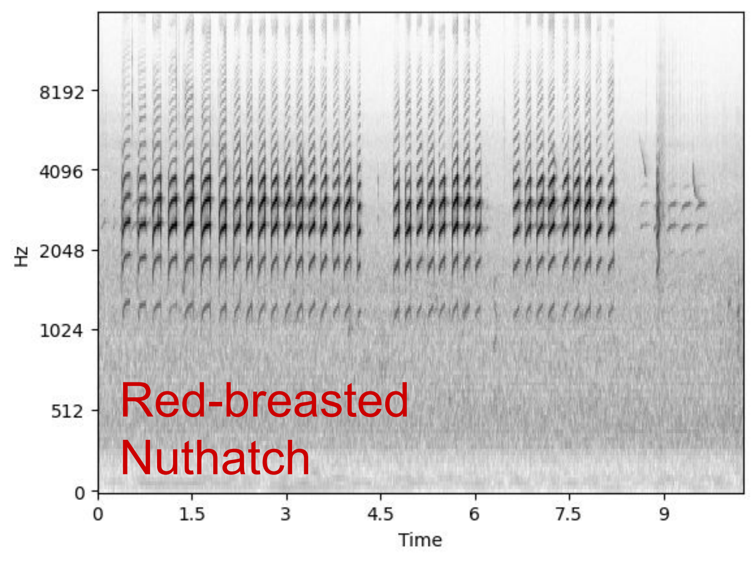


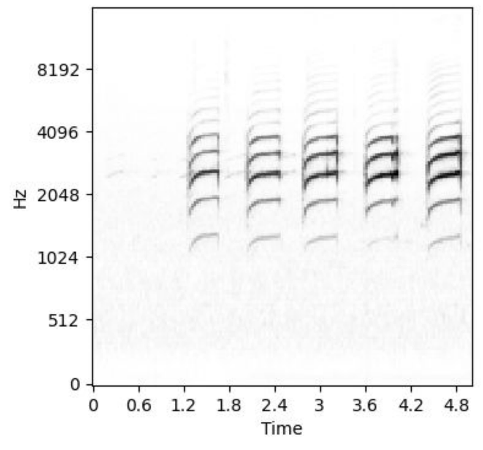
Red-breasted Nuthatch

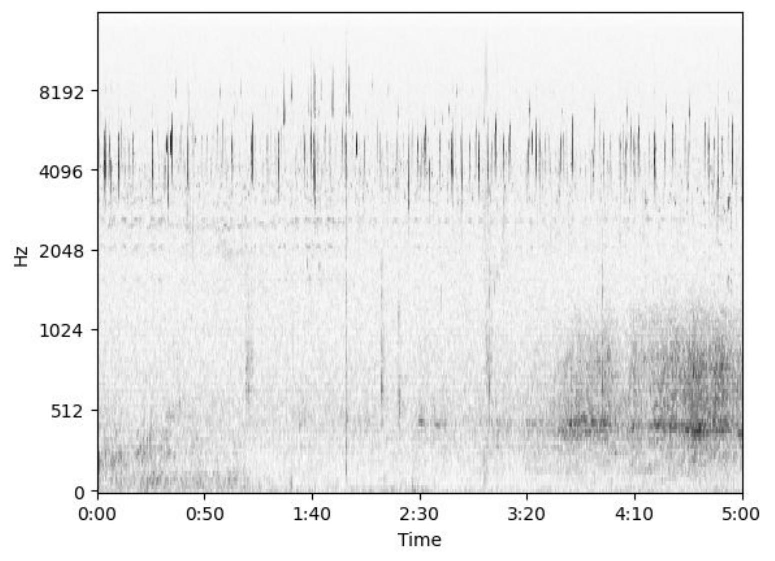


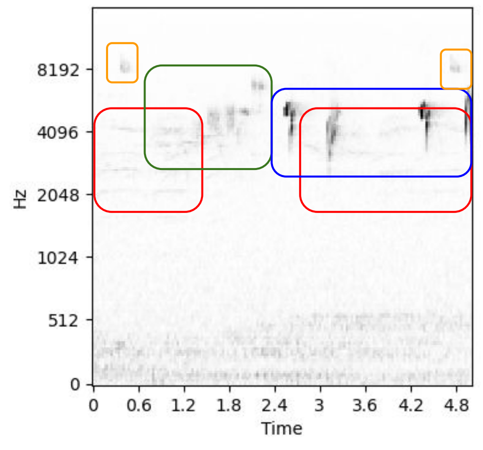





Red-breasted Nuthatch
Hz
8192
4096
2048
1024
512
0
0
1.5
3
4.5
6
7.5
9
Time


Hz
8192
4096
2048
1024
512
0
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time


Hz
8192
4096
2048
1024
512
0
0:00
0:50
1:40
2:30
3:20
4:10
5:00
Time


Hz
8192
4096
2048
1024
512
0
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time

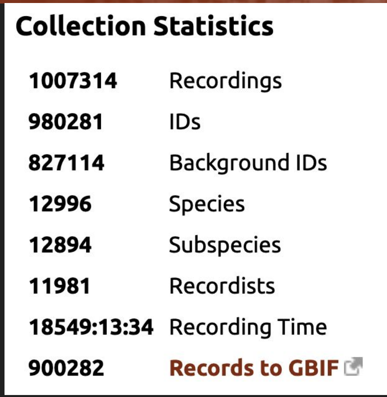
Focal Recordings

Recordings of a **target species**, usually captured by a human with a microphone.

- Generally 1s – 5m in length.
- Other species may be present in a given recording.
- iNaturalist is ~⅓ the size of XC, but has a wider variety of taxa.
- Key for training foundation models.

xeno-canto

Sharing wildlife sounds from around the world

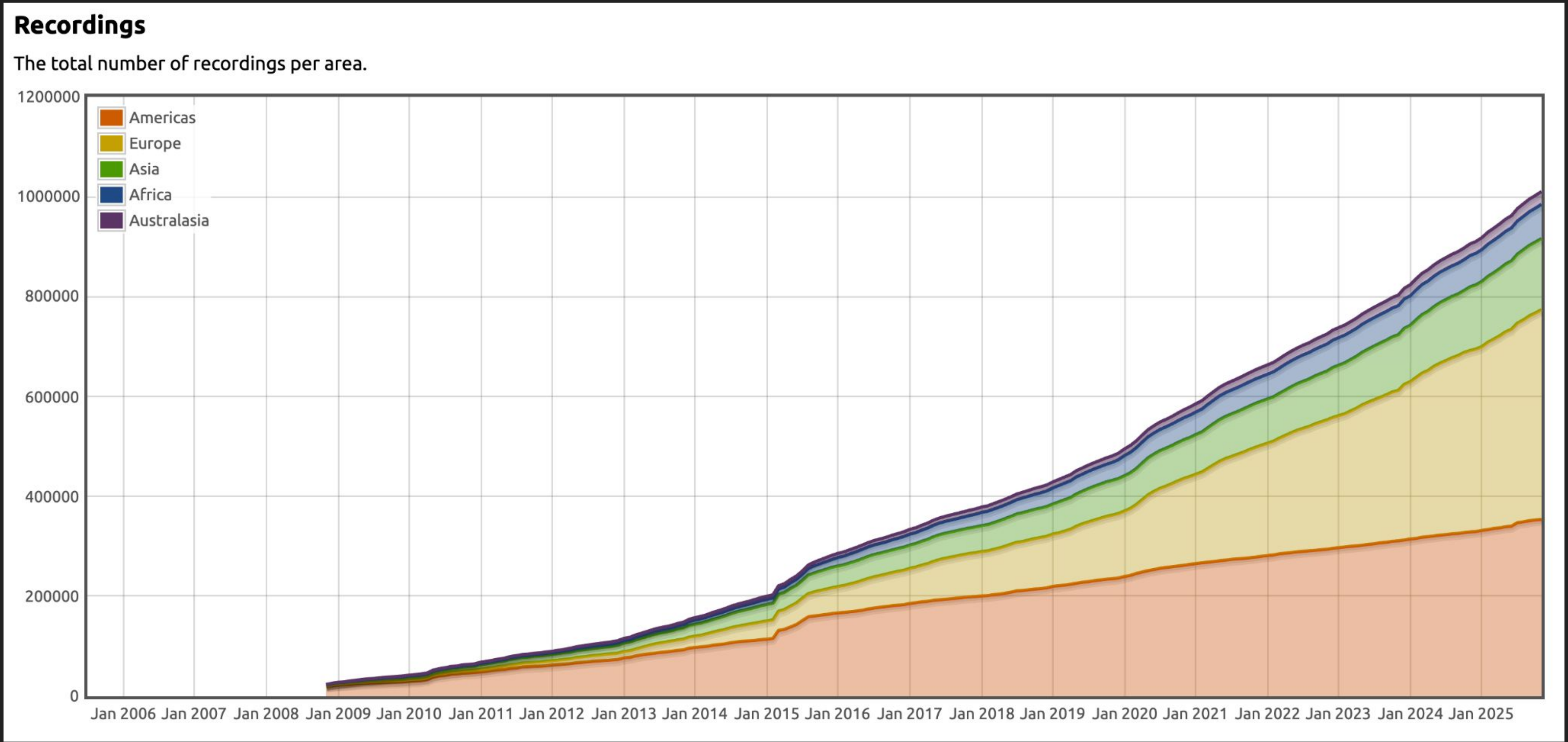


Collection Statistics

1007314	Recordings
980281	IDs
827114	Background IDs
12996	Species
12894	Subspecies
11981	Recordists
18549:13:34	Recording Time
900282	**Records to GBIF**

iNaturalist

Xeno-Canto Recordings by Location



Recordings

The total number of recordings per area.

Focal Training Data Challenges

However, it's not perfect!

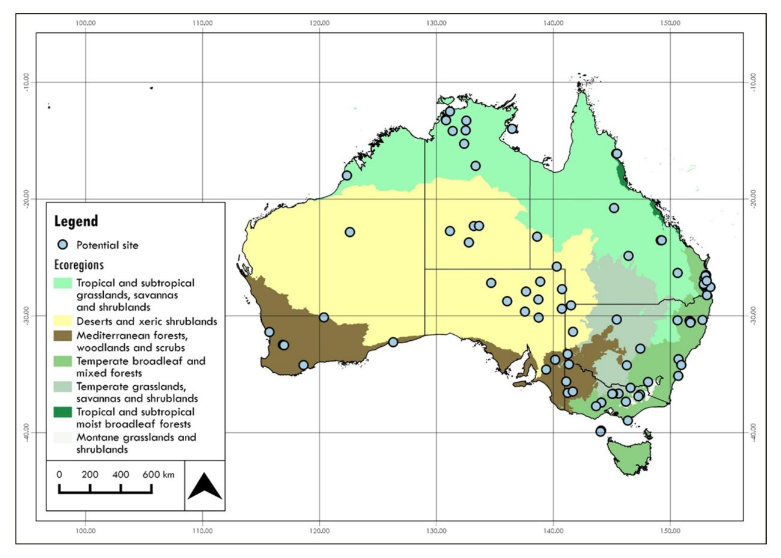
- Extreme **label imbalance**.
- **Uneven distribution** of data, with more focus on NorthAm+Europe.
- "**Weakly labeled**" - No annotations for exactly where the target occurs in file.
- **High SNR** - Tend to be very clean compared to PAM data.
- **Human biased**. May tend to contain:
  - More 'diagnostic' sounds rather than difficult to classify sounds.
  - In global north, common birds are over-represented (convenience factor).
  - In global south, common birds are under-represented (tourist factor).

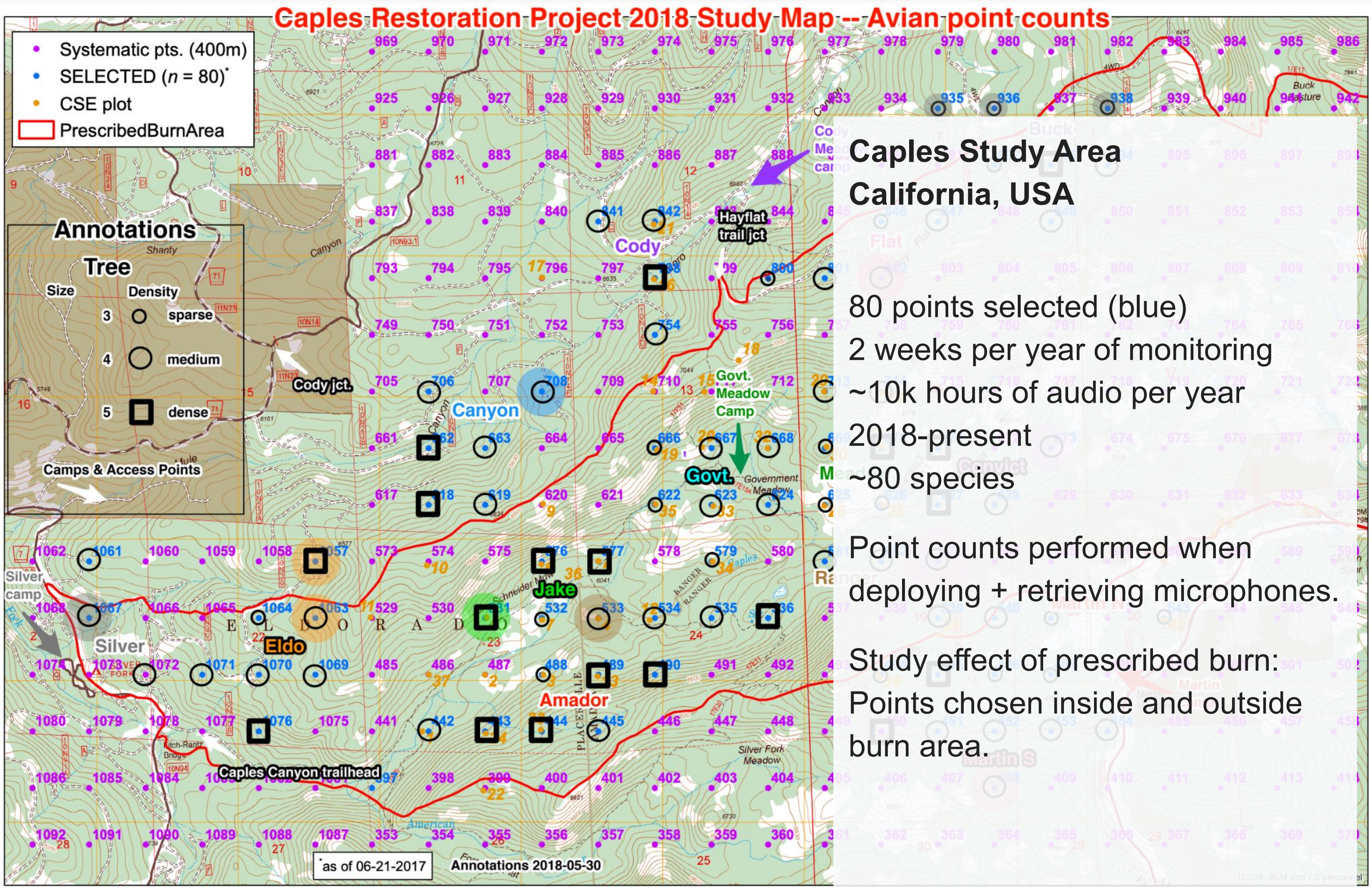
Certain birds in **India** with **the most training recordings**
had very **poor results** in the 2024 BirdCLEF competition… Can you guess why?

Passive Acoustic Datasets

- Hang 10-1000 microphones from trees, collect arbitrarily large amounts of audio.
- Usually few/no ground truth labels.
- Usually **1k-1M hours of audio**. Larger datasets can be hard to share, expensive to store.
- Typically **50-500 bird species** in a dataset, depending on deployment size + region…

Australian
Acoustic
Observatory


100.00
110.00
120.00
130.00
140.00
150.00
-10.00
-20.00
-30.00
-40.00
Legend
○ Potential site
Ecoregions
🟩 Tropical and subtropical grasslands, savannas and shrublands
🟨 Deserts and xeric shrublands
🟫 Mediterranean forests, woodlands and scrubs
🟩 Temperate broadleaf and mixed forests
🟩 Temperate grasslands, savannas and shrublands
🟩 Tropical and subtropical moist broadleaf forests
⬜ Montane grasslands and shrublands
0 200 400 600 km



Caples Study Area
California, USA

80 points selected (blue)
2 weeks per year of monitoring
~10k hours of audio per year
2018-present
~80 species

Point counts performed when deploying + retrieving microphones.

Study effect of prescribed burn:
Points chosen inside and outside burn area.

Fully Annotated Passive Datasets

Annotations on a subset of a passive acoustic dataset.
Very difficult to produce!

Key questions:

- Which species were labeled?
- How much data was labeled?
- How are unknowns handled?

BirdCLEF competitions provide a variety of datasets from around the world, but others exist, too!

Main Terrestrial Bioacoustic Benchmarks

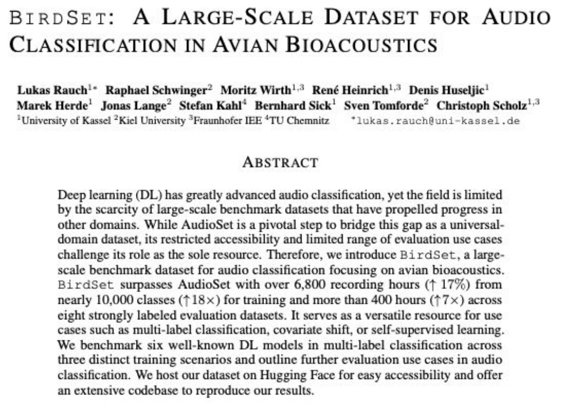
- **BirdSet**
  - Mostly collects past **BirdCLEF** datasets.
  - Adds a couple other bird datasets.
  - Fully-annotated data.
- World Annotated Bird Acoustic Dataset
  - aka, **WABAD**
  - More recent dataset, developed to evaluate BirdNet (which trained on all of the Birdset data).
  - Quite large, with annotated data from around the world.

- BEnchmark of ANimal Sounds (**BEANS**)
  - Earth Species Project
  - Collects a set of detection and classification tasks.
  - Multi-taxa.

BirdCLEF Competitions

BirdCLEF measures the ability of machine learning classifiers to operate across a **domain shift** on **geographically specific** passive acoustic monitoring data.

Produces new datasets from different regions of the world each year, by working with local community.



BIRDSET: A LARGE-SCALE DATASET FOR AUDIO CLASSIFICATION IN AVIAN BIOACOUSTICS

Lukas Rauch[1*] Raphael Schwinger[2] Moritz Wirth[1,3] René Heinrich[1,3] Denis Huseljic[1]
Marek Herde[1] Jonas Lange[2] Stefan Kahl[4] Bernhard Sick[1] Sven Tomforde[2] Christoph Scholz[1,3]
[1]University of Kassel [2]Kiel University [3]Fraunhofer IEE [4]TU Chemnitz *lukas.rauch@uni-kassel.de

ABSTRACT

Deep learning (DL) has greatly advanced audio classification, yet the field is limited by the scarcity of large-scale benchmark datasets that have propelled progress in other domains. While AudioSet is a pivotal step to bridge this gap as a universal-domain dataset, its restricted accessibility and limited range of evaluation use cases challenge its role as the sole resource. Therefore, we introduce BirdSet, a large-scale benchmark dataset for audio classification focusing on avian bioacoustics. BirdSet surpasses AudioSet with over 6,800 recording hours (↑ 17%) from nearly 10,000 classes (↑18×) for training and more than 400 hours (↑7×) across eight strongly labeled evaluation datasets. It serves as a versatile resource for use cases such as multi-label classification, covariate shift, or self-supervised learning. We benchmark six well-known DL models in multi-label classification across three distinct training scenarios and outline further evaluation use cases in audio classification. We host our dataset on Hugging Face for easy accessibility and offer an extensive codebase to reproduce our results.

BirdCLEF Datasets (Kaggle):

- 2025 - Colombia multi-taxa
- 2024 - India Western Ghats
- 2023 - Mount Kenya
- 2022 - Hawai`i
- 2021 - Multidataset!
    - Colombia
    - Costa Rica
    - Sapsucker Woods (New York)
    - Sierra Nevada
- 2020 - Caples Creek (California)
    - "Cornell Birdcall Identification" challenge.

Before Kaggle:

- 2020 - Sapsucker Woods (New York)
- 2019 - Sapsucker Woods + Colombia
- 2018 - Sapsucker Woods + Colombia
- 2017 - Amazon Soundscapes
- 2015-16 - Xeno-Canto Brazil
- 2014 - Xeno-Canto Brazil

evaluating models

Classifier scores

- Bird species identification is a multiclass, multilabel problem.
  - Multiclass: many classes.
  - Multilabel: 0+ labels per example.
  - Usually classification tasks are multiclass and *single* label.

- Easy way forward:
  Consider each classifier output head as an independent binary classifier.

- Most generally, we obtain a numerical **score** for each class on each example.
  We want positive classes to have **higher scores** than negative classes...

Sample ID	Gull Score	Hawk Score	Starling Score	Primary Prediction
001	**0.92**	0.03	0.05	Gull
002	0.12	**0.85**	0.03	Hawk
003	0.02	0.08	**0.90**	Starling
004	**0.45**	**0.48**	0.07	Hawk (Low Confidence)
005	0.10	**0.42**	**0.48**	Starling (Low Confidence)
006	0.15	0.12	0.10	None
007	0.05	0.04	0.06	None

Per-Example and Per-Class Metrics

We may also have labels for the various examples; color them green when positive…

- Metrics are generally computed and per-**example** or per-**class** and then averaged.

- If we choose a **threshold** (globally, or for each class), we can compute traditional binary classification metrics for each class.
  - Precision, Recall, F1 score

- We can compute "**Top-1 Accuracy**" - whether the top scoring class in each row is correct. (Q: What can go wrong here?)

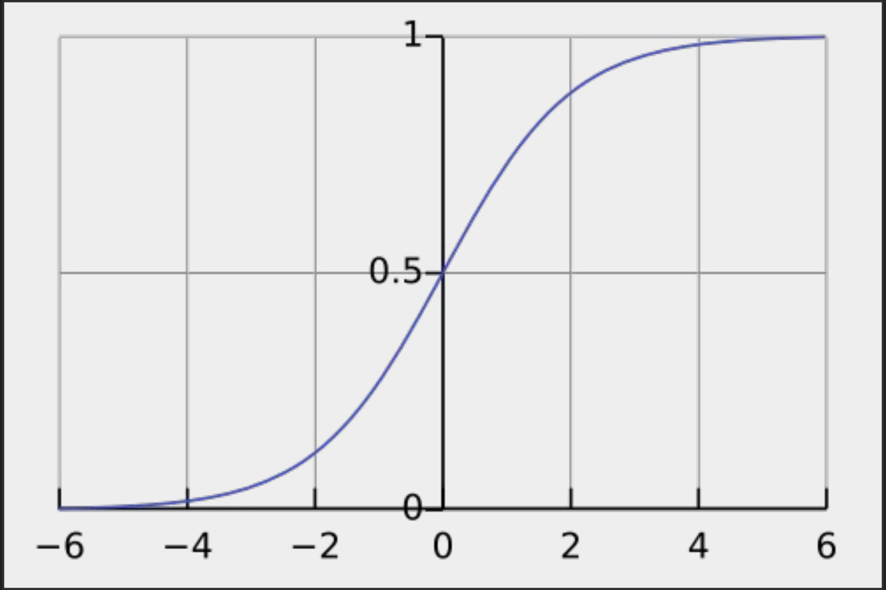
Sample ID	Gull Score	Hawk Score	Starling Score	Primary Prediction
001	**0.92**	0.03	0.05	Gull
002	0.12	**0.85**	0.03	Hawk
003	0.02	0.08	**0.90**	Starling
004	**0.45**	**0.48**	0.07	Hawk (Low Confidence)
005	0.10	**0.42**	**0.48**	Starling (Low Confidence)
006	0.15	0.12	0.10	None
007	0.05	0.04	0.06	None

Classifier Scores: Logit Scale

- Usually we train classifiers to produce real numbers on the **logit scale**.
- The outputs are then transformed by the logistic/sigmoid function to a "probability":

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} = 1 - \sigma(-x).$$

- Then our classifier scores are arbitrary real numbers, but *usually* between +-10.

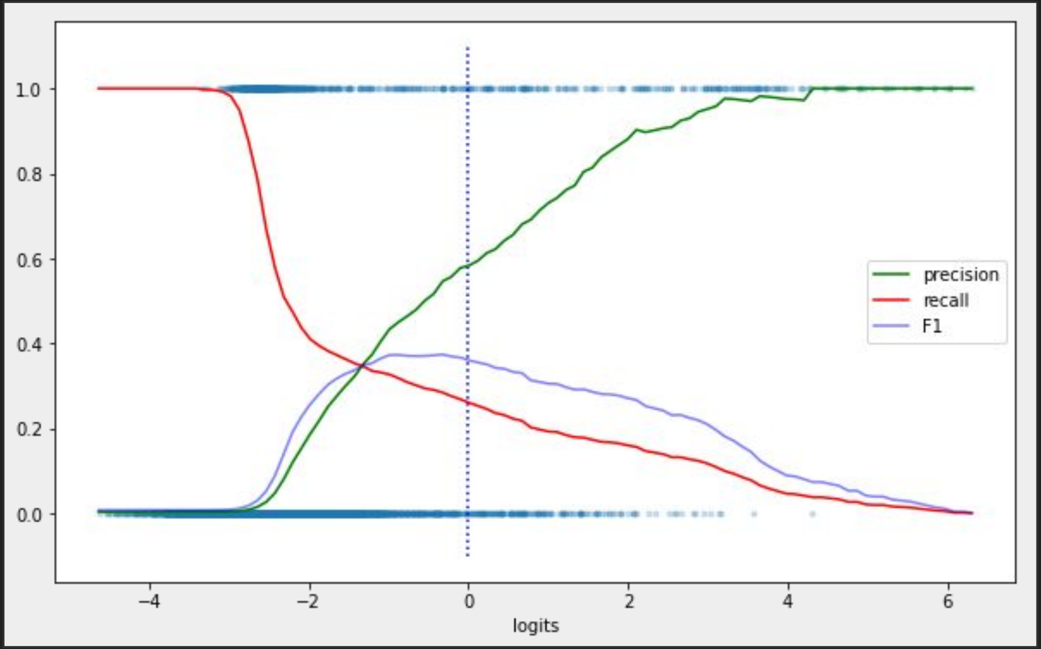


Threshold Metrics

Selecting a threshold converts the scores to binary decisions, and we can then compute:

- **Precision**: TP / (TP + FP)
  How many predicted positives are correct?
- **Recall**: TP / #P = TP / (TP + FN)
  How many true positives do we find?
- **F1**: Harmonic mean of Prec and Recall.
  2 * (Prec + Recall) / (Prec * Recall)

Changing the threshold changes the metrics!



Threshold Selection

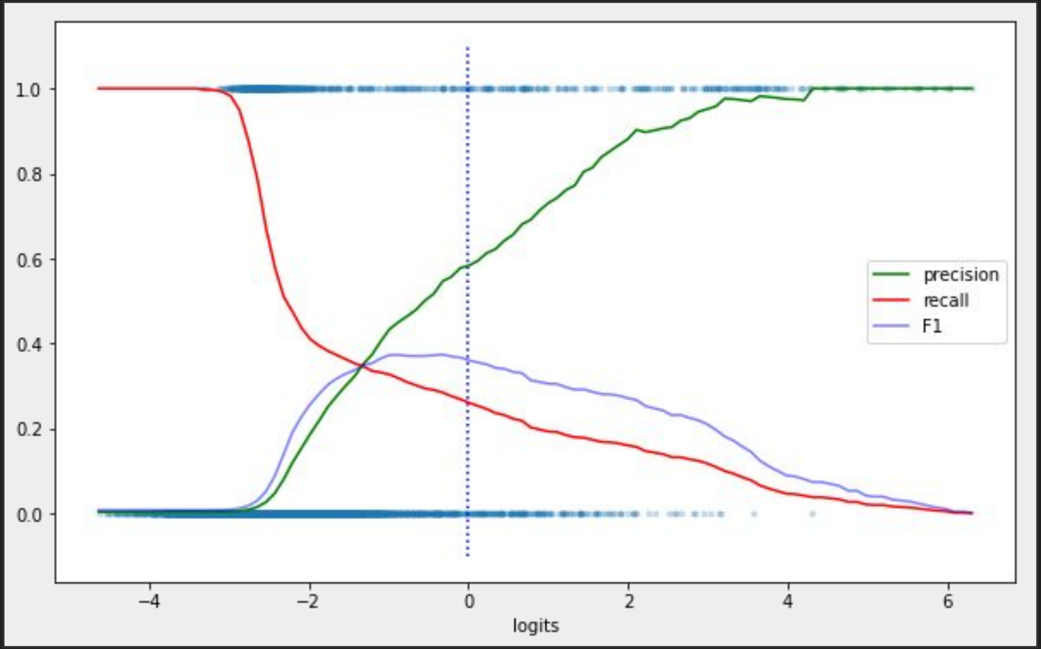
Choosing a threshold depends on how the model will be used downstream.

False negatives are very hard to measure accurately in large datasets!

This means that threshold selection is often driven by **targeting False Positive Rate** and then performing some data validation to find an appropriate threshold.

Some downstream modeling approaches (eg, occupancy models) have a very hard time with false positives.

Also, manual review of lots of FPs is tedious.



How Inference Really Works in Biodiversity Monitoring

In passive acoustics, often people only want to know if a (bird) species is **present at a site** in a **particular time period**.

For example, we might want to verify that a bird appears on three consecutive days to prove that it is inhabiting (rather than just passing through) a site.

Then we can take all of the scores for each species in a site/time slice, and apply human **review of the top K examples** until a positive example is found.

This produces a verified **trophy recording**.

This process **does not involve a threshold**!

The classifier is being used as an information retrieval mechanism.

Threshold-Free Metrics

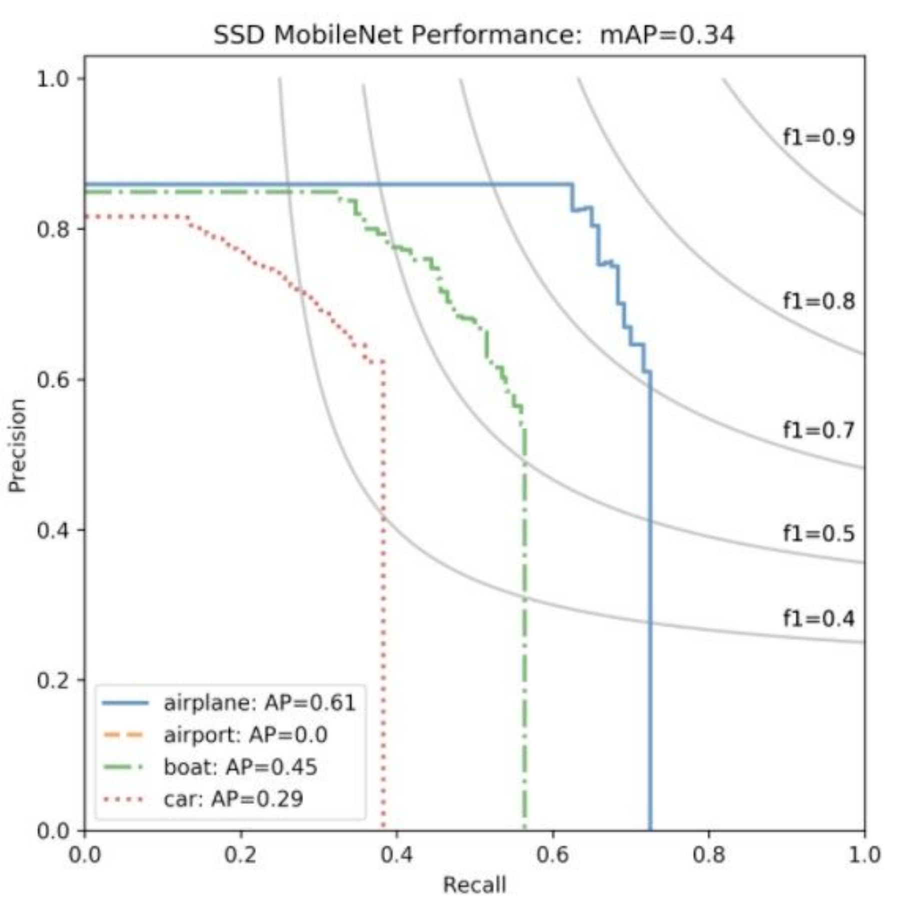
There are two main threshold-free statistics:

- **Mean Average Precision** (mAP, PR-AUC)
- **ROC-AUC**

These are surprisingly mathematically rich!

"Obviously" threshold-invariant, since these metrics **integrate over all thresholds**.

Can be any number between 0.0 and 1.0.



Average Precision

$$\text{AveP} = \int_0^1 p(r)dr$$

$$\text{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

$$\text{AveP} = \frac{\sum_{k=1}^{n} P(k) \times \text{rel}(k)}{\text{total number of relevant documents}}$$

High-level mAP definition:

- For each threshold, plot (precision, recall).
- Compute the area under this curve.

Then average over classes.

In practice: Use a finite sum over the labeled data.

P(k) = precision at kth highest example

rel(k) = 1 iff the kth document is relevant (positive)

Example: AP for Hawk

- Sort by score.
- Compute P(k) for each position.
- Write r(k) as an additional column.

$$\operatorname{AveP} = \frac{\sum_{k=1}^{n} P(k) \times \operatorname{rel}(k)}{\text{total number of relevant documents}}$$

Sample ID	Gull Score	Hawk Score	Starling Score	Primary Prediction
001	**0.92**	0.03	0.05	Gull
002	0.12	**0.85**	0.03	Hawk
003	0.02	0.08	**0.90**	Starling
004	**0.45**	**0.48**	0.07	Hawk (Low Confidence)
005	0.10	**0.42**	**0.48**	Starling (Low Confidence)
006	0.15	0.12	0.10	None
007	0.05	0.04	0.06	None

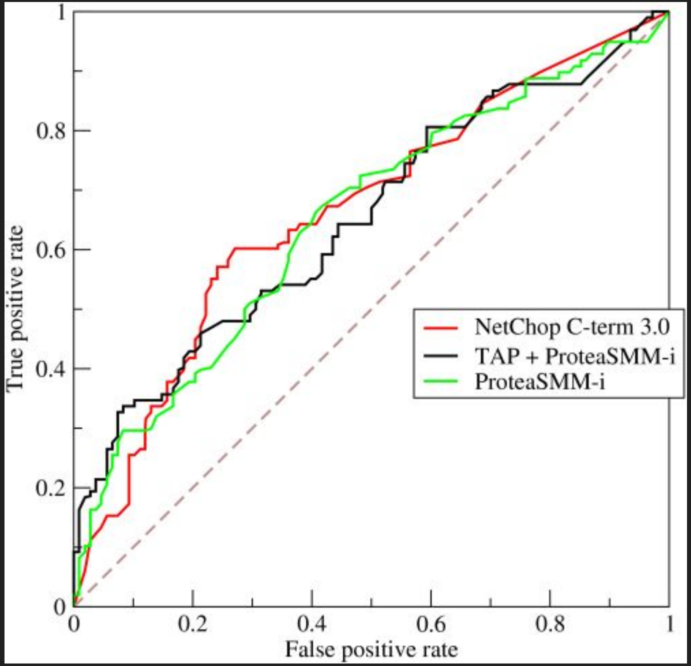
Hawk	r(k)	P(k)	r(k)*P(k)
0.85	**1**	1	1
0.48	**1**	1	1
0.42	**0**	2/3	0
0.12	**0**	1/2	0
0.08	**1**	3/5	3/5
0.04	**0**	1/2	0
0.03	**0**	3/7	0

ROC-AUC

Similar to PR-AUC, ROC-AUC is initially defined as the area under the curve when plotting the false positive rate vs true positive rate at each threshold.

TPR = #TP / #P
FPR = #FP / #N

Usually ~0.5 (up to noise) with a completely random model.
So generally varies from 0.5-1.0.

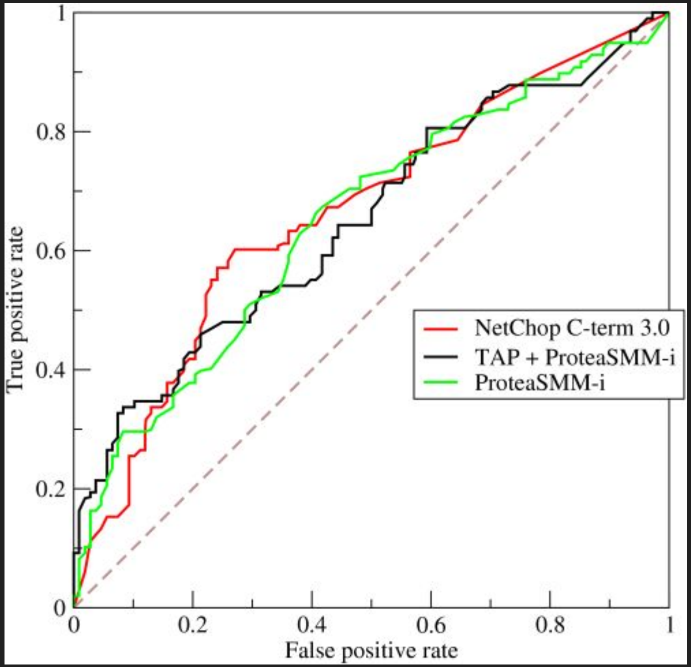


ROC-AUC - Probabilistic Definition

The objectively best definition of ROC-AUC is not area under curve, though!

ROC-AUC is (equivalently):
The **probability** that a uniformly-random **positive example has a higher score** than a uniformly random negative example.

ie, P(score(+) > score(-)), or (shorter) **P(+>-)**.



Example: ROC-AUC for Hawk

- Collect all the positive and negative scores.
- For each positive score, count the proportion of negative scores it is greater than: P(+k > -)
- Take the mean of P(+k > -).

Hawk Neg	Hawk Pos	$P(+_k > -)$
0.42	**0.85**	**1**
0.12	**0.48**	**1**
0.04	0.08	1/2
0.03		

Observations…

Both PR and ROC-AUC measure how well the positives float to the top:
Generally want green scores above red scores.

Neither metric cares what the actual scores are, just their relative position.

For PR, we see that a single high-scoring negative example 'caps' the P(k) for all other examples.

- In other words, PR is sensitive to "lazy" false negatives in the ground truth data.
- This is a big problem because annotating passive acoustic datasets is extremely difficult.

But there's another problem with PR…

Hawk	r(k)	P(k)	r(k)*P(k)
0.86	**0**		0
0.85	**1**	1/2	1/2
0.48	**1**	2/3	2/3
0.42	**0**		0
0.12	**0**		0
0.08	**1**	1/2	1/2
0.04	**0**		0
0.03	**0**		0

Impact of Label Balance

Synthetic data

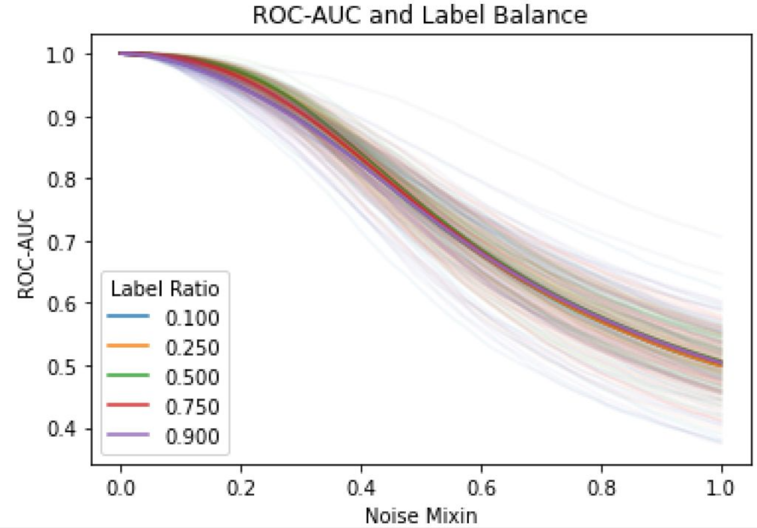
- Label each data point 1 or -1 for pos, neg.
- Choose a 'noise level' x in range(0, 1).
- Set model scores:
  x * N(0, 1) + (1 - x) * label

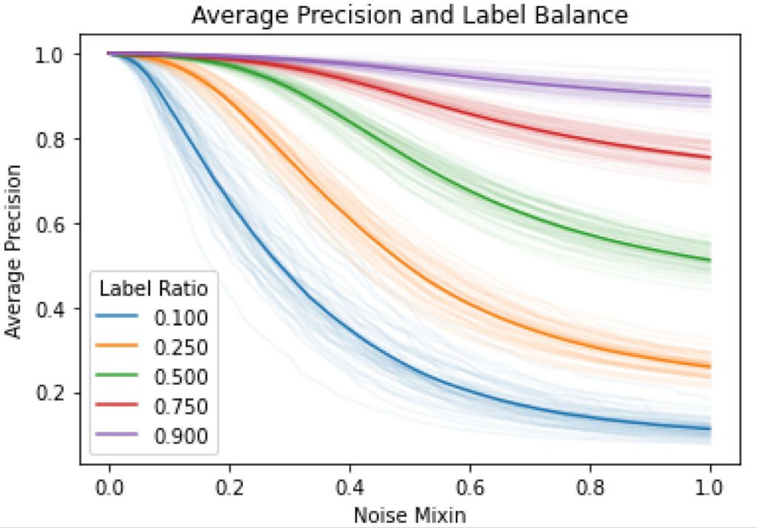
Compute AP and ROC-AUC for the model.

Changing the label balance demonstrates that:

- ROC-AUC is **invariant to label balance** (also obvious from the definition).
- PR-AUC depends on label balance.

As a result, PR-AUC **cannot compare different class qualities** if label balance is different per-class.


ROC-AUC and Label Balance
ROC-AUC
Noise Mixin
Label Ratio


Average Precision and Label Balance
Average Precision
Noise Mixin
Label Ratio