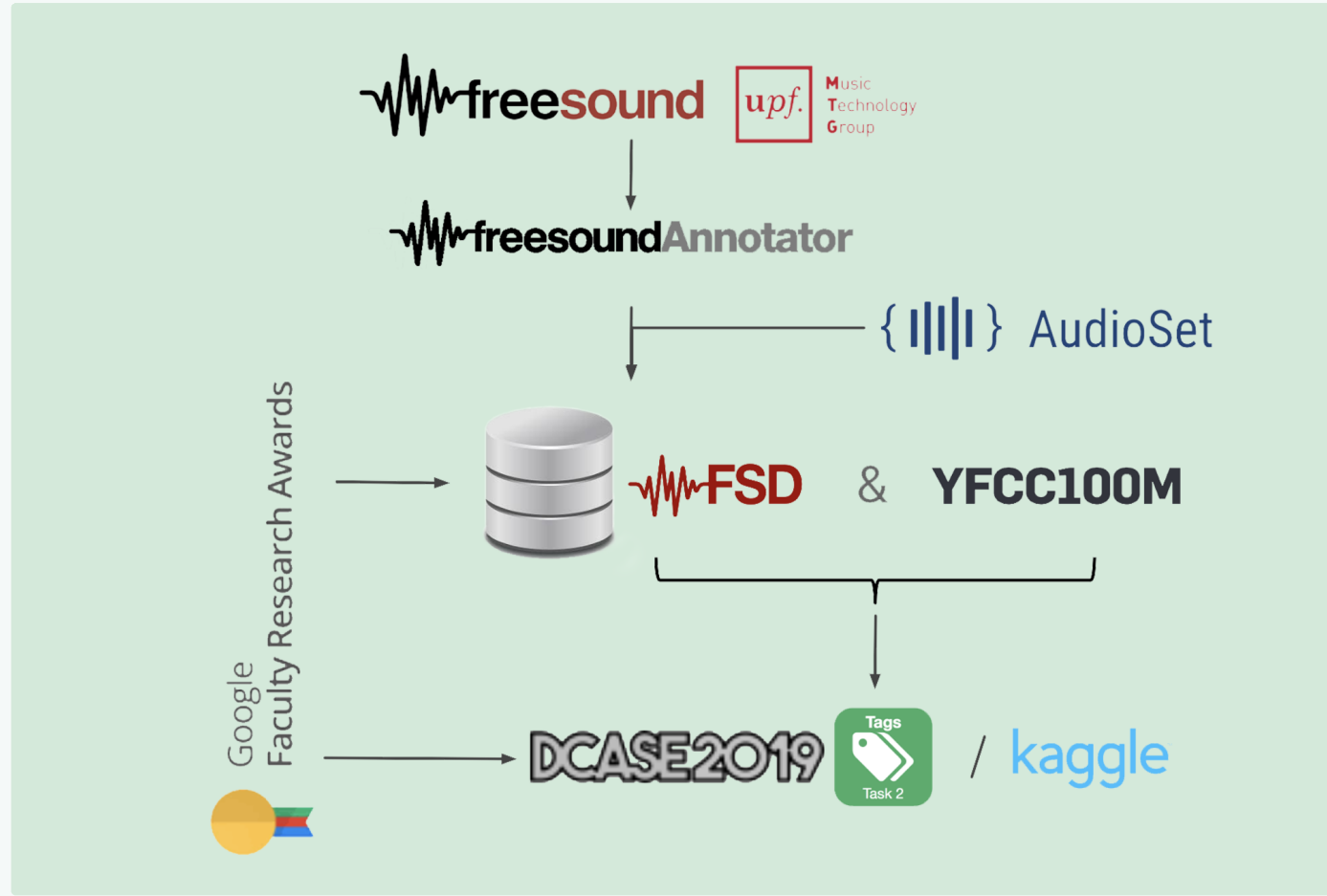


### T2 Audio tagging with noisy labels and minimal supervision

Coordinators: Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis  
 Results: <https://tinyurl.com/yjja>

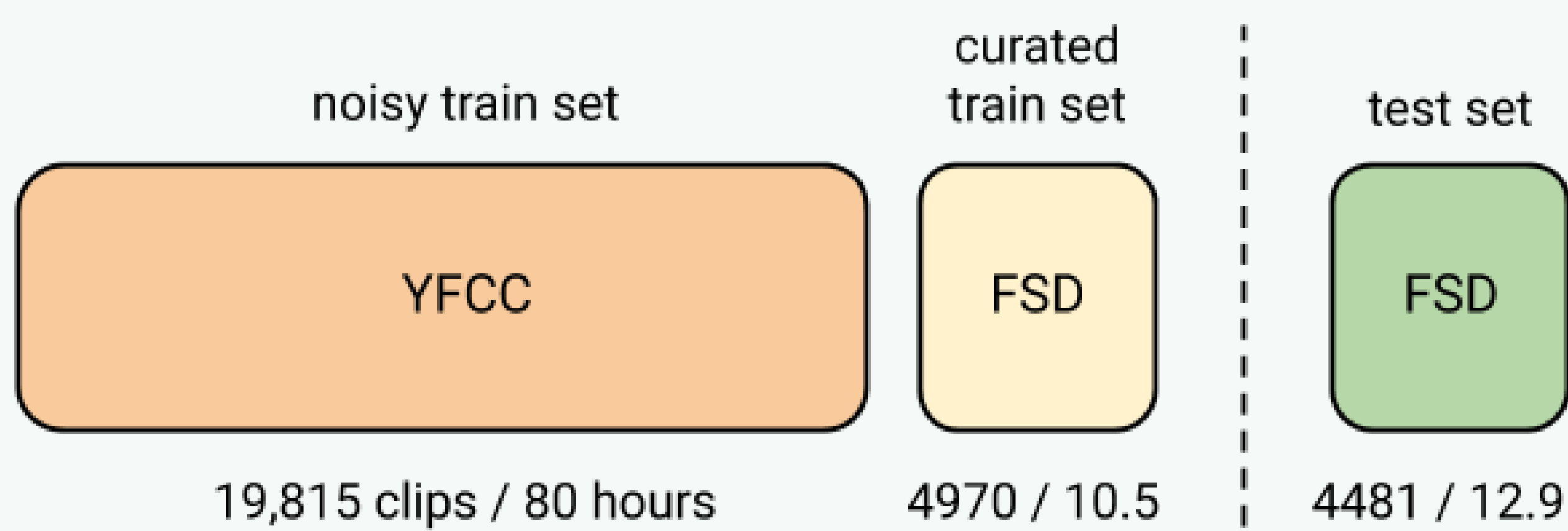
#### Motivation



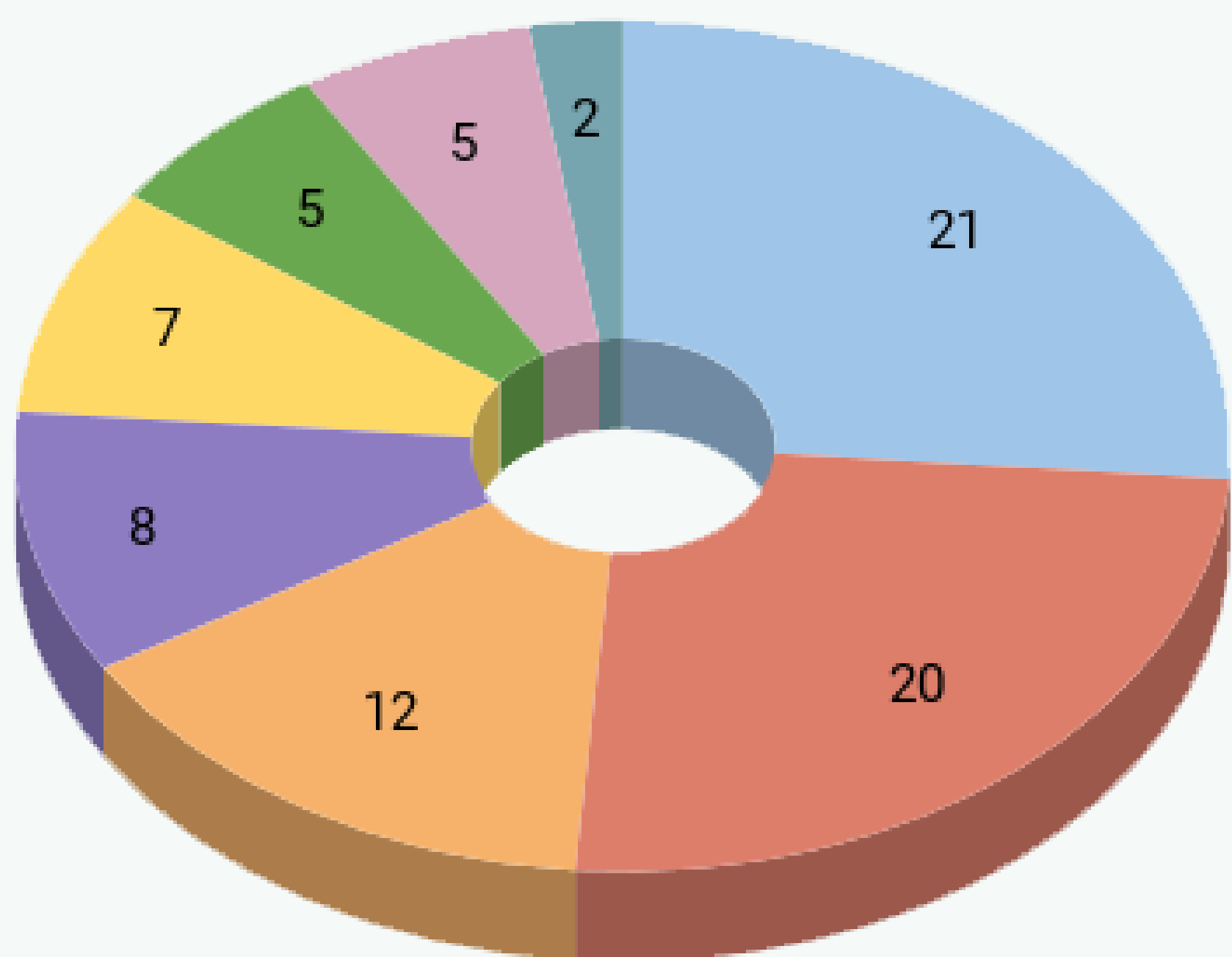
- General-purpose sound event recognizers
- follow-up of DCASE Task 1, but:
  - double number of categories
  - much more data
  - multi-class multi-label

#### Dataset: FSDKaggle2019

- classes, over hours
- labels/clip
- variable length s
- weak labels of varying reliability
- FSD: human-labeled
- YFCC: machine-labeled (noisy)

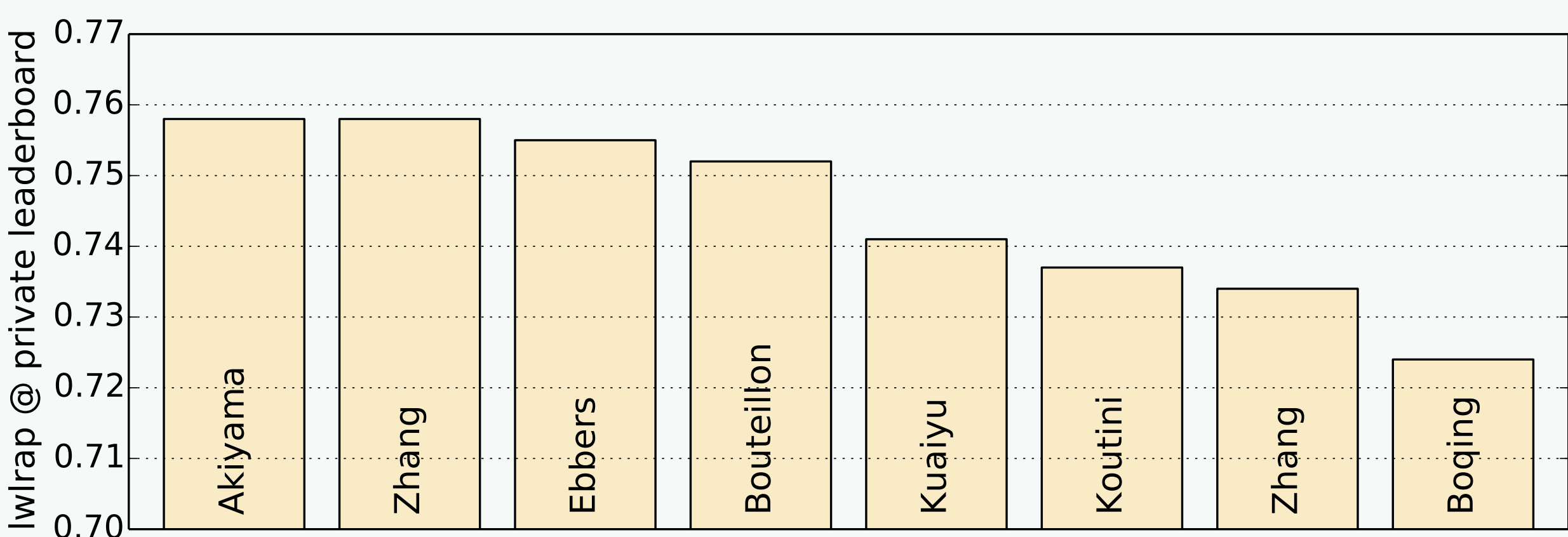


#### Class distribution



- Human sounds (e.g. speech, applause)
- Domestic sounds (e.g. microwave oven, toilet flush)
- Musical instrument (e.g. accordion, acoustic guitar)
- Vehicles (e.g. car passing by, motorcycle)
- Animal sounds (e.g. cat meow, dog bark)
- Natural sounds (e.g. fire crackle, raindrop)
- Materials (e.g. glass shatter, fill (with liquid))
- Mechanisms (printer, fan)

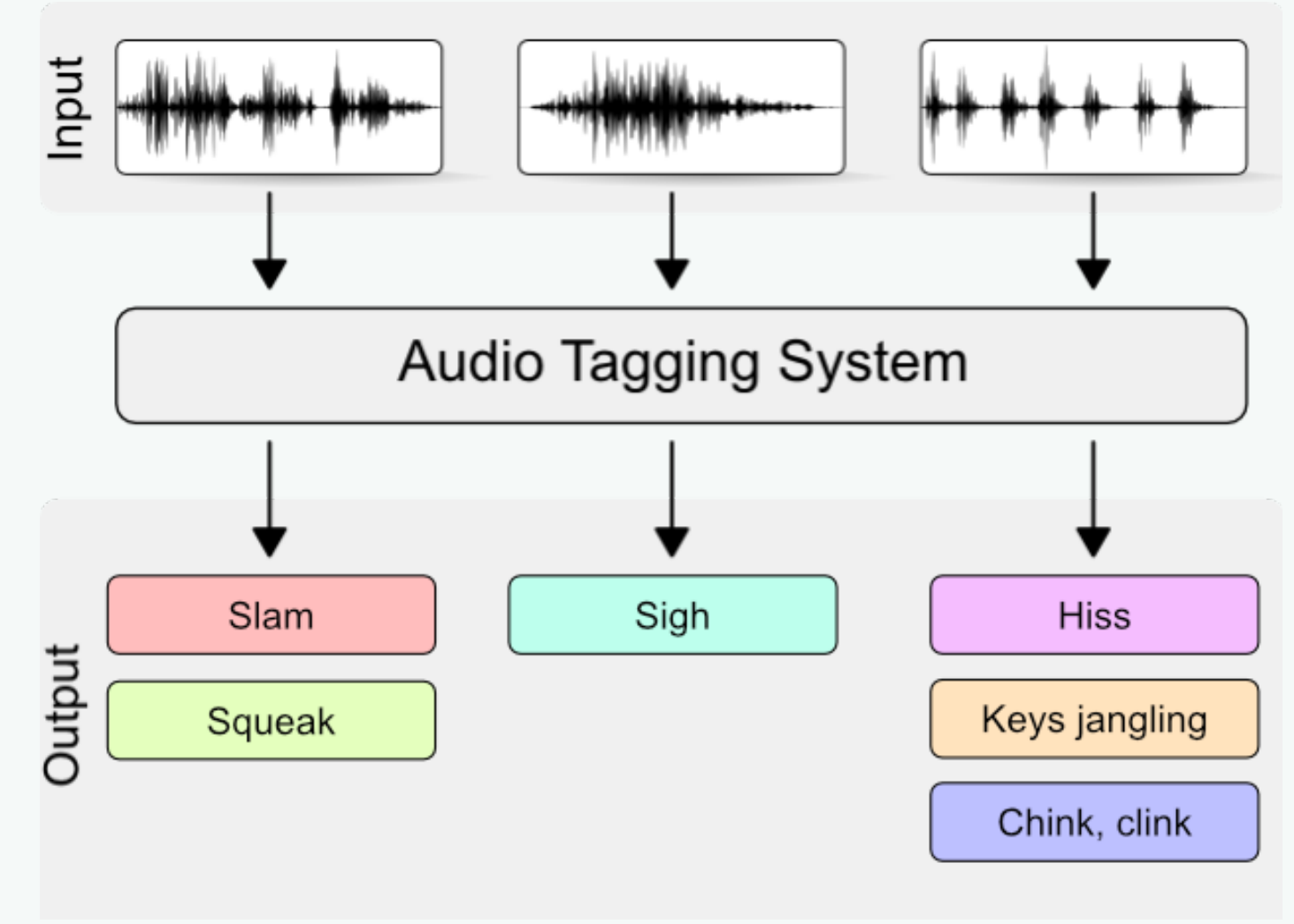
#### Results, Top-8



System	Features	Classifier	lwrap
Akiyama	log-mel energies, waveform	CNN, ensemble	.
Zhang	log-mel energies, CQT	CNN, RNN, ensemble	.
Ebbers	log-mel energies	CRNN, ensemble	.
Bouteillon	log-mel energies	CNN	.
Kuaiyu	log-mel energies	CNN, ensemble	.
Koutini	log-mel energies	CNN, RFR, ensemble	.
Zhang	log-mel energies, PCEN	CNN, ensemble	.
Boqing	log-mel energies	CNN, ensemble	.

#### Task Description

- Goal:** predict labels among diverse categories
- multi-label audio tagging
- many noisy labels & limited supervision
- domain mismatch
- Freesound - Flickr



#### Task Setup & Some Numbers

- Kaggle kernels-only competition: all systems run in Kaggle's servers with scores computed on a hidden test set
- teams ( of them submitting systems to DCASE) & entries
- maximum of daily submissions
- Judges Award to foster novel, problem-specific, efficient approaches
- top-winners & Judges Award winner are required to publish the code

#### Metric: label-weighted label-ranking average precision (lwrap)

For sample  $s$  & class label  $c$ :

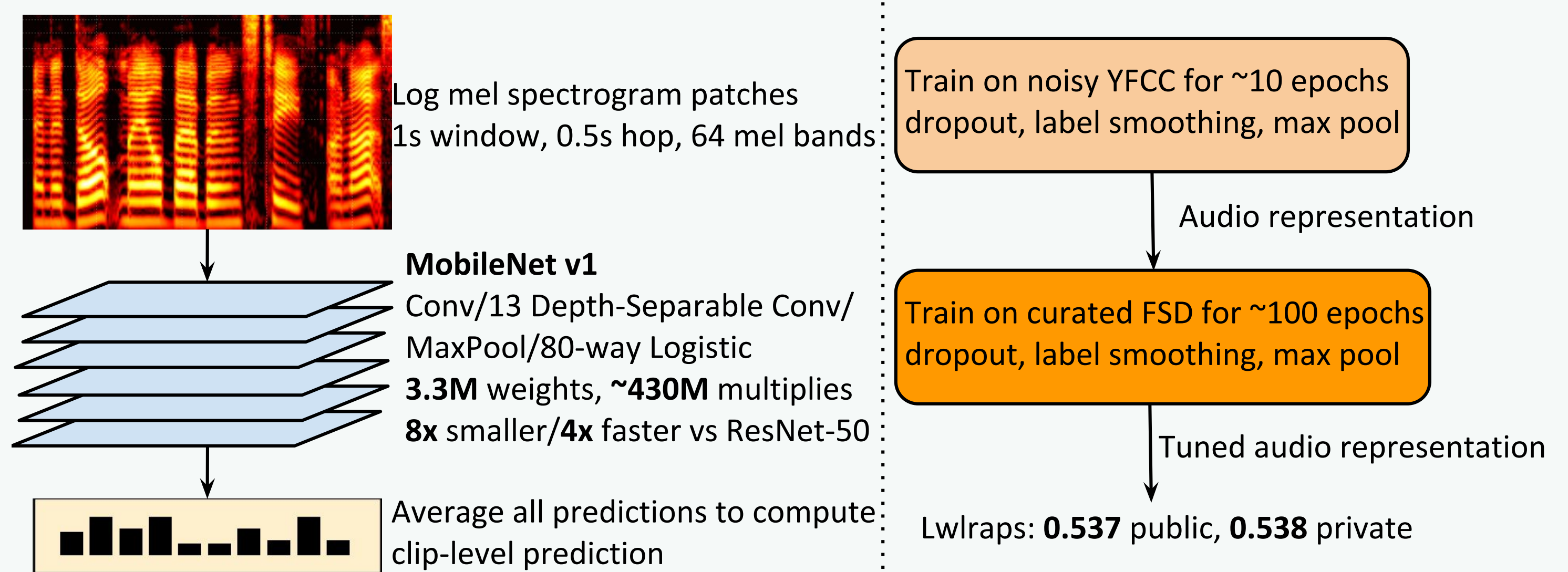
- $C(s)$ : set of true class labels
- $Lab(s, r)$ : class label at rank  $r$
- $Rank(s, c)$ : rank of class  $c$
- $[\cdot]$ : if argument true, else

$$Prec(s, c) = \frac{1}{Rank(s, c)} \sum_{r=1}^{Rank(s, c)} \mathbf{1}[Lab(s, r) \in C(s)]$$

$$lwrap = \frac{1}{\sum_s |C(s)|} \sum_s \sum_{c \in C(s)} Prec(s, c)$$

- $Prec(s, c)$ : label-ranking precision
- $lwrap$  = average  $Prec$  over all labels

#### Baseline System



#### Adopted techniques

- Log-mel energies**, waveform, CQT, ...
- Mainly **CNN/CRNN**: VGG, DenseNet, ResNe(X)t, Shake-Shake, Frequency-Aware CNNs, Squeeze-and-Excitation, MobileNet
- Heavy usage of **ensembles** ( ): several nets or snapshot learning. Aggregation of predictions or shallow meta-learning
- Exploiting **curated** train set: mixup, SpecAugment, SpecMix, TTA, ...
- Label noise**: semi-supervised learning, instance selection, multi-task learning, loss functions, per-class loss weighting, stochastic weight averaging, adaptive-weighting of noisy samples, MixMatch, ...

#### Open Knowledge: discussion forum & sharing kernels

