

# **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>

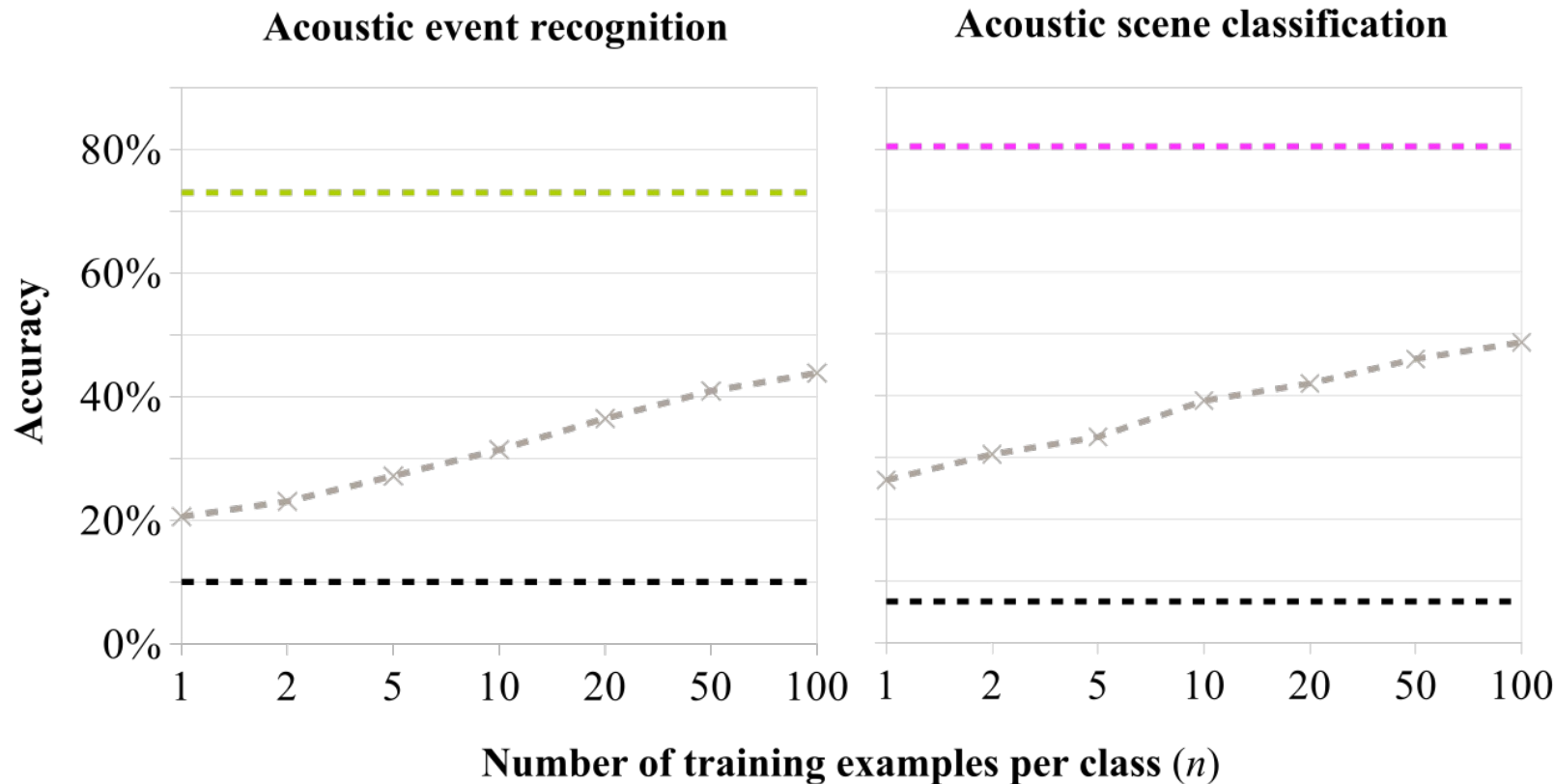
# Methodology

# 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, ...*

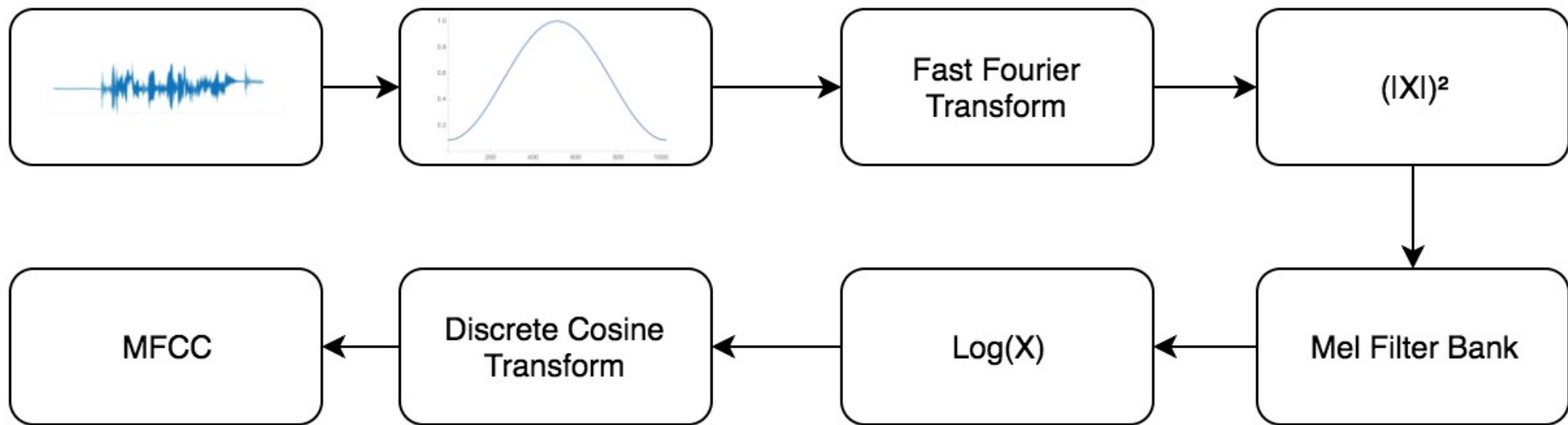
# Evaluation

*The MFCC's + nearest neighbor baseline*



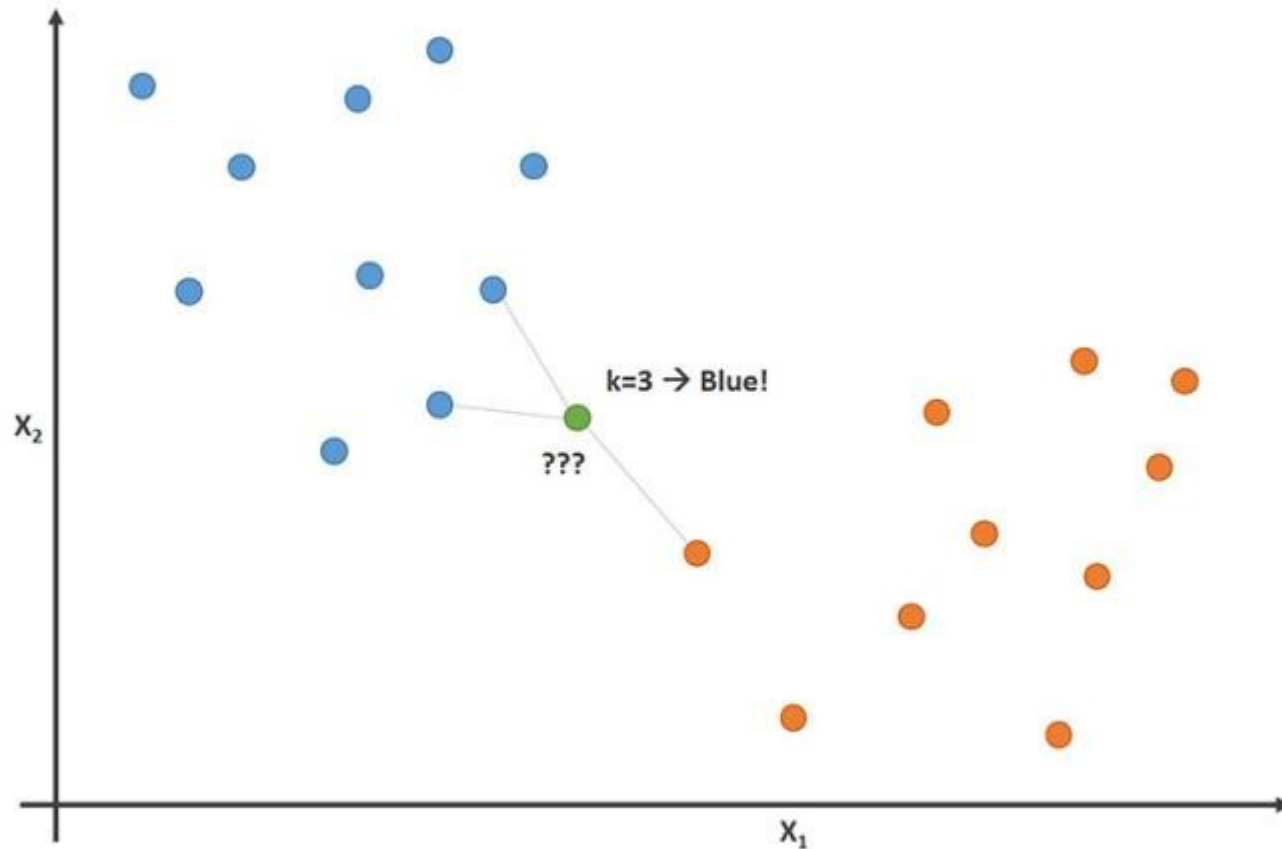
# Evaluation

*The **MFCC**'s + nearest neighbor baseline*



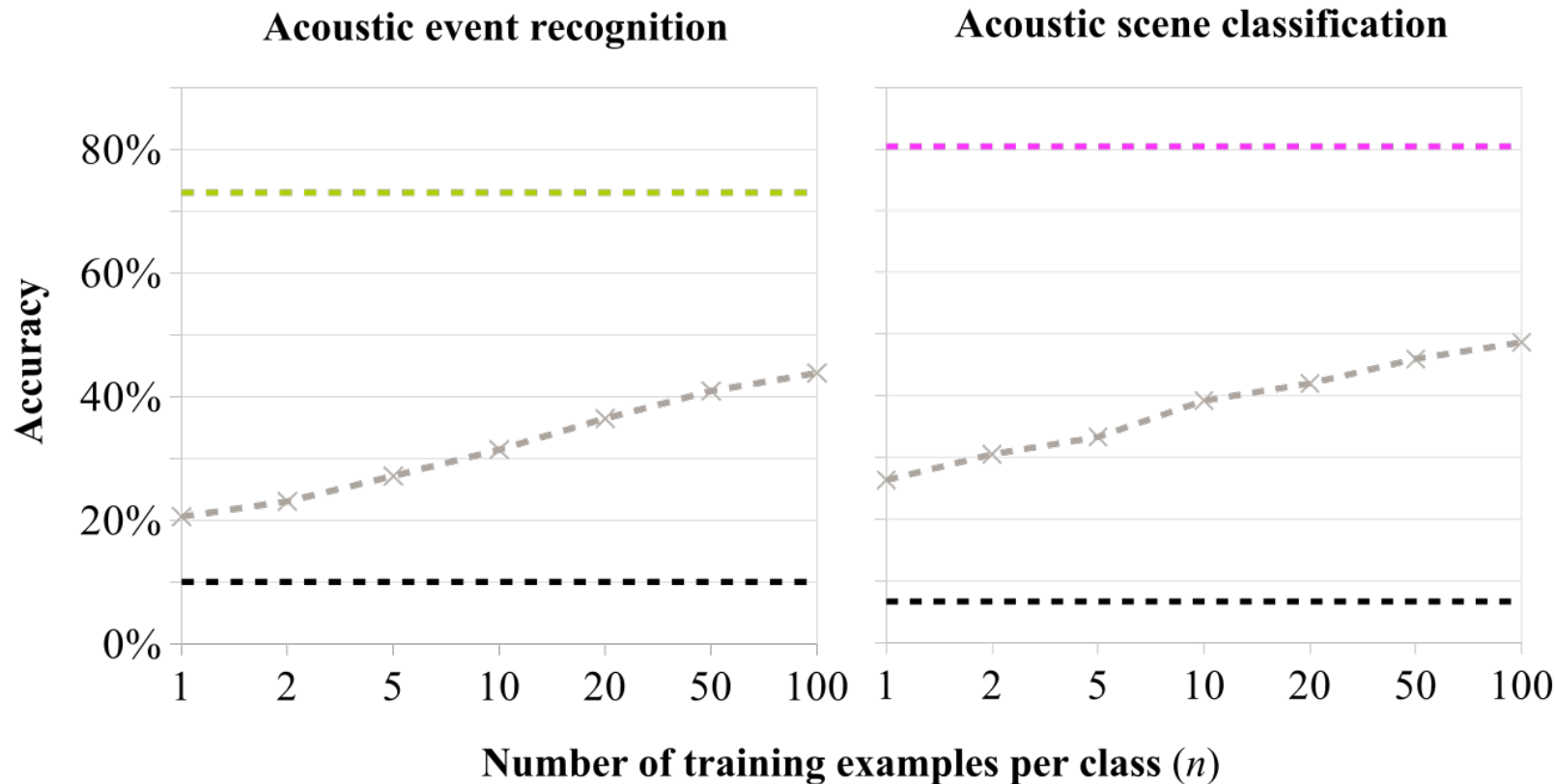
# Evaluation

*The MFCC's + **nearest neighbor** baseline*



# Evaluation

*The MFCC's + nearest neighbor baseline*



Regularized models

Transfer learning



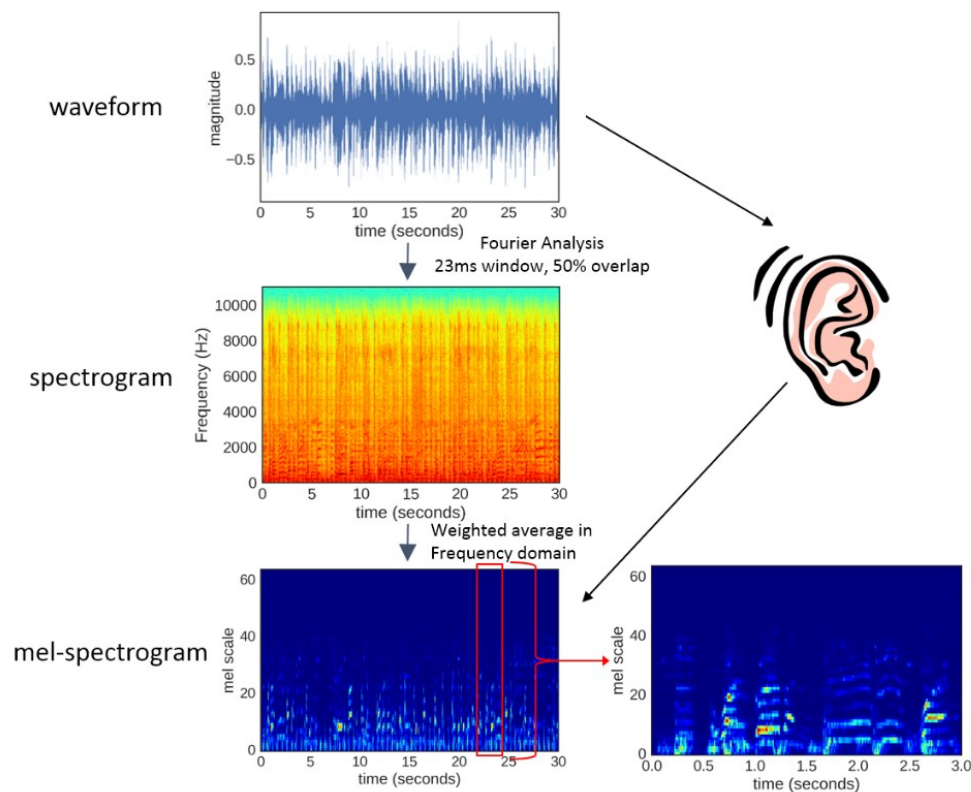
Regularized models

Transfer learning

# **Regularized models**

# Input

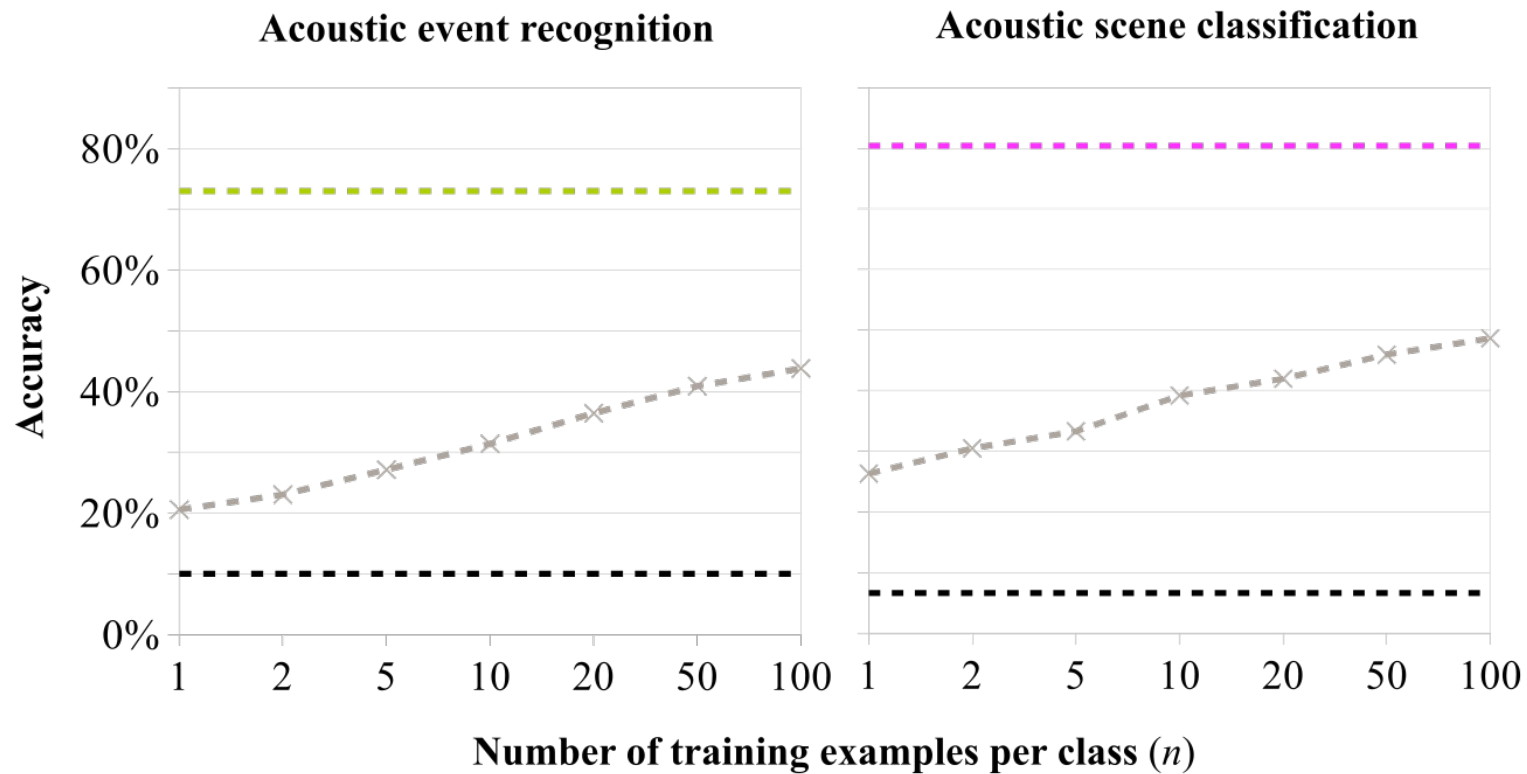
**Input:** log-mel spectrogram of 128 bins x 3 sec (128 frames)

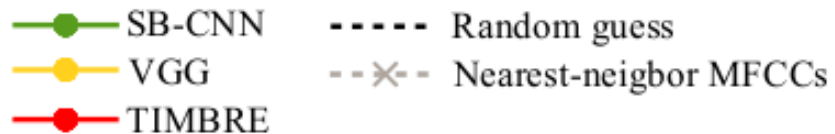


# Regularized models

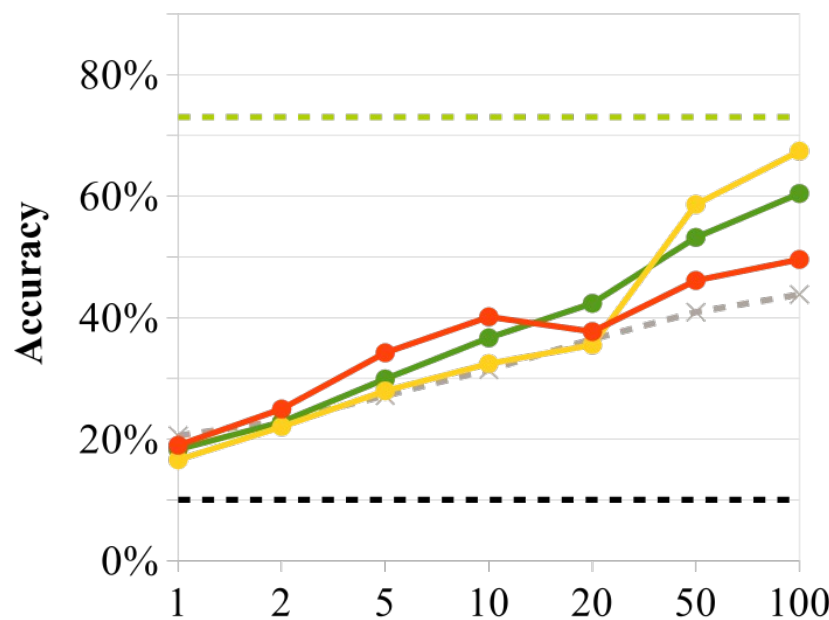
- **SB-CNN: 250k** parameters
  - Inspired by AlexNet's computer vision architecture
  - *3 CNN layers (5x5) with max-pool + dense layer + softmax*
- **VGG: 50k** parameters
  - Yet another computer vision architecture
  - *5 CNN layers (3x3) with max-pool (2x2) + softmax*
- **TIMBRE: 10k** parameters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

- Random guess
- - x - - Nearest-neighbor MFCCs

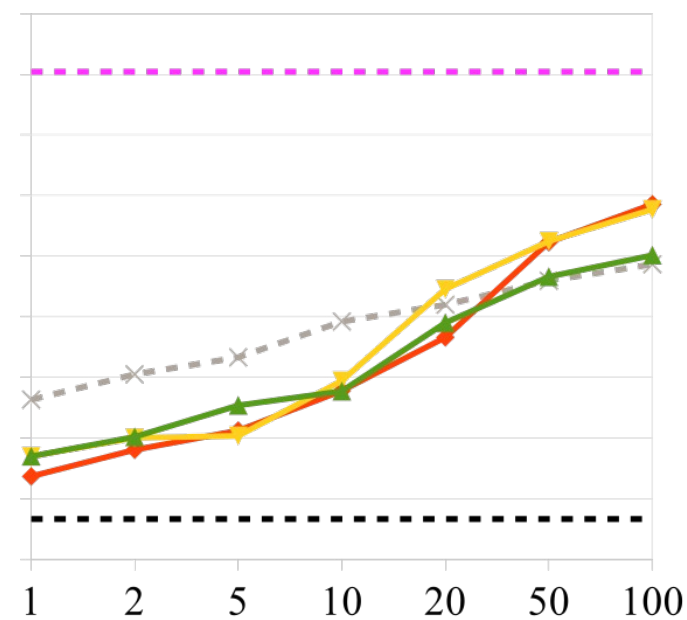




**Acoustic event recognition**



**Acoustic scene classification**



Number of training examples per class ( $n$ )

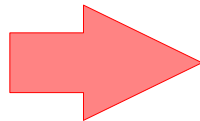
Regularized models

Transfer learning

**Transfer learning**

# Transfer learning

**Pretrain with  
source task**



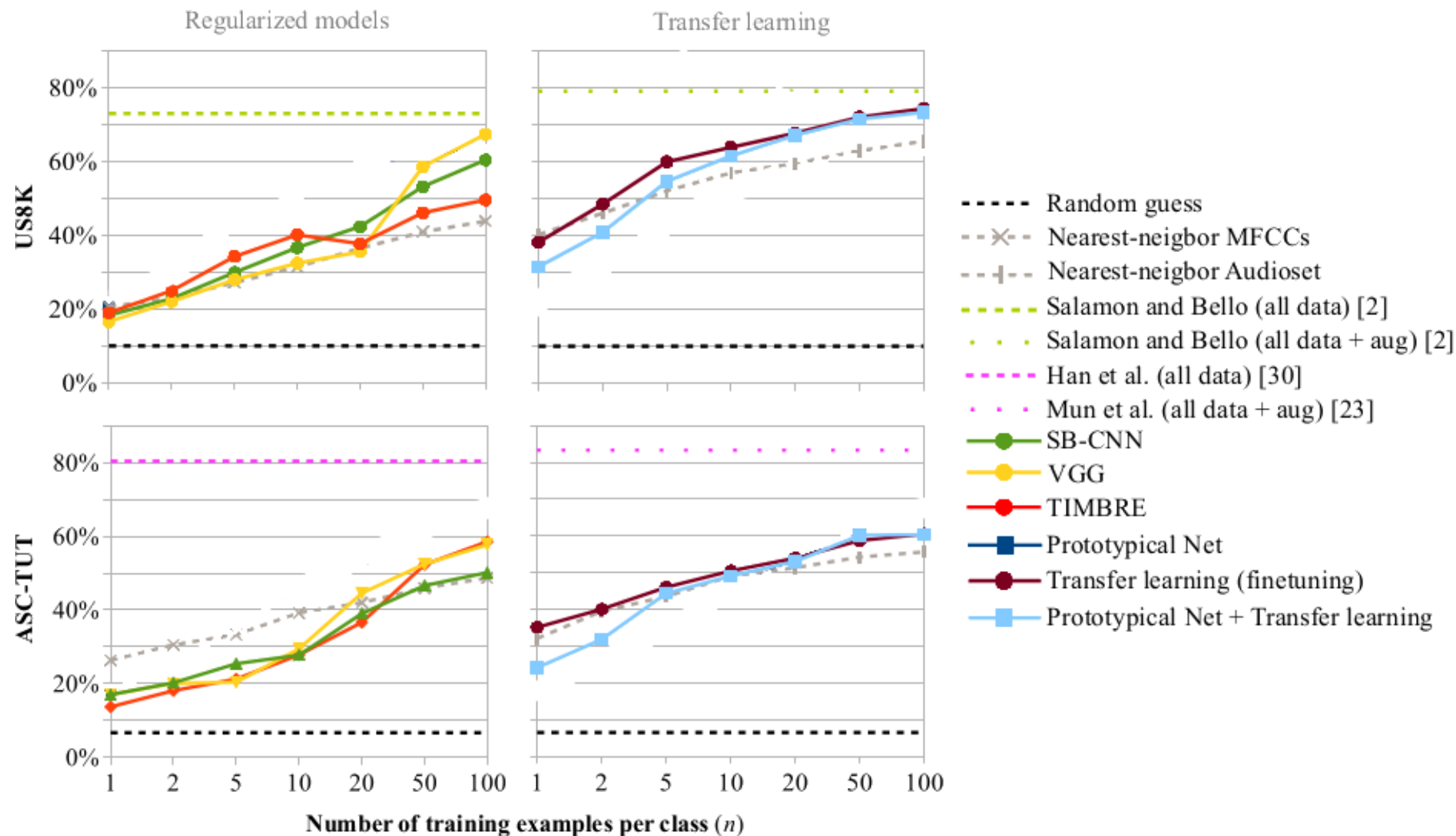
**Finetune with  
target task(s)**

**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 (3×3)  
with max-pool layers (2×2) +  
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!

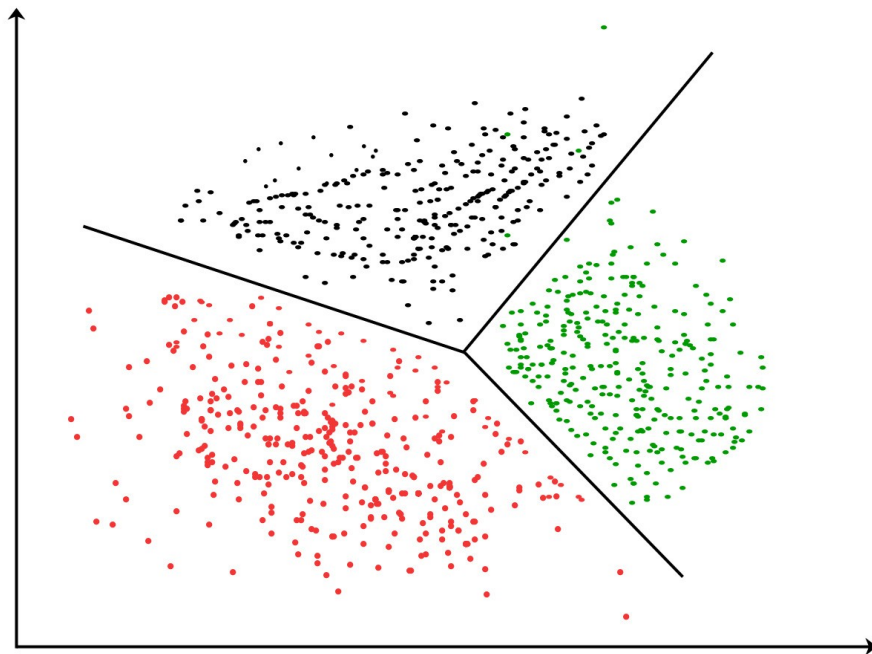
# Demonstration

# 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

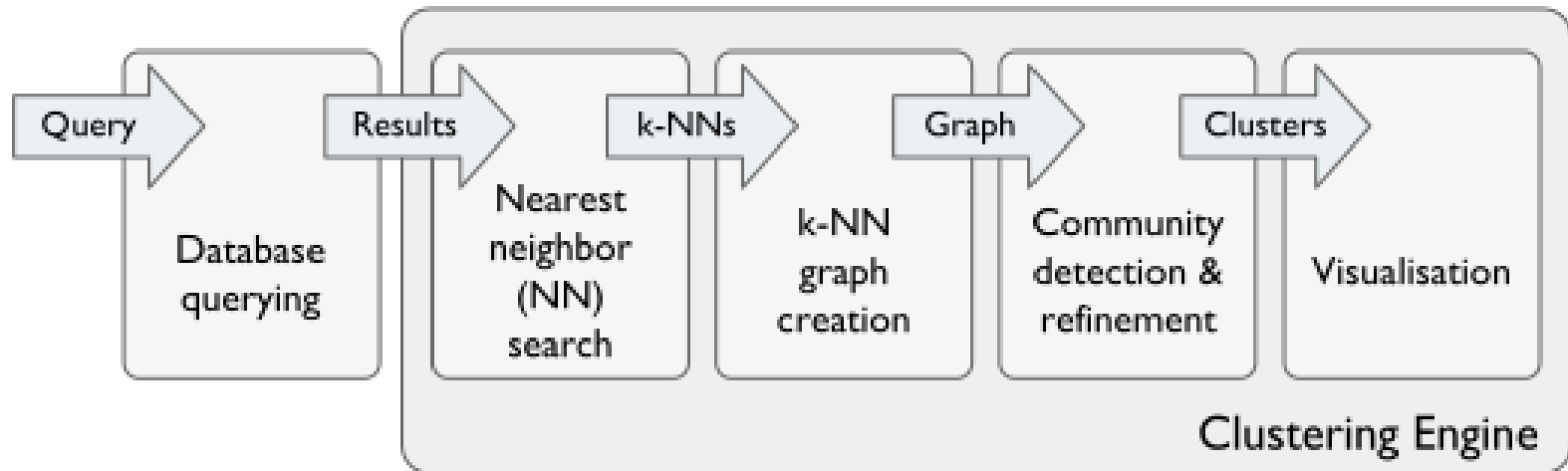
<https://freesound.org>



A screenshot of the Freesound website's search results page for the query "wind". The page has a light gray header with a search bar containing "wind", a dropdown menu set to "Automatic by relevance", and a "search" button. Below the header, there's a navigation bar with "previous" and "next" links, and a pagination bar showing "1 2 3 4 5 6 7 ... 355 | 9925 sounds". The main content area displays a list of search results. Each result includes a small audio waveform icon, the sound title, a star rating, the uploader's name, the upload date, the number of downloads, and the number of comments. The first result is "Soft Wind" by florianreichelt, uploaded on February 18th, 2019, with 5421 downloads and 10 comments. The second is "Wind Through Trees 3b" by spoonbender, uploaded on August 16th, 2014, with 36046 downloads and 52 comments. The third is "Wind blow, mouth.wav" by Wesselorg, uploaded on November 5th, 2017, with 11022 downloads and 15 comments. The fourth is "short wind noise" by michimuc2, uploaded on August 2nd, 2013, with 8889 downloads and 89 comments. To the right of the search results, there are two sections: "licenses" and "tags". The "licenses" section lists "Attribution (4815)", "Attribution Noncommercial (1260)", "Creative Commons 0 (3678)", and "Sampling+ (172)". The "tags" section lists various tags like "air", "ambience", "atmosphere", "background", "birds", "blowing", "calm", "city", "field-recording", "forest", "general-noise", "holland", "nature", "noise", "rain", "soundscape", "spring", "storm", "traffic", "trees", "water", "waves", "weather", "white-noise", "wind", and "windy".

# 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!**