Data Exploration

Nikolina Lalic, Elias Bürger, Niklas Angert, Matthias Gander
Team Attempt

Image Source: https://www.pngarts.com/explore/1105

Annotator agreement

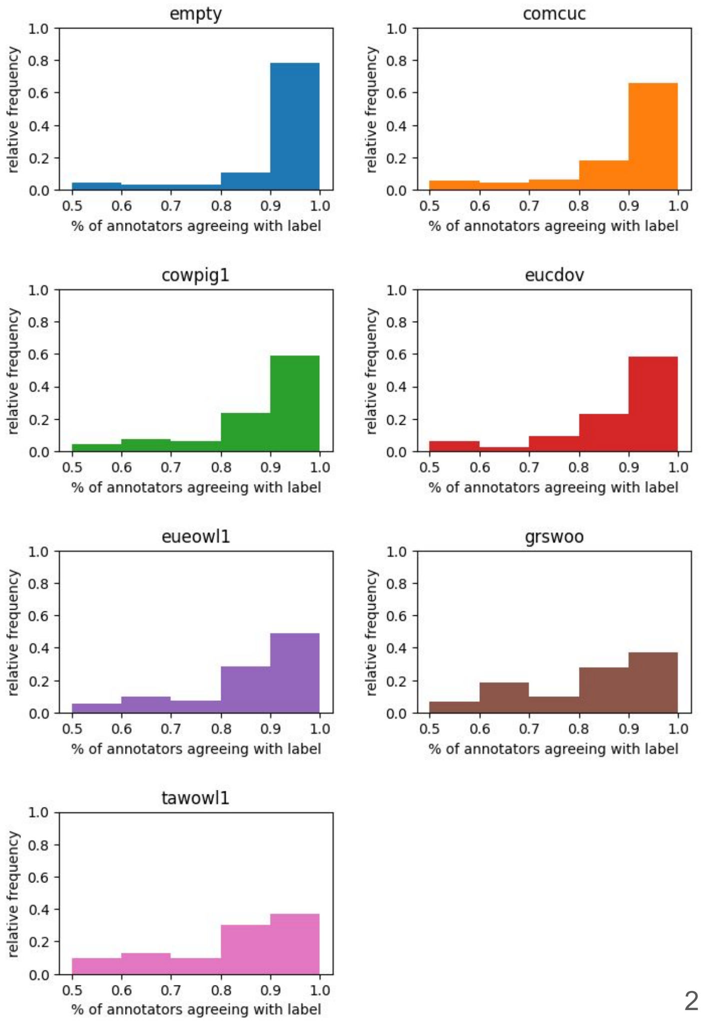
- **Ideal:** most mass in the last bin (e.g. empty)
- **Actual:** big uncertainty (e.g. tawowl1)

→ Agreement is class dependent

→ Might be a useful (indirect: e.g. make sample frequency dependent on agreement) training parameter

→ Some birds may be easier to distinguish

→ Diagrams do not tell us about random chance of agreement

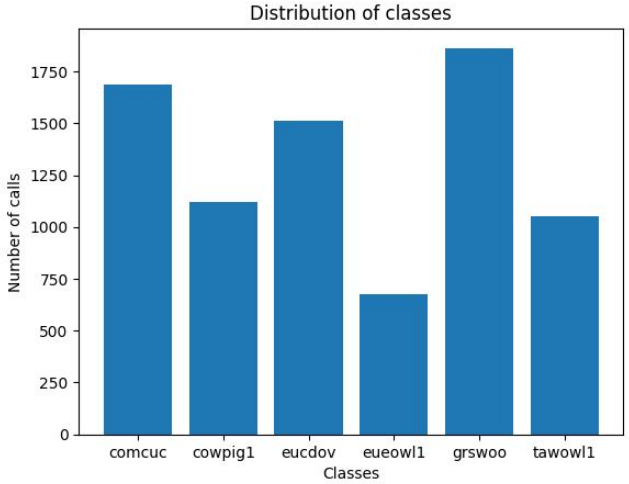


<u>Diagrams explained:</u> the rightmost bar in cowpig1 tells us that given the final label is cowpig1, roughly 60% of the time 90 to 100% of the annotators agree in hearing this bird

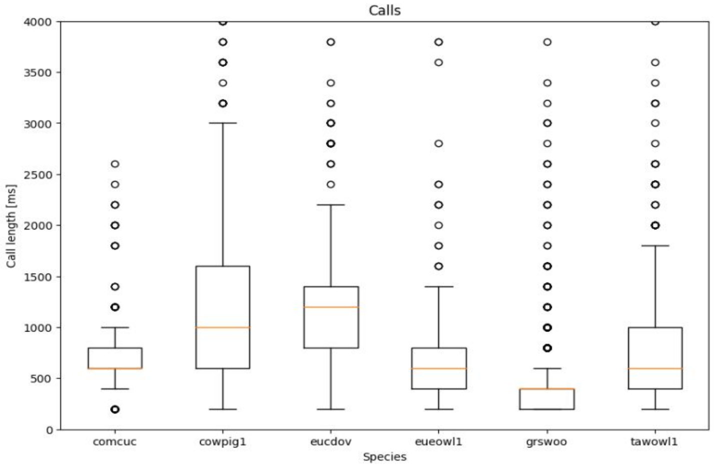
Label characteristics

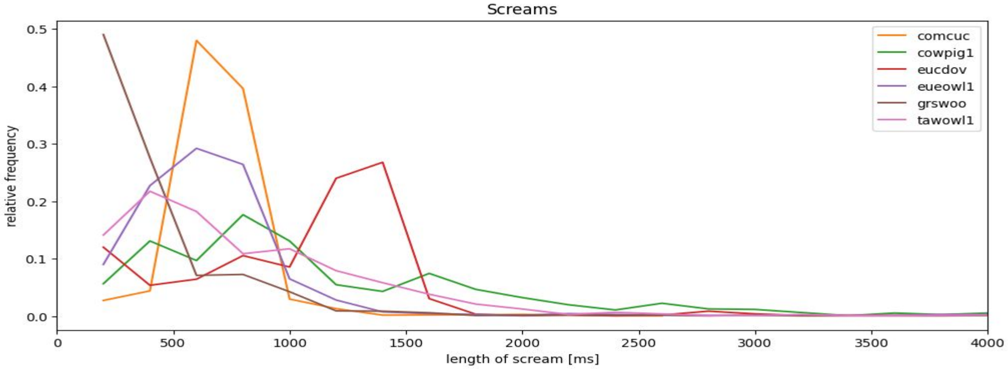
average durations:

```
('comcuc', 691)
('cowpig1', 1598)
('eucdov', 1083)
('eueowl1', 685)
('grswoo', 548)
('tawowl1', 804)
```



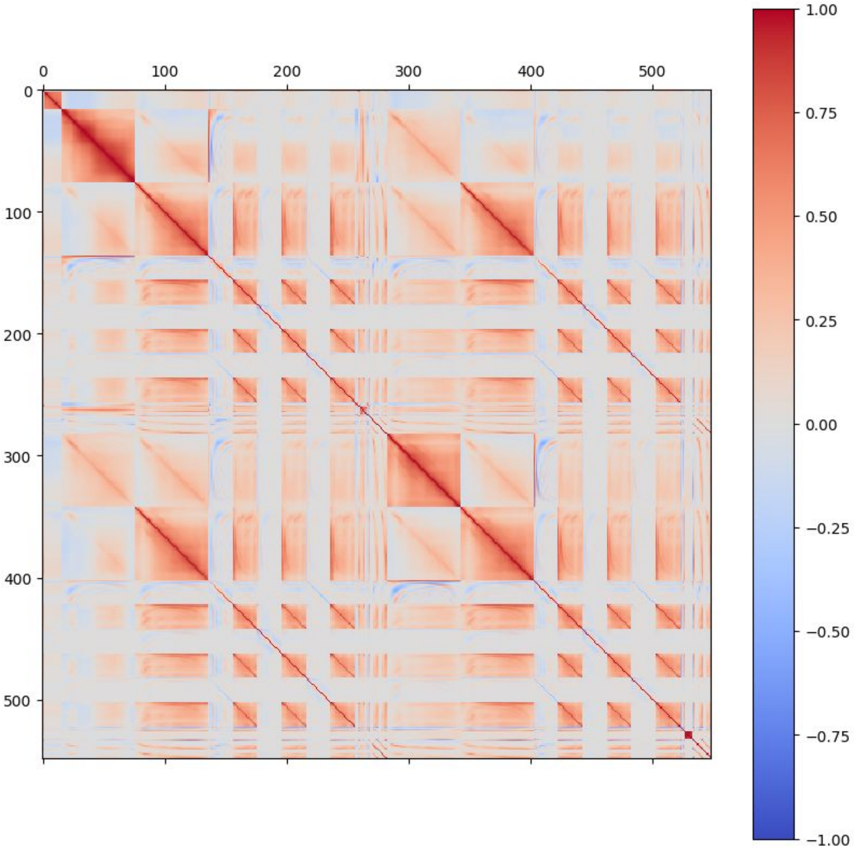
- Number of calls, mean and variance are computed
- Number of annotated calls: eueowl1 - lowest, grswoo - highest
- **Intra-class** variance: cowpig1 - high, comcuc - low
- **Inter-class** variance: grswoo - low





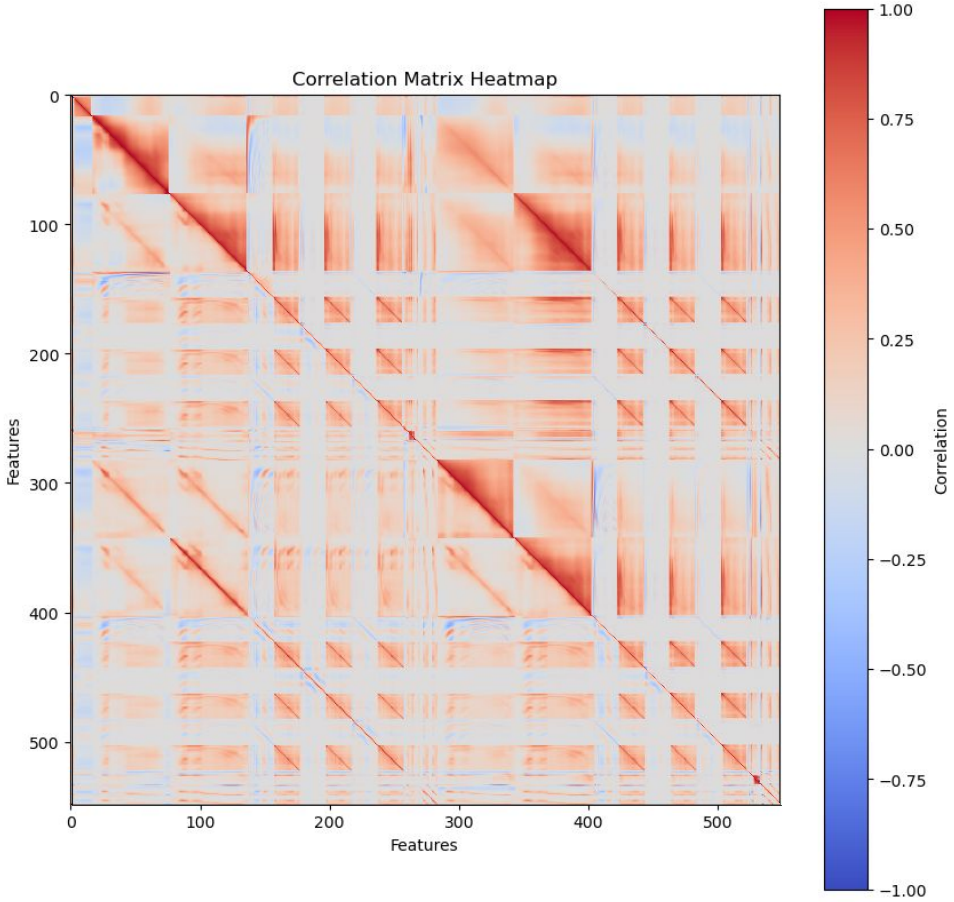
Feature characteristics

- Calculate the **Pearson Correlation coefficient** between all features
- Visualize them in a heatmap
- Areas of **high correlation** highlighted in red
- Negative correlation highlighted in blue
- Typically feature groups share high correlation
  - Raw_melspect_mean (top left corner)
  - cln_melspect_mean and cln_melspect_std



Feature characteristics

- To illustrate the difference of feature correlation between birds we show the correlation heatmap for two different birds split along the diagonal
- Top half: grswoo
- Bottom half: comcuc



Feature/Label agreement

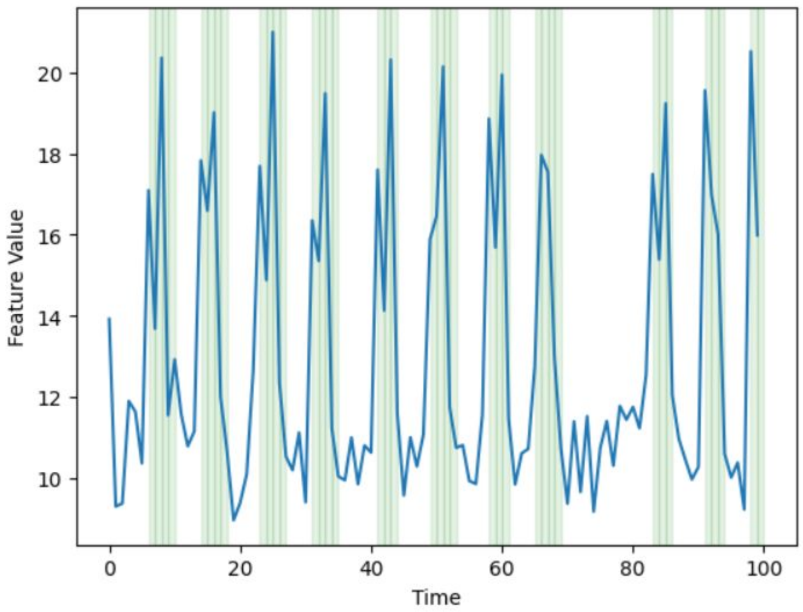
- Calculate the **Pearson Correlation coefficient** between Labels and Features
- **Rank** them from best to worst
- Highest coefficient ➔ **good feature** to do **Classification** with (depending on Class/Bird)
- **Different birds** need **different features** to do Classification, so we need to calculate this for each bird independently

```
Class 1

Highest correlation
cln_contrast_mean_3:     0.6925317556411079     Feature:537
cln_melspect_mean_9:     0.6688152653571825     Feature:291
raw_melspect_std_8:      0.6539925851966788     Feature:84
cln_melspect_std_8:      0.6481760752549925     Feature:350
raw_contrast_std_3:      0.6389625666737243     Feature:278
raw_contrast_mean_3:     0.6361956393552644     Feature:271
cln_melspect_mean_10:    0.6300679170590937     Feature:292
raw_melspect_mean_9:     0.6258599465179724     Feature:25
cln_melspect_mean_8:     0.6182878006430244     Feature:290
cln_contrast_std_3:      0.6155290093619367     Feature:544

Class 2

Highest correlation
raw_contrast_mean_3:     0.7285215322298139     Feature:271
raw_melspect_mean_6:     0.6433970111529351     Feature:22
cln_melspect_mean_6:     0.6396913678055655     Feature:288
cln_contrast_mean_3:     0.6331997796142523     Feature:537
cln_melspect_mean_5:     0.6206590039520823     Feature:287
raw_melspect_mean_5:     0.615191561711744      Feature:21
cln_melspect_mean_7:     0.6150777626772981     Feature:289
raw_melspect_mean_7:     0.609734786261283      Feature:23
cln_melspect_mean_4:     0.5675091346737678     Feature:286
cln_melspect_mean_8:     0.5295964256185339     Feature:290
```

Table explained: We take a look at the first and the second Class/Bird which is the comcuc and the cowpig1. We show the top 10 features, their correlation coefficient between Label and Features and the corresponding feature index. (Whole Dataset was used.)

Feature/Label agreement

- As an **Example** we look at the best feature from the previous slide. To visualize this we take a 20 second fragment of the **comcuc** and visualize the feature over time.
- Green marks the time frames when the bird is heard.
- We can see that everytime there is a **spike** in the feature value, the **bird is heard**.



Graphic explained: We take a closer look at the best feature of class 1 from the previous slide (Feature Index: 537, cln_contrast_mean_3). The X-axis shows the time (20 seconds -> 1 = 20ms) and the Y-axis shows the corresponding feature value. Green parts show the 20ms frames when the bird is heard.

Consequences

- We know more about the dataset itself (e.g. it’s distribution)
- We got a better idea about which features are **useful for classification**
- We know more details about the birds (e.g. the length of a bird call) which could also be useful for classification later
- We could now clean up the dataset by **excluding features** that don't provide any useful information in order to **deal with less data**
- Different species share a lot of features that are highly correlating with the label, single features don't provide enough information for classification