Detection and Classification of Acoustic Scenes and Events

ICASSP 2019 Tutorial

Tuomas Virtanen
Professor

Annamaria Mesaros
Assistant Professor

Toni Heittola
Researcher

AUDIO RESEARCH GROUP

Audio Research Group
Tampere University

http://arg.cs.tut.fi/

Outline

Session 1: Machine learning approach

14:00 - 15:20

- Problem definition, motivation, applications
- General machine learning approach
- Sound classification with Python
- Task specific processing
- Datasets, evaluation, reproducible research
- Questions & answers

Session 2: Advanced methods

15:50 - 17:00

- Sound event detection with Python
- Real-life challenges and solutions
- Future perspectives
- Summary
- Questions & answers

Machine learning approach

Session 1

Outline

Introduction

General machine learning approach

Sound Classification with Python

Task specific processing

Datasets, evaluation, reproducible research

Questions & answers

Introduction

Information in everyday soundscapes

1. Entire scene
    - Birthday party, busy street, home, etc.

⇒ **Acoustic scene classification**

2. Individual sources
    - Car, beep, dog barking, etc.

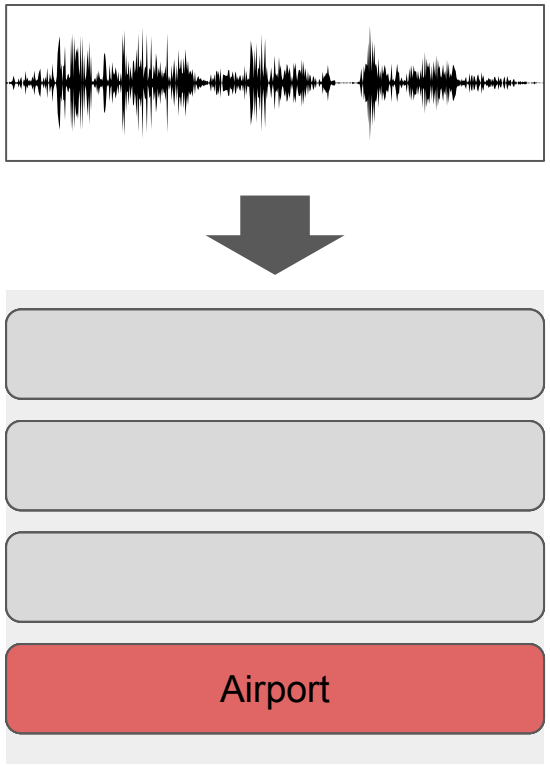
⇒ **Sound event detection**

Acoustic scene classification

- A whole acoustic scene is characterized with **one label**
- Example scene labels:

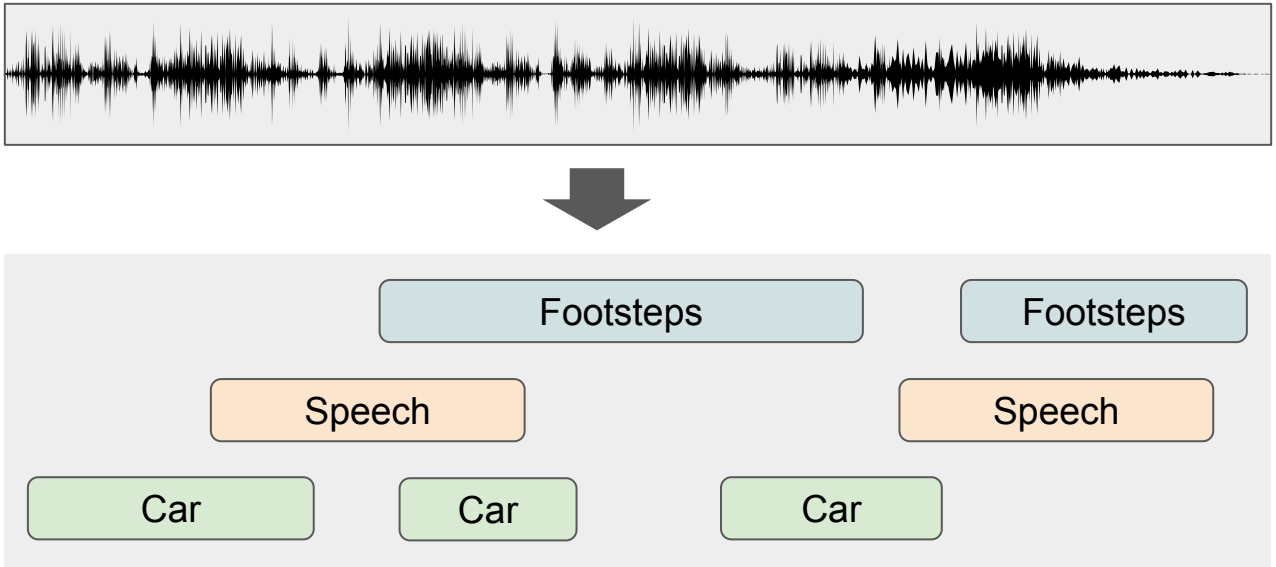
Airport
Indoor shopping mall
Metro station
Pedestrian street
Public square
Street with medium level of traffic
In tram
In bus
In metro
Urban park
Cafe
Restaurant
In car

City center
Forest path
Grocery store
Home
Lakeside beach
Library
Metro station
Office
Residential area
In train
Busy street
Open air market
Quiet street



Sound event detection

- Estimating start and end times of target sound class(es) ⇒ **Detection**
- Possible to have multiple classes to be detected, which can be overlapping



Example sound event labels

Baby crying
Glass breaking
Gunshot
Train horn
Air horn
Car alarm
Reversing beeps
Ambulance siren
Police car siren
Civil defense siren
Screaming
Bicycle
Skateboard
Car passing by
Bus
Truck
Motorcycle

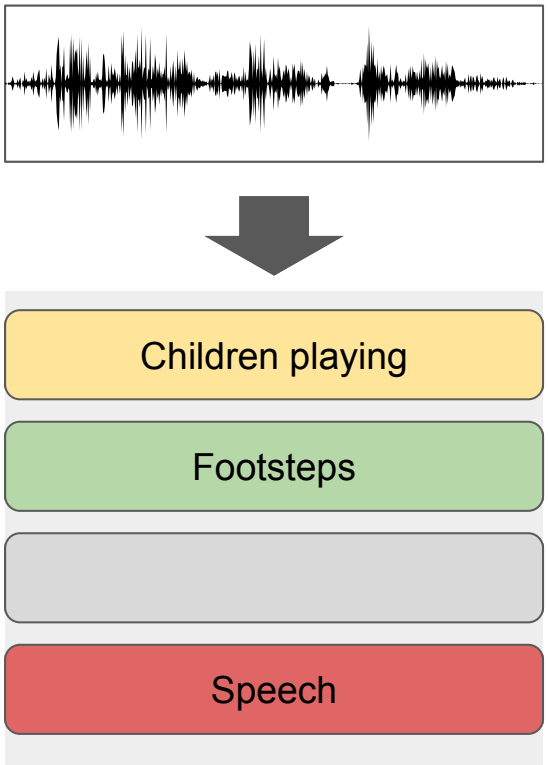
Train
Speech
Dog
Cat
Alarm/bell/ringing
Dishes
Frying
Blender
Running water
Vacuum cleaner
Electric
Shaver/toothbrush
Tearing
Shatter
Gunshot, gunfire
Fireworks
Writing

Computer keyboard
Scissors
Microwave oven
Keys jangling
Drawer open or close
Squeak
Knock
Telephone
Saxophone
Oboe
Flute
Clarinet
Acoustic guitar
Tambourine
Glockenspiel
Gong
Snare drum

Bass drum
Hi-hat
Electric piano
Harmonica
Trumpet
Violin, fiddle
Double bass
Cello
Chime
Cough
Laughter
Applause
Finger snapping
Fart
Burping, eructation
Bark
Meow

Tagging / weak labels

- No temporal information
- Multilabel classification: multiple classes can be active simultaneously



Applications

- Context-aware devices
- Acoustic monitoring
- Assistive technologies

Applications: Context aware devices

- Examples: hearing aids, smartphones, other devices changing the processing mode depending on context
- Autonomous cars, robots, etc. reacting to events in an environment

Applications: Acoustic monitoring

Examples: baby cry monitoring, window breakage, dog barking monitoring, bird sound detection, incident detection in tunnels, machine condition monitoring, environmental noise monitoring etc.

Applications: Assistive technologies

Example: automatic captioning of acoustic events in videos, multimedia information retrieval

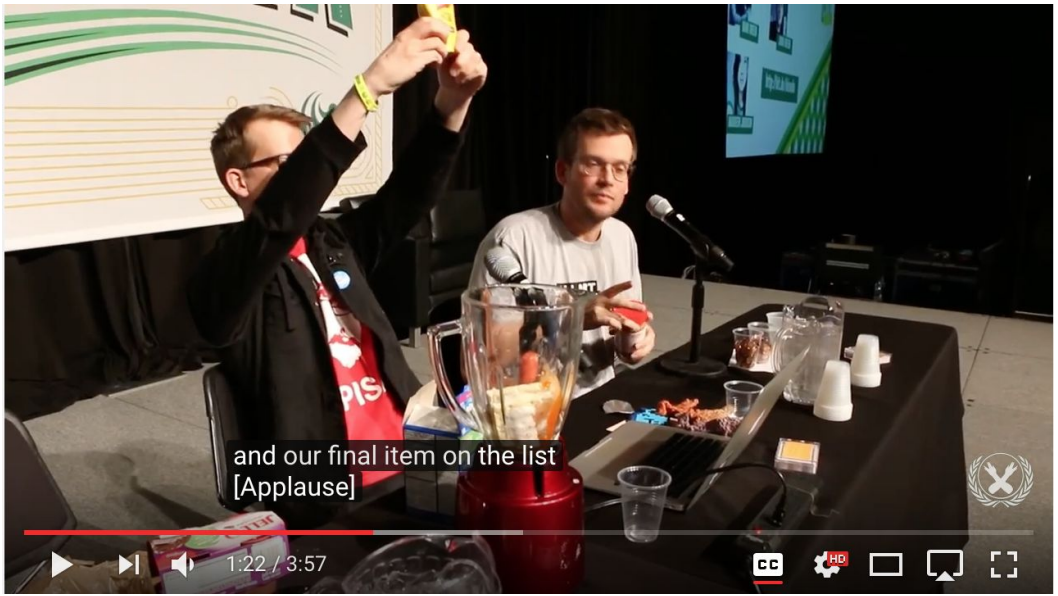


Photo from a video of vlogbrothers / CC BY

Comparison to other audio processing fields

- Speech analysis and recognition
- Music information retrieval

Similarities

- Acoustic properties
  - Harmonic, transient, noise-like sounds
  - Additive sources, convolutive mixing
- Similar acoustic features can be used
  - E.g. Spectral features, log-mel energies
- Classification tools
  - CNNs, FNNs, RNNs, GMMs, HMMs, etc.

Differences (1/2)

- No established taxonomy of events and scenes
  - Each application has different target scene and event classes

- In typical applications target sounds far away from microphone
  - Transfer function from source to microphone
  - Low SNR because of other competing sources

$$x(t) = \sum_{n} s_n(t) \star h_n(t)$$

Differences (2/2)

- Environmental sounds in general have less structure in comparison to speech and music
  - Many independent sources
  - Sources with many different types of acoustic characteristics

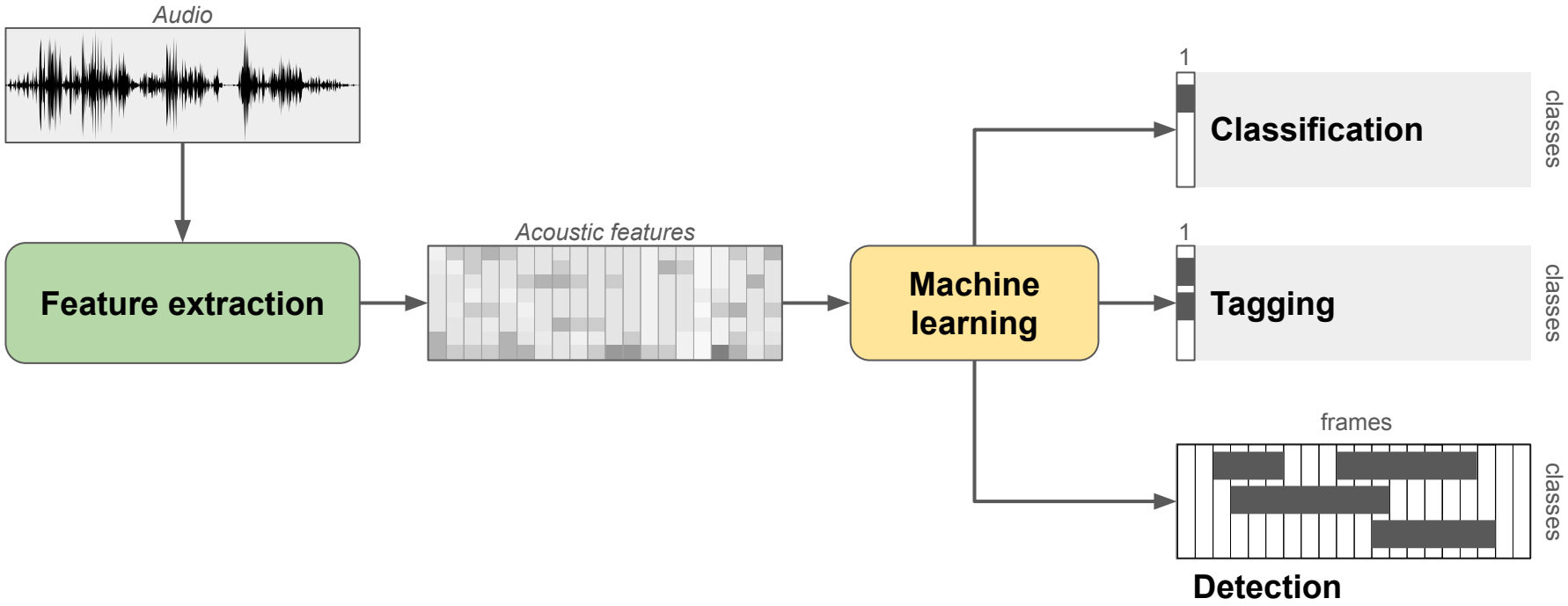
- Available datasets still smaller in comparison to speech and music datasets

General machine learning approach

General machine learning approach

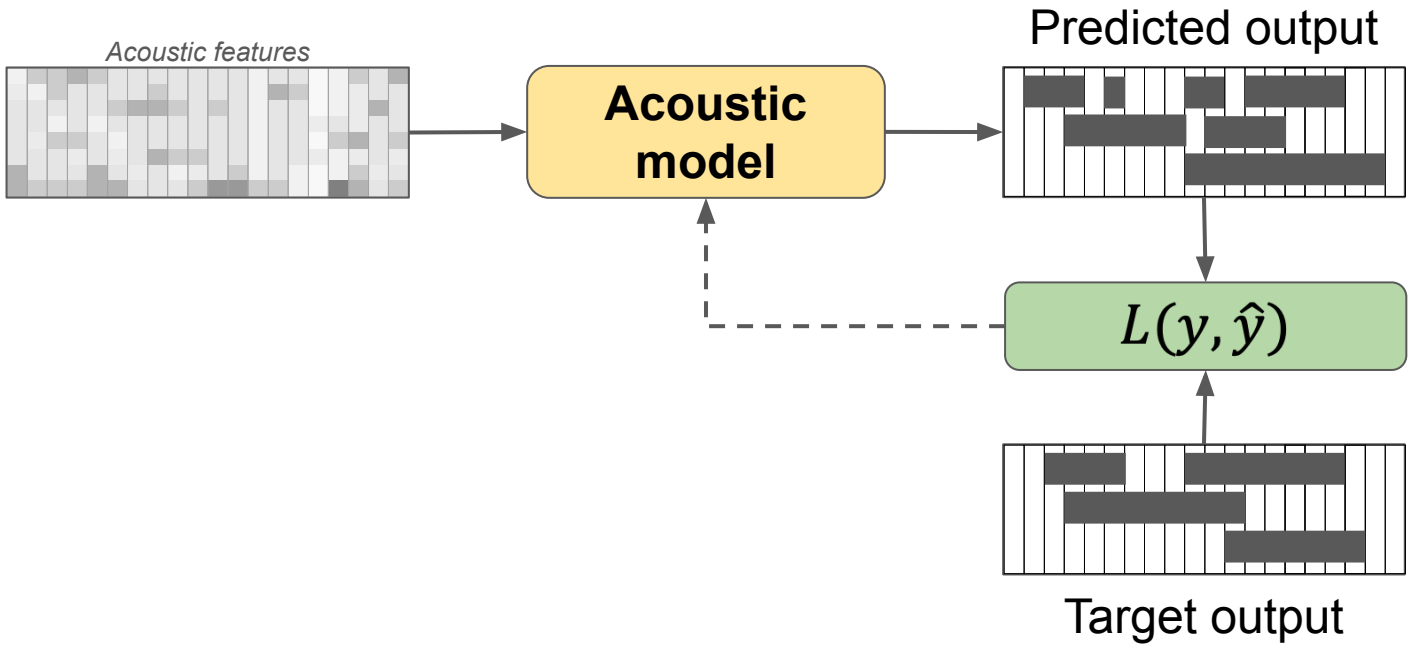
- Based on supervised learning
- Set of possible sound classes defined in advance
- Need for annotated training material from all the classes
  - Audio recordings and its class annotations
- Algorithms that find mapping between training examples (audio) and labels (annotations)

General machine learning approach

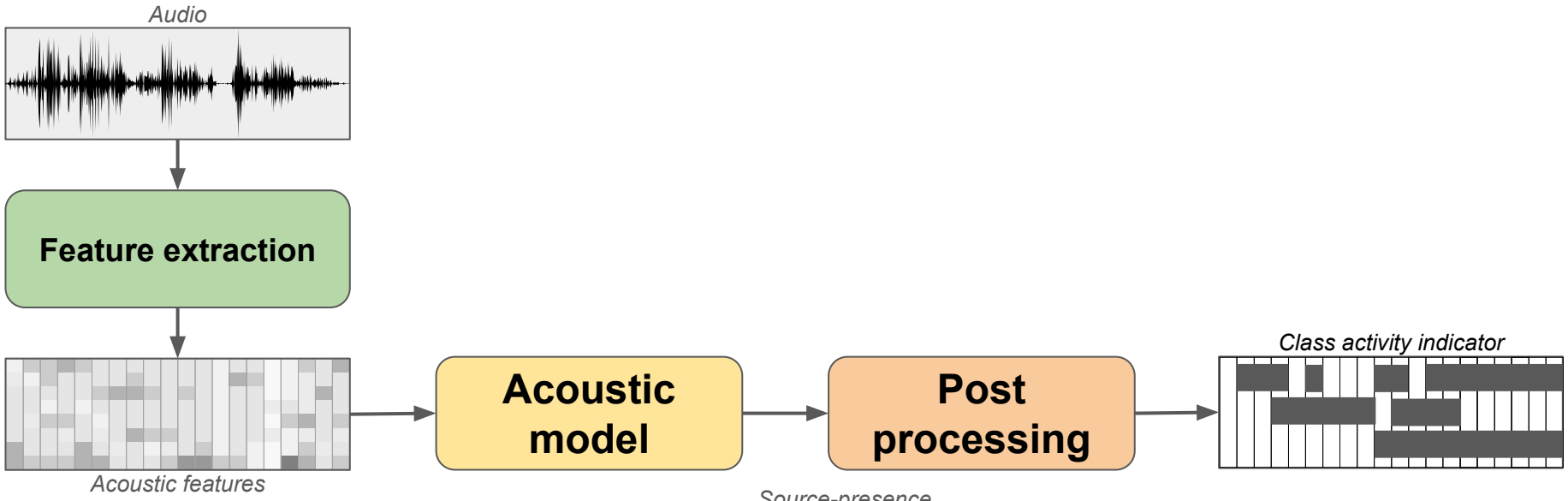


Training stage

Optimize acoustic model parameters to minimize a loss between predicted vs. target output

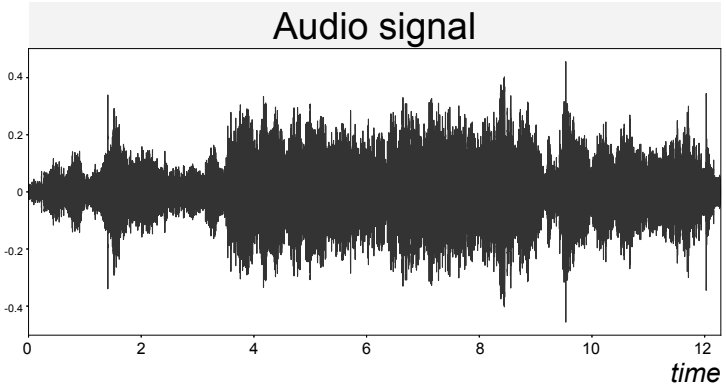


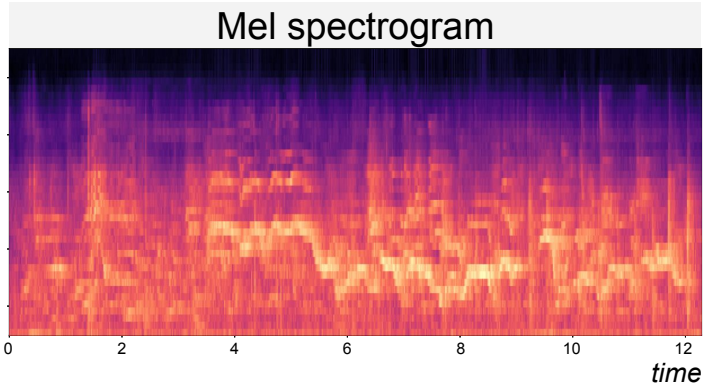
Test stage


Audio
Feature extraction
Acoustic features
Acoustic model
Source-presence probabilities
Post processing
Class activity indicator

Acoustic features

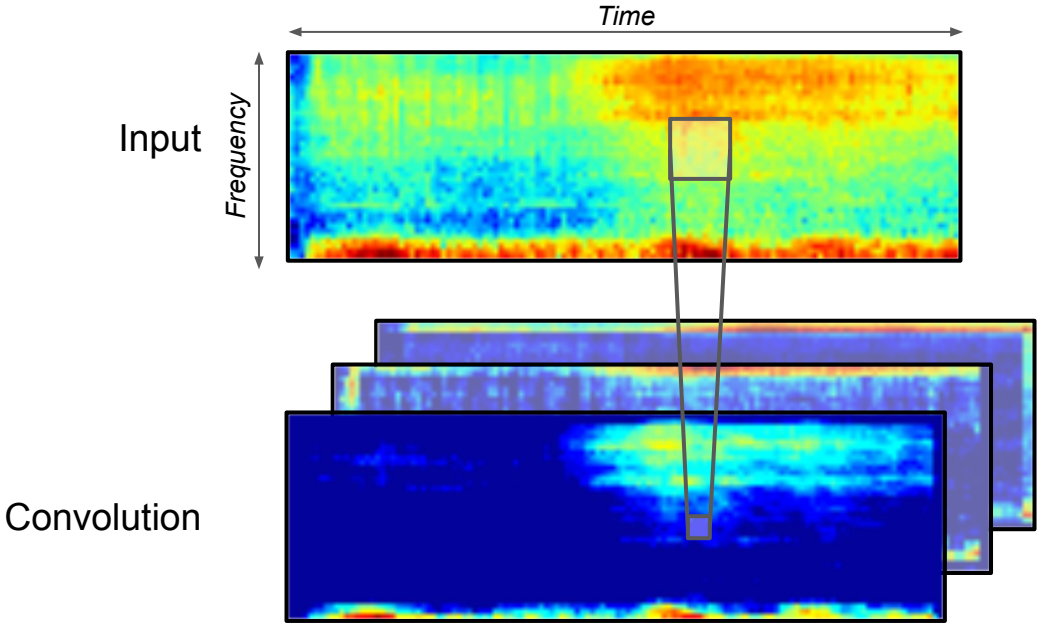
- Signals typically represented in the spectral domain
- Mel spectrogram (log of energies in mel bands) a commonly used representation
- Can use machine learning to extract more high-level features





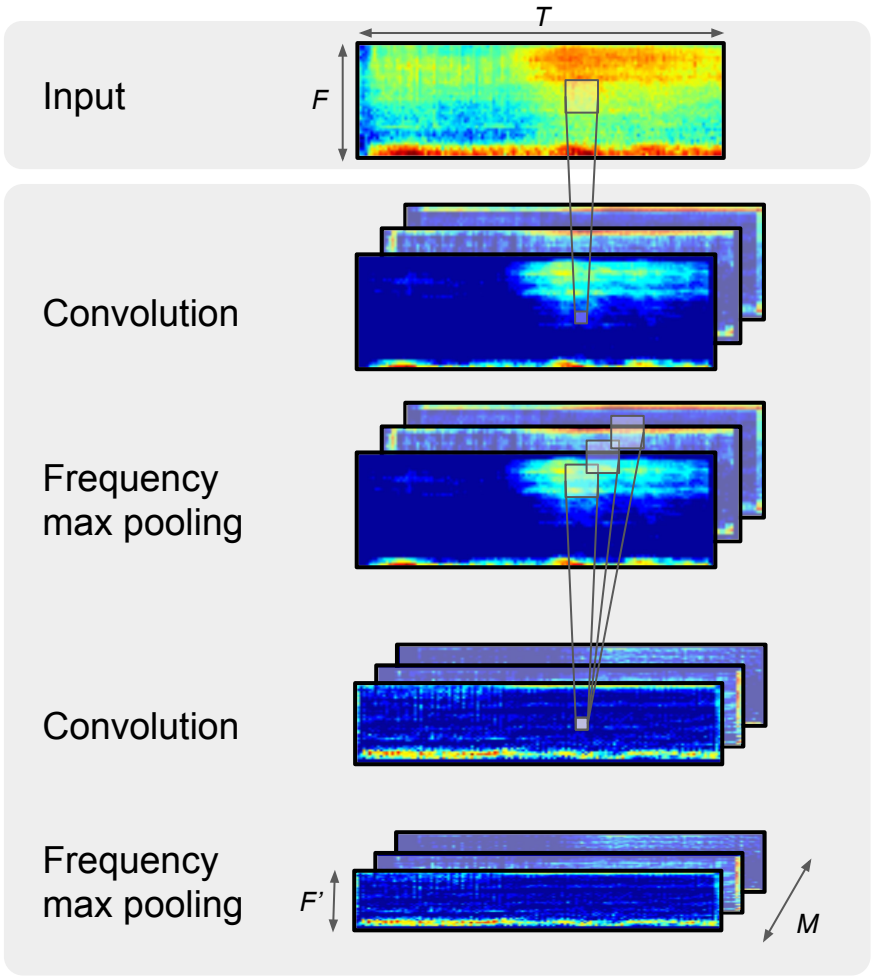
Convolutional neural networks

Layers of convolutions allow learning time-frequency filters to automatically find relevant representations

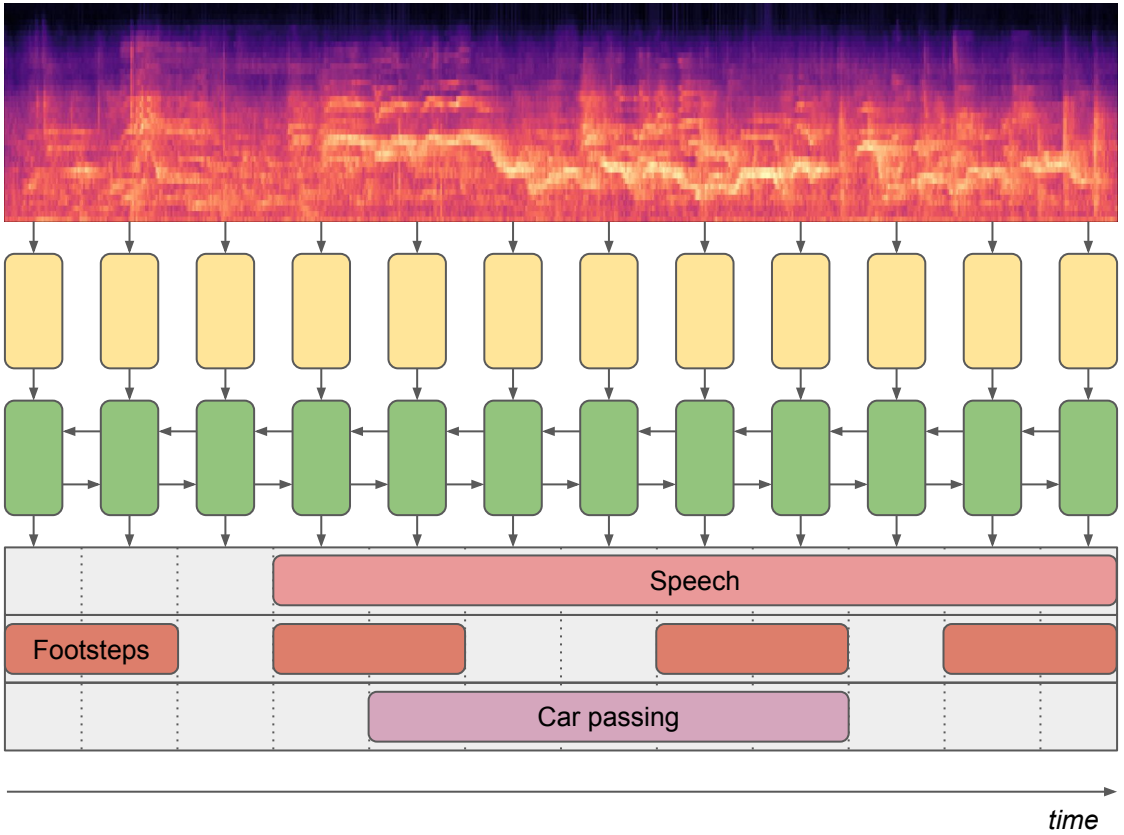


CNN

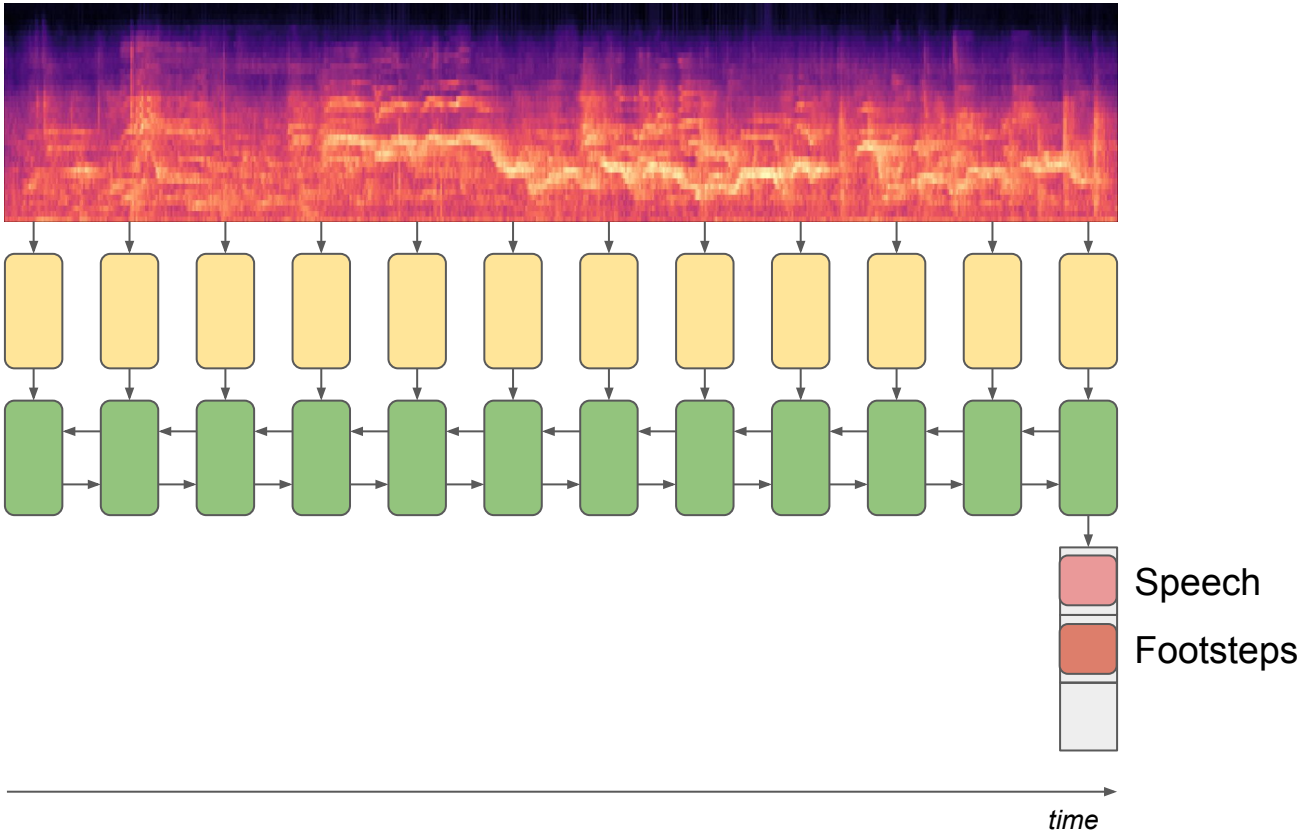
- Pooling allows learning shift-invariant features
- Multiple CNN layers allows learning higher-level features



Recurrent neural networks: sequence to sequence

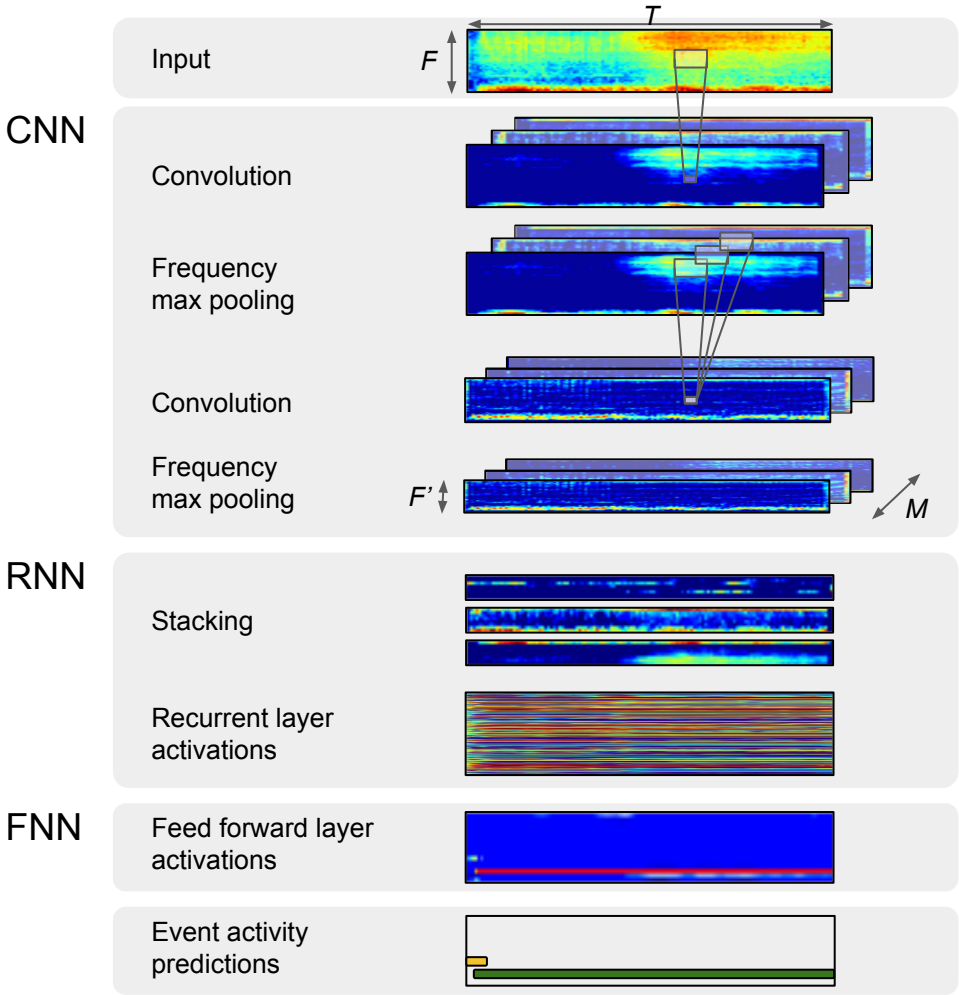


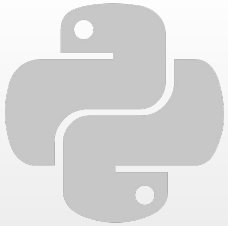
Recurrent neural networks: sequence to vector



End-to-end learning

- Possible to combine different processing units, e.g. CNNs and RNNs
- The whole network is optimized simultaneously
- Example: convolutional recurrent neural network





Sound Classification with Python

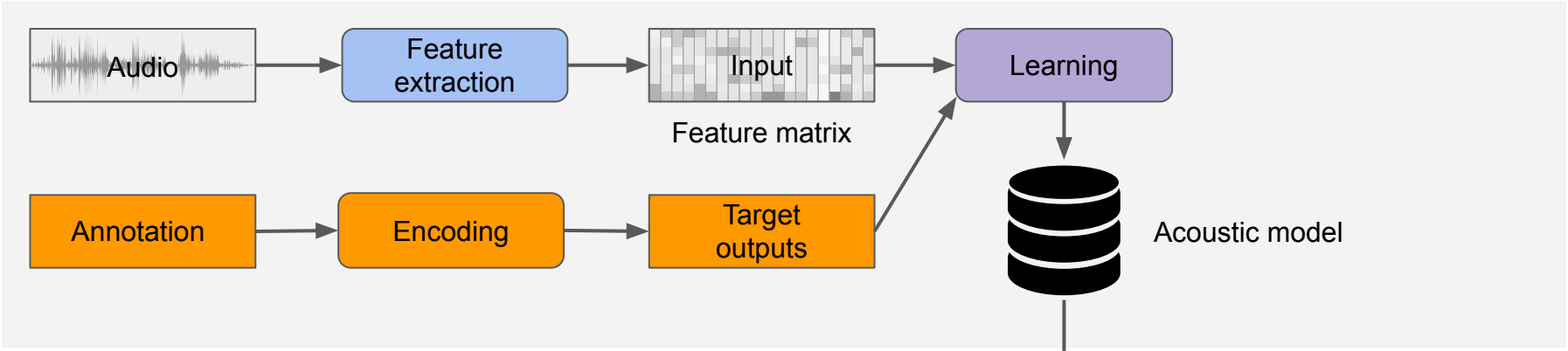
Jupyter notebooks:

https://github.com/toni-heittola/icassp2019-tutorial

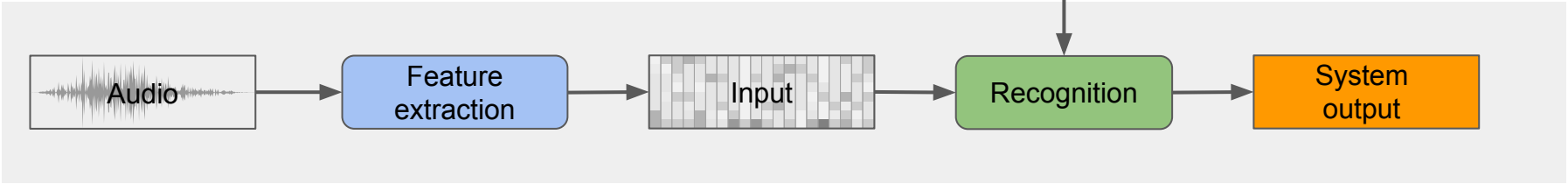
Task specific processing

General system architecture

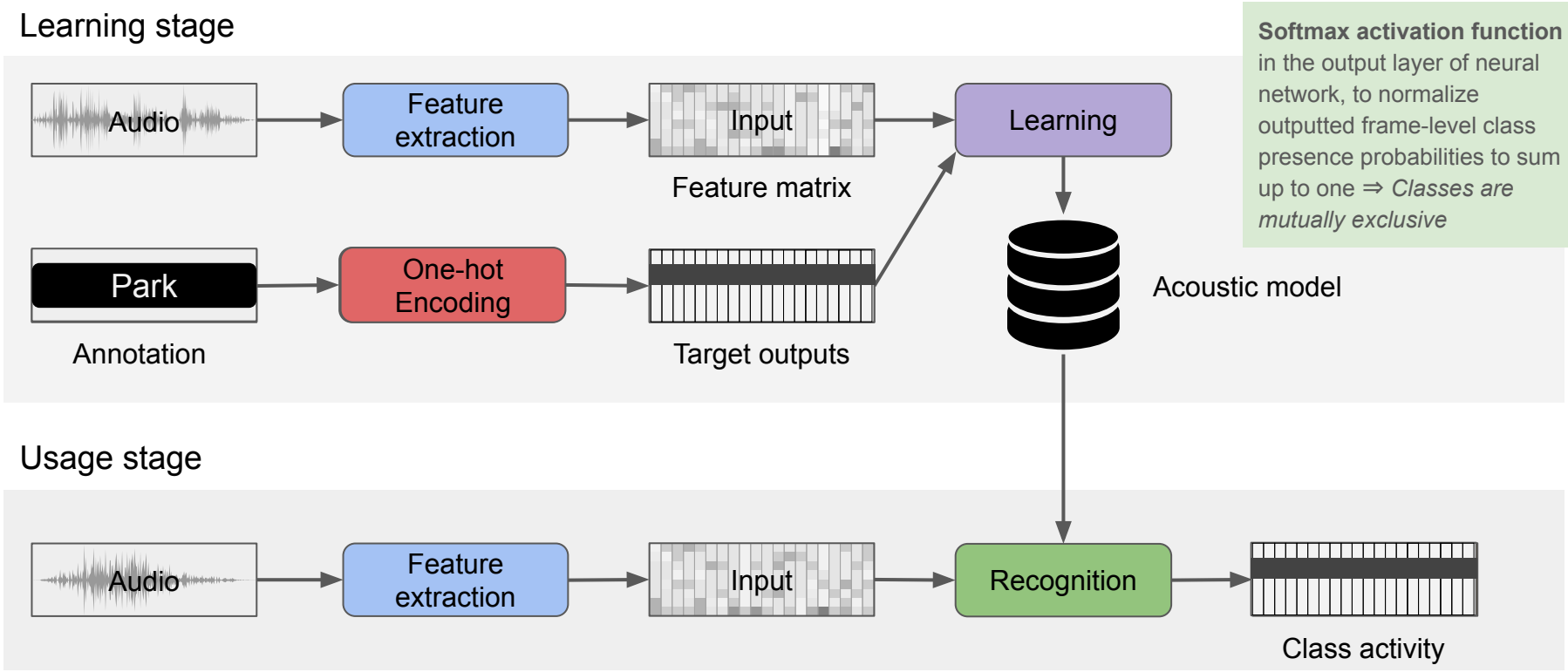
Learning stage



Usage stage

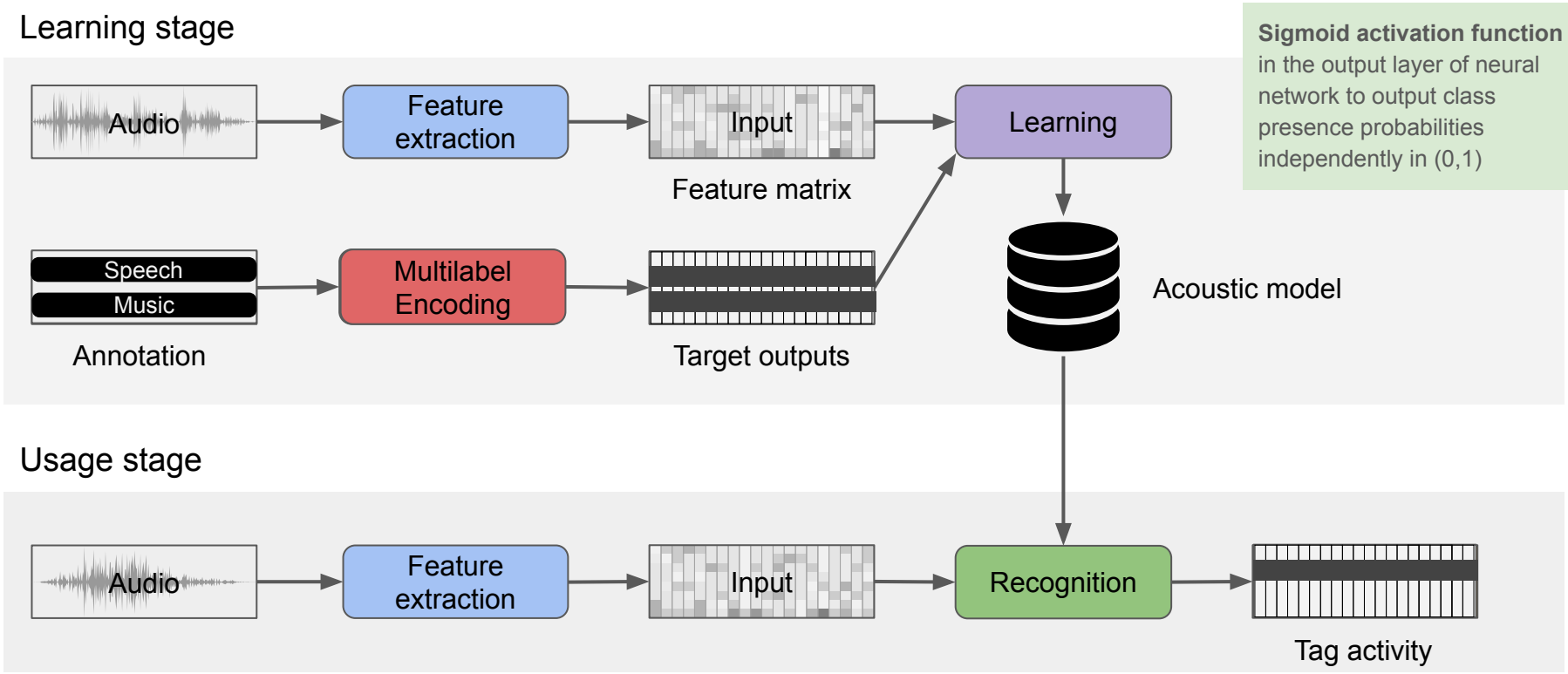


Sound classification (single label classification)



Softmax activation function in the output layer of neural network, to normalize outputted frame-level class presence probabilities to sum up to one ⇒ *Classes are mutually exclusive*

Audio tagging (multi label classification)

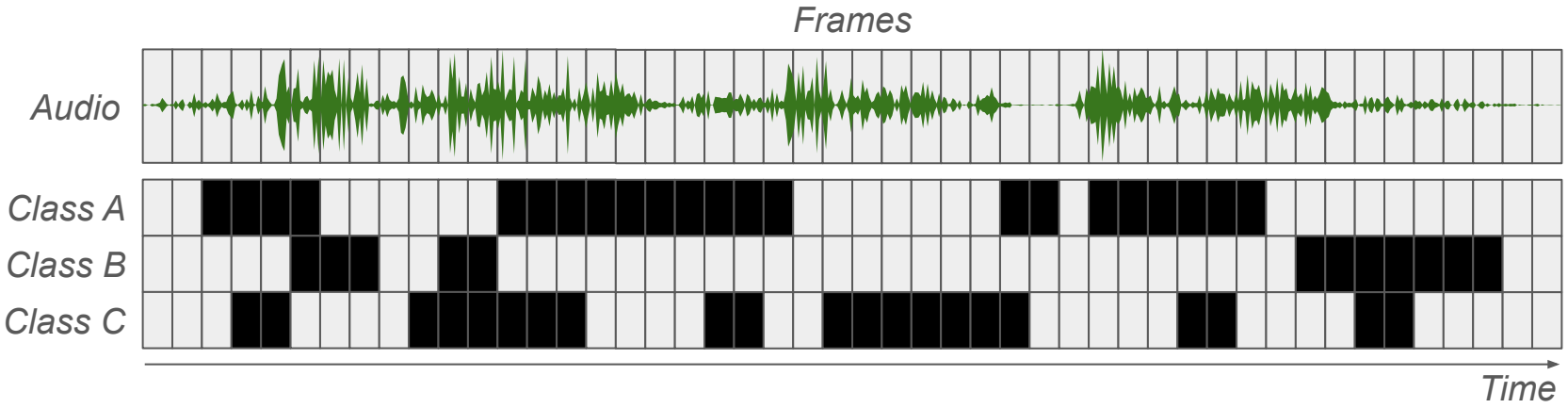


Sigmoid activation function in the output layer of neural network to output class presence probabilities independently in (0,1)

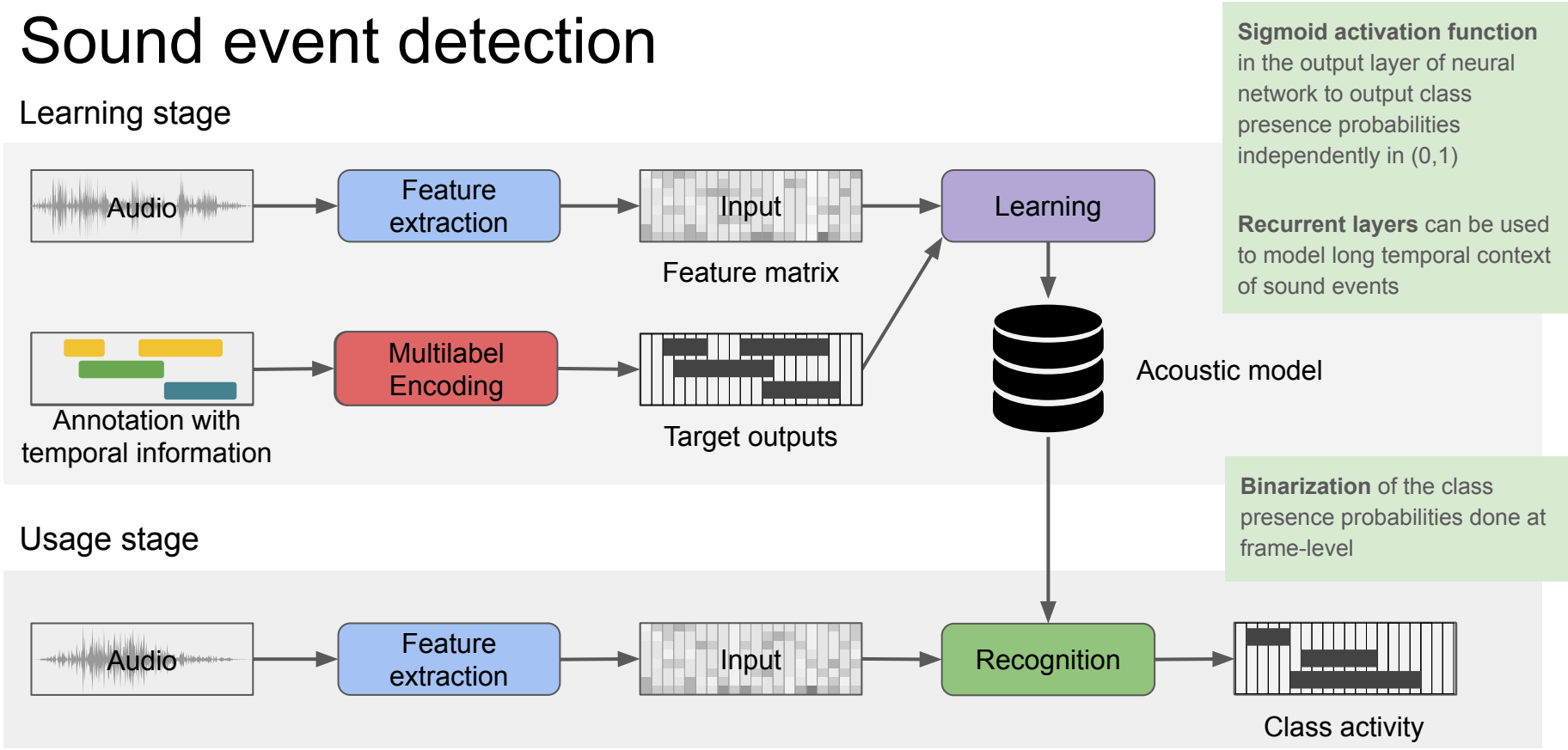
Sound event detection

Temporal activity estimated along with class labels

Multi-label classification of short consecutive audio frames, using contextual information from consecutive frames



Sound event detection



Sigmoid activation function in the output layer of neural network to output class presence probabilities independently in (0,1)

Recurrent layers can be used to model long temporal context of sound events

Binarization of the class presence probabilities done at frame-level

Datasets and evaluation

Datasets

Datasets for supervised learning

Audio

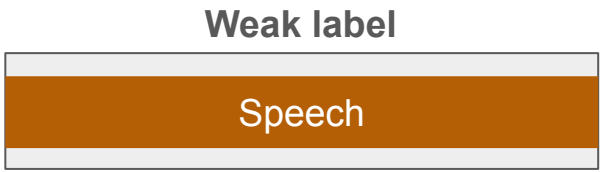
- **Coverage** – all categories relevant to the task
- **Variability** – examples with variable conditions of generation, recording, etc.
- **Size** – many examples; class balance if possible

Labels

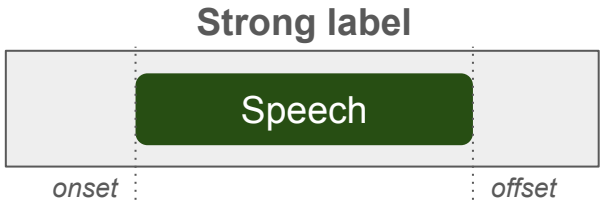
- **Representation** – allow understanding of the sound properties
- **Non-ambiguity** – one-to-one correspondence between sound and label

Labels for sound scenes and events

- **Acoustic scene labels** – description of the scene
  - Meaningful clue for identifying it: e.g. park, office, meeting
- **Sound event labels** – description of the sound as perceived by humans
  - Highly subjective (vocabulary)
  - **Everyday listening** – interpretation of the sound in terms of its source vs. **musical listening** – interpretation of the sound in terms of its acoustic qualities



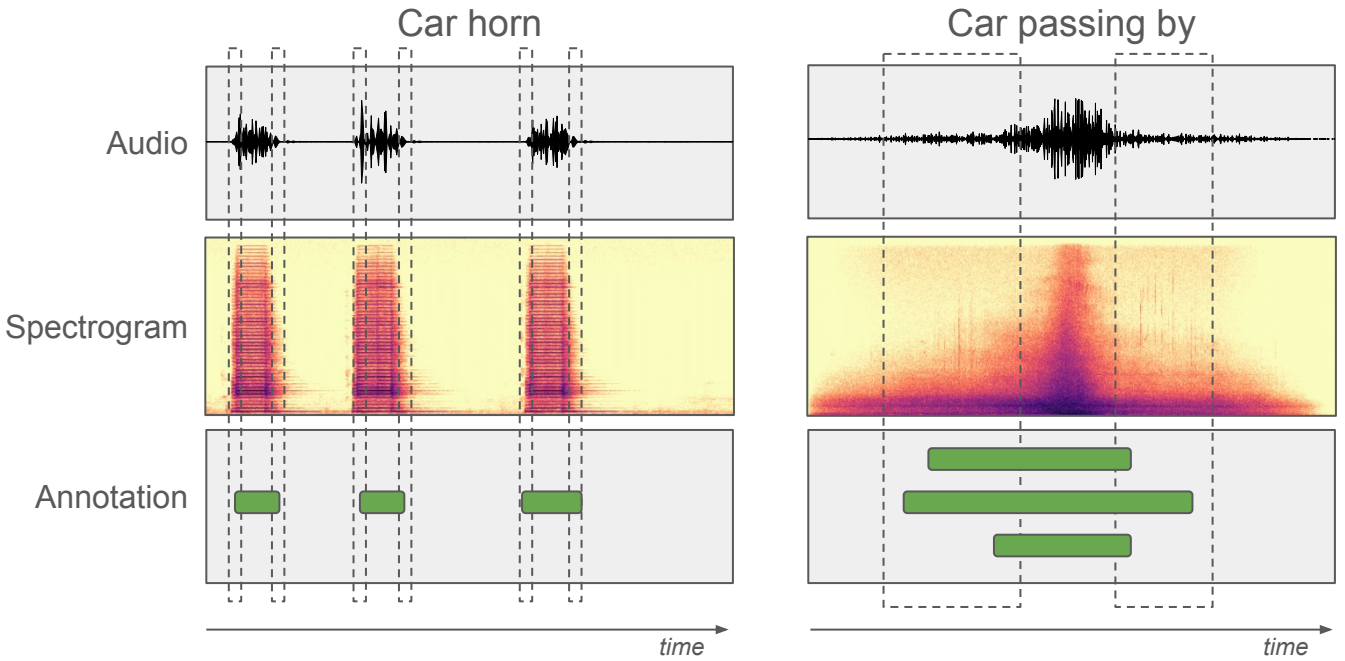
Event considered to be active throughout the segment



Event onset and offset is annotated to the timeline

Onset and offset ambiguity

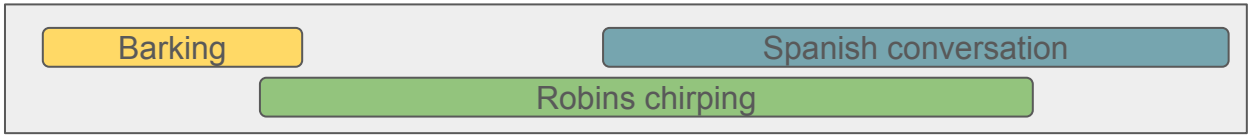
Boundaries of the sound event are not always obvious ⇒ subjectivity!



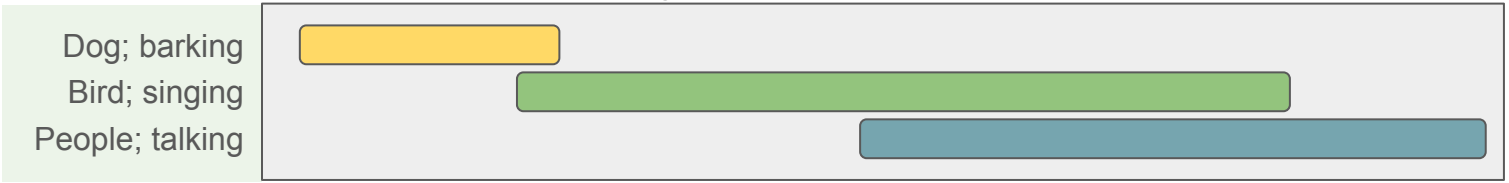
Multiple possible onset and offset positions

Types of annotations for sound events

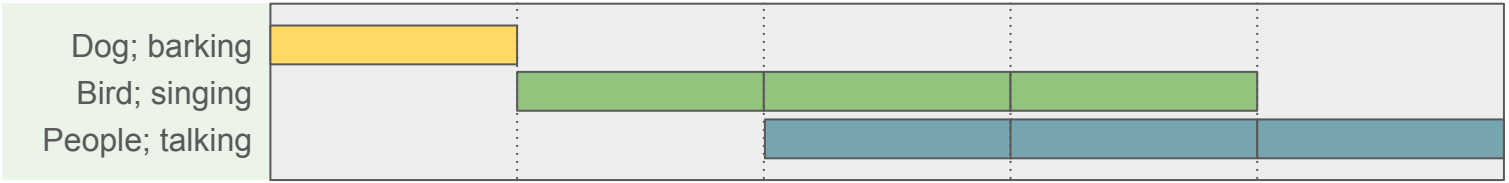
Free segmentation and labeling



Free segmentation and pre-selected labels



Pre-segmented audio and pre-selected labels



Decreasing annotation effort

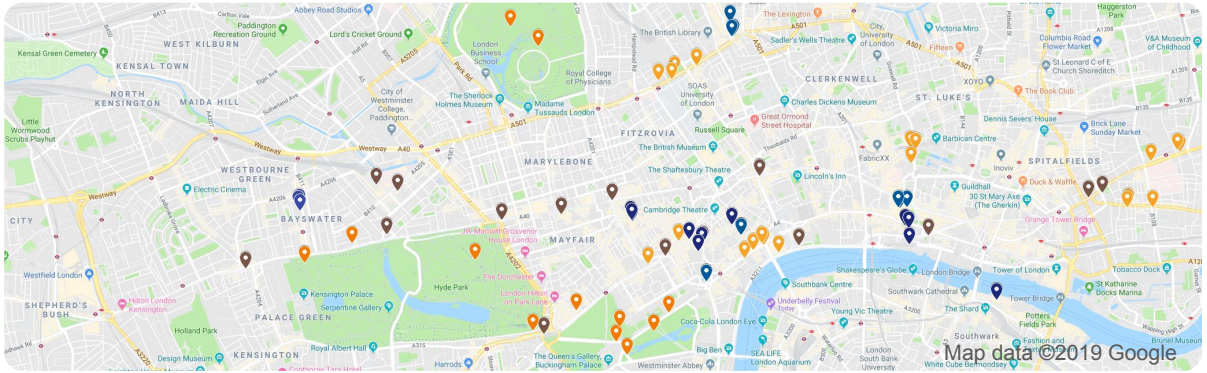
Examples of datasets

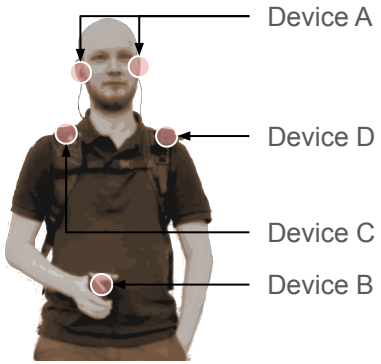
Name		Task	Data	Class#	Comments
TUT Acoustic Scenes 2017	*Mesaros et al.*	ASC	13h	15	Real-life recordings, recorded in a **single** country
TAU Urban Acoustic Scenes 2019	*Mesaros et al.*	ASC	40h	10	Real-life recordings, recorded in **multiple** countries
TUT Sound Events 2017	*Mesaros et al.*	SED	1.5h	6	Real-life recordings with **manual annotations**
Urban-SED	*Salamon et al.*	SED	30h	10	**Synthetically** generated audio material
CHiME-Home	*Foster et al.*	Tagging	6.5h	7	Real-life recordings from **domestic environment** with manual annotations
Freesound Dataset 2019	*Fonseca et al.*	Tagging	90h	80	Curated / **verified** annotations (10h) and **noisy** crowdsourced annotations (80h)
AudioSet	*Google*	Tagging	5000h	527	Youtube videos annotated with weak labels, **automatically tagged**, partially verified

A more comprehensive list of openly available datasets can be found at: http://www.cs.tut.fi/~heittolt/datasets

TAU Urban Acoustic Scenes 2019

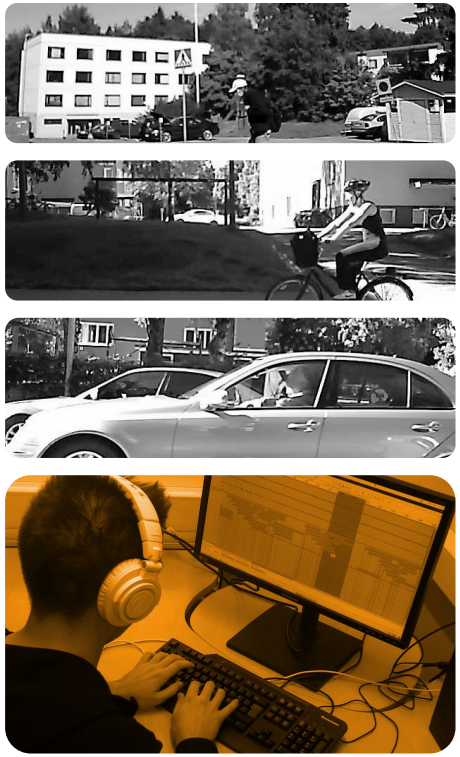
- 10 classes, predefined labels
- 12 large European cities, multiple locations per acoustic scene
- Binaural recordings, multiple devices simultaneously (high-quality and mobile devices)
- Recordings checked for private content





TUT Sound Events 2017

- Street scenes, Finland (city center, residential area)
- Manual annotation: structured labels (noun+verb) but open vocabulary
- Selected most frequent sound events related to human presence and traffic
- Original labels merged by the sound source:
  - *“car passing by”, “car engine running”,“car idling”* ⇒ *“car”*
  - sounds produced by buses and trucks ⇒ *“large vehicle”*



Evaluation Metrics

Introduction

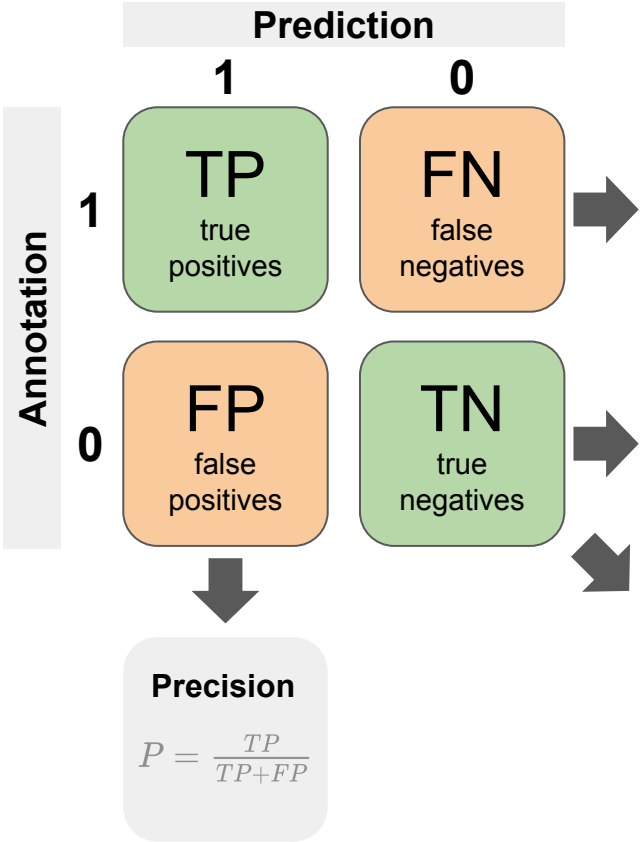
How do we measure system performance?

Common metrics in machine learning / pattern recognition problems:

- Accuracy (ACC)
- F-score, Precision (P), Recall (R)
- Error rate (ER)
- Average precision (AP) and Mean average precision (mAP)
- Receiver operating characteristic (ROC) curve and corresponding area under the curve (AUC)
- Equal error rate (ERR)

All applicable to classification and tagging

Contingency table



True Positive Rate
Sensitivity
Recall

$$R = \frac{TP}{TP+FN}$$

False Positive Rate

$$FPR = \frac{FP}{FP+FN}$$

Specificity

$$Specificity = \frac{TN}{FP+FN}$$

Accuracy

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

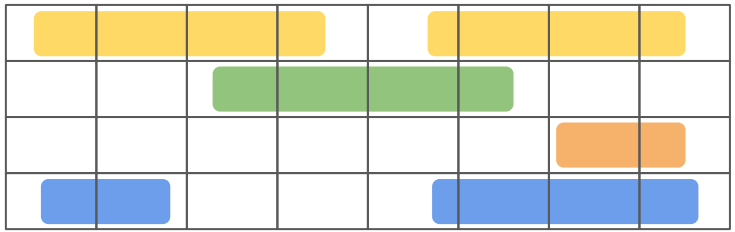
F-score

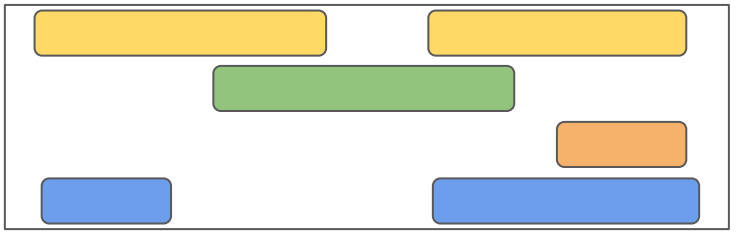
$$F = \frac{2 \cdot P \cdot R}{P+R}$$

Evaluating sound event detection

Two different ways of measuring performance [1]:

- **Segment-based metrics**: system output and reference are compared in short time segments
- **Event-based metrics:** system output and reference are compared event by event

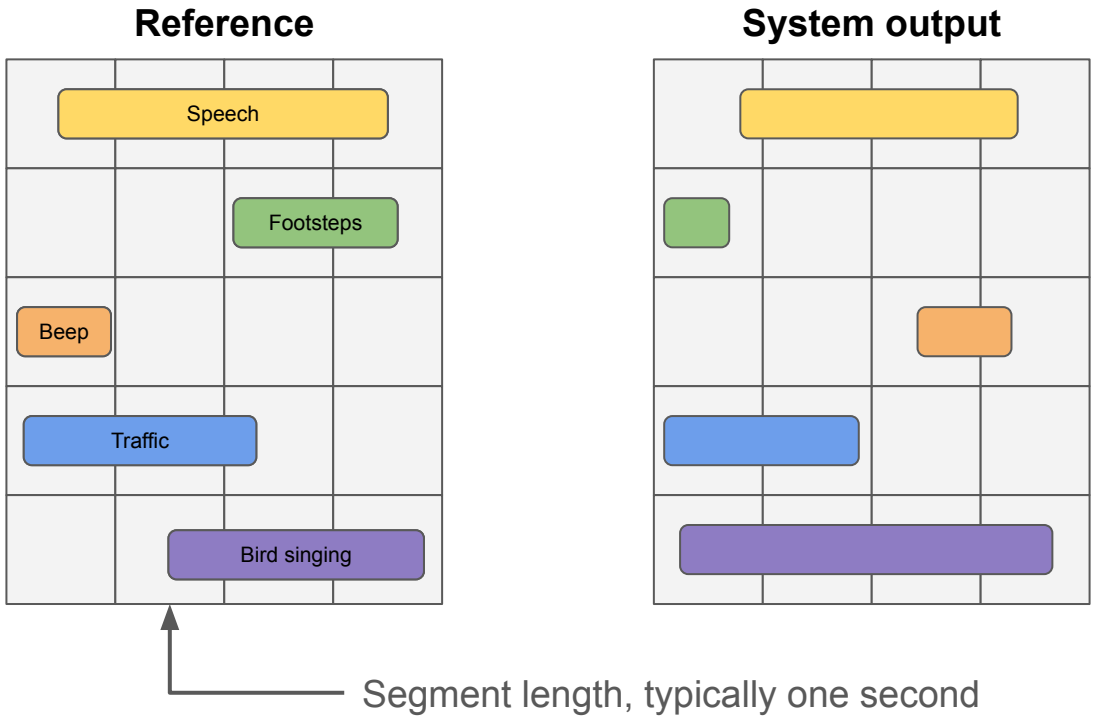




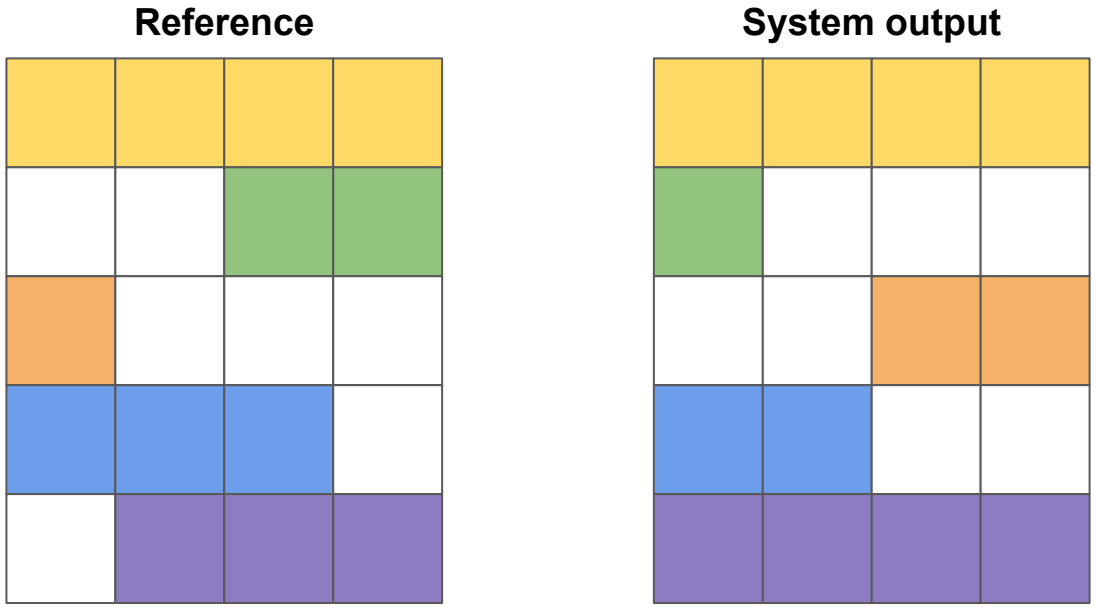
Intermediate statistics defined accordingly

[1] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. *Metrics for polyphonic sound event detection. Applied Sciences*, 6(6):162, 2016

Segment-based evaluation: example



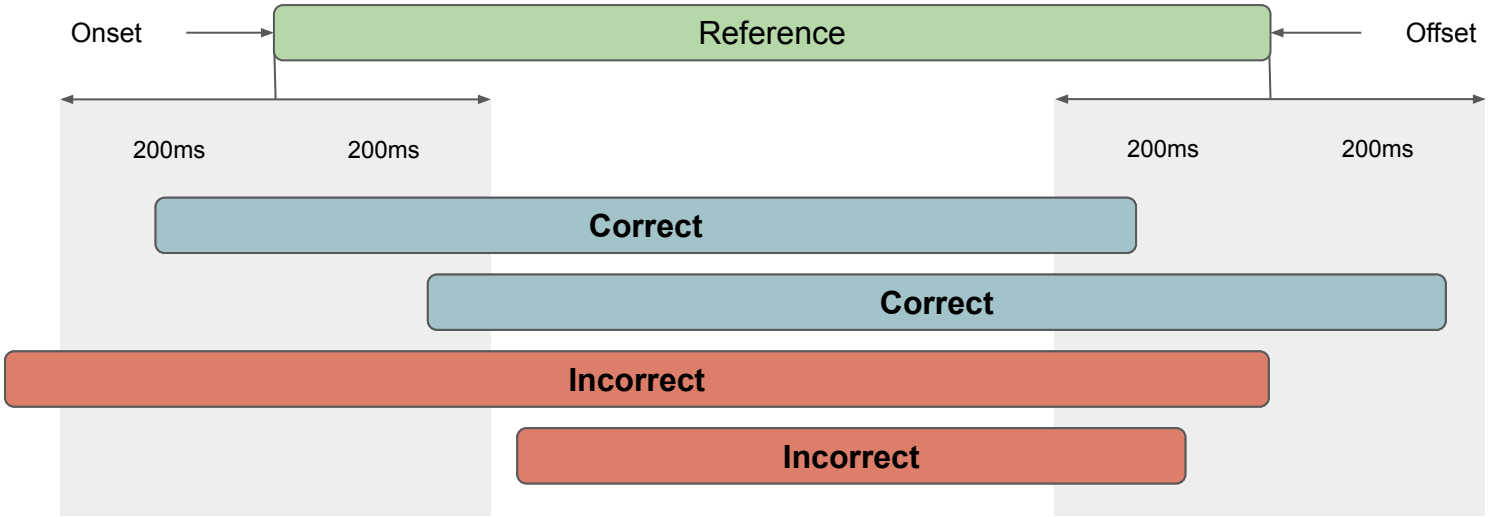
Transform event activity into same time resolution



Segment by segment comparison: TP, FP, TN, FN

Event-based evaluation

Tolerate a small misalignment (e.. 200 ms for onset, and 200 ms or half length for offset)



Metrics used in sound event detection

- F-score (segment-based, 1 second)
- Error Rate: measures the amount of errors in terms of
  - *substitutions* (S) – joint occurrence of a false positive and a false negative
  - *insertions* (I) – false positives unaccounted for in S
  - *deletions* (D) – false negatives unaccounted for in S
  - segment-based:

$$ER = \frac{\sum S(k) + \sum D(k) + \sum I(k)}{\sum N(k)}$$

- Choice of class-wise or instance-wise averaging

Which metric is best?

	Advantages	Disadvantages
Accuracy	Simple measure of the ability of the system to take the correct decision	Influenced by the class balance: ● for rare classes (i.e., where TP+FN is small), a system can have a high proportion of true negatives even if it makes no correct predictions, leading to a paradoxically high accuracy metric
F-score	Widely known and easy to understand	Choice of averaging scheme is especially important: ● In **instance-based averaging**, large classes dominate small classes ● In **class-based averaging**, one needs to ensure presence of all classes in the test data to avoid recall to be undefined
Error Rate	Parallel to established metrics in speech recognition and speaker diarization evaluation	A score rather than a percentage: ● Can be over 1.0 in cases when the system makes more errors than correct predictions ● Interpretation difficult, considering that it is trivial to obtain an error rate of 1 by outputting no active events

Evaluation pitfalls

- Segments from the same recording or location are highly correlated!
  - When the dataset contains short segments of long recordings, all segments originating from the same recording should be in one subset (train or test)
  - Sound events from the same recording or location are likely produced by the same physical source
  - Synthetic data: use different instances for train and test mixtures
- Cross-validation setup carefully constructed to avoid contamination (use location information for guiding the train/test/validation split)
- Statistical significance – related to data size

Reproducible research

Reproducible research

Use an open dataset or publish your own dataset:

- Datasets available in services like *zenodo.org, ieee-dataport.org, archive.org*
- Datasets introduced with a scientific paper, baseline system, cross-validation setup

Release your system:

- Release the code to allow reproducing results from your publications (e.g. GitHub)

Report your results in uniform way (same as other publications using the same dataset):

- Use same cross-validation setup as others
- Use established metric implementations (e.g. in Python scikit-learn, sed_eval)

DCASE CHALLENGE

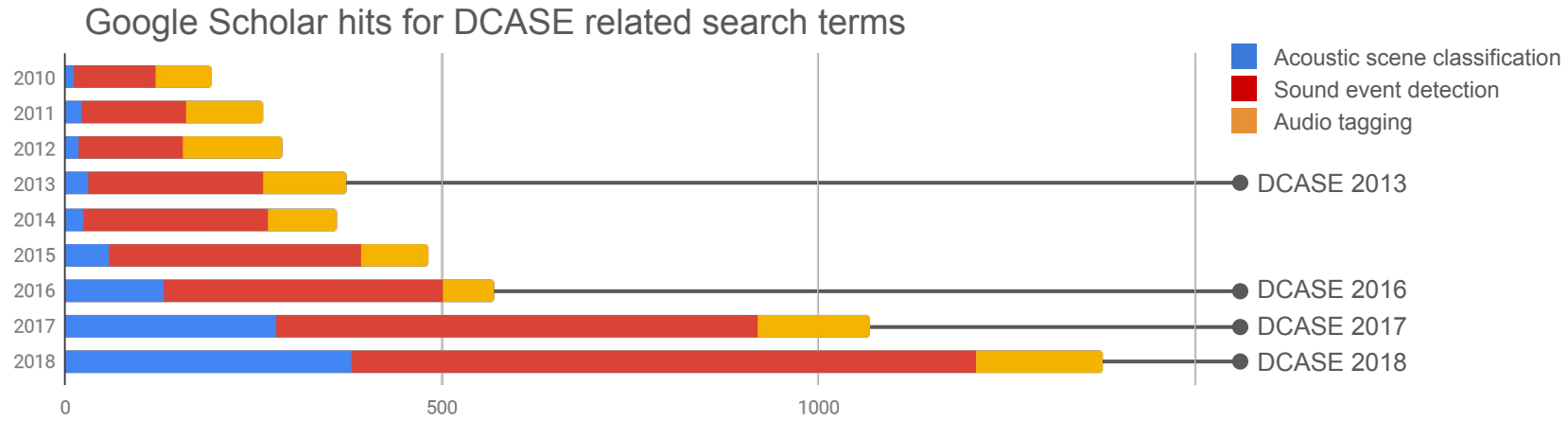
evaluation campaign

Scope of the challenge

- Aim to provide open data for researchers to use in their work
- Encourage reproducible research
- Attract new researchers into the field
- Create reference points for performance comparison

Outcome

- Development of state of the art methods
- Many new open datasets
- Rapidly growing community of researchers



Challenge tasks 2013 - 2019

Classical tasks:

- **Acoustic scene classification** – textbook example of supervised classification (2013-2019) with increasing amount of data and acoustic variability; mismatched devices (2018, 2019); open set classification (2019)
- **Sound event detection** – synthetic audio (2013-2016), real-life audio (2013-2017), rare events (2017), weakly labeled training data (2017-2019)
- **Audio tagging** – domestic audio, smart cars, Freesound, urban (2016-2019)

Novel openings:

- **Bird detection** (2018) – mismatched training and test data, generalization
- **Multichannel** audio classification (2018)
- Sound event **localization** and detection (2019)

Questions & Answers

Advanced methods

Session 2

Outline

Sound event detection with Python

Real-life challenges and solutions

Future perspectives

Questions & Answers

Sound event detection with Python

Jupyter notebooks:

https://github.com/toni-heittola/icassp2019-tutorial

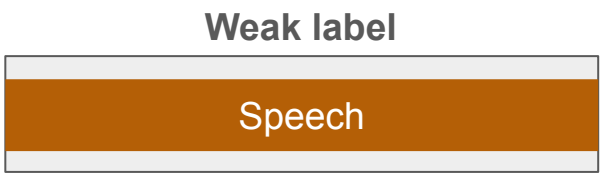
Real-life challenges and solutions

Weak labels

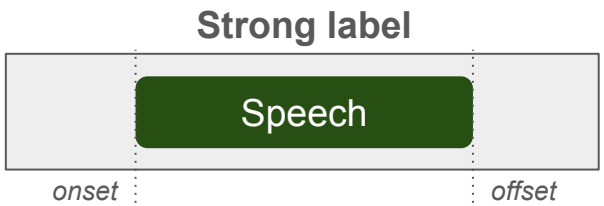
Problem: Obtaining strong labels is very expensive

Solution: Use weak labels in training (weakly supervised learning)

Key issue: Systems must cope with the weak labels during the learning process



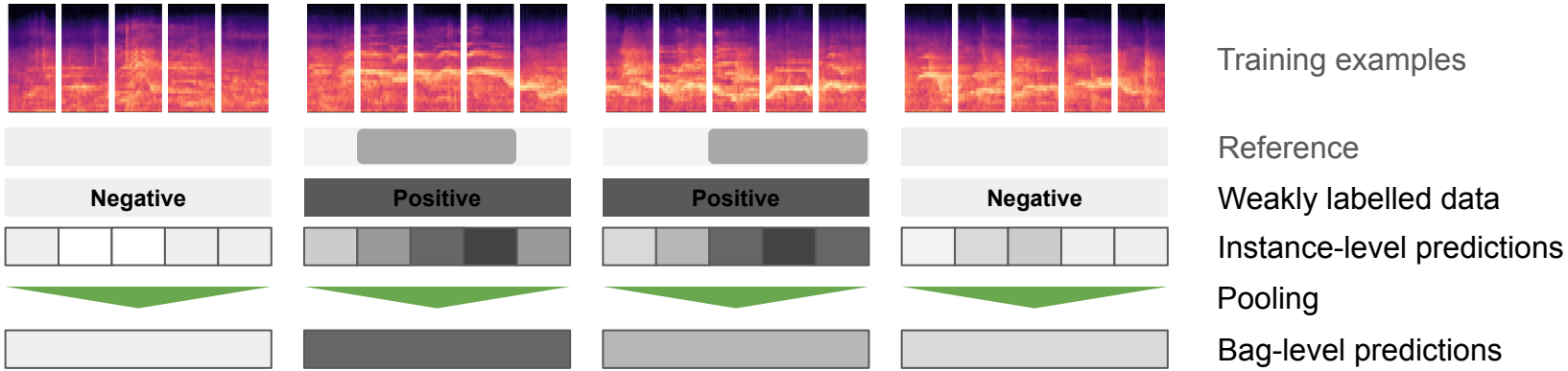
Event considered to be active throughout the segment



Event onset and offset is annotated to the timeline

Weakly supervised learning: multi-instance learning

- Training instances (frames) are arranged into **bags** (segments / clips)
- Label is attached to bag, rather than individual instances within
  - **Negative** bags contains only negative instances ⇒ **pure**
  - **Positive** bags can contain negative and positive instances ⇒ **impure**
- **Learning**:
  - Neural network predicts the probability for class at instance-level
  - **Pooling function** aggregates instance-level information into bag-level
  - Loss is minimized at bag-level during the training



Weak labels

Approaches:

- Multi-instance learning
- Label refinement
- Attention-based networks

Disadvantages: Evaluation still requires strongly labelled data

Advantages: Possibility of using large amount of data for training

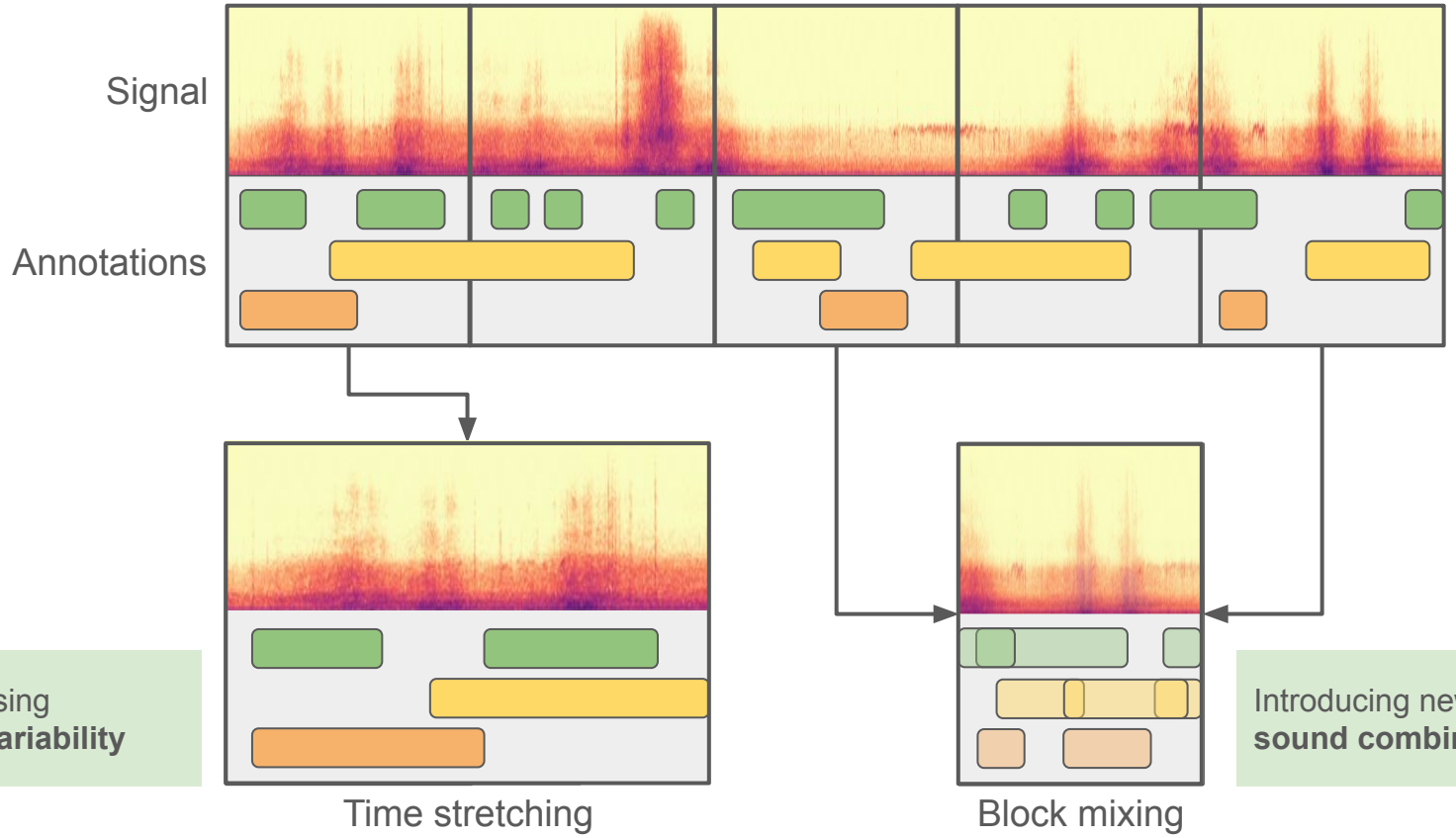
Data augmentation

Problem: Scarcity of data for specific problems

Solution: Modification of available data such that it mimics having larger and more acoustically diverse data

Key issue: Producing realistic and useful data

Data augmentation: reusing existing data



Increasing
data variability

Introducing new
sound combinations

Data augmentation

Approaches:

- Time-stretching, pitch shifting, dynamic range compression, equalization
- Convolution with various impulse responses to simulate various microphones and acoustic environments
- Sub-frame time shifting and random block mixing
- Simulating set of noise conditions by adding background noise while varying SNR

Disadvantage: Hard to mimic the complexity of real recordings

Advantage: Many useful combinations possible

Transfer learning

Problem: High-complexity models need huge amounts of data

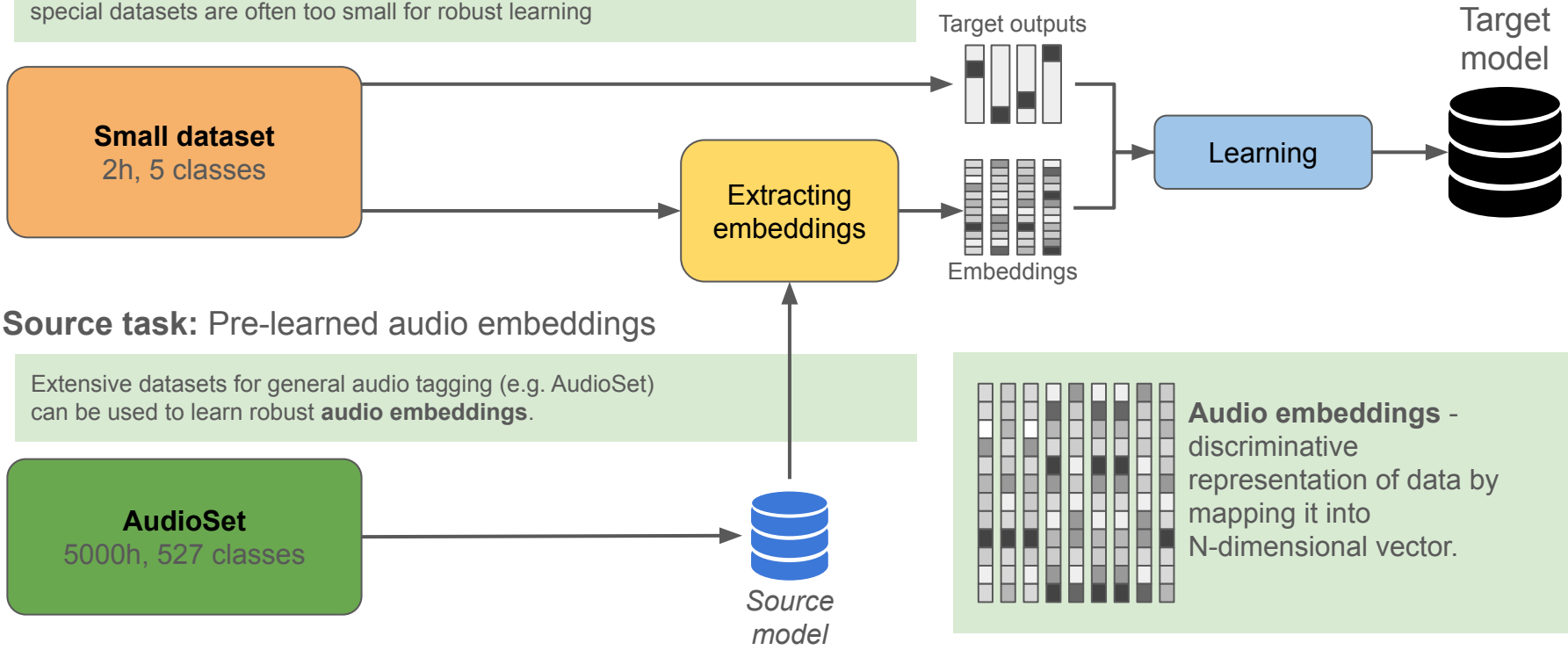
Solution: Use pre-trained system that already “knows” a lot from other domain; transfer neural network structure and weights from the source task to solve the target task

Key issue: Identify transferable knowledge

Transfer learning: classifier with small target dataset

Target task: Classification of agricultural machinery

It is time consuming to collect **extensive** dataset for specific tasks. Because of this, special datasets are often too small for robust learning



Source task: Pre-learned audio embeddings

Extensive datasets for general audio tagging (e.g. AudioSet) can be used to learn robust **audio embeddings**.

Audio embeddings - discriminative representation of data by mapping it into N-dimensional vector.

Transfer learning

Approaches:

- Using pre-trained model or specifically developed source model as a starting point for the target model; use fully or partially the source model
- Using source model as feature extractor: extract embeddings and use them as input when learning target model

Disadvantage: No guarantee that it works; in some cases can make the learning process even harder (negative transfer)

Advantage: Many pre-trained models available, enables including large amount of knowledge into learning process with minimal computational power

Data crowdsourcing

Problem: Annotation process is time consuming, especially for large datasets

Solution: Crowdsourcing of both audio and labels or just labels

Key issue: Systems must cope with labels noisiness and unreliability

Data crowdsourcing: label noise

- **Web audio** enables rapid dataset collection
  - Large amounts of user generated audio material available (Youtube / Freesound)
  - Labels can be inferred from user generated metadata ⇒ **noisy** labels
  - Example: **AudioSet** consists of 5000h labelled audio (527 classes), label error is above 50% for 18% of the classes

- Effect: increased complexity of learned models; decreased performance
- Can be handled at various stages of a system:
  - **Data**: Use label verification after each learning step to gradually verify the data (data relabelling)
  - **Learning**: Use noise-robust loss functions which are relying on model predictions more and more as learning progress instead of noisy labels (soft bootstrapping)

Data crowdsourcing

Approaches:

- Annotations with crowdsourcing services; postprocess to get less noisy labels
- Collect audio from web services and handle label noise during the learning

Disadvantage: Noisy labels, usually only feasible for weak labels; for evaluation, verified labels still necessary

Advantage: Fast access to large amount of annotated data

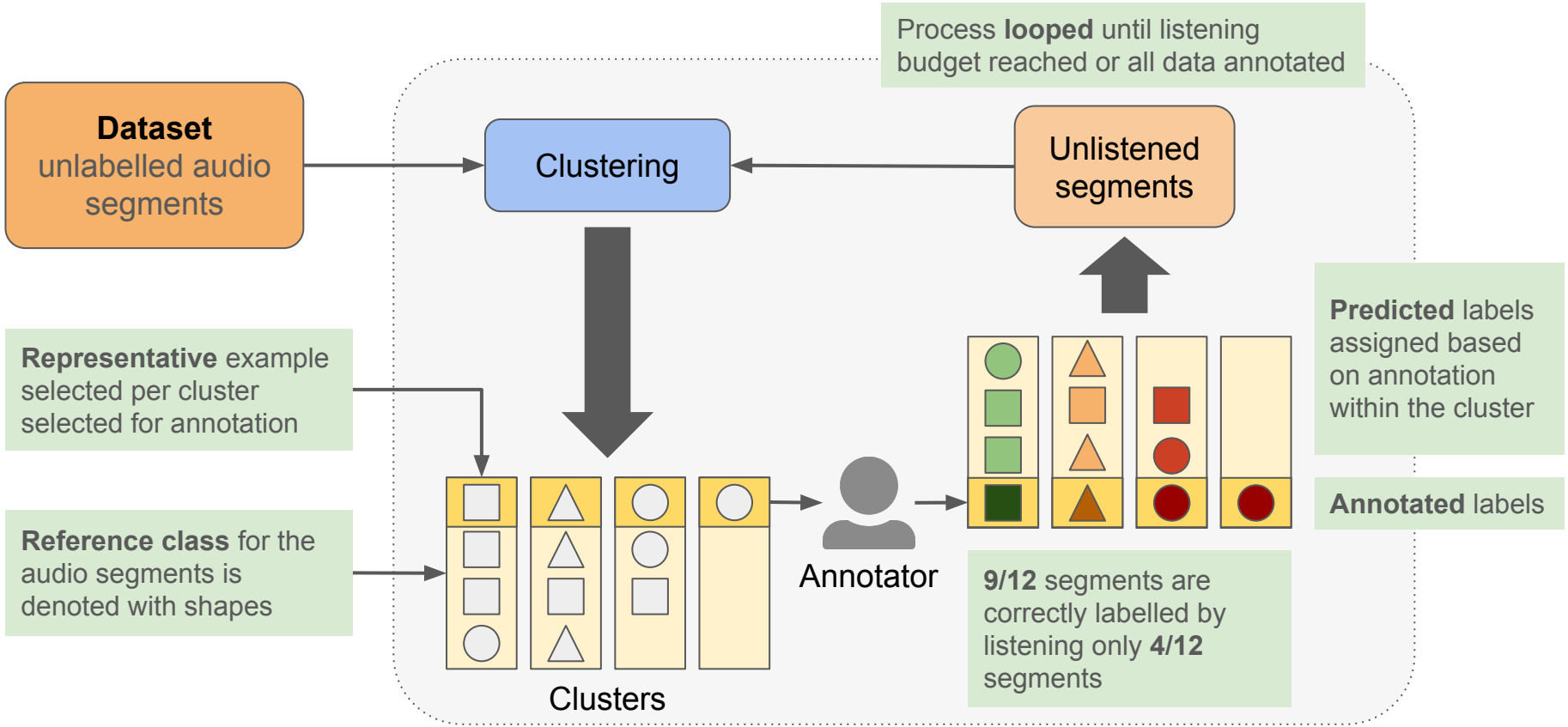
Limited annotation budget

Problem: Manual annotation is time consuming, requires extensive listening

Solution: Automatically select key examples for annotation

Key issue: How to select representative examples for manual annotation

Limited annotation budget: active learning



Limited annotation budget

Approaches:

- Active learning
- Semi-supervised learning

Disadvantage: Works best with classification; difficult for more complex tasks

Advantage: For very large datasets respectable accuracy can be achieved with relatively small listening budget

Future perspectives

Future research directions

- Structured class labels, taxonomies
- Spatial audio (localization, tracking, separation of sources)
- Audio + video + other modalities
- Joint data collection platforms
- Robust classification
- Weakly labeled data
- Crowdsourcing
- Transfer learning
- Active learning

Challenges

Fragility of deep learning:

How to predict when the methods are going to work or fail?

Privacy and personal data:

How to handle in data collection, how to prevent misuse of the methods?

Summary

- Scene classification and sound event detection: research fields with several potential applications
- Technical challenges: robust classification, dealing with overlapping sounds, reverberation, weak and noisy labels
- Practical & scientific challenges: acquisition of annotated data, robust use of data to help generalization
- Convolutional recurrent networks can be applied to a wide variety of different tasks
- Public evaluation campaigns allow comparison of different methods and reproducible research

Publication channels

DCASE WORKSHOP

Workshop on Detection and Classification of Acoustic Scenes and Events:

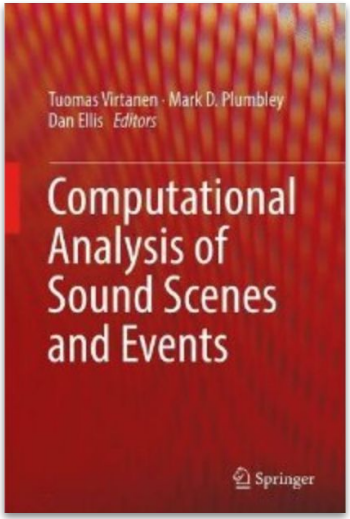
- Topics: tasks, methods, resources, applications, and evaluation
- DCASE 2019 Workshop: 25-26 Oct. 2019, NY (paper submission deadline: 12th July 2019)
- DCASE 2019 Challenge (submission deadline: 10 June 2019)

Audio and signal processing journals: IEEE/ACM TASLP

Conferences: ICASSP, WASPAA, IWAENC

Special sessions in signal processing conferences: EUSIPCO, MMSP, IJCNN

References



T. Virtanen, M. D. Plumbley, D. Ellis (eds).
Computational Analysis of Sound Scenes and Events.
Springer, 2018.

Contributors

Researchers at Audio Research Group / Tampere University
DCASE organizers

Questions & Answers