

# **LAB: Audio Transfer Learning with Scikit-learn and Tensorflow**

Jordi Pons and Xavier Favory

<https://github.com/jordipons/sklearn-audio-transfer-learning>

# **About us**

# Music Technology Group

- Research group at Universitat Pompeu Fabra
  - Research on **music** and **audio technologies**
  - Founded 25 years ago by his current director Xavier Serra
  - Four labs:
    - Audio Signal Processing Lab (led by Xavier Serra)
    - Music Information Research Lab (led by Emilia Gómez)
    - Music and Multimodal Interaction Lab (led by Sergi Jordà)
    - Music and Machine Learning Lab (led by Rafael Ramírez)
- + info at [www.mtg.upf.edu](http://www.mtg.upf.edu) and @mtg\_upf

# Jordi Pons

- **Researcher** at Dolby
  - **PhD candidate** at the Music Technology Group
  - Also worked at:
    - Telefónica Research (Barcelona)
    - Pandora Radio (Bay Area, USA)
    - German Hearing Center (Hannover, Germany)
    - IRCAM (Paris, France)
- + info at [www.jordipons.me](http://www.jordipons.me) and at @jordiponsdotme

# Xavier Favory

- Music & Web technologies enthusiast
- **PhD candidate** at the Music Technology Group
- **ENSEA** (École nationale supérieure de l'électronique et de ses applications)
- **Master** of engineering at ENSEA (École Nationale Supérieure de l'Électronique et de ses Applications), Cergy.
- **Master** Acoustics, Signal processing, Informatics applied to Music (ATIAM), IRCAM , Paris.

# Today's plan

# Today's schedule

- Setup: download audio data and pre-trained model
- Introduction to Audio Transfer Learning
- Demonstration
- Questions
- **Hands-on example: transfer learning with neural networks**
- **Start competition!**

# **Audio transfer learning**

# Training neural audio classifiers with few data

## HOW?

- **Strong regularization**
  - We assess the limitations of the standard deep learning pipeline
- **Transfer learning**
  - Enables to leverage external sources of audio data

+ info in our paper:

<https://arxiv.org/abs/1810.10274>

# Methodology

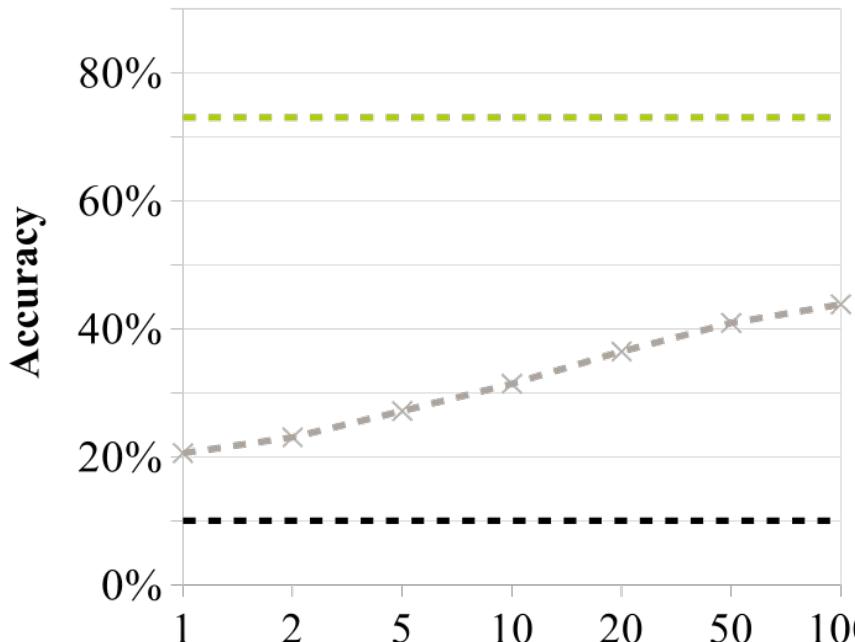
# Targeted tasks and our data

- **Acoustic Event Recognition** (US8K dataset)
  - 8,732 urban sounds
  - **10 classes**: *car horn, children playing, dog bark, gun shot, siren, ...*
  - 10 folds
- **Acoustic Scene Classification** (ASC-TUT dataset)
  - 4,680 training audio segments
  - 1,620 evaluation audio segments
  - **15 classes**: *park, home, office, train, bus, ...*

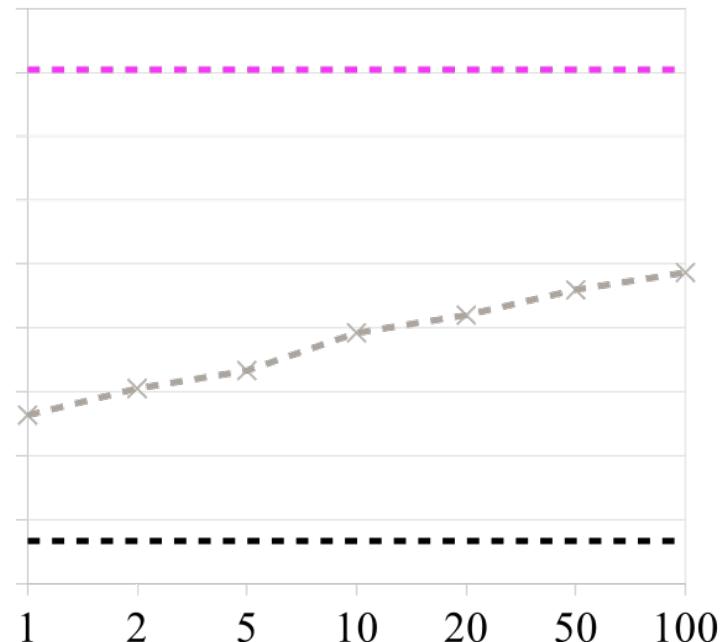
# Evaluation

*The MFCC's + nearest neighbor baseline*

Acoustic event recognition



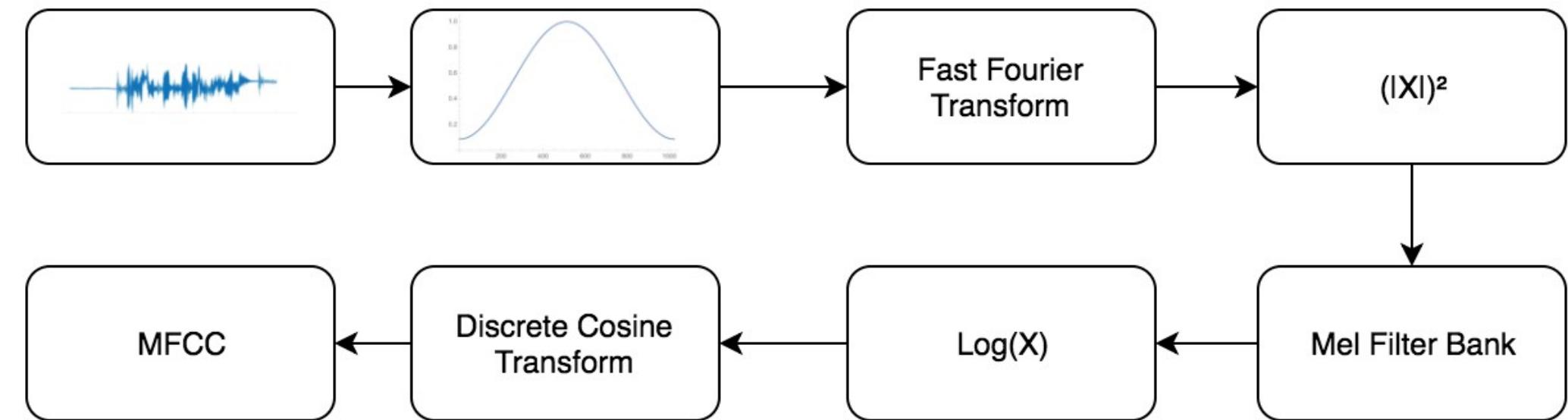
Acoustic scene classification



Number of training examples per class ( $n$ )

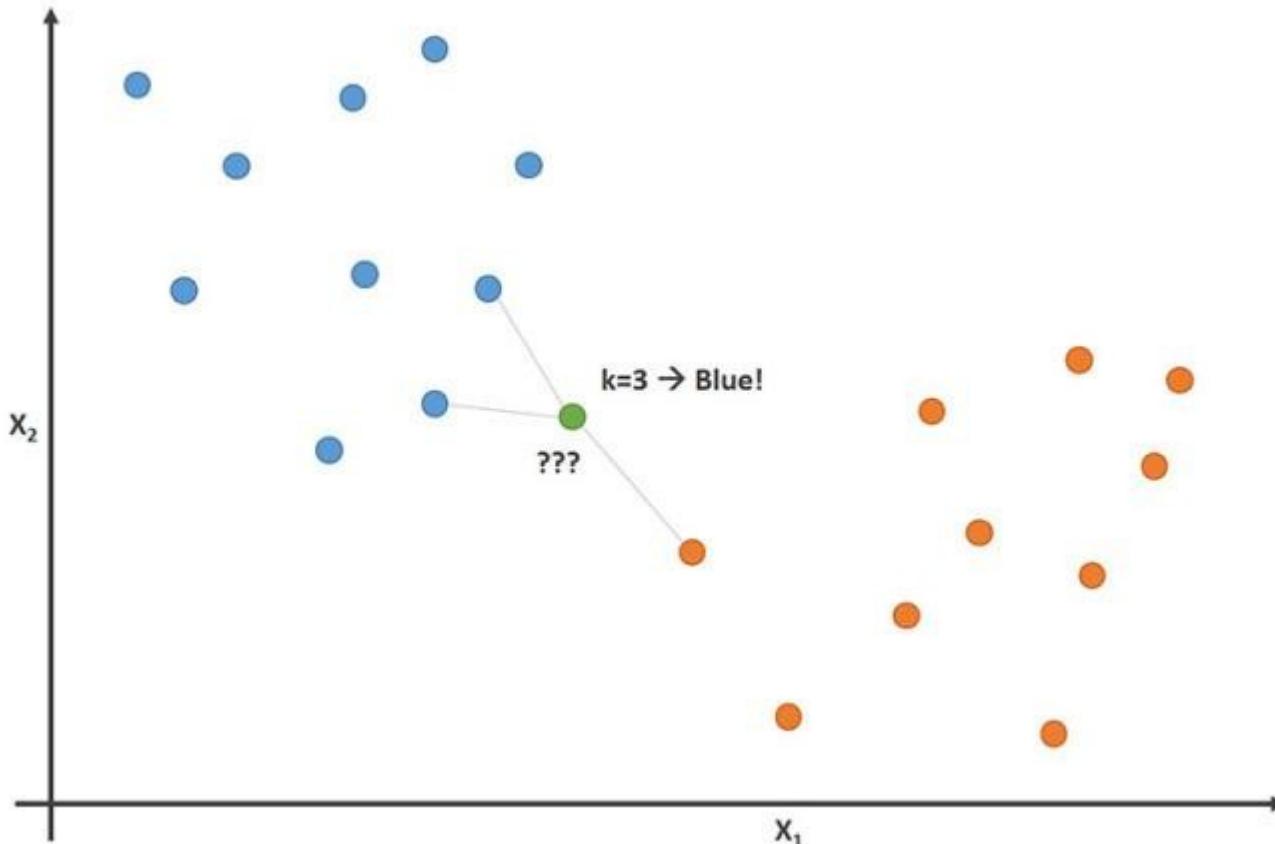
# Evaluation

*The MFCC's + nearest neighbor baseline*



# Evaluation

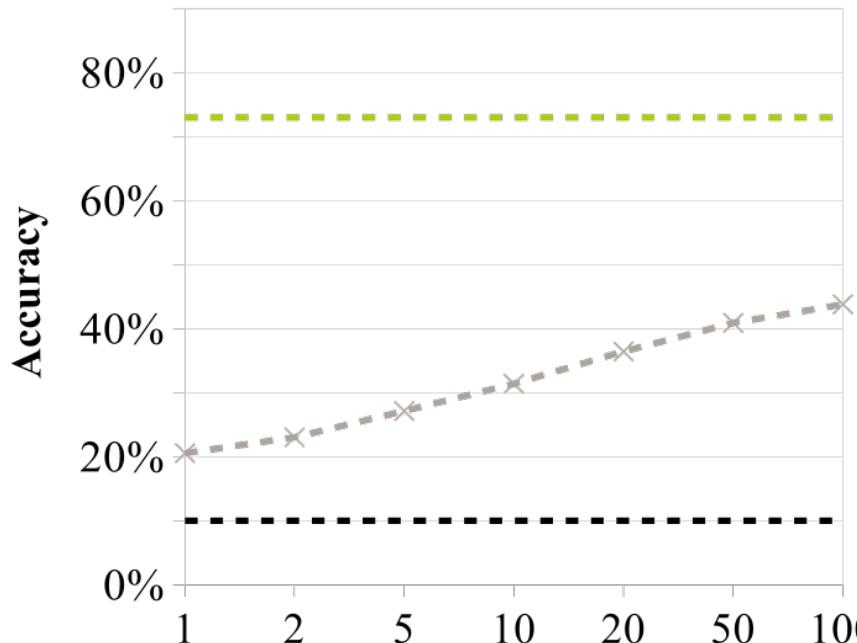
*The MFCC's + nearest neighbor baseline*



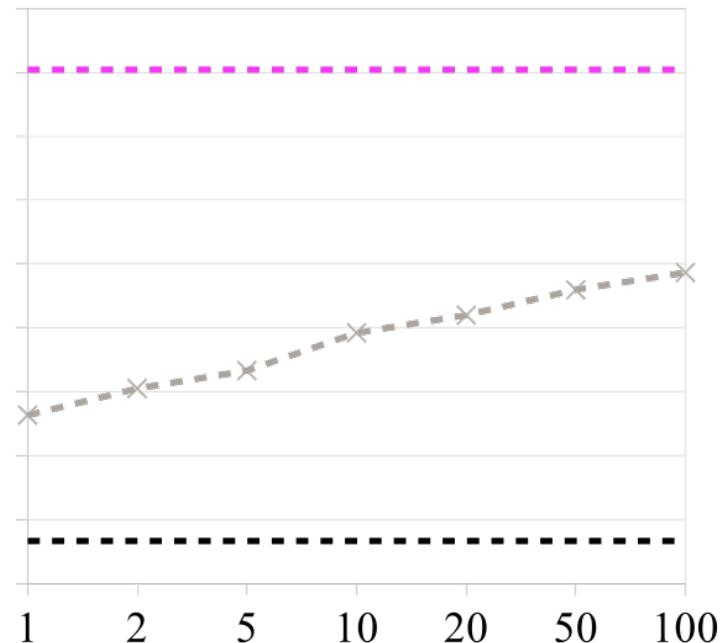
# Evaluation

*The MFCC's + nearest neighbor baseline*

Acoustic event recognition



Acoustic scene classification



Number of training examples per class ( $n$ )

Regularized models

Transfer learning

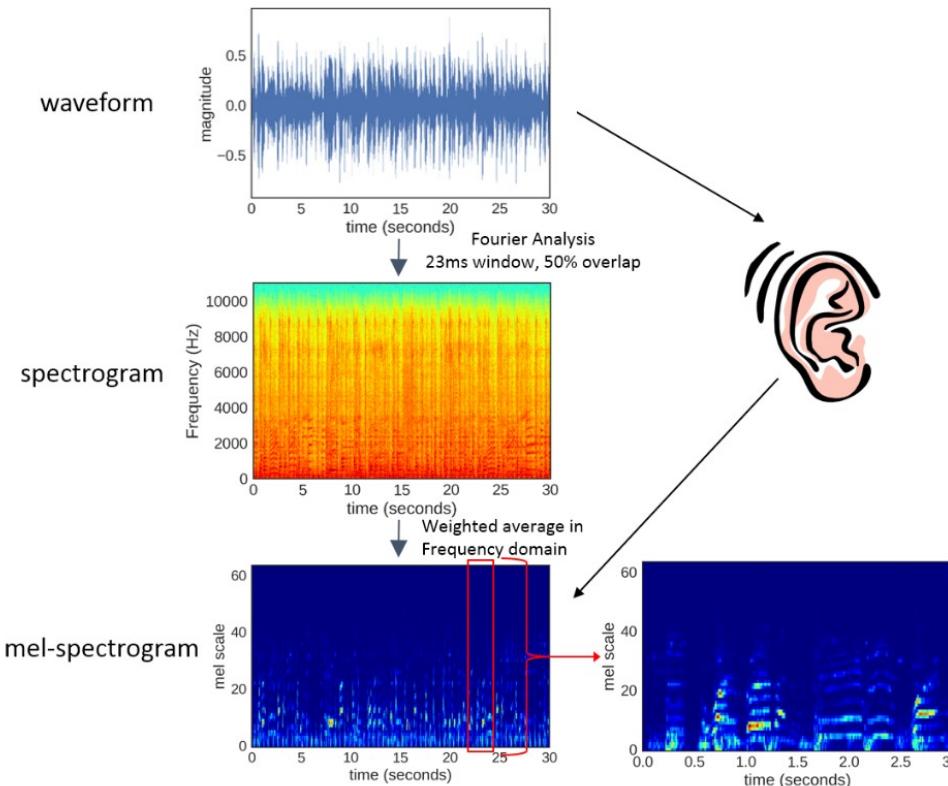
Regularized models

Transfer learning

# **Regularized models**

# Input

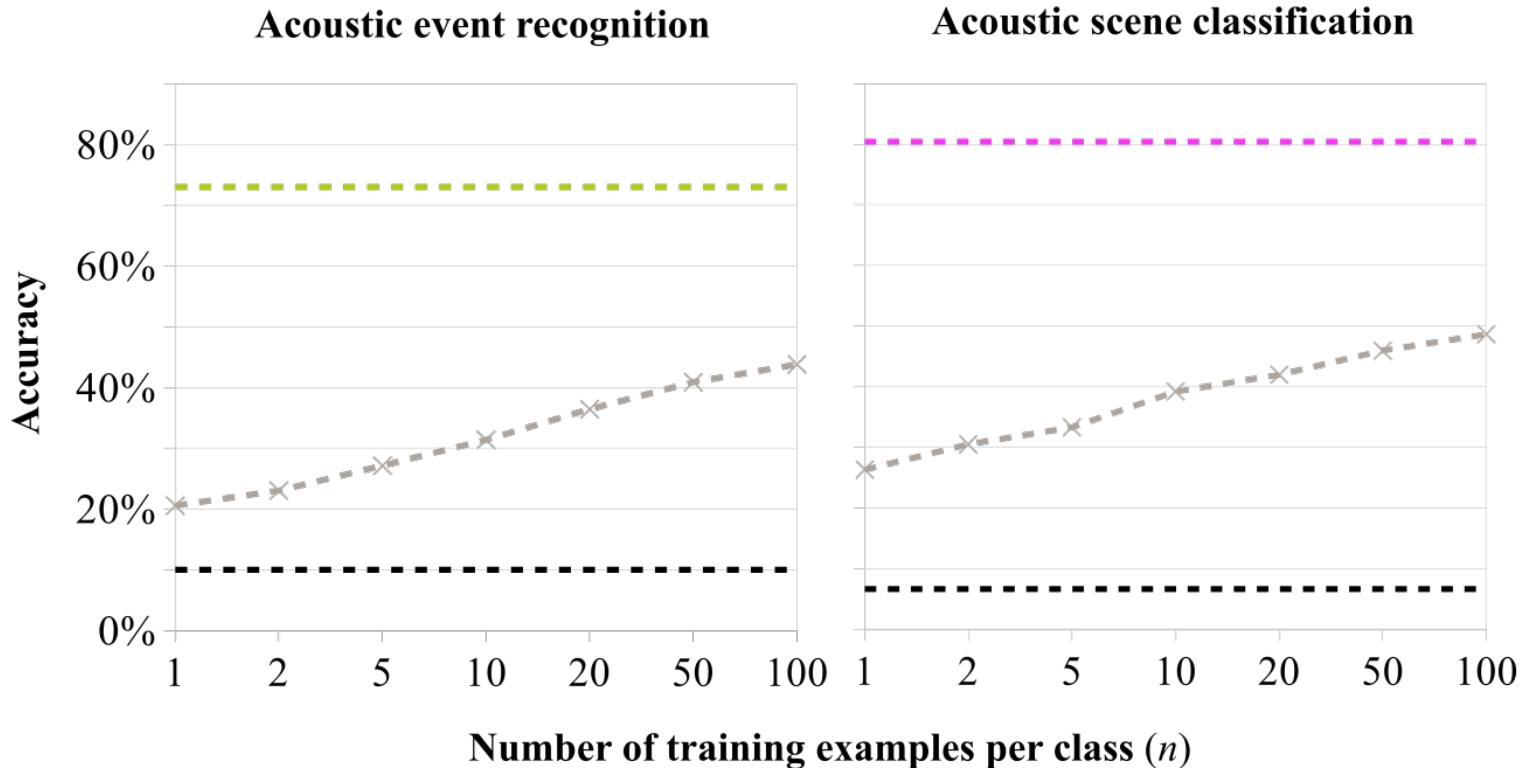
**Input:** log-mel spectrogram of 128 bins x 3 sec (128 frames)



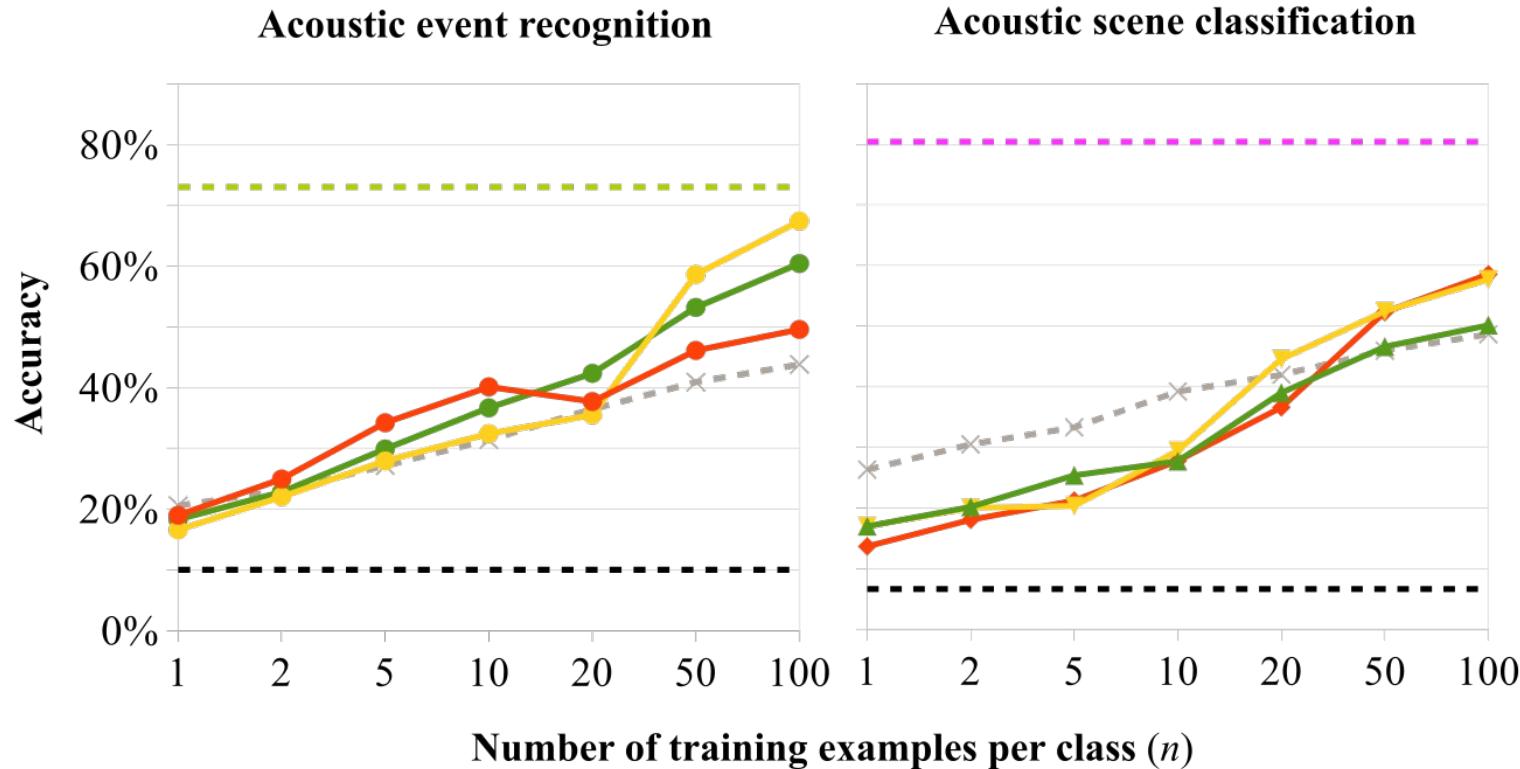
# Regularized models

- **SB-CNN**: **250k** parameters
  - Inspired by AlexNet's computer vision architecture
  - *3 CNN layers (5x5) with max-pool + dense layer + softmax*
- **VGG**: **50k** parameters
  - Yet another computer vision architecture
  - *5 CNN layers (3x3) with max-pool (2x2) + softmax*
- **TIMBRE**: **10k** parameters
  - The smallest CNN one can imagine for learning timbral traces
  - *1 CNN layer (vertical filters 108x7) with maxpool + softmax*

----- Random guess  
-- $\times$ -- Nearest-neigbor MFCCs



SB-CNN      Random guess  
VGG      Nearest-neigbor MFCCs  
TIMBRE



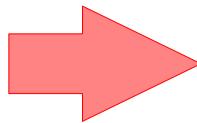
Regularized models

Transfer learning

**Transfer learning**

# Transfer learning

Pretrain with  
source task



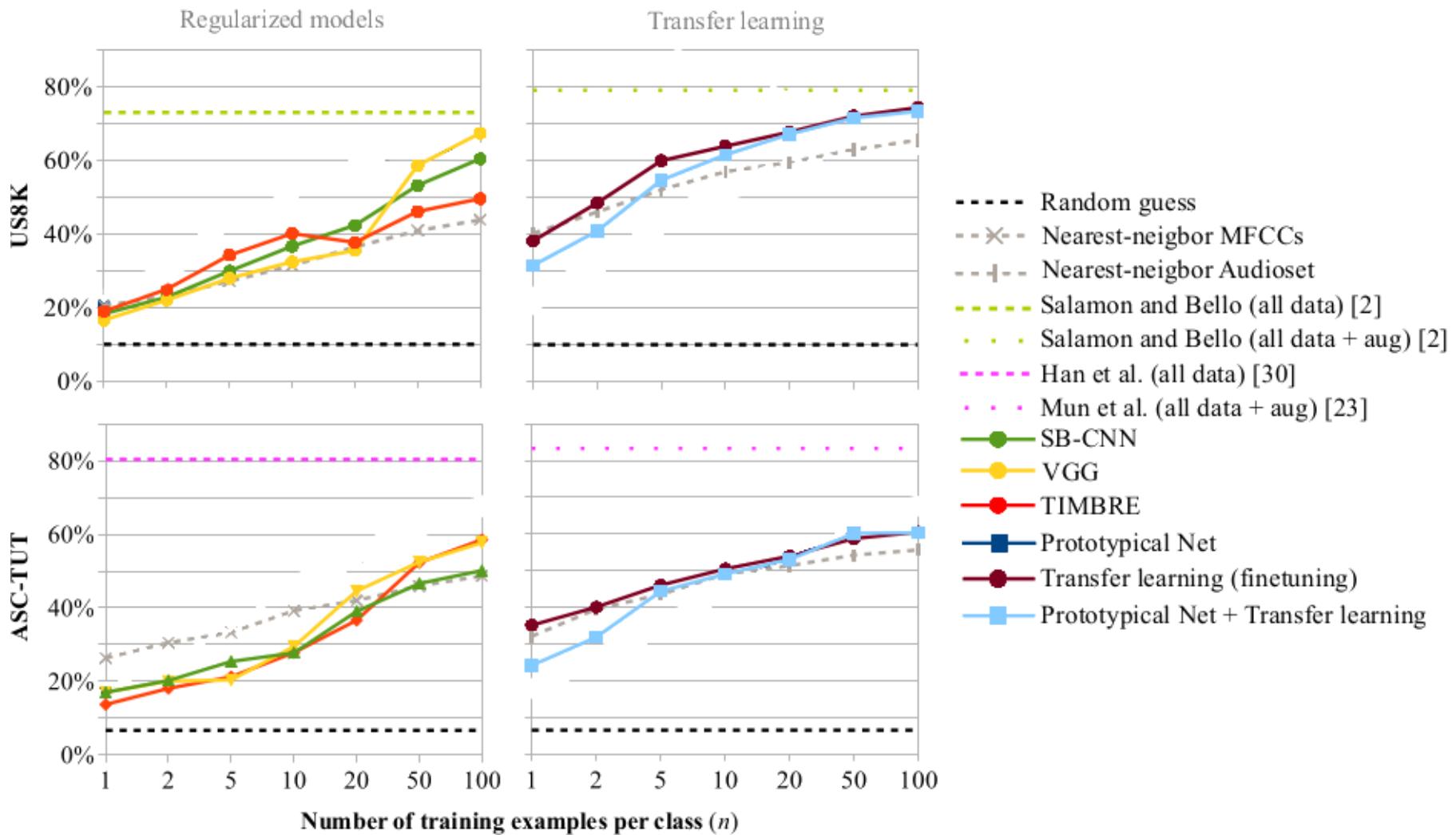
Finetune with  
target task(s)

**AudioSet dataset**  
(acoustic event recognition)  
2M Youtube audios

**US8K dataset**  
(acoustic event recognition)

**ASC-TUT dataset**  
(acoustic scene classification)

Pre-trained **VGGish** on AudioSet:  
6 CNN layers ( $3 \times 3$ )  
with max-pool layers ( $2 \times 2$ ) +  
3 dense layers (4096, 4096, 128)



# **Conclusions**

# Training neural audio classifiers with few data

- **Strong regularization**
  - We assess the limitations of the standard deep learning pipeline
- **Transfer learning**
  - Enables to leverage external sources of audio data
  - Be careful with source and target tasks. They need to be similar!

Let's try it!

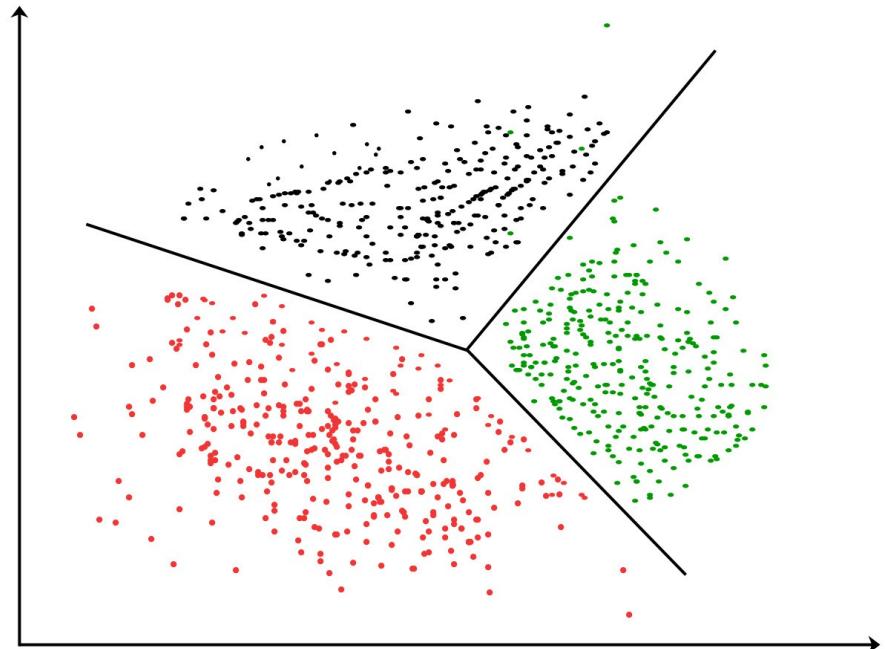
# Demonstration

# Using Deep Embeddings for Unsupervised Tasks

- **Clustering:**

- Unsupervised classification which consists in organising similar objects in groups called **clusters**
- exploratory data analysis
- Requires a notion of **similarity**

What about using **AudioSet embeddings**?



# Using Deep Embeddings for Unsupervised Tasks

- **Freesound**

- Sound sharing platform
- > 400 000 sounds
- + 1 000 sounds per week
- Creative Commons licenses

<https://freesound.org>



A screenshot of the Freesound search interface. The search bar at the top contains the word "wind". Below the search bar, there are buttons for "Automatic by relevance" and "search". A "Show advanced search options" link is also present. The search results page shows a list of sound clips related to "wind". Each result includes a thumbnail image, the title, a star rating, the upload date, download count, comment count, and a "more" button. To the right of the results, there are sections for "licenses", "tags", and "type".

**licenses**

- Attribution (4815)
- Attribution Noncommercial (1260)
- Creative Commons 0 (3678)
- Sampling+ (172)

**tags**

air ambience ambience  
ambient atmosphere  
atmospheric background-sound  
birds blowing calm city field-  
recording forest general-noise  
holland nature noise rain  
soundscape spring storm traffic trees  
water waves weather white-noise  
wind windy winter

**type**

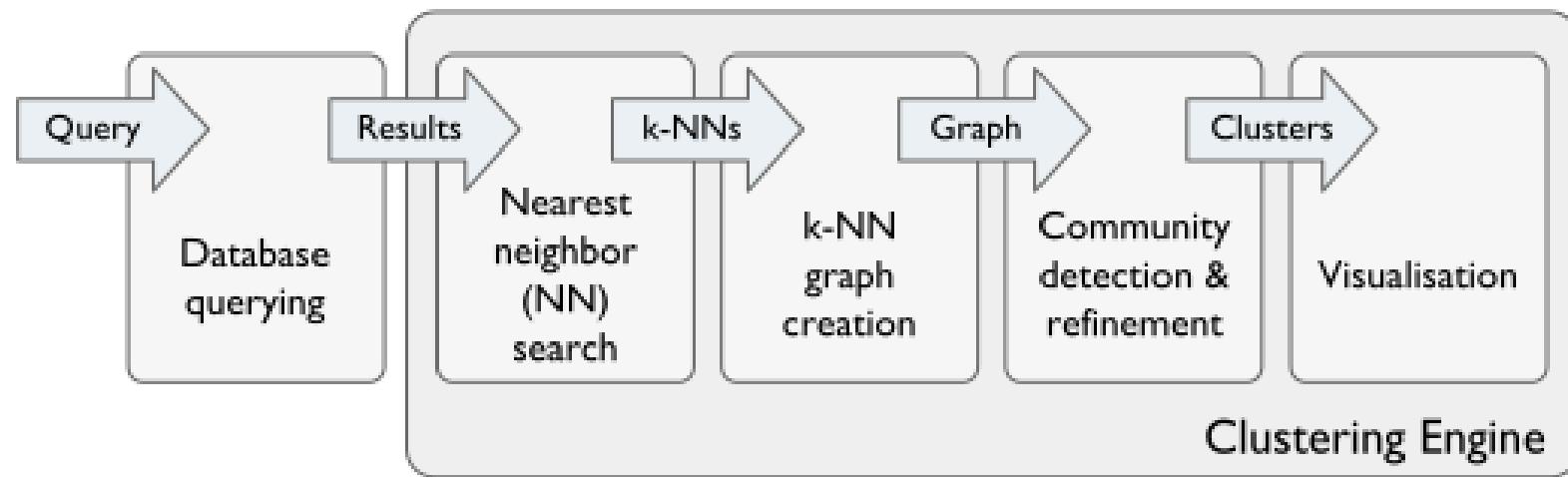
- aiff (503)
- flac (1195)
- m4a (76)
- mp3 (949)
- ogg (176)
- wav (7026)

**Search Results for "wind":**

- Soft Wind** (4 stars) - uploaded by florianreichelt on February 18th, 2019. 5421 downloads, 10 comments. Tags: Wind, white-noise, Soft, soundscape, atmospheric, ambience, ambient, background, wind, atmos. 1 more result in the same pack "Wind Sounds".
- Wind Through Trees 3b** (4 stars) - uploaded by spoonbender on August 16th, 2014. 36046 downloads, 52 comments. Tags: Recordings, Field-Recording, Nature, Leaves, Northumberland. 3 more results in the same pack "Wind through trees".
- Wind blow, mouth.wav** (4 stars) - uploaded by Wesselorg on November 5th, 2017. 11022 downloads, 15 comments. Tags: wind, blow, mouth, air, blowing, blow. OWI wind Wind mouth air blowing blow.
- short wind noise** (3 stars) - uploaded by michimuc2 on August 2nd, 2013. 8889 downloads, 89 comments. Tags: a short wind noise which I made myself, yeah an it's fuckin amazin.

# Using Deep Embeddings for Unsupervised Tasks

- **Search Result Clustering**



DEMO

# **Questions?**

# Today's schedule

- Setup: download audio data and pre-trained model
- Introduction to Audio Transfer Learning
- Demonstration
- Questions
- **Hands-on example: transfer learning with neural networks**
- **Start competition!**