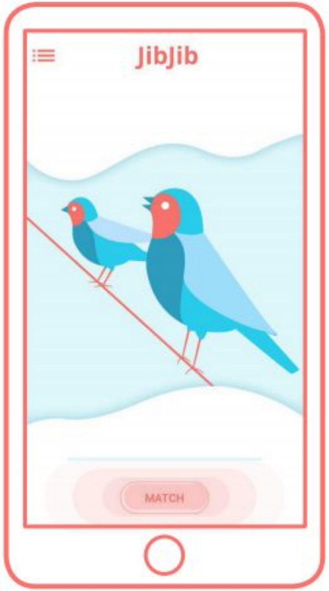
LEAI #5 - Workshop

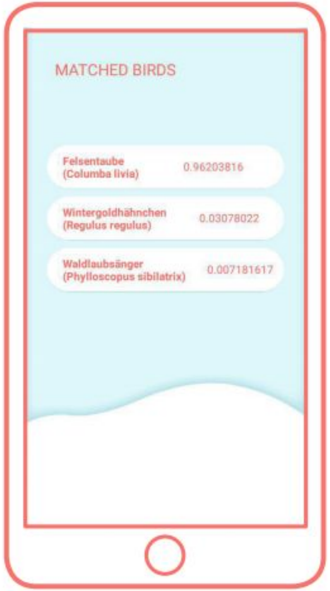
JibJib - A bird song classifier

JibJib

- Entry for Coding Davinci OST 2018 Hackathon
- "Shazam for Birds"
- Android App ←→ Backend ←→ TensorFlow Model
- Google Play Store: JibJib
- https://github.com/gojibjib

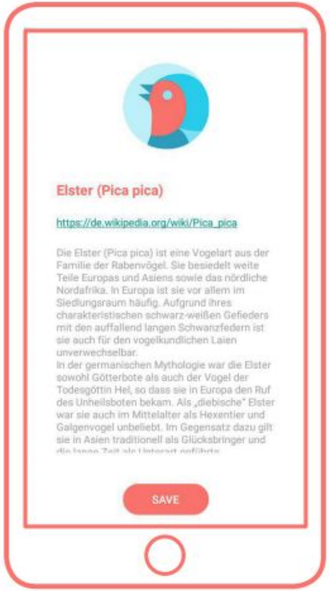
JibJib





MATCHED BIRDS

Felsentaube (Columba livia)	0.96203816
Wintergoldhähnchen (Regulus regulus)	0.03078022
Waldlaubsänger (Phylloscopus sibilatrix)	0.007181617



Elster (Pica pica)

https://de.wikipedia.org/wiki/Pica_pica

Die Elster (Pica pica) ist eine Vogelart aus der Familie der Rabenvögel. Sie besiedelt weite Teile Europas und Asiens sowie das nördliche Nordafrika. In Europa ist sie vor allem im Siedlungsraum häufig. Aufgrund ihres charakteristischen schwarz-weißen Gefieders mit den auffallend langen Schwanzfedern ist sie auch für den vogelkundlichen Laien unverwechselbar.

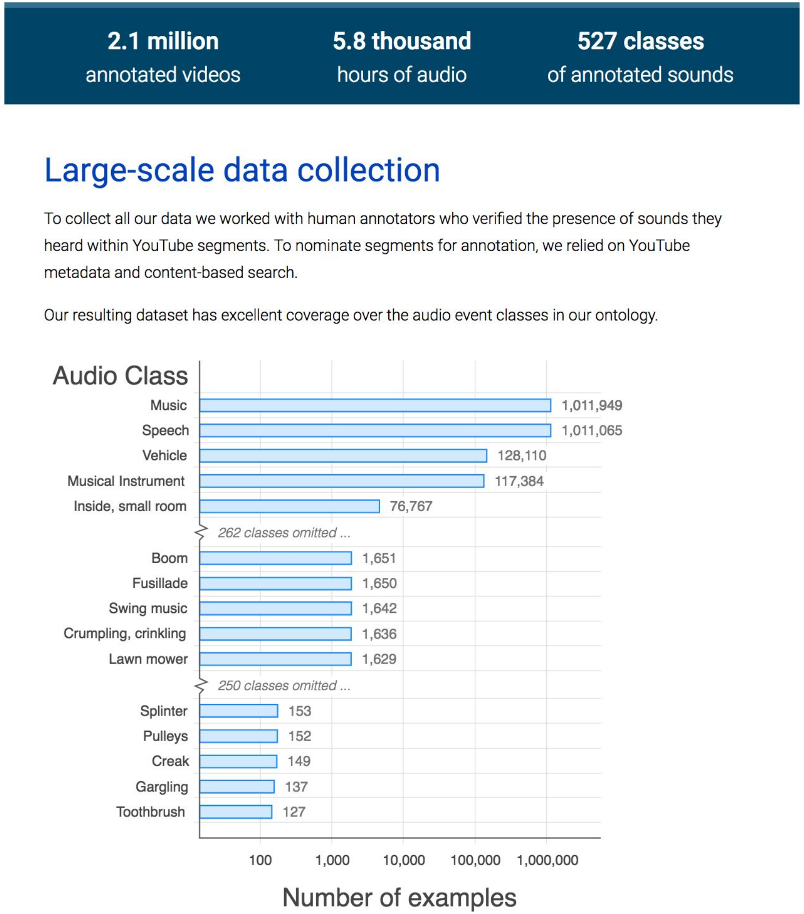
In der germanischen Mythologie war die Elster sowohl Götterbote als auch der Vogel der Todesgöttin Hel, so dass sie in Europa den Ruf des Unheilsboten bekam. Als „diebische" Elster war sie auch im Mittelalter als Hexentier und Galgenvogel unbeliebt. Im Gegensatz dazu gilt sie in Asien traditionell als Glücksbringer und die lange Zeit als Unterart geführte

Today

- Us:
    - Model basics & training
    - Data collection
    - Backend API
    - Model serving
- You: ask any questions you like!

Model Training

AudioSet



2.1 million annotated videos

5.8 thousand hours of audio

527 classes of annotated sounds

Large-scale data collection

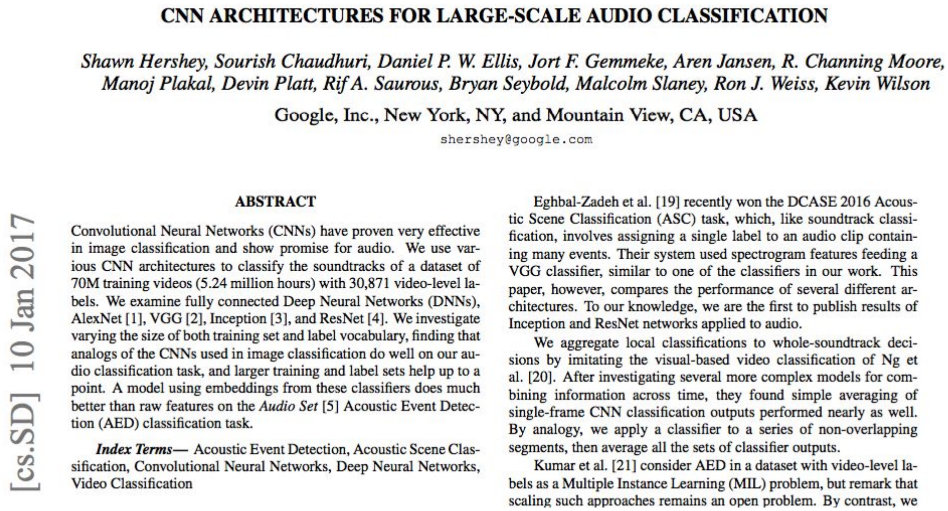
To collect all our data we worked with human annotators who verified the presence of sounds they heard within YouTube segments. To nominate segments for annotation, we relied on YouTube metadata and content-based search.

Our resulting dataset has excellent coverage over the audio event classes in our ontology.

https://research.google.com/audioset

Hershey et al. (2017)

https://ai.google/research/pubs/pub45611



CNN ARCHITECTURES FOR LARGE-SCALE AUDIO CLASSIFICATION

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, Kevin Wilson

Google, Inc., New York, NY, and Mountain View, CA, USA

shershey@google.com

[cs.SD] 10 Jan 2017

ABSTRACT

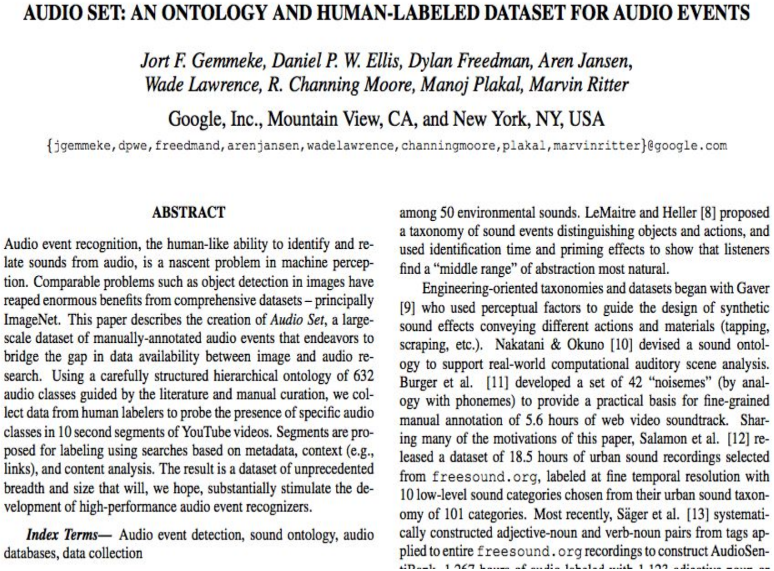
Convolutional Neural Networks (CNNs) have proven very effective in image classification and show promise for audio. We use various CNN architectures to classify the soundtracks of a dataset of 70M training videos (5.24 million hours) with 30,871 video-level labels. We examine fully connected Deep Neural 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 classification task, and larger training and label sets help up to a point. A model using embeddings from these classifiers does much better than raw features on the *Audio Set* [5] Acoustic Event Detection (AED) classification task.

Index Terms— Acoustic Event Detection, Acoustic Scene Classification, Convolutional Neural Networks, Deep Neural Networks, Video Classification

Eghbal-Zadeh et al. [19] recently won the DCASE 2016 Acoustic Scene Classification (ASC) task, which, like soundtrack classification, involves assigning a single label to an audio clip containing many events. Their system used spectrogram features feeding a VGG classifier, similar to one of the classifiers in our work. This paper, however, compares the performance of several different architectures. To our knowledge, we are the first to publish results of Inception and ResNet networks applied to audio.

We aggregate local classifications to whole-soundtrack decisions by imitating the visual-based video classification of Ng et al. [20]. After investigating several more complex models for combining information across time, they found simple averaging of single-frame CNN classification outputs performed nearly as well. By analogy, we apply a classifier to a series of non-overlapping segments, then average all the sets of classifier outputs.

Kumar et al. [21] consider AED in a dataset with video-level labels as a Multiple Instance Learning (MIL) problem, but remark that scaling such approaches remains an open problem. By contrast, we



AUDIO SET: AN ONTOLOGY AND HUMAN-LABELED DATASET FOR AUDIO EVENTS

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, Marvin Ritter

Google, Inc., Mountain View, CA, and New York, NY, USA

{jgemmeke,dpwe,freedmand,arenjansen,wadelawrence,channingmoore,plakal,marvinritter}@google.com

ABSTRACT

Audio event recognition, the human-like ability to identify and relate sounds from audio, is a nascent problem in machine perception. Comparable problems such as object detection in images have reaped enormous benefits from comprehensive datasets – principally ImageNet. This paper describes the creation of *Audio Set*, a large-scale dataset of manually-annotated audio events that endeavors to bridge the gap in data availability between image and audio research. Using a carefully structured hierarchical ontology of 632 audio classes guided by the literature and manual curation, we collect data from human labelers to probe the presence of specific audio classes in 10 second segments of YouTube videos. Segments are proposed for labeling using searches based on metadata, context (e.g., links), and content analysis. The result is a dataset of unprecedented breadth and size that will, we hope, substantially stimulate the development of high-performance audio event recognizers.

Index Terms— Audio event detection, sound ontology, audio databases, data collection

among 50 environmental sounds. LeMaitre and Heller [8] proposed a taxonomy of sound events distinguishing objects and actions, and used identification time and priming effects to show that listeners find a "middle range" of abstraction most natural.

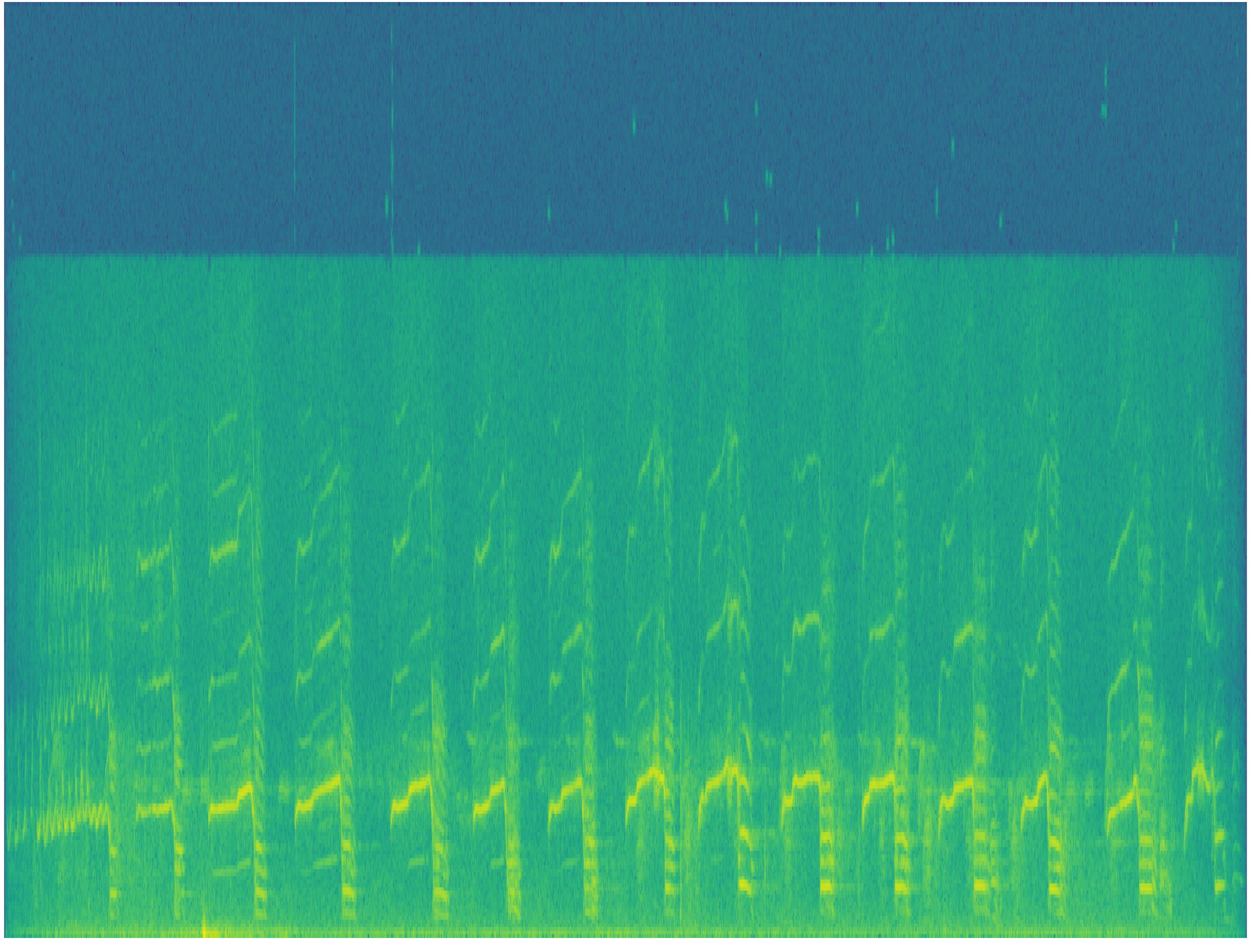
Engineering-oriented taxonomies and datasets began with Gaver [9] who used perceptual factors to guide the design of synthetic sound effects conveying different actions and materials (tapping, scraping, etc.). Nakatani & Okuno [10] devised a sound ontology to support real-world computational auditory scene analysis. Burger et al. [11] developed a set of 42 "noisemes" (by analogy with phonemes) to provide a practical basis for fine-grained manual annotation of 5.6 hours of web video soundtrack. Sharing many of the motivations of this paper, Salamon et al. [12] released a dataset of 18.5 hours of urban sound recordings selected from freesound.org, labeled at fine temporal resolution with 10 low-level sound categories chosen from their urban sound taxonomy of 101 categories. Most recently, Säger et al. [13] systematically constructed adjective-noun and verb-noun pairs from tags applied to entire freesound.org recordings to construct AudioSentiBank. 1,267 hours of audio labeled with 1,123 adjective-noun or

Genmeke et al. (2017)

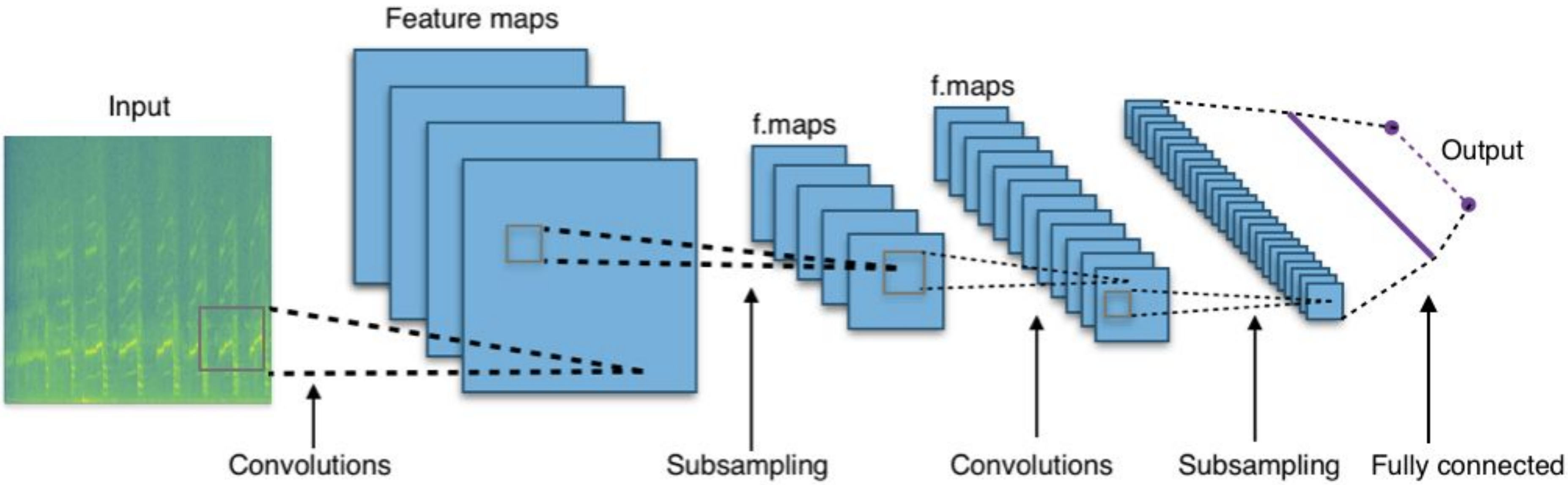
https://ai.google/research/pubs/pub45857

Spectrogram

- **Wikipedia:** *In sound processing, a spectrogram is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency*
- Mel frequency analysis for better representation of sound spectrum
- Visual representation of the spectrum of frequencies that vary in time (voice print)



LEAI #5 -- Alexander Knipping & Sebastian Biermann


Feature maps
Input
f.maps
f.maps
Output
Convolutions
Subsampling
Convolutions
Subsampling
Fully connected

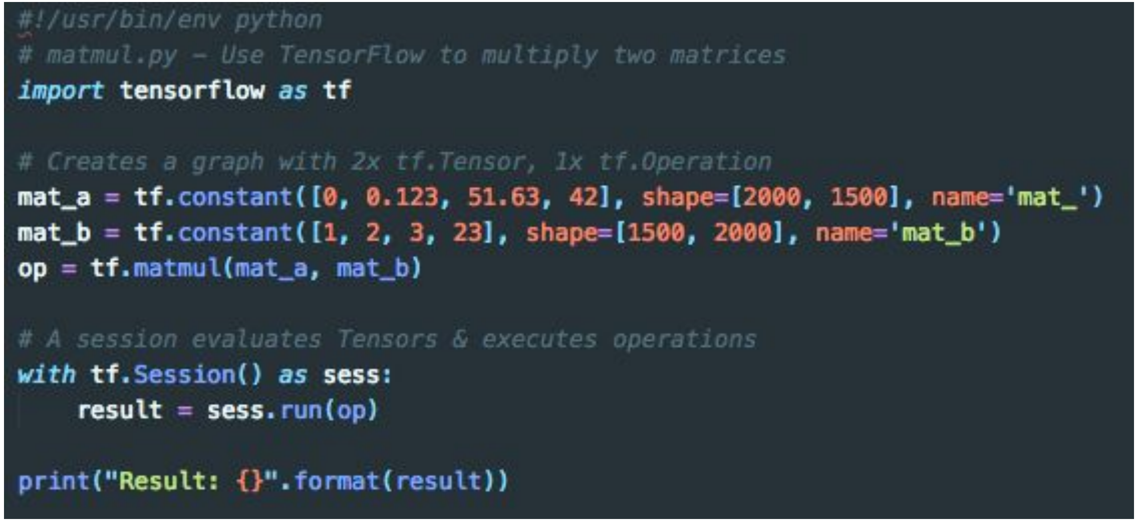
VGG-like TensorFlow model

- VGG model with 11 weight layers
- Input size of 96x64 for log mel spectrograms (instead of 224x224 for RGB images)
- Only four blocks of convolution and pooling/subsampling layers
- 128-wide fully connected layer as compact embedding layer

Crash Course: TensorFlow

- "Open-source software library for **dataflow programming**"
- Release by Google in November 2015 (Apache 2.0)
- Dataflow: programming model for parallel computing
    - **Nodes**: units of computation
    - **Edges**: data consumed or produced by nodes
- → https://tensorflow.org/guide/graphs

Crash Course: TensorFlow

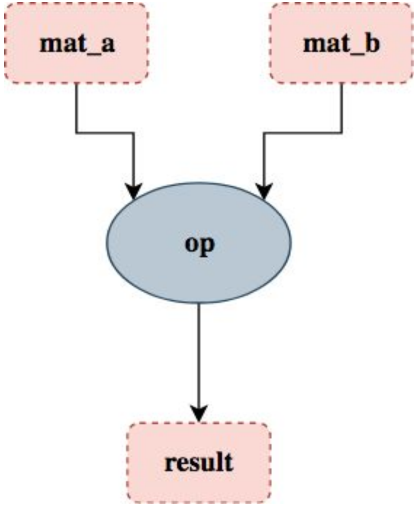


```python
#!/usr/bin/env python
# matmul.py - Use TensorFlow to multiply two matrices
import tensorflow as tf

# Creates a graph with 2x tf.Tensor, 1x tf.Operation
mat_a = tf.constant([0, 0.123, 51.63, 42], shape=[2000, 1500], name='mat_')
mat_b = tf.constant([1, 2, 3, 23], shape=[1500, 2000], name='mat_b')
op = tf.matmul(mat_a, mat_b)

# A session evaluates Tensors & executes operations
with tf.Session() as sess:
    result = sess.run(op)

print("Result: {}".format(result))
```



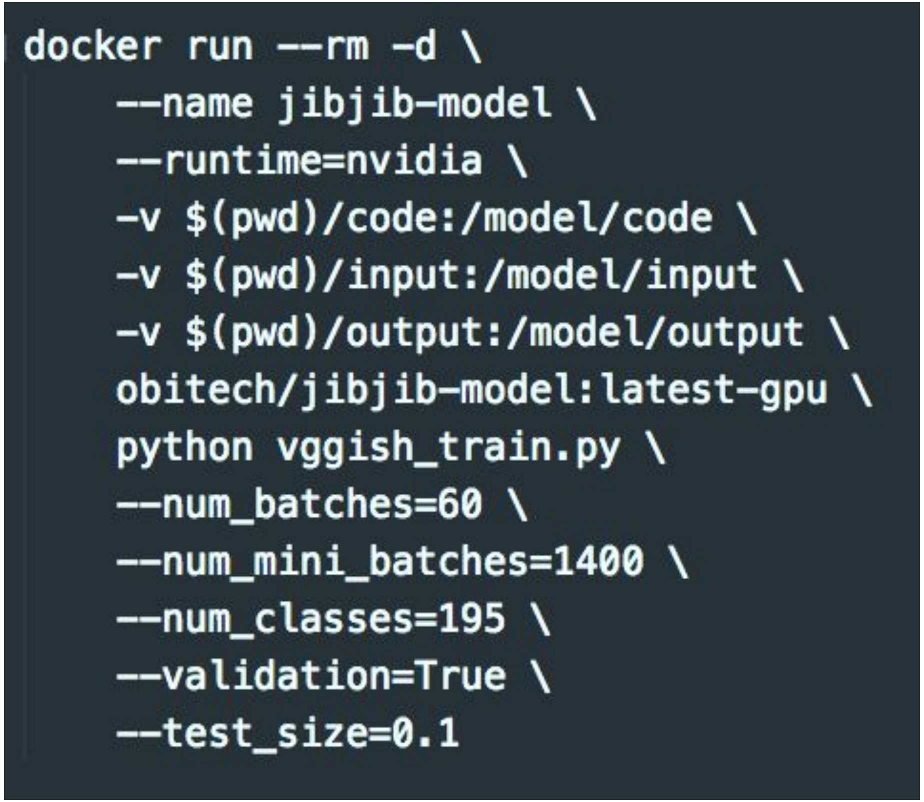
Model Training w/ Docker

Input:

- Dataset

Output:

- Logs, metrics
- Model.ckpt
- mappings.pickle

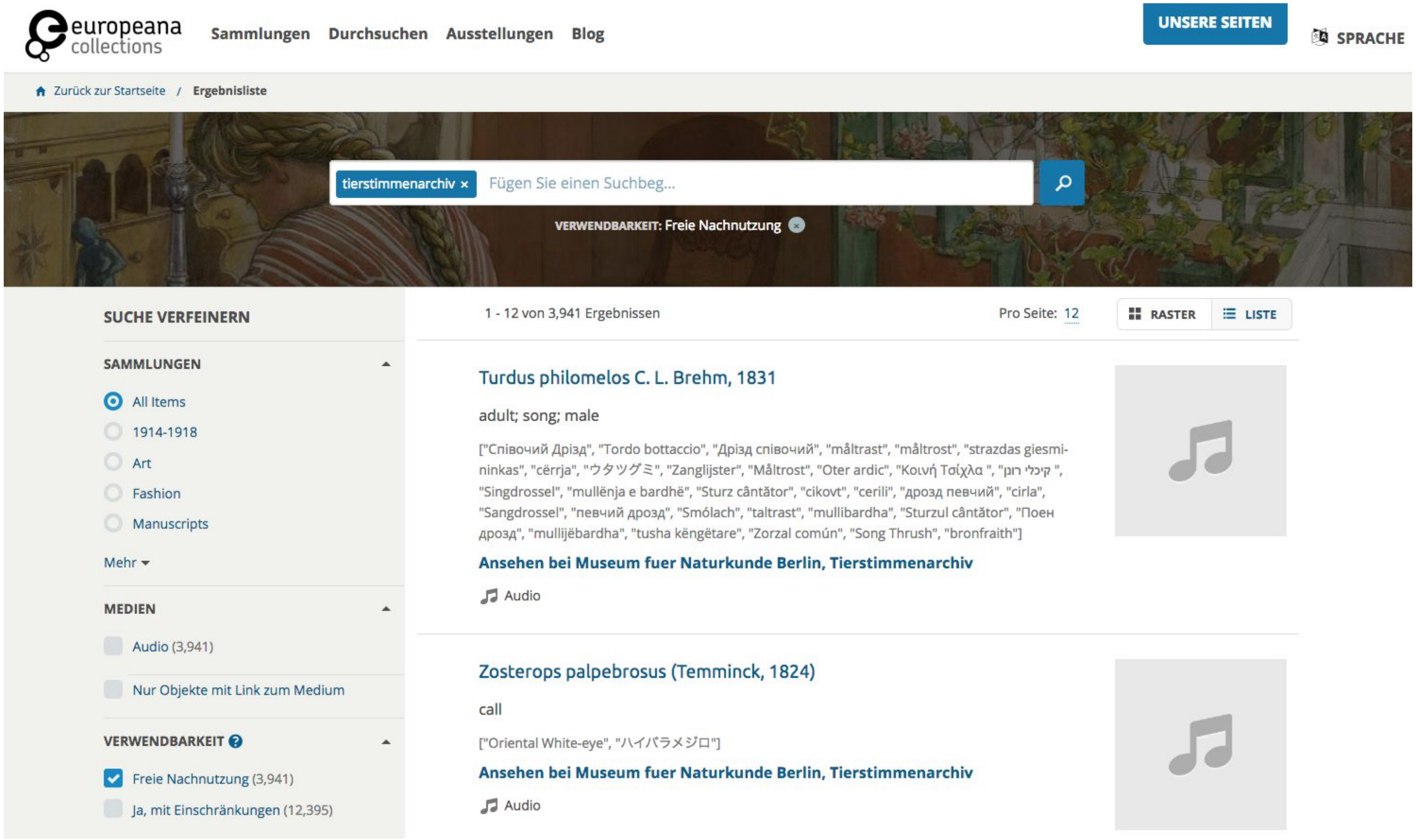


```
docker run --rm -d \
    --name jibjib-model \
    --runtime=nvidia \
    -v $(pwd)/code:/model/code \
    -v $(pwd)/input:/model/input \
    -v $(pwd)/output:/model/output \
    obitech/jibjib-model:latest-gpu \
    python vggish_train.py \
    --num_batches=60 \
    --num_mini_batches=1400 \
    --num_classes=195 \
    --validation=True \
    --test_size=0.1
```

Data Preparation

Initial Dataset

- Museum of Natural History Berlin (MFN) data set for CDV OST 2018
- 4000 openly licensed out of 120k from tierstimmenarchiv.de
- Accessible through Europeana (metadata API)
- Questions:
  - How many different birds?
  - How many files per bird?
  - How to download them?



europeana collections

Sammlungen Durchsuchen Ausstellungen Blog

UNSERE SEITEN

SPRACHE

Zurück zur Startseite / Ergebnisliste

tierstimmenarchiv × Fügen Sie einen Suchbeg...

VERWENDBARKEIT: Freie Nachnutzung ×

SUCHE VERFEINERN

SAMMLUNGEN

- ● All Items
- ○ 1914-1918
- ○ Art
- ○ Fashion
- ○ Manuscripts

Mehr

MEDIEN

- □ Audio (3,941)
- □ Nur Objekte mit Link zum Medium

VERWENDBARKEIT ?

- ☑ Freie Nachnutzung (3,941)
- □ Ja, mit Einschränkungen (12,395)

1 - 12 von 3,941 Ergebnissen

Pro Seite: 12

RASTER LISTE

Turdus philomelos C. L. Brehm, 1831

adult; song; male

["Співочий Дрізд", "Tordo bottaccio", "Дрізд співочий", "måltrast", "måltrost", "strazdas giesmininkas", "cërrja", "ウタツグミ", "Zanglijster", "Måltrost", "Oter ardic", "Κοινή Τσίχλα ", "קיכלי רונן ", "Singdrossel", "mullënja e bardhë", "Sturz cântător", "cikovt", "cerili", "дрозд певчий", "cirla", "Sangdrossel", "певчий дрозд", "Smólach", "taltrast", "mullibardha", "Sturzul cântător", "Поен дрозд", "mullijëbardha", "tusha këngëtare", "Zorzal común", "Song Thrush", "bronfraith"]

Ansehen bei Museum fuer Naturkunde Berlin, Tierstimmenarchiv

Audio

Zosterops palpebrosus (Temminck, 1824)

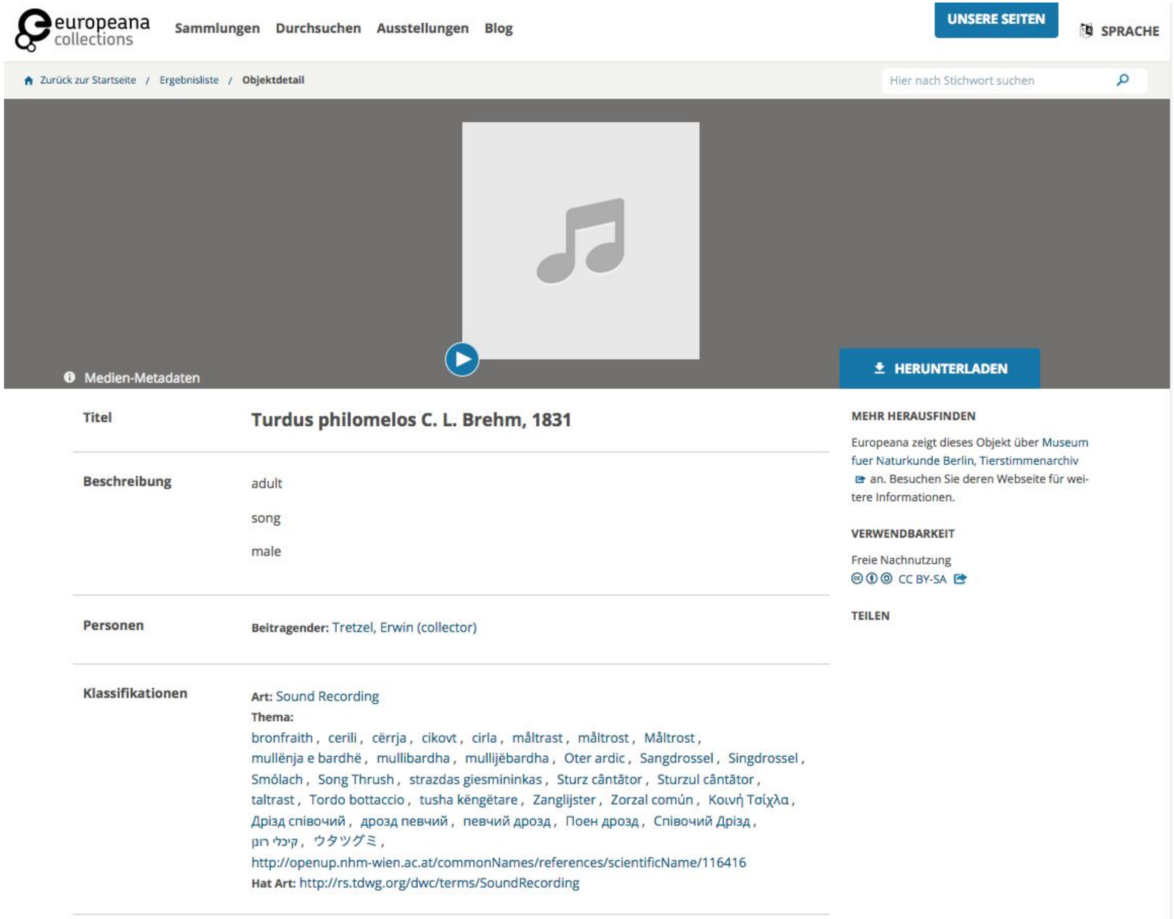
call

["Oriental White-eye", "ハイバラメジロ"]

Ansehen bei Museum fuer Naturkunde Berlin, Tierstimmenarchiv

Audio

https://europeana.eu



Titel Turdus philomelos C. L. Brehm, 1831

Beschreibung adult

song

male

Personen **Beitragender:** Tretzel, Erwin (collector)

Klassifikationen **Art:** Sound Recording

Thema:

bronfraith , cerili , cërrja , cikovt , cirla , måltrast , måltrost , Måltrost , mullënja e bardhë , mullibardha , mullijëbardha , Oter ardic , Sangdrossel , Singdrossel , Smólach , Song Thrush , strazdas giesmininkas , Sturz cântător , Sturzul cântător , taltrast , Tordo bottaccio , tusha këngëtare , Zanglijster , Zorzal común , Κοινή Τσίχλα , Дрізд співочий , дрозд певчий , певчий дрозд , Поен дрозд , Співочий Дрізд , קיכלי רונן , ウタツグミ ,

http://openup.nhm-wien.ac.at/commonNames/references/scientificName/116416

Hat Art: http://rs.tdwg.org/dwc/terms/SoundRecording

MEHR HERAUSFINDEN

Europeana zeigt dieses Objekt über Museum fuer Naturkunde Berlin, Tierstimmenarchiv an. Besuchen Sie deren Webseite für weitere Informationen.

VERWENDBARKEIT

Freie Nachnutzung

CC BY-SA

TEILEN

https://europeana.eu

Initial Dataset

1. Write Europeana API Client
   - https://github.com/gojibjib/gopeana
2. Use the client to dump URLs + names into JSON
   - https://github.com/gojibjib/voice-grabber/blob/master/data_grabber/data_grabber.go
3. Iterate over JSON to download files
   - https://github.com/gojibjib/voice-grabber/blob/master/file_grabber/file_grabber.go

Initial Dataset

Total size	6GB
Total # of files	3843
Unique classes	1189
Files / bird	3,23

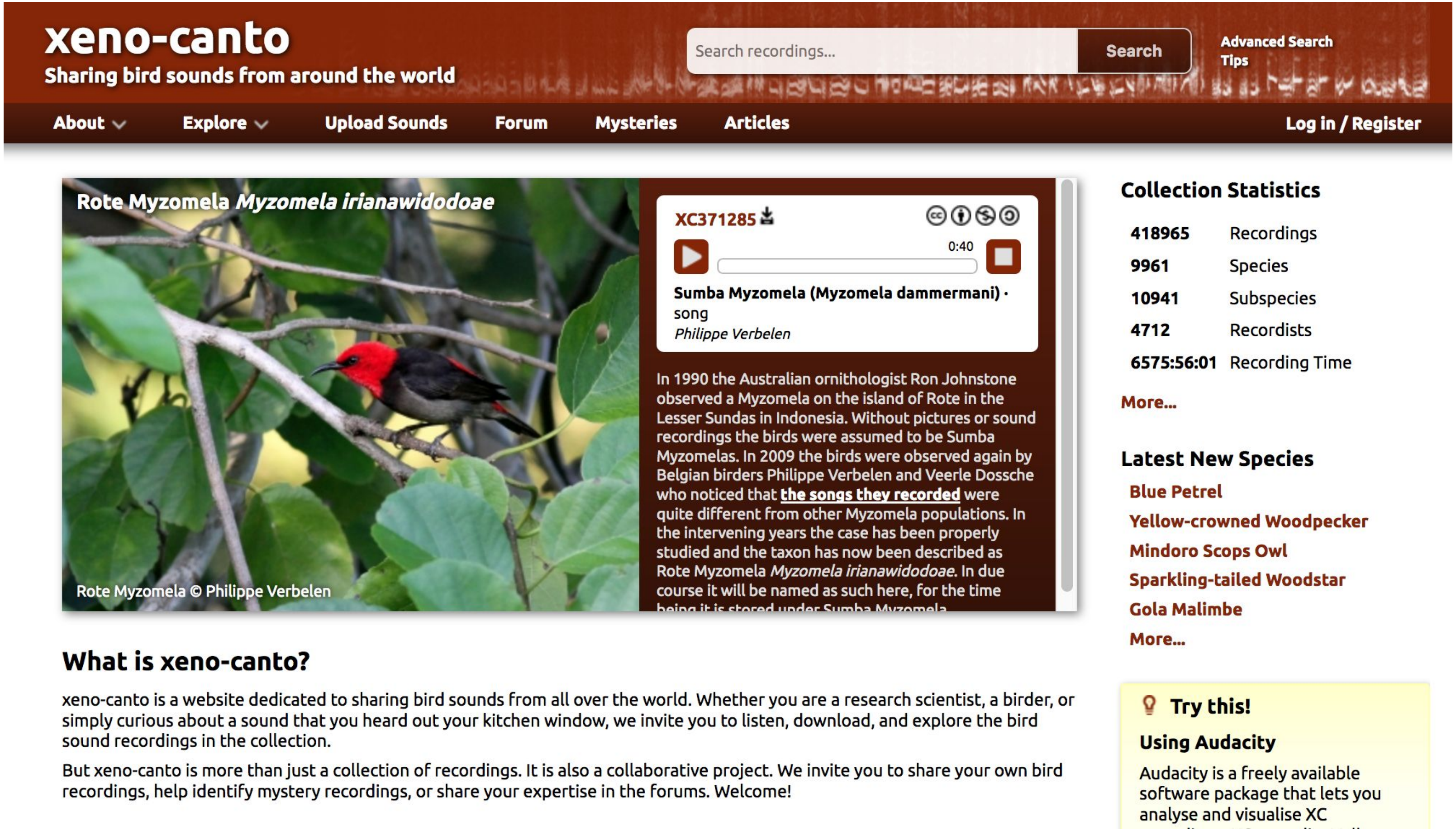
Wikipedia

Idea: check which bird has a German Wikipedia entry

1. Query Wikipedia API to retrieve descriptions for all birds
2. Throw out birds without German entry
3. Update JSON

→ https://github.com/gojibjib/voice-grabber/blob/master/info_grabber/wiki_grabber.py

Total # of files	3400
Unique classes	800
Files / bird	4,25



Rote Myzomela *Myzomela irianawidodoae*

Rote Myzomela © Philippe Verbelen

XC371285

0:40

Sumba Myzomela (Myzomela dammermani) · song

Philippe Verbelen

In 1990 the Australian ornithologist Ron Johnstone observed a Myzomela on the island of Rote in the Lesser Sundas in Indonesia. Without pictures or sound recordings the birds were assumed to be Sumba Myzomelas. In 2009 the birds were observed again by Belgian birders Philippe Verbelen and Veerle Dossche who noticed that **the songs they recorded** were quite different from other Myzomela populations. In the intervening years the case has been properly studied and the taxon has now been described as Rote Myzomela *Myzomela irianawidodoae*. In due course it will be named as such here, for the time being it is stored under Sumba Myzomela

Collection Statistics

418965	Recordings
9961	Species
10941	Subspecies
4712	Recordists
6575:56:01	Recording Time

More...

Latest New Species

- **Blue Petrel**
- **Yellow-crowned Woodpecker**
- **Mindoro Scops Owl**
- **Sparkling-tailed Woodstar**
- **Gola Malimbe**
- **More...**

What is xeno-canto?

xeno-canto is a website dedicated to sharing bird sounds from all over the world. Whether you are a research scientist, a birder, or simply curious about a sound that you heard out your kitchen window, we invite you to listen, download, and explore the bird sound recordings in the collection.

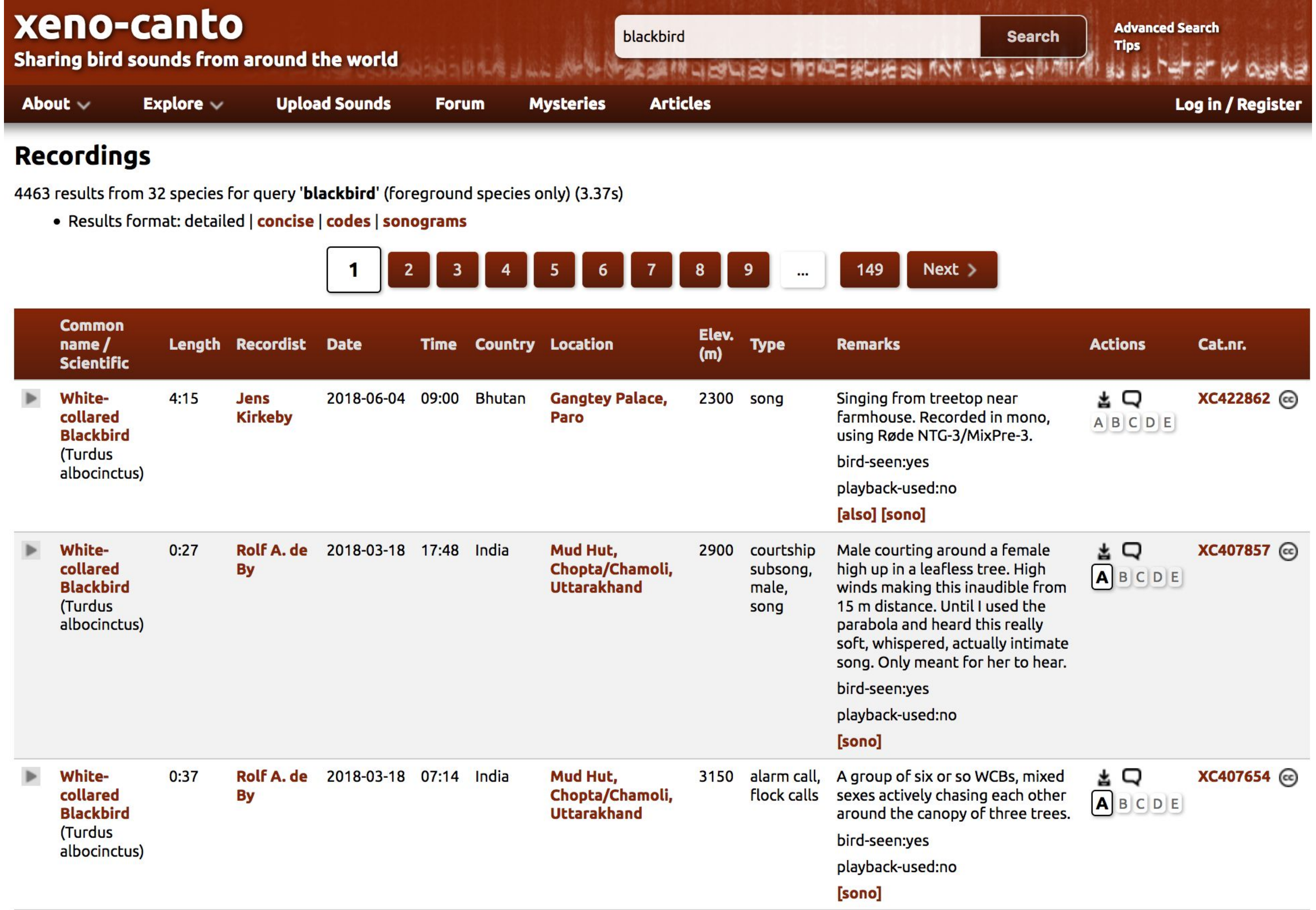
But xeno-canto is more than just a collection of recordings. It is also a collaborative project. We invite you to share your own bird recordings, help identify mystery recordings, or share your expertise in the forums. Welcome!

Try this!

Using Audacity

Audacity is a freely available software package that lets you analyse and visualise XC

http://xeno-canto.org



Recordings

4463 results from 32 species for query '**blackbird**' (foreground species only) (3.37s)

- Results format: detailed | **concise** | **codes** | **sonograms**

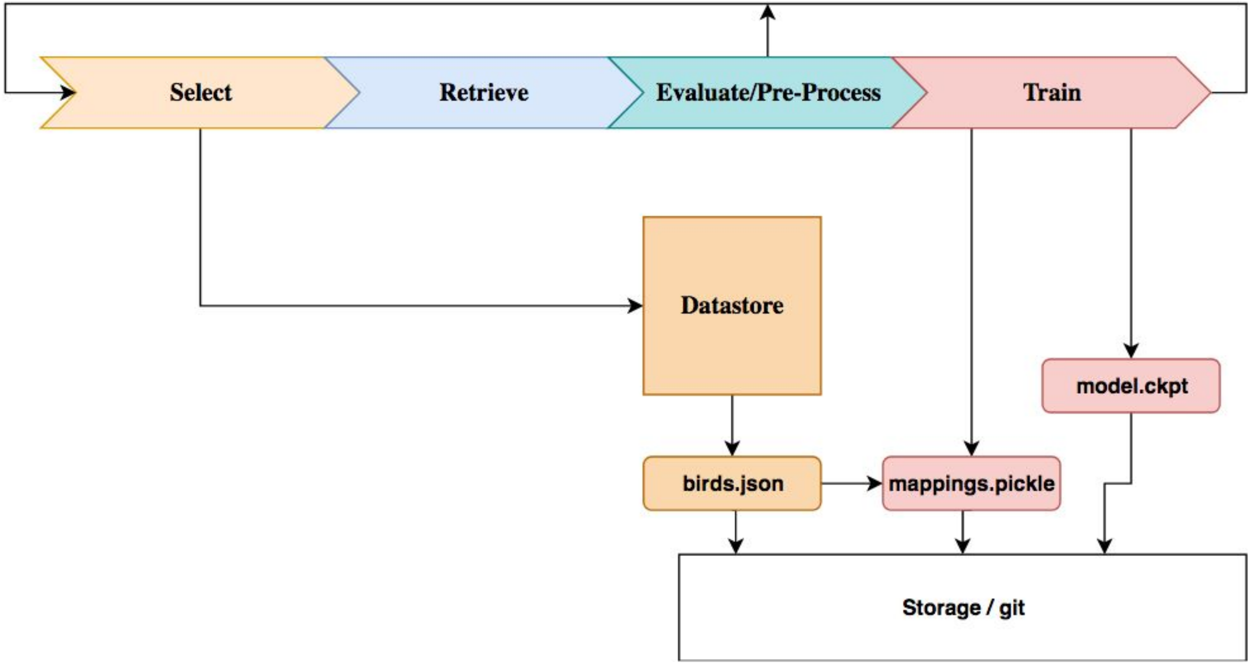
	Common name / Scientific	Length	Recordist	Date	Time	Country	Location	Elev. (m)	Type	Remarks	Actions	Cat.nr.
▶	**White-collared Blackbird** (Turdus albocinctus)	4:15	**Jens Kirkeby**	2018-06-04	09:00	Bhutan	**Gangtey Palace, Paro**	2300	song	Singing from treetop near farmhouse. Recorded in mono, using Røde NTG-3/MixPre-3. bird-seen:yes playback-used:no **[also] [sono]**	A B C D E	**XC422862**
▶	**White-collared Blackbird** (Turdus albocinctus)	0:27	**Rolf A. de By**	2018-03-18	17:48	India	**Mud Hut, Chopta/Chamoli, Uttarakhand**	2900	courtship subsong, male, song	Male courting around a female high up in a leafless tree. High winds making this inaudible from 15 m distance. Until I used the parabola and heard this really soft, whispered, actually intimate song. Only meant for her to hear. bird-seen:yes playback-used:no **[sono]**	A B C D E	**XC407857**
▶	**White-collared Blackbird** (Turdus albocinctus)	0:37	**Rolf A. de By**	2018-03-18	07:14	India	**Mud Hut, Chopta/Chamoli, Uttarakhand**	3150	alarm call, flock calls	A group of six or so WCBs, mixed sexes actively chasing each other around the canopy of three trees. bird-seen:yes playback-used:no **[sono]**	A B C D E	**XC407654**

http://xeno-canto.org

Final Dataset

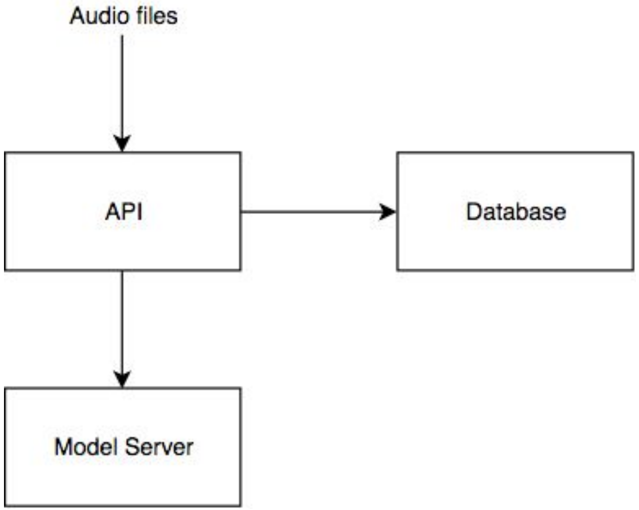
Total size of dataset	120GB
Total # of files	80.000
Unique classes	194
Files / bird	412,4

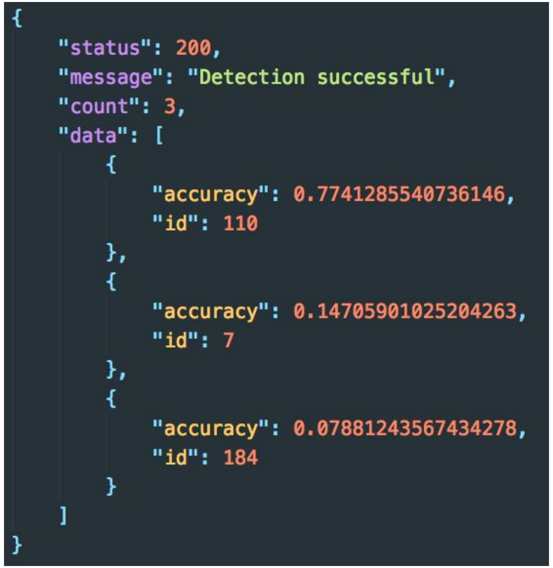
Data Science Workflow



Backend Architecture

Backend Architecture





```
{
    "status": 200,
    "message": "Detection successful",
    "count": 3,
    "data": [
        {
            "accuracy": 0.7741285540736146,
            "id": 110
        },
        {
            "accuracy": 0.14705901025204263,
            "id": 7
        },
        {
            "accuracy": 0.07881243567434278,
            "id": 184
        }
    ]
}
```

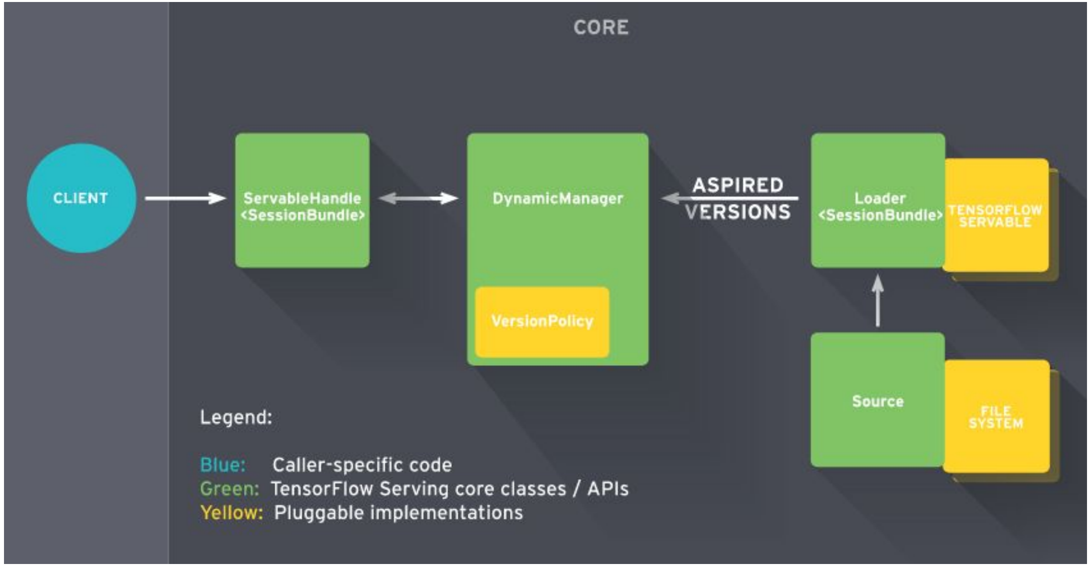
Model Server

1. Load the model (= construct the graph)
2. Accept audio file
3. Convert to .wav
4. Convert .wav to spectrogram to tf.Tensor()
5. Run tf.Session() ?
6. Parse output Tensor
7. Send response

TensorFlow Serving

- Models represented as **servables**
- **Managers** handling loading, serving, unloading of servables
- Lifecycle management, versioning, A/B testing
- Client-Server architecture to query model
- Able to handle 100k QPS / core
- → https://arxiv.org/pdf/1712.06139.pdf

Tensorflow Serving



Legend:

Blue: Caller-specific code
Green: TensorFlow Serving core classes / APIs
Yellow: Pluggable implementations

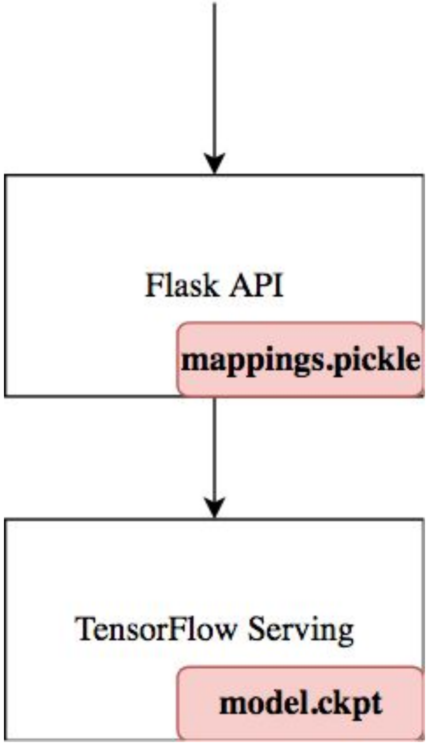
Model Server

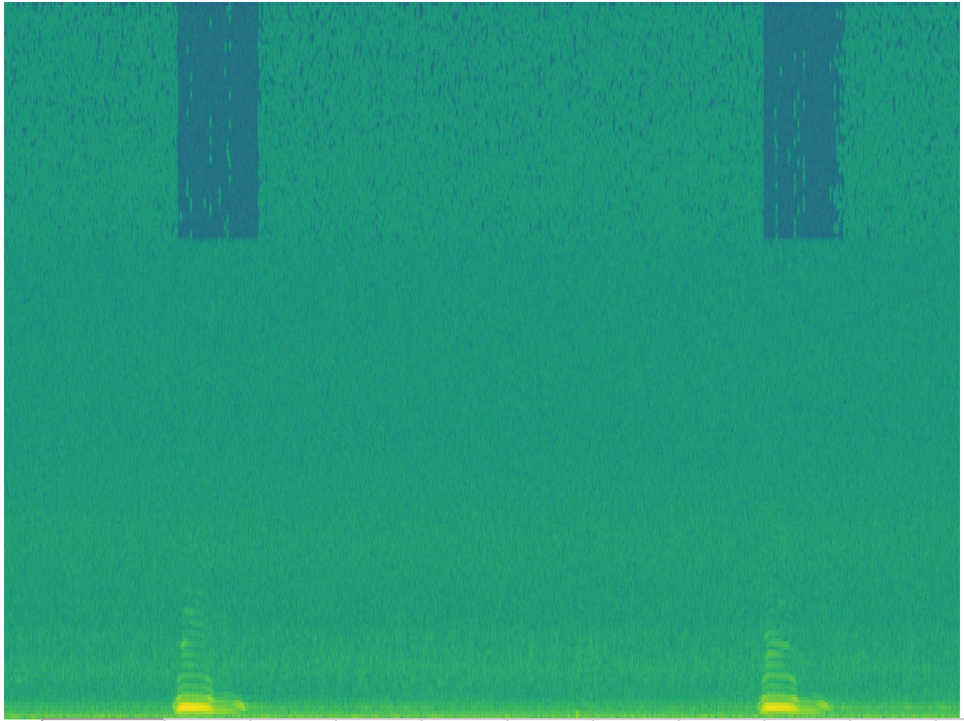
Thin Python Flask API

- Converts audio file
- Constructs request Tensor + converts to protobuf
- Parses response Tensorf from TF Serving

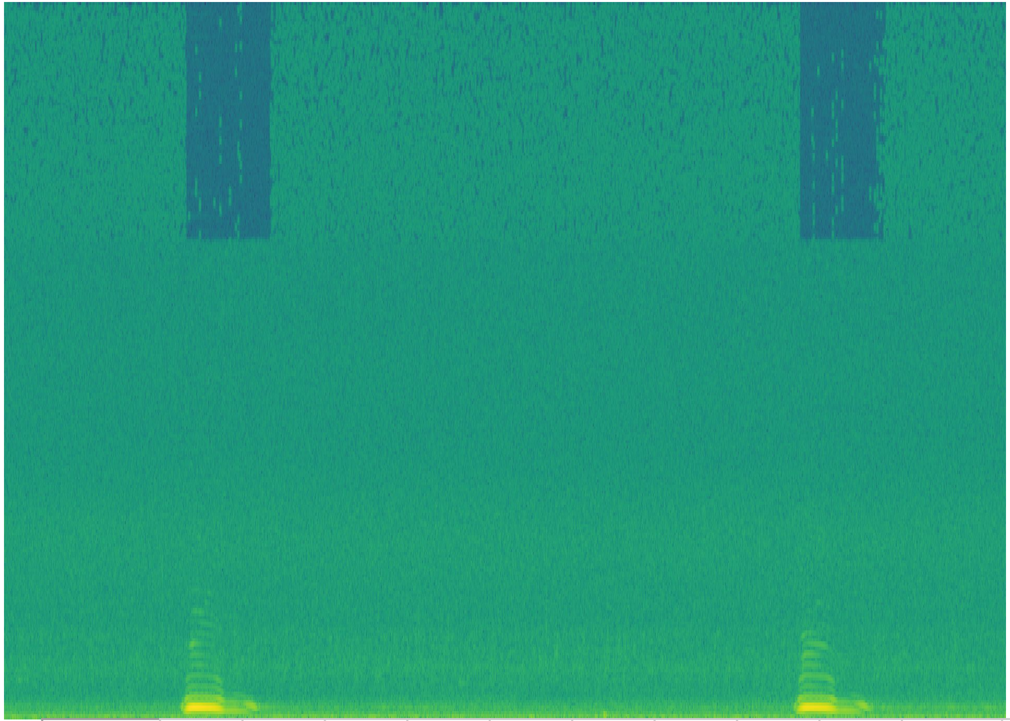
TF Serving

- Holds model in protobuf format
- Handles inference

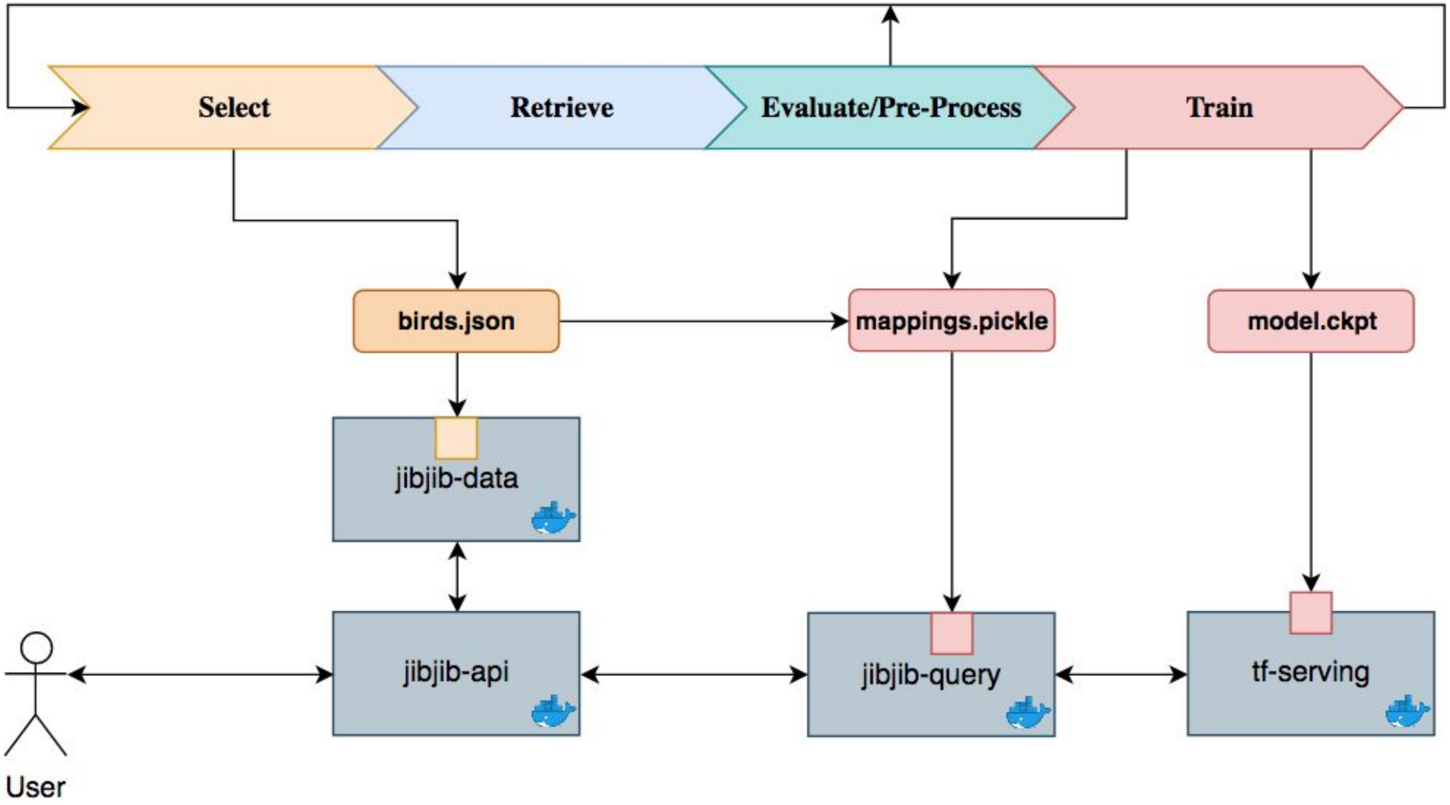




Class 1 (actual class)	0.8	0.04	0.2	0.08	0.04	0.12	0.12	0.8	0.08	0.12	**2.4**
Class 2	0.4	0.8	0.4	0.04	0.6	0.04	0.08	0.4	0.04	0.08	**2.88**



Class 1 (actual class)	0.008	0.000512	0.000064	0.001728	0.001728	0.512	0.000512	0.001728	0.008	0.000512	1.038336
Class 2	0.064	0.000064	0.216	0.000064	0.000512	0.064	0.000064	0.000512	0.064	0.000064	0.921216


Select
Retrieve
Evaluate/Pre-Process
Train
birds.json
mappings.pickle
model.ckpt
jibjib-data
jibjib-api
jibjib-query
tf-serving
User

Overview

Thank you!

https://github.com/gojibjib

Google Play Store: JibJib

gojibjib@gmail.com