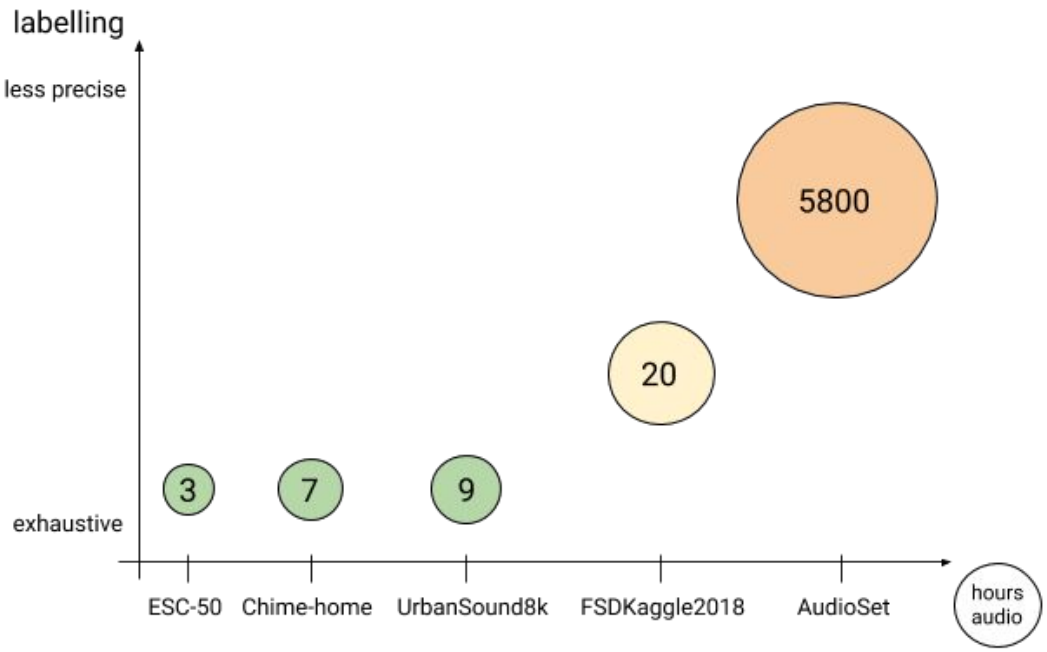
Learning Sound Event Classifiers from Web Audio with Noisy Labels

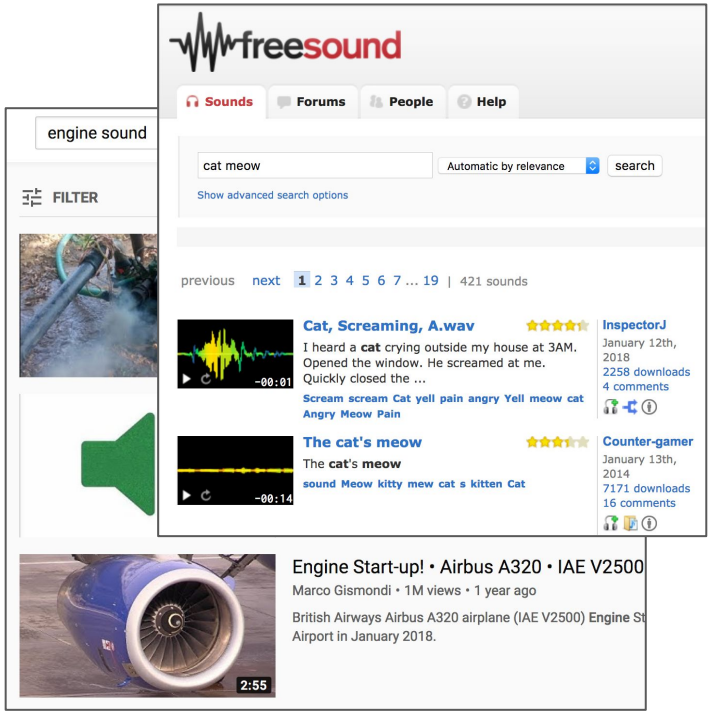
Eduardo Fonseca[1], Manoj Plakal[2], Daniel P. W. Ellis[2],
Frederic Font[1], Xavier Favory[1], and Xavier Serra[1]

1 Universitat Pompeu Fabra Barcelona
MTG Music Technology Group
2 Google

Label noise in sound event classification

- Labels that fail to properly represent acoustic content in audio clip
- Why is label noise relevant?

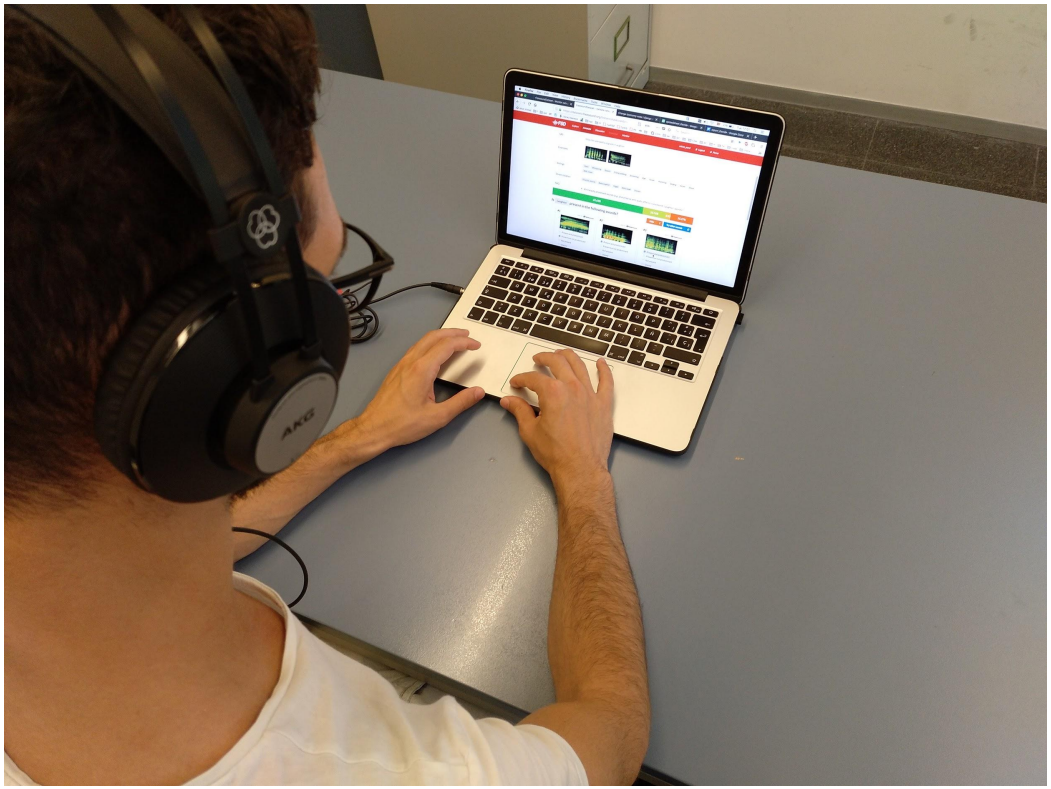




- Label noise effects: performance decrease / increased complexity

How to mitigate label noise?

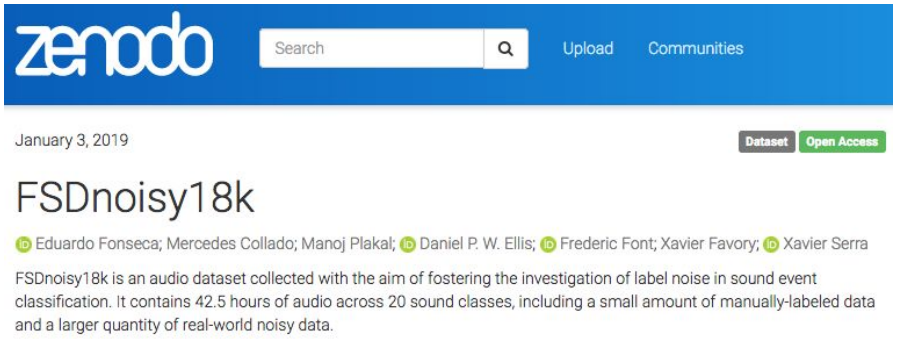
How to mitigate label noise?



How to mitigate label noise?

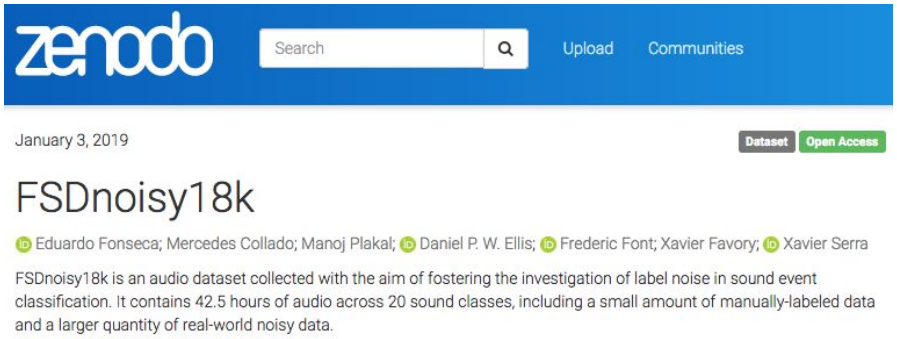
Our contributions

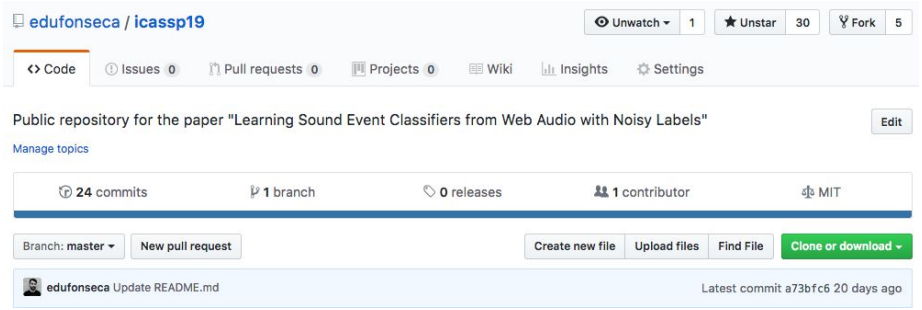
1. FSDnoisy18k: a dataset to foster label noise research



Our contributions

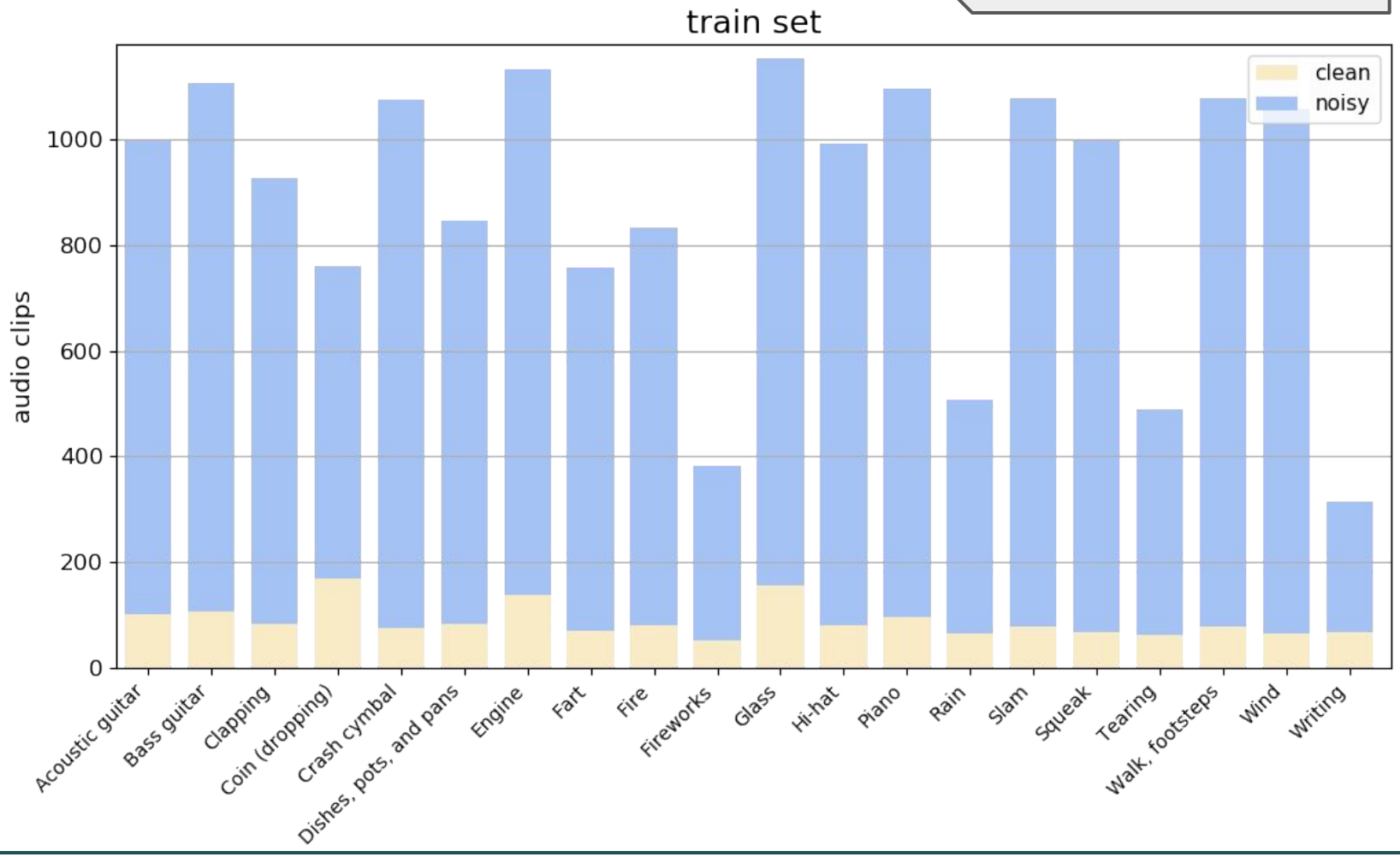
1. FSDnoisy18k: a dataset to foster label noise research
2. CNN baseline system
3. Evaluation of noise-robust loss functions





FSDnoisy18k

- 20 classes
- 18k audio clips
- 42.5 hours of audio

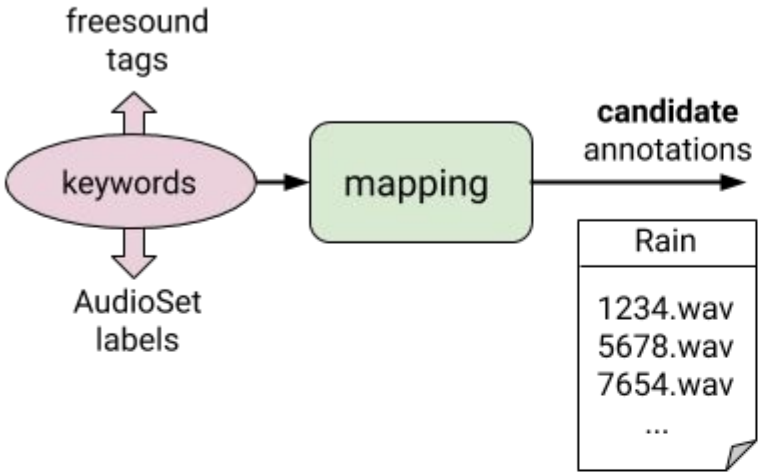


FSDnoisy18k: creation

- Freesound
  - → Audio content & metadata (tags)
- AudioSet Ontology
  - → 20 classes (labels)

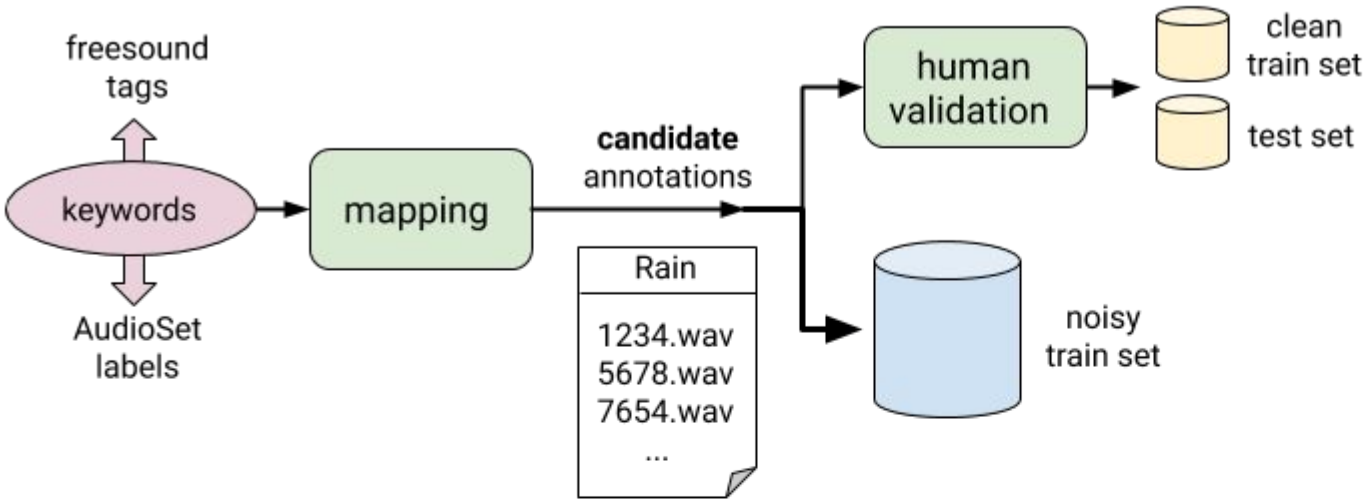
FSDnoisy18k: creation

- Freesound
  - → Audio content & metadata (tags)
- AudioSet Ontology
  - → 20 classes (labels)



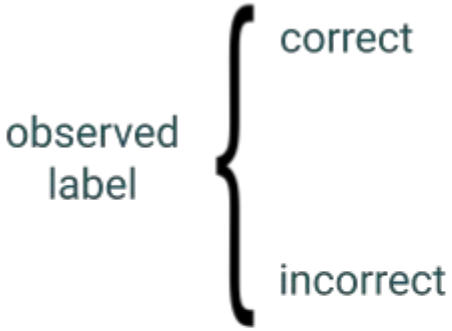
FSDnoisy18k: creation

- Freesound
  - → Audio content & metadata (tags)
- AudioSet Ontology
  - → 20 classes (labels)



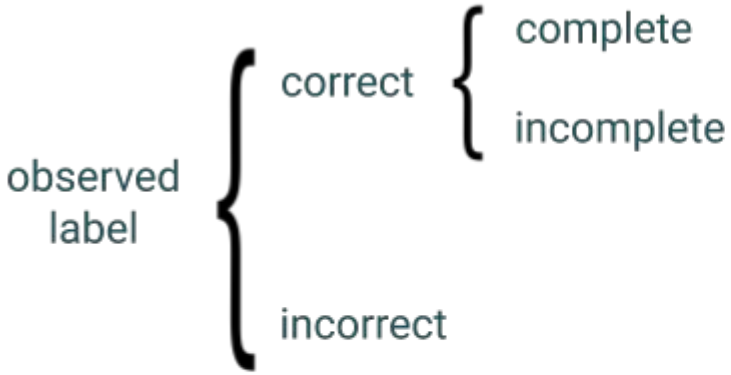
Types of label noise

- singly-labeled data



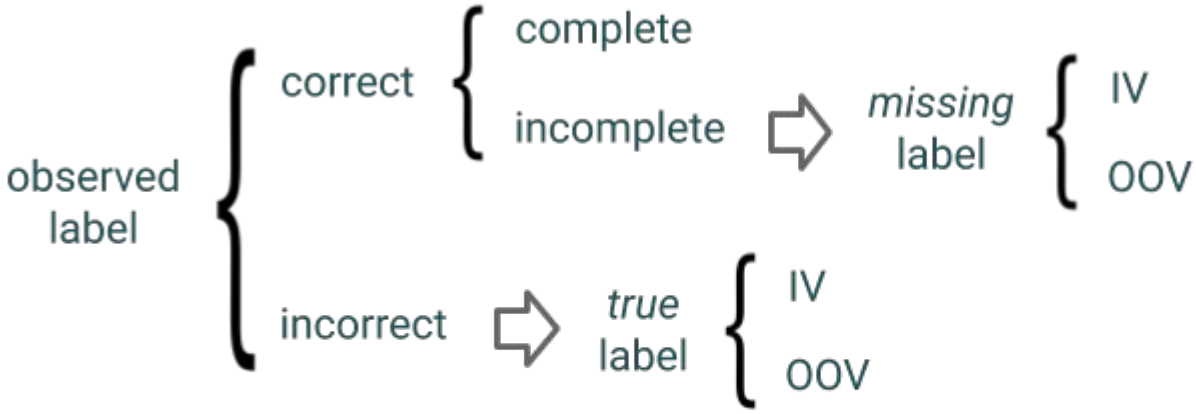
Types of label noise

- singly-labeled data



Types of label noise

- singly-labeled data



- in-vocabulary (IV): events that are part of our target class set (closed-set)
- out-of-vocabulary (OOV): events not covered by the class set (open-set)

Examples: clip #1

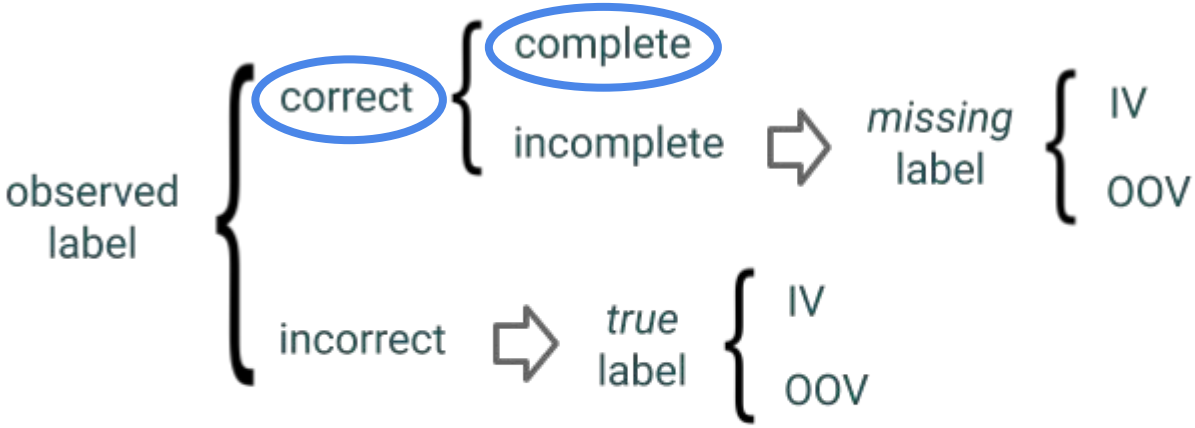
Observed label from the vocabulary:

Acoustic guitar / Bass guitar / Clapping / Coin (dropping) / Crash cymbal / Dishes, pots, and pans / Engine / Fart / Fire / Fireworks / Glass / Hi-hat / Piano / Rain / Slam / Squeak / Tearing / **Walk, footsteps** / Wind / Writing

Examples: clip #1

<u>True</u> label from the vocabulary:

Acoustic guitar / Bass guitar / Clapping / Coin (dropping) / Crash cymbal / Dishes, pots, and pans / Engine / Fart / Fire / Fireworks / Glass / Hi-hat / Piano / Rain / Slam / Squeak / Tearing / **Walk, footsteps** / Wind / Writing



Examples: clip #2

<u>Observed</u> label from the vocabulary:

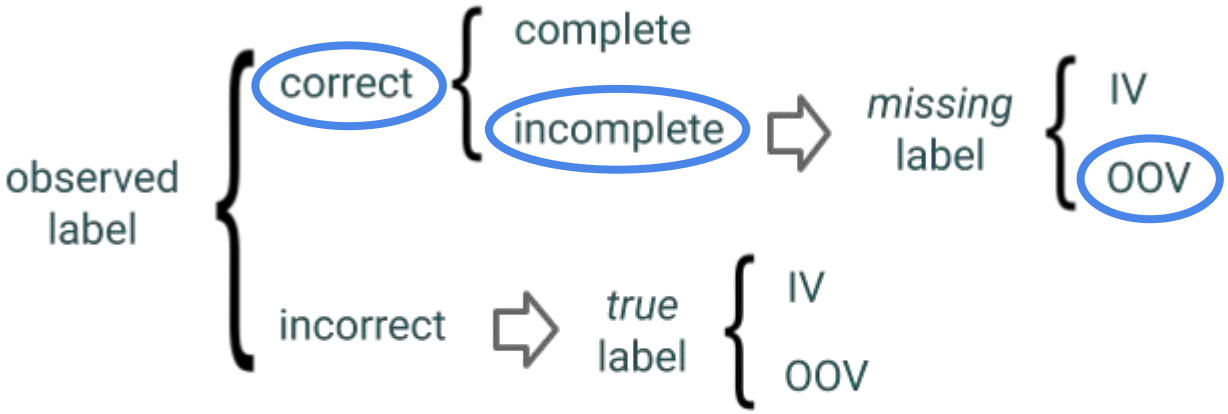
Acoustic guitar / Bass guitar / Clapping / Coin (dropping) / Crash cymbal / Dishes, pots, and pans / Engine / Fart / Fire / Fireworks / Glass / Hi-hat / Piano / Rain / Slam / Squeak / Tearing / **Walk, footsteps** / Wind / Writing

Examples: clip #2

True label from the vocabulary:

Acoustic guitar / Bass guitar / Clapping / Coin (dropping) / Crash cymbal / Dishes, pots, and pans / Engine / Fart / Fire / Fireworks / Glass / Hi-hat / Piano / Rain / Slam / Squeak / Tearing / **Walk, footsteps** / Wind / Writing

Missing labels: male speech / laughter / children shouting / chirp, tweet / chatter



Examples: clip #3

Observed label from the vocabulary

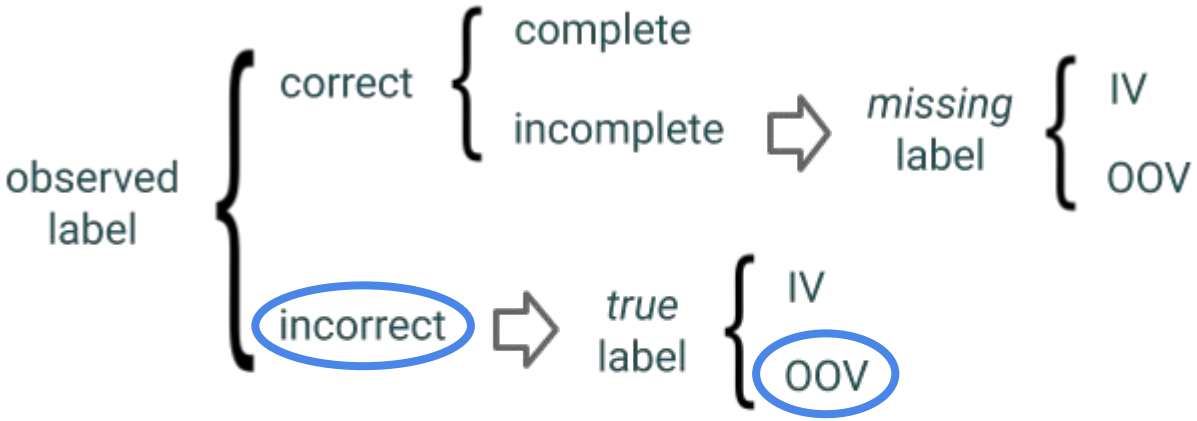
Acoustic guitar / Bass guitar / Clapping / Coin (dropping) / Crash cymbal / Dishes, pots, and pans / Engine / Fart / Fire / Fireworks / Glass / Hi-hat / Piano / Rain / Slam / Squeak / Tearing / **Walk, footsteps** / Wind / Writing

Examples: clip #3

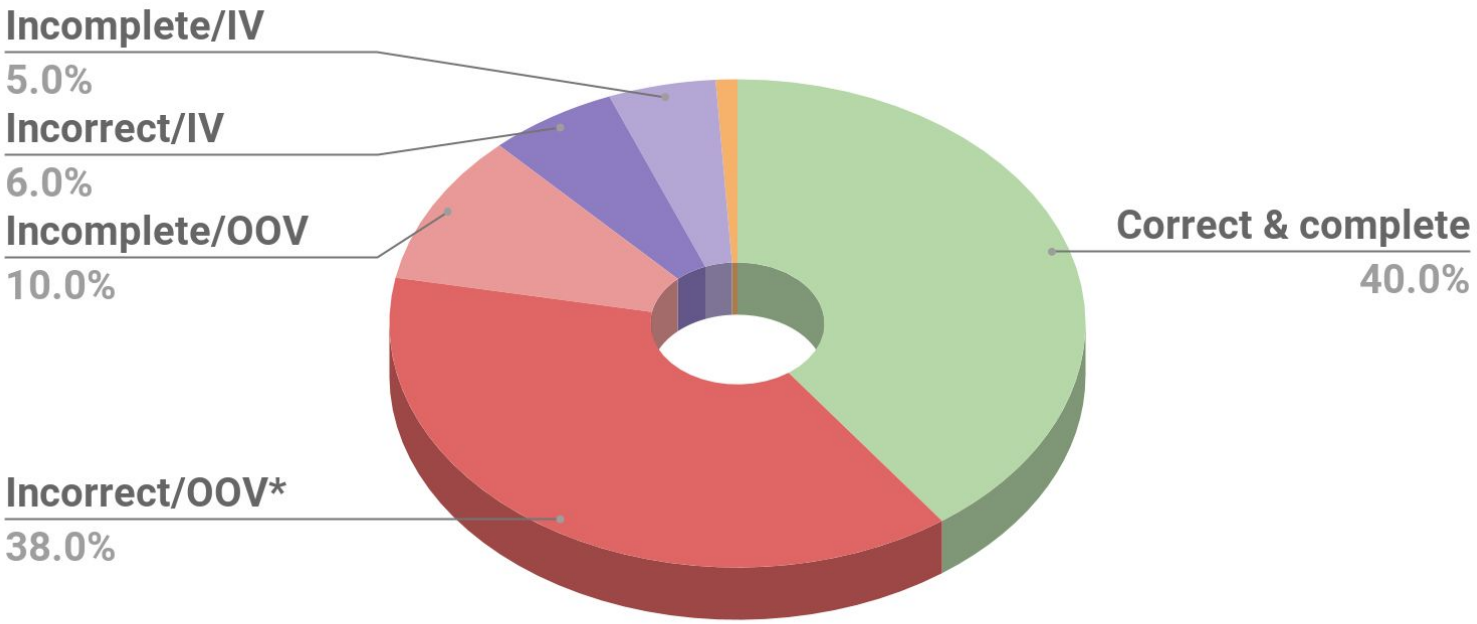
True label from the vocabulary:

Acoustic guitar / Bass guitar / Clapping / Coin (dropping) / Crash cymbal / Dishes, pots, and pans / Engine / Fart / Fire / Fireworks / Glass / Hi-hat / Piano / Rain / Slam / Squeak / Tearing / ~~Walk, footsteps~~ / Wind / Writing

True label: electronic music



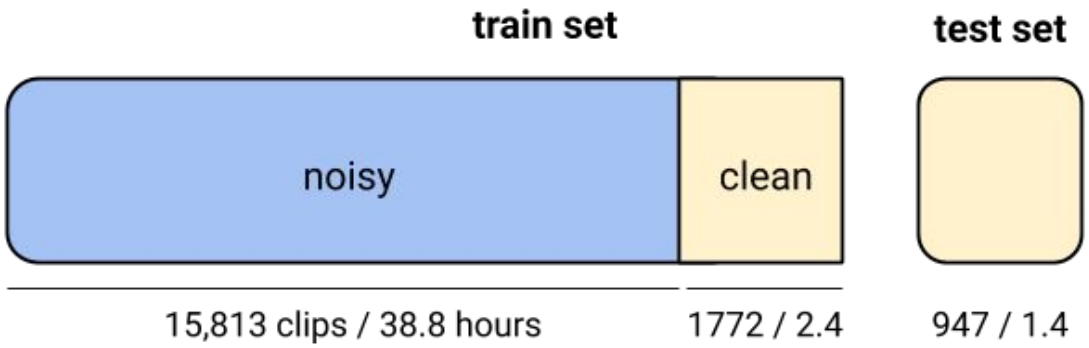
Label noise distribution in FSDnoisy18k



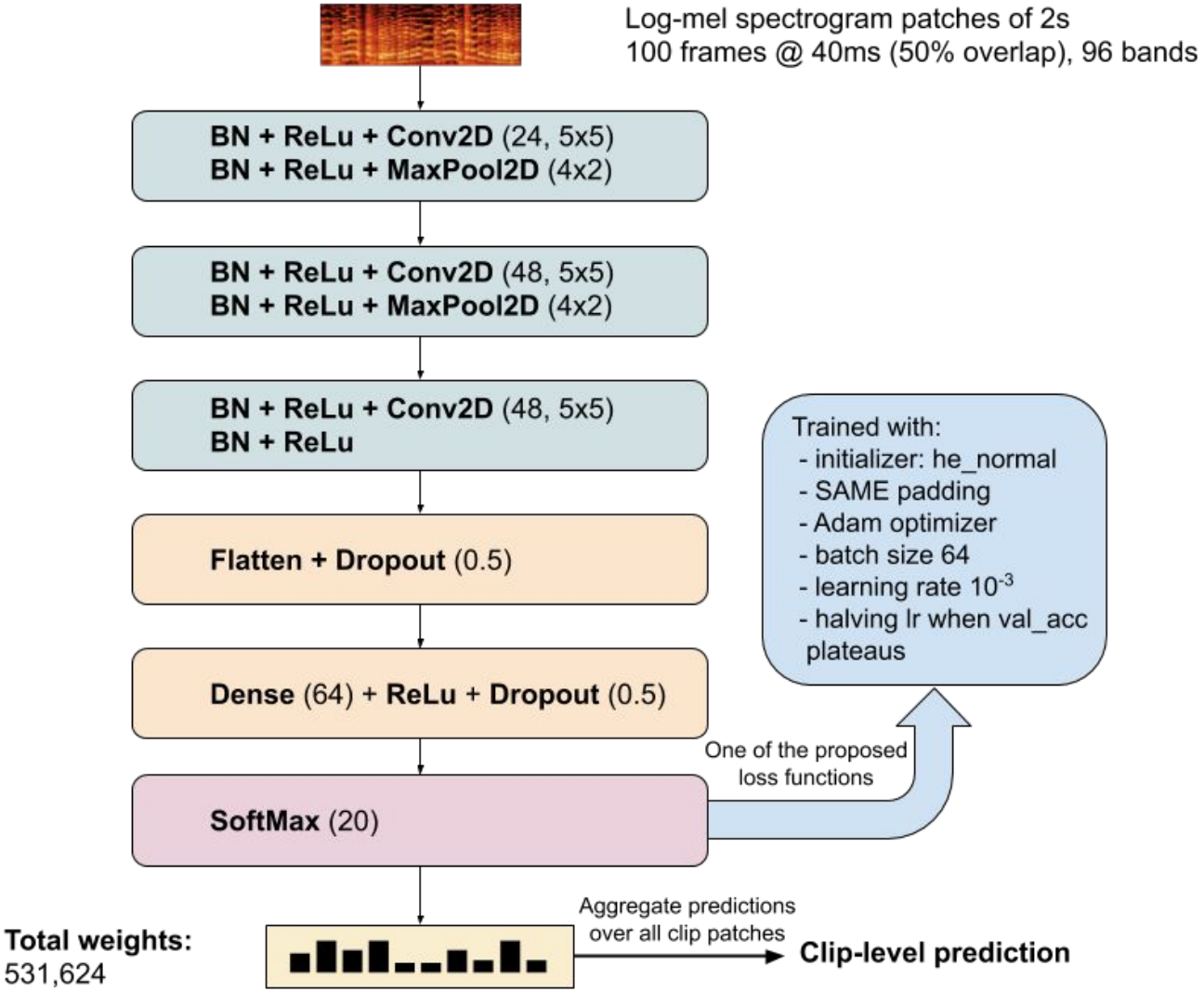
- most frequent types of label noise: OOV
- *some clips are incorrectly labeled, but still similar in terms of acoustics

FSDnoisy18k

- 20 classes / 18k clips / 42.5 h
- singly-labeled data
- variable clip duration: 300ms - 30s
- proportion train_noisy / train_clean = 90% / 10%
- per-class varying degree of types and amount of label noise
- expandable
- http://www.eduardofonseca.net/FSDnoisy18k/



CNN baseline system



Noise-robust loss functions

- Why?
  - → model-agnostic / minimal intervention / efficient

Noise-robust loss functions

- Why?
  - → model-agnostic / minimal intervention / efficient
- Default loss function in multi-class setting: Categorical Cross-Entropy (CCE)

$$\mathcal{L}_{cce} = -\sum_{k=1}^{K} y(k) \log(p(k))$$

predictions

target labels

Noise-robust loss functions

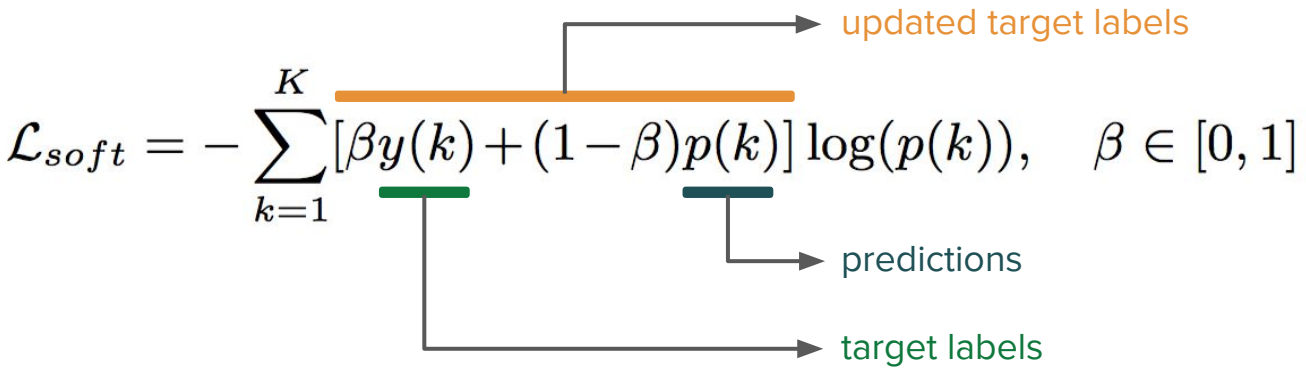
- Why?
  - → model-agnostic / minimal intervention / efficient
- Default loss function in multi-class setting: Categorical Cross-Entropy (CCE)

$$\mathcal{L}_{cce} = -\sum_{k=1}^{K} y(k) \log(p(k))$$

- CCE is sensitive to label noise: emphasis on *difficult* examples (weighting)
  - → beneficial for clean data
  - → detrimental for noisy data

Noise-robust loss functions

- Soft bootstrapping
  - → dynamically update target labels based on model's current state
  - → updated target label: convex combination



Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, Andrew Rabinovich, ***Training Deep Neural Networks on Noisy Labels with Bootstrapping***. In ICLR 2015

Noise-robust loss functions

- $\mathcal{L}_q$ loss intuition
  - CCE: sensitive to noisy labels (weighting)
  - Mean Absolute Error (MAE): $\mathcal{L}_{mae} = \sum_{k=1}^{K} |y(k) - p(k)|$
    - avoid weighting
    - difficult convergence

Zhilu Zhang and Mert Sabuncu, ***Generalized cross entropy loss for training deep neural networks with noisy labels***. In NeurIPS 2018

Noise-robust loss functions

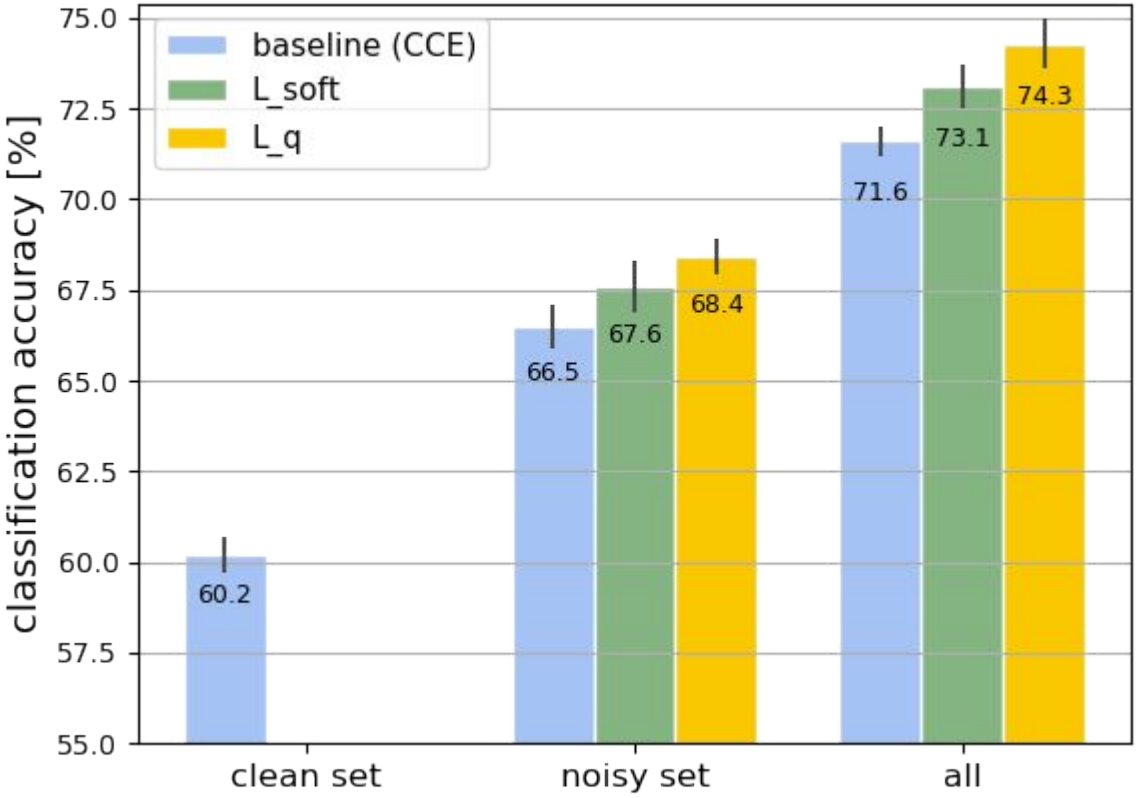
- $\mathcal{L}$q loss intuition
  - CCE: sensitive to noisy labels (weighting)
  - Mean Absolute Error (MAE): $\mathcal{L}_{mae} = \sum_{k=1}^{K} |y(k) - p(k)|$
    - avoid weighting
    - difficult convergence
- $\mathcal{L}$q loss is a generalization of CCE and MAE:
  - negative Box-Cox transformation of softmax predictions

$$\mathcal{L}_q = \frac{1 - (\sum_{k=1}^{K} y(k)p(k))^q}{q}, \quad q \in (0, 1]$$

  - $q = 1 \rightarrow \mathcal{L}$q = MAE ; $q \rightarrow 0 \rightarrow \mathcal{L}$q = CCE

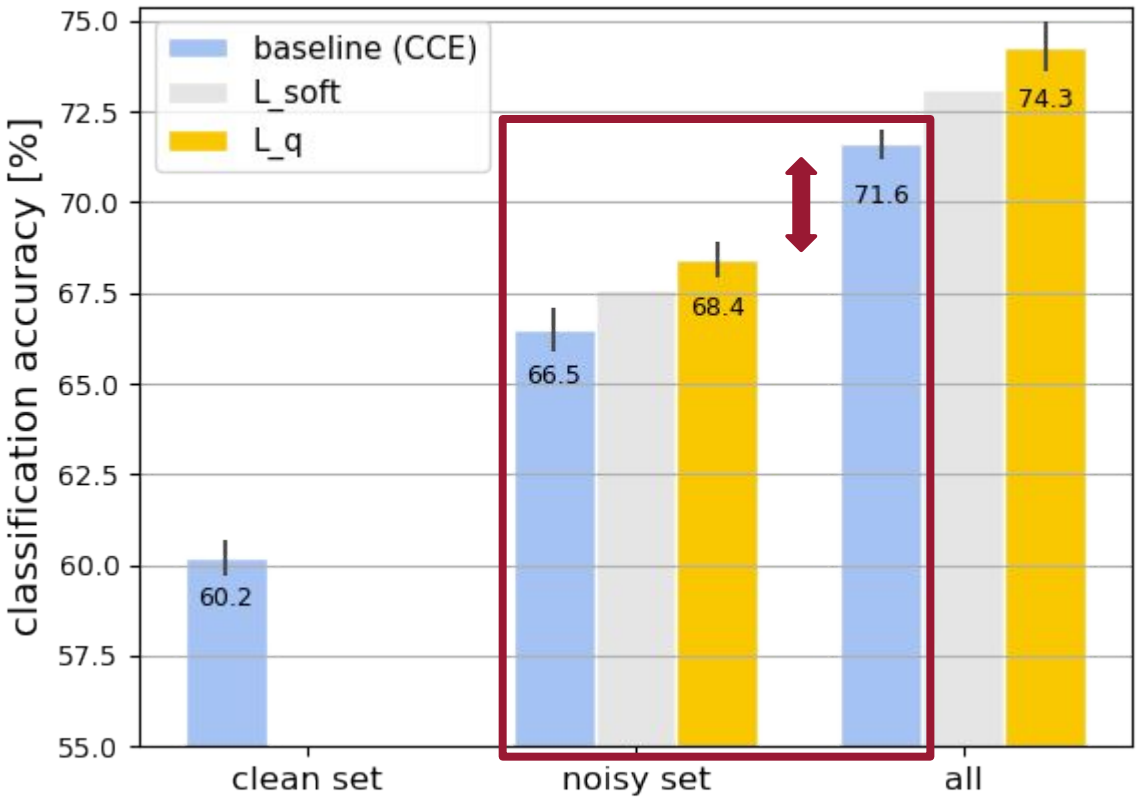
Zhilu Zhang and Mert Sabuncu, ***Generalized cross entropy loss for training deep neural networks with noisy labels***. In NeurIPS 2018

Experiments



- supervision by user-provided tags can be useful for sound event classification
- $\mathcal{L}_q$ works well for sound classification tasks with OOV (and some IV) noises

Experiments



- boost by using $\mathcal{L}_q$ on noisy set: 1.9% (**little** engineering effort)
- boost by adding curated data to noisy set: 5.1% (**significant** manual effort)

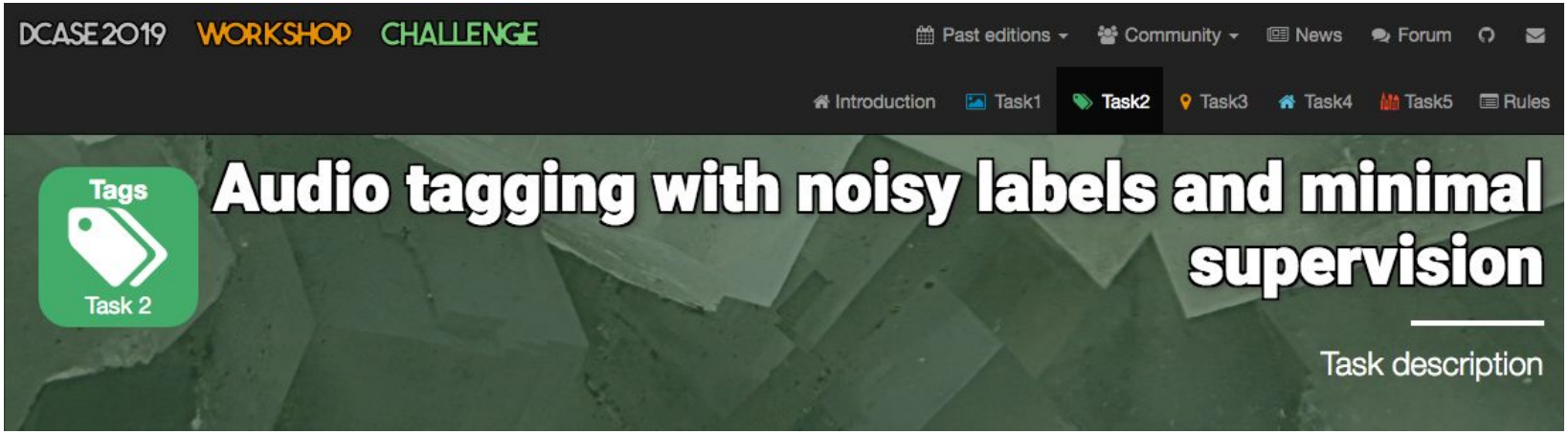
Summary & takeaways

- **FSDnoisy18k**
  - → open dataset for investigation of label noise
  - → 20 classes / 18k clips / 42.5 h / singly-labeled data
  - → small amount of manually-labelled data and a large amount of noisy data
  - → label noise characterization

- **CNN baseline system**
  - → large amount of Freesound audio & tags feasible for training sound recognizers

- **Noise-robust loss functions**
  - → efficient way to improve performance in presence of noisy labels
  - → $\mathscr{L}_q$ is top-performing loss

If you are interested in label noise...



Learning Sound Event Classifiers from Web Audio with Noisy Labels

Thank you!

http://www.eduardofonseca.net/FSDnoisy18k/

https://zenodo.org/record/2529934

https://github.com/edufonseca/icassp19

Eduardo Fonseca[1], Manoj Plakal[2], Daniel P. W. Ellis[2], Frederic Font[1], Xavier Favory[1], and Xavier Serra[1]

1 Universitat Pompeu Fabra Barcelona — MTG Music Technology Group

2 Google

Why this vocabulary?

- data availability
- classes “suitable” for the study of label noise
  - classes described with tags also used for other audio materials
    - Bass guitar, Crash cymbal, Engine, …
  - field-recordings: several sound sources expected
    - only the most predominant(s) tagged: Rain, Fireworks, Slam, Fire, …
  - pairs of related classes:
    - Squeak & Slam / Wind & Rain