# Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations

**Mark Cartwright[1]**, Ayanna Seals[1], Justin Salamon[1], Alex Williams[2], Stefanie Mikloska[2], Duncan MacConnell[1], Edith Law[2], Juan Bello[1], Oded Nov[1]

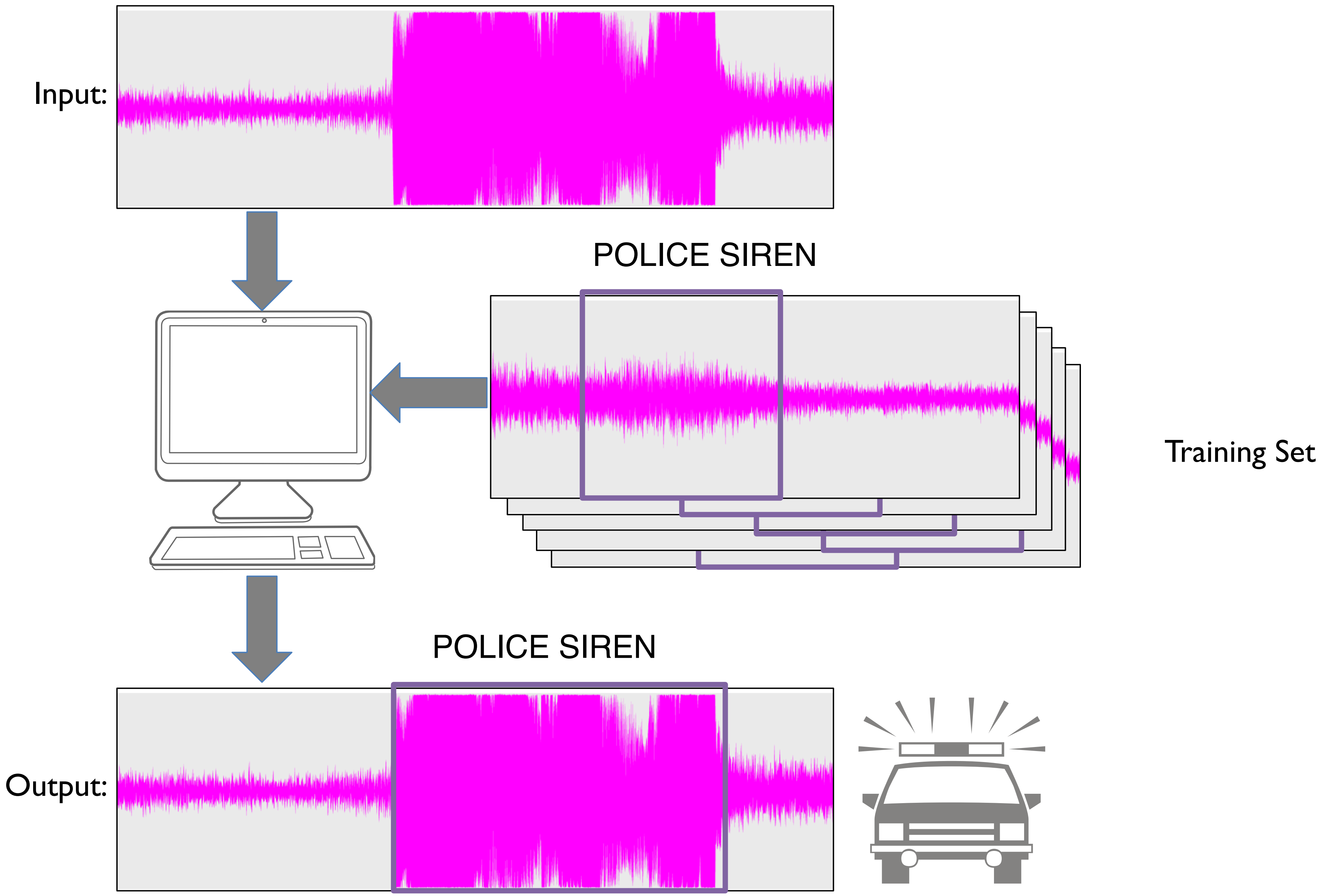
1. New York University
2. University of Waterloo

# SONYC

## **S**ounds of **N**ew **Y**ork **C**ity

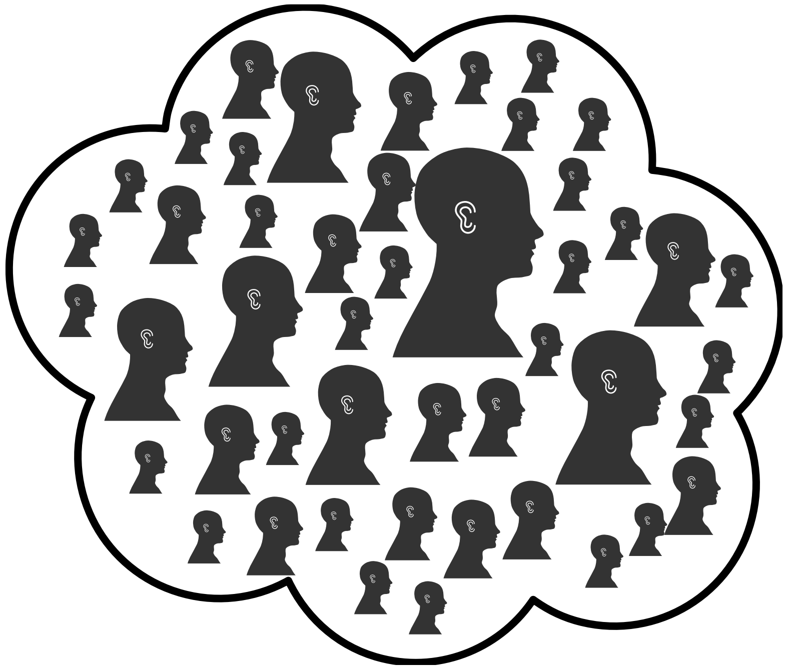
A cyber-physical system powered by an acoustic sensor network that aims to **monitor**, **analyze**, and **mitigate** urban noise pollution.

# Audio Annotation of Sound-Event Detection



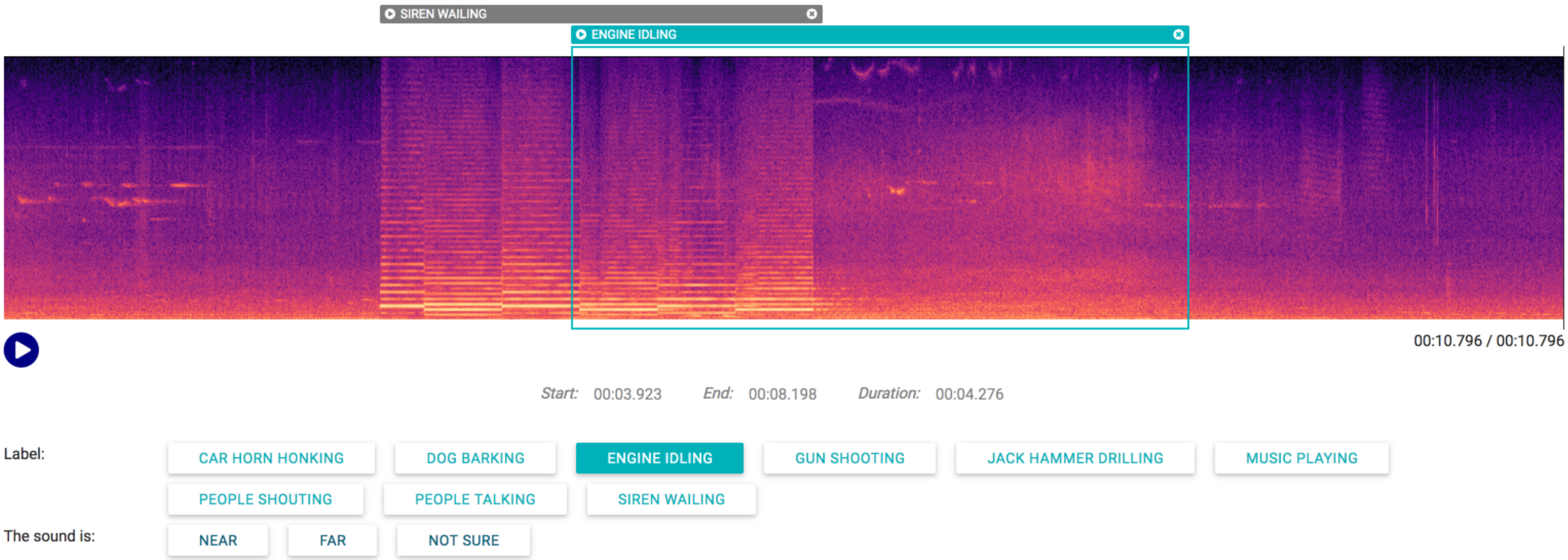
# Research Questions

- Which sound visualization aid yields the highest quality crowdsourced audio annotations?
- What limitations can we expect from crowdsourced audio annotations as a function of soundscape complexity?
- What is the trade-off between reliability and redundancy in crowdsourced audio annotation?



# The Audio Annotator

Configured with the spectrogram visualization:

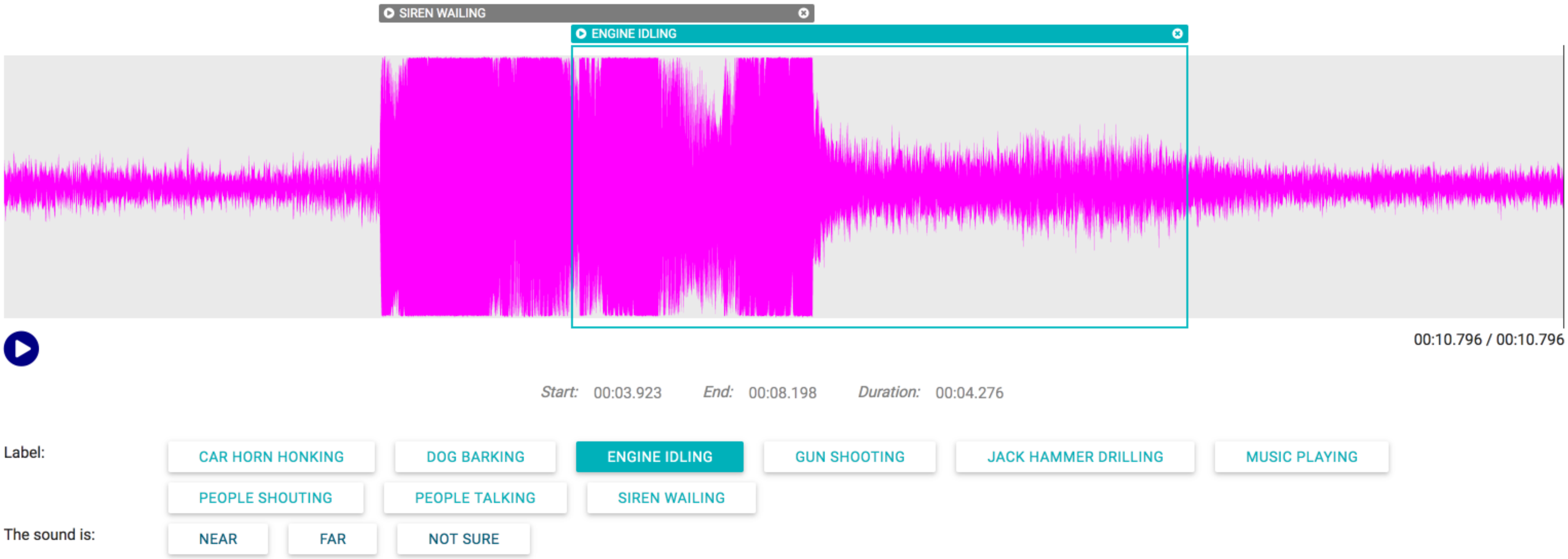


*Start:* 00:03.923 *End:* 00:08.198 *Duration:* 00:04.276

github.com/CrowdCurio/audio-annotator

# The Audio Annotator

Configured with the waveform visualization:



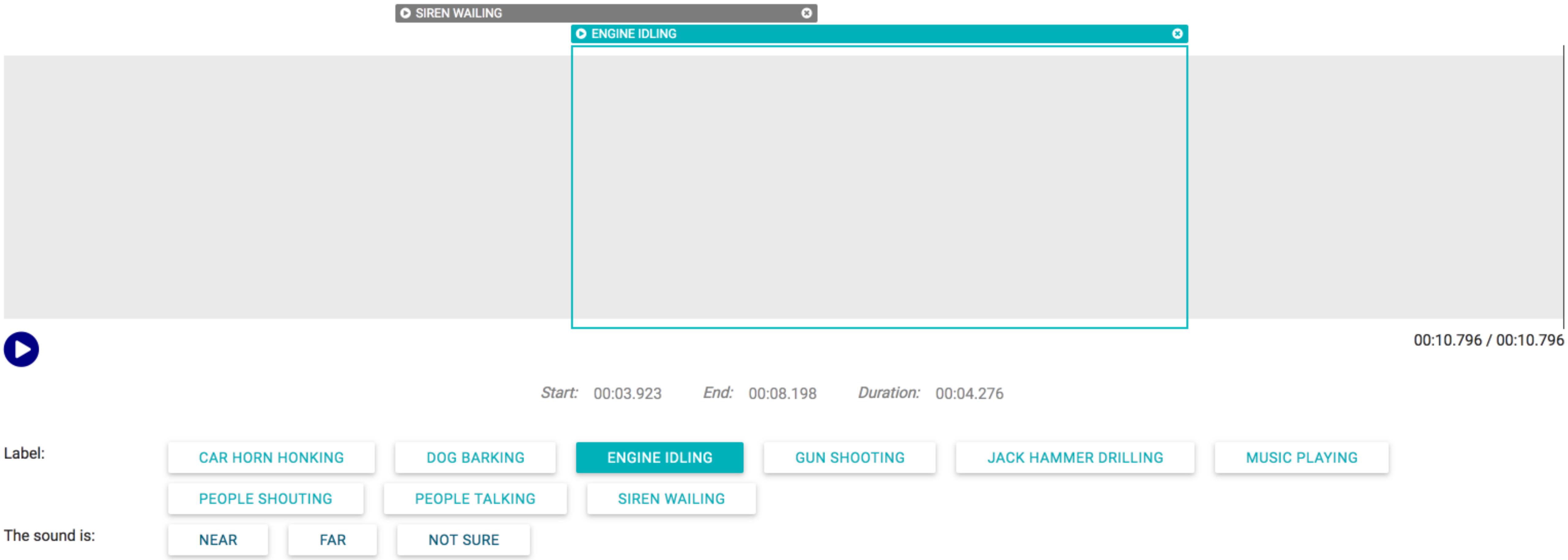
00:10.796 / 00:10.796

*Start:* 00:03.923 *End:* 00:08.198 *Duration:* 00:04.276

github.com/CrowdCurio/audio-annotator

# The Audio Annotator

Configured without a visualization:



00:10.796 / 00:10.796

*Start:* 00:03.923 *End:* 00:08.198 *Duration:* 00:04.276

github.com/CrowdCurio/audio-annotator

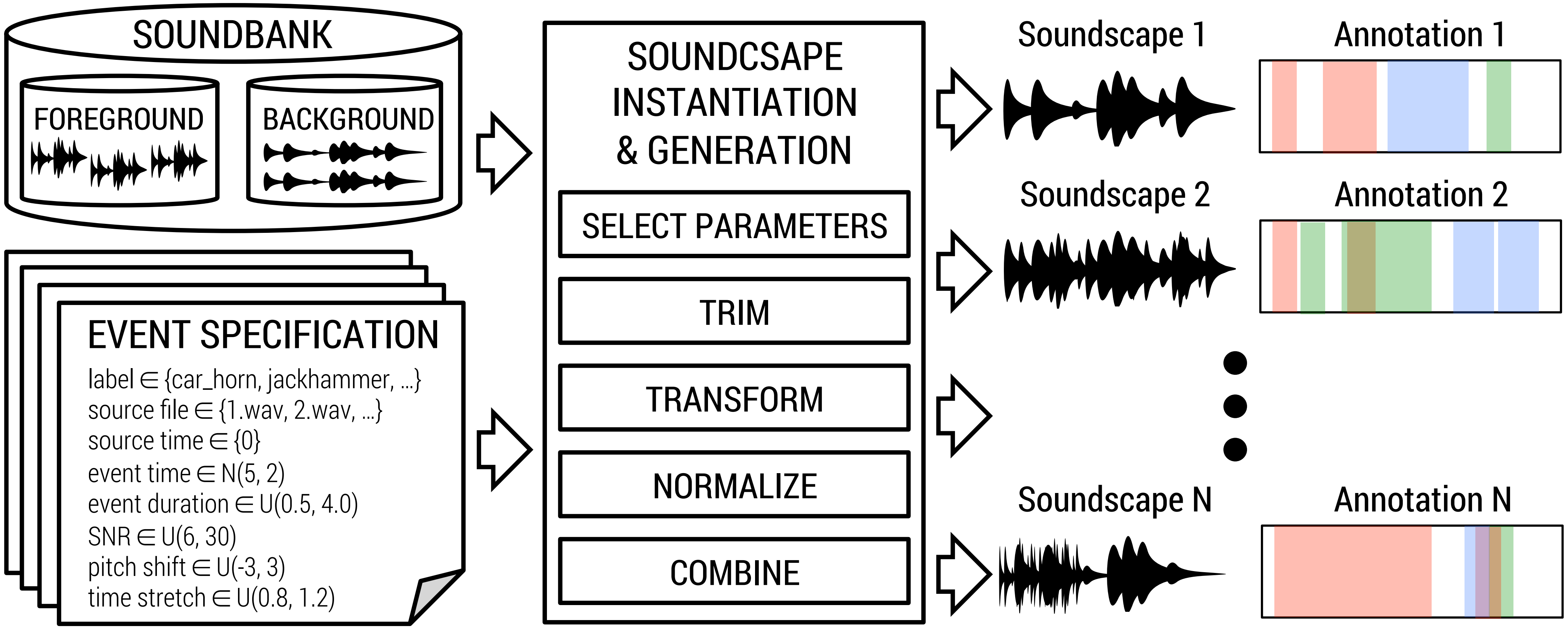
CrowdCurio.

Fostering Curiosity Through Science.

crowdcurio.com

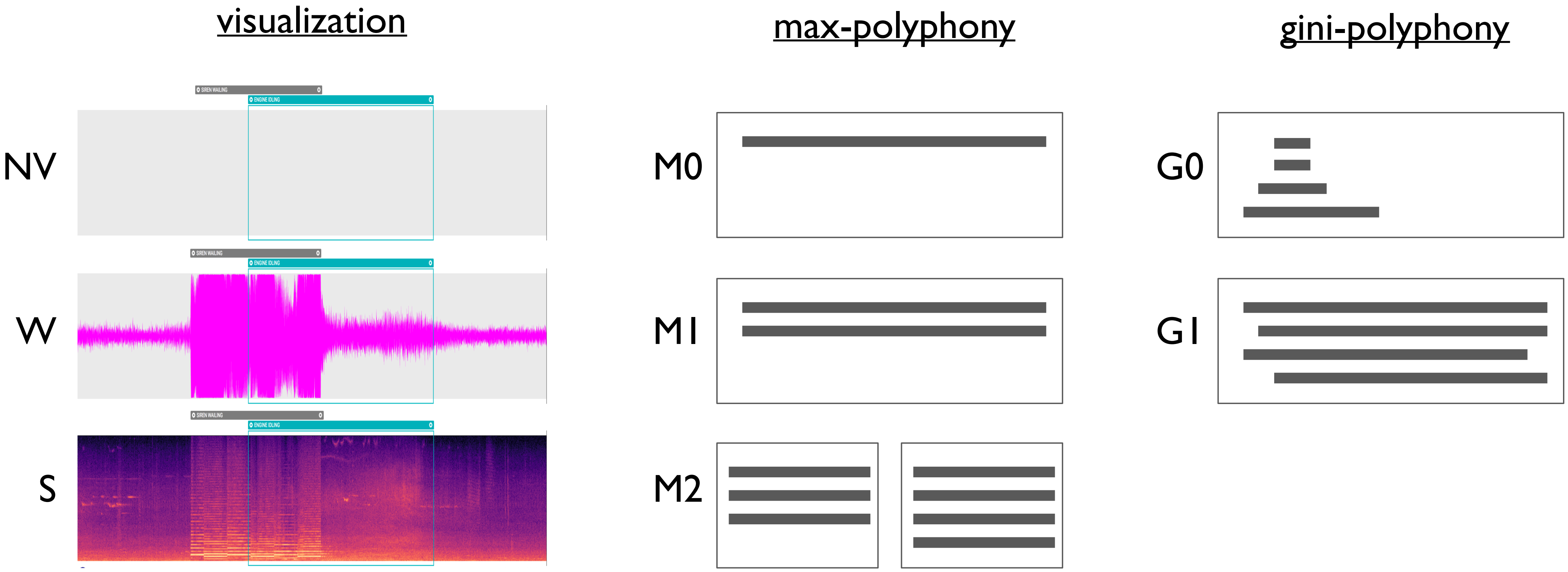
# Scaper: Soundscape Synthesis

- Open source python library for soundscape synthesis (WASPAA 2017)
- github.com/justinsalamon/scaper



# Experiment

- 3 x 3 x 2 between-subjects factorial design:



visualization

max-polyphony

gini-polyphony

NV

W

S

M0

M1

M2

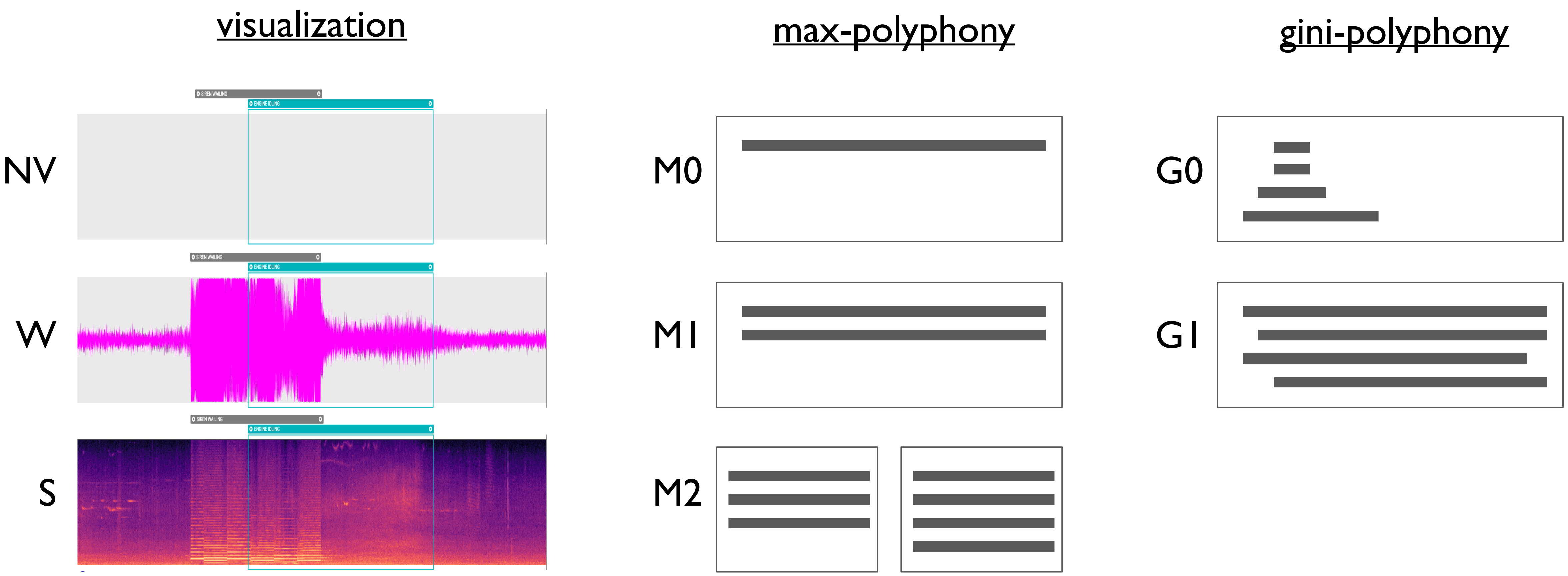
G0

G1

- Soundscape examples:
  M0G0 M0G1 M2G0 M2G1

# Experiment

- 3 x 3 x 2 between-subjects factorial design:



visualization

max-polyphony

gini-polyphony

NV

W

S

M0

M1

M2

G0

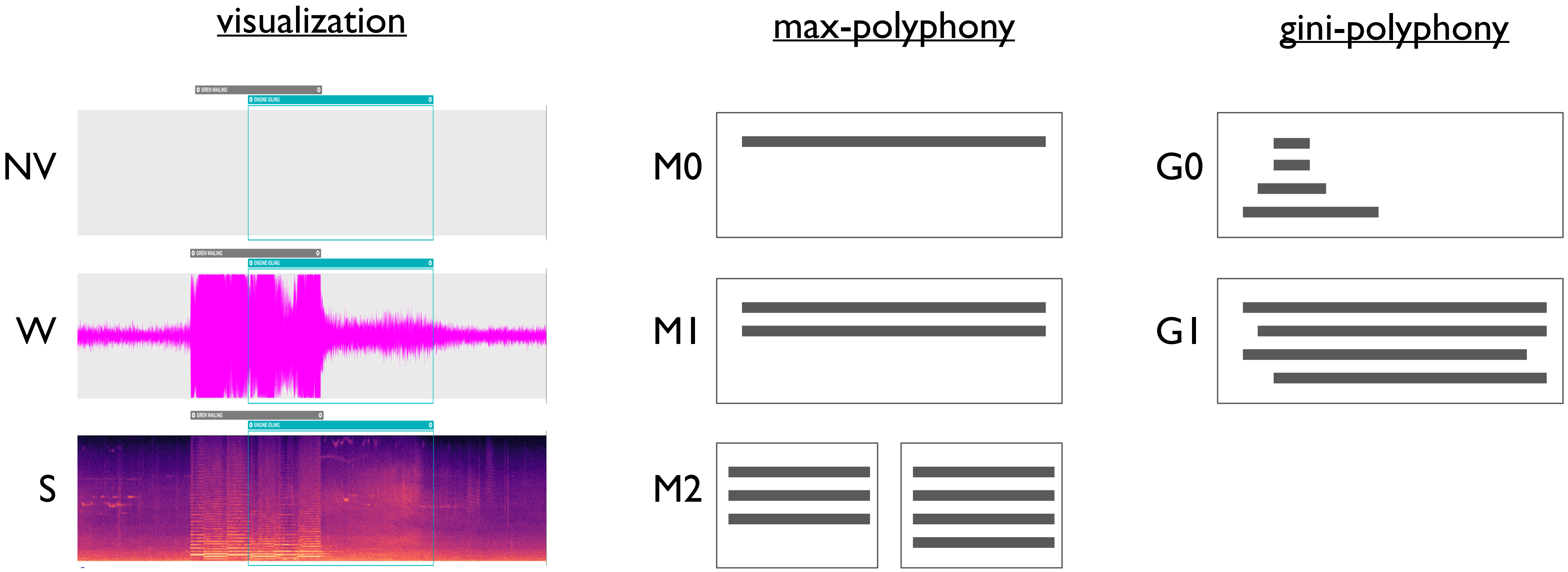
G1

- Soundscape examples:

M0G0 M0G1 M2G0 M2G1

# Experiment

- 3 x 3 x 2 between-subjects factorial design:



visualization

max-polyphony

gini-polyphony

NV

W

S

M0

M1

M2

G0

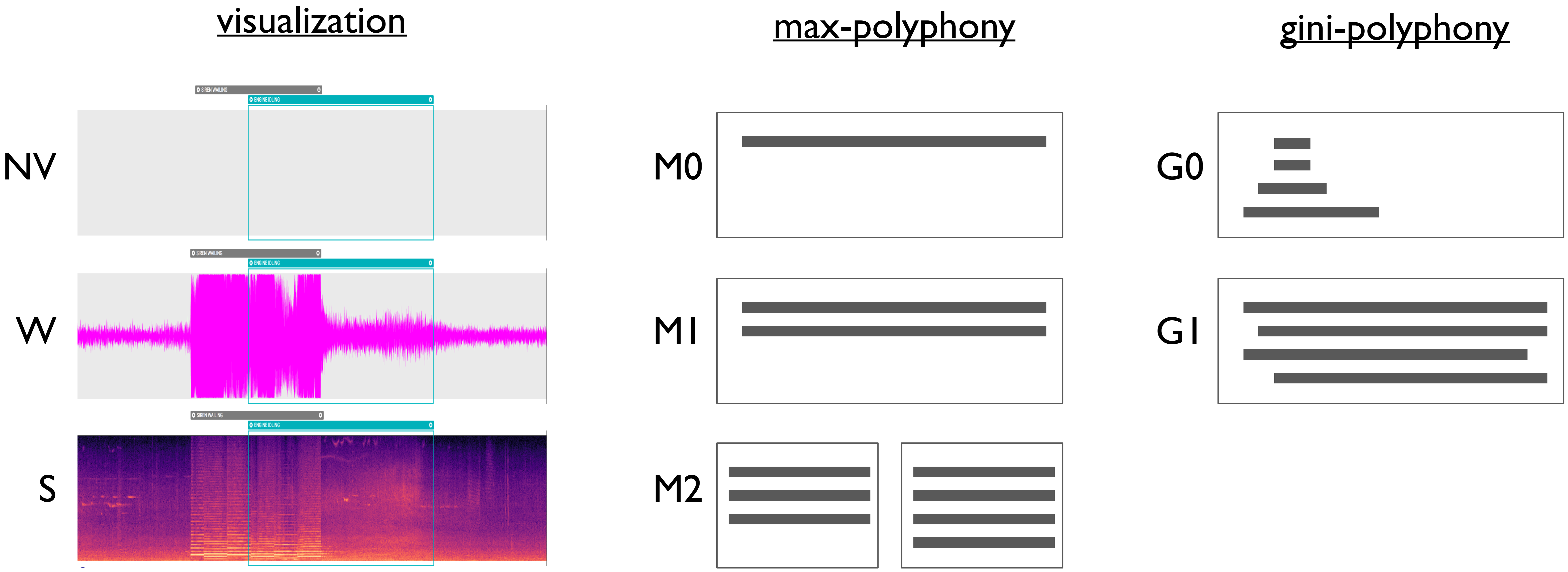
G1

- Soundscape examples:

M0G0 M0G1 M2G0 M2G1

# Experiment

- 3 x 3 x 2 between-subjects factorial design:



- Soundscape examples:
  M0G0 M0G1 M2G0 M2G1

# Experiment

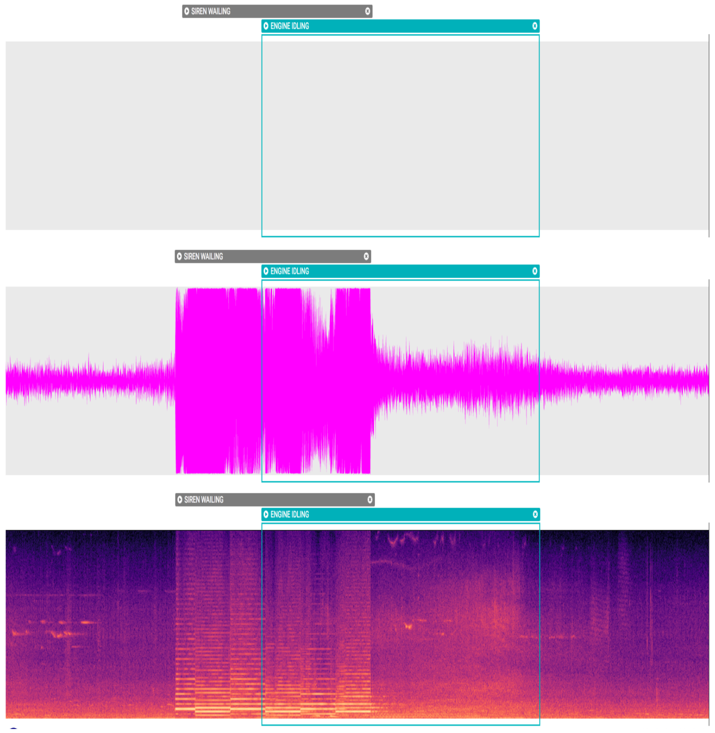
- 3 x 3 x 2 between-subjects factorial design:

visualization

NV

W

S

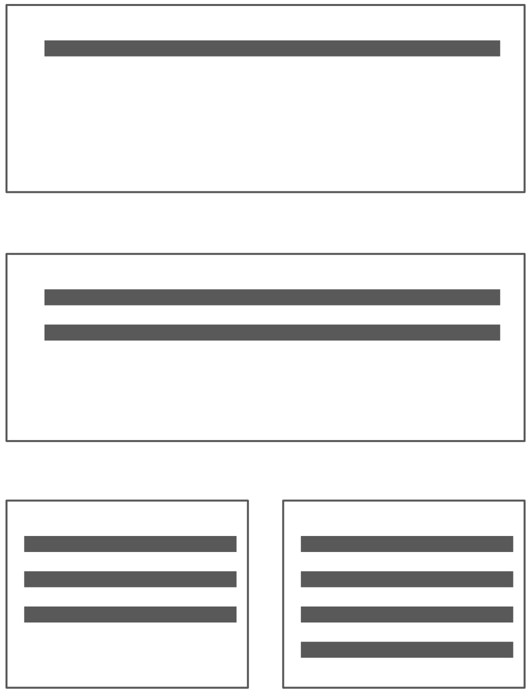


max-polyphony

M0

M1

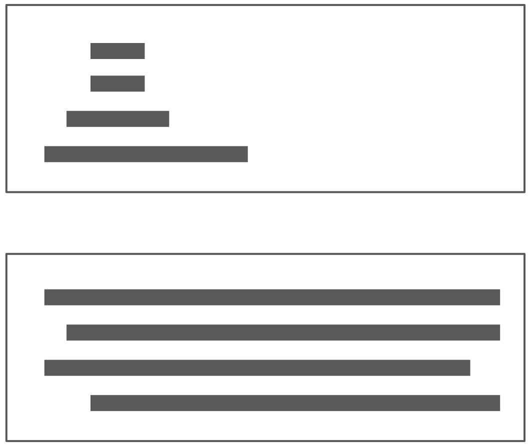
M2



gini-polyphony

G0

G1



- Soundscape examples:
  M0G0 M0G1 M2G0 M2G1

# Experiment

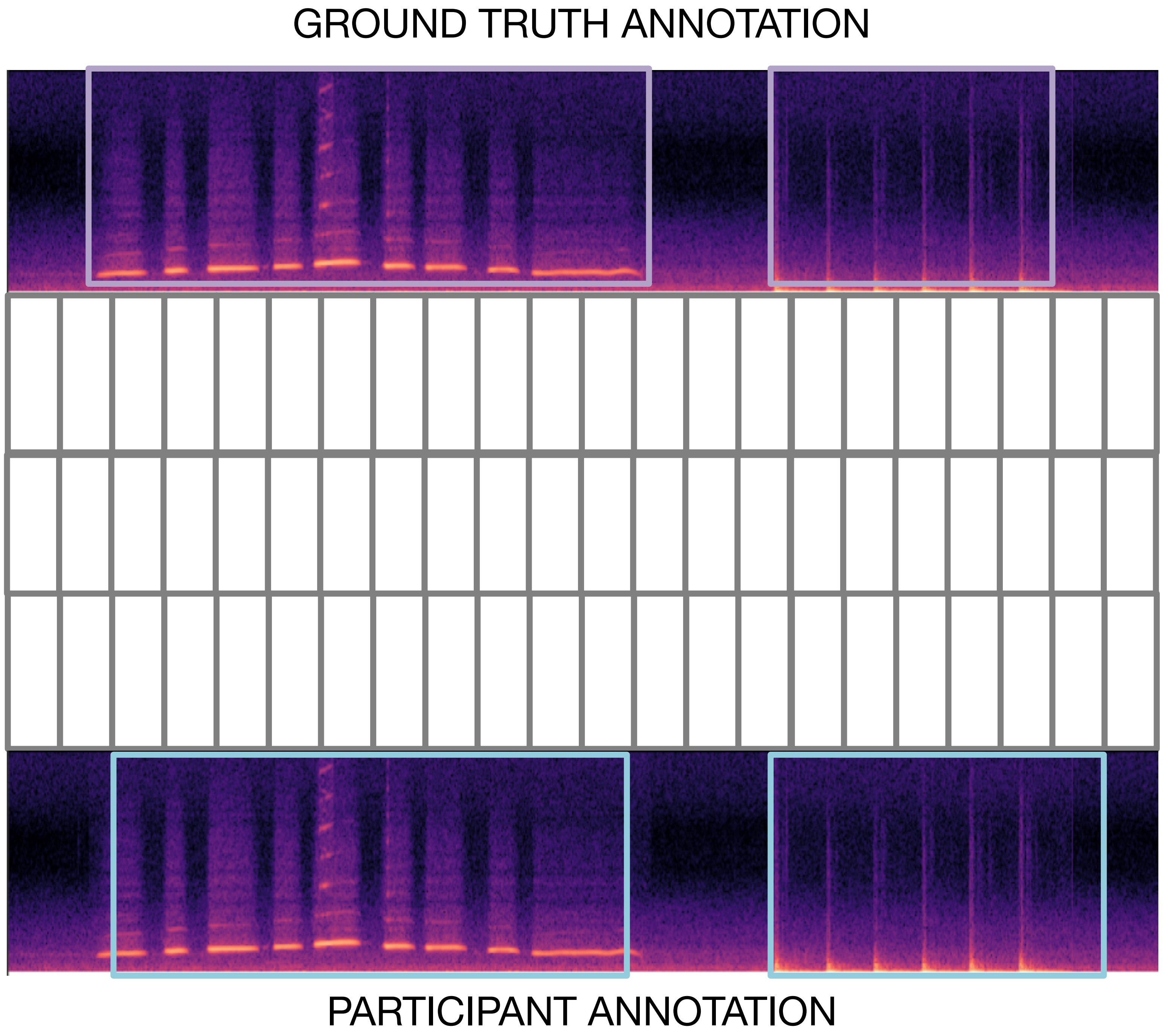
- 10 s synthesized urban soundscapes (i.e. audio stimuli)
- Classes: *car horn honking, dog barking, engine idling, gun shooting, jack hammer drilling, music playing, people shouting, people talking, siren wailing*
- 30 replications / 540 participants from Mechanical Turk
- 10 soundscapes per complexity condition (i.e. max- x gini-polyphony pair)
- Counterbalanced ordering of soundscapes
- Ran on the CrowdCurio platform

# Participant Tasks

- Hearing screening
- Pre-task questionnaire
- Tutorial video
- Practice annotation task
- Series of 10 annotation tasks
- Post-task questionnaire

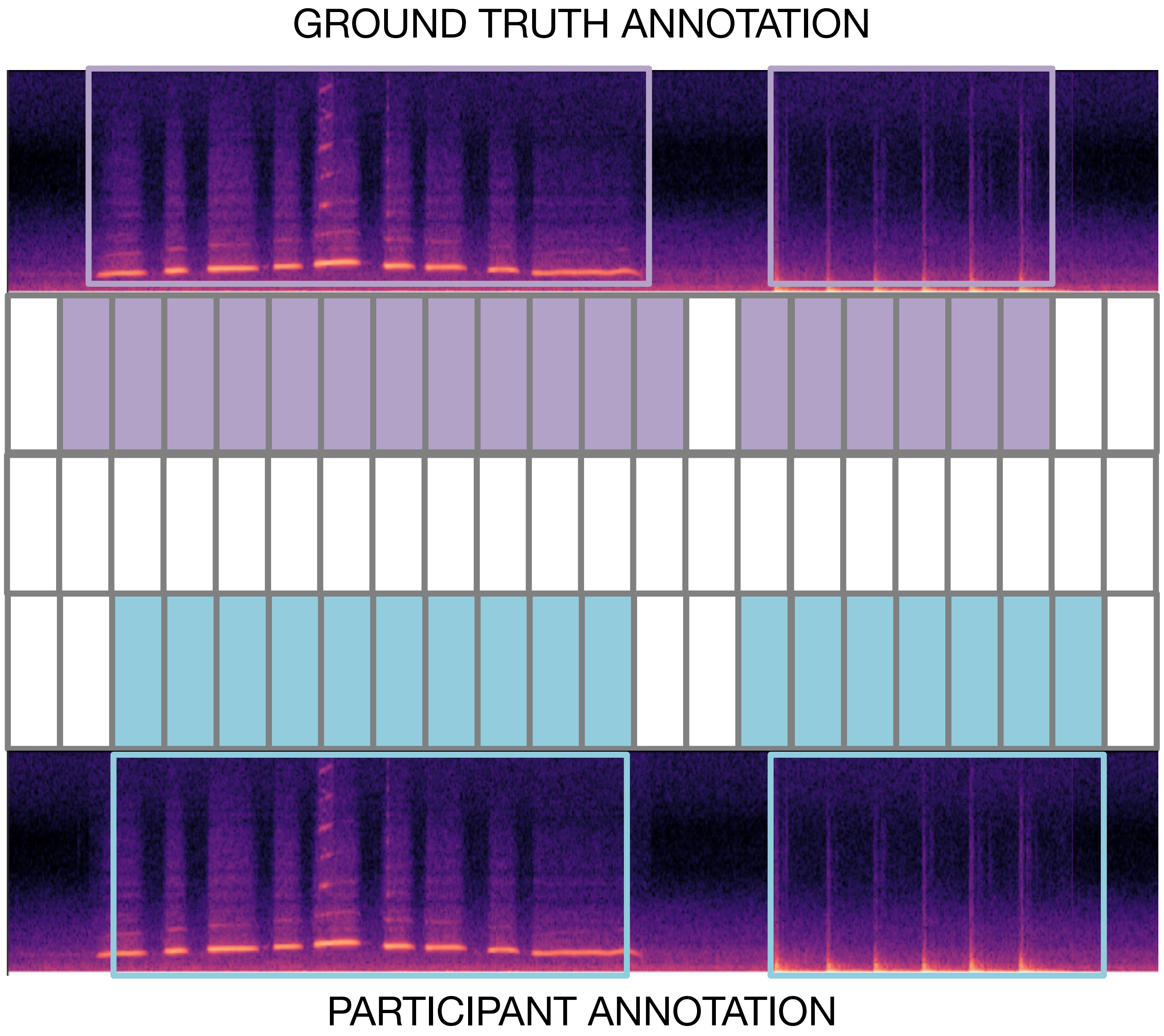
# Frame-based Evaluation

- Segment signal into 100ms frames.



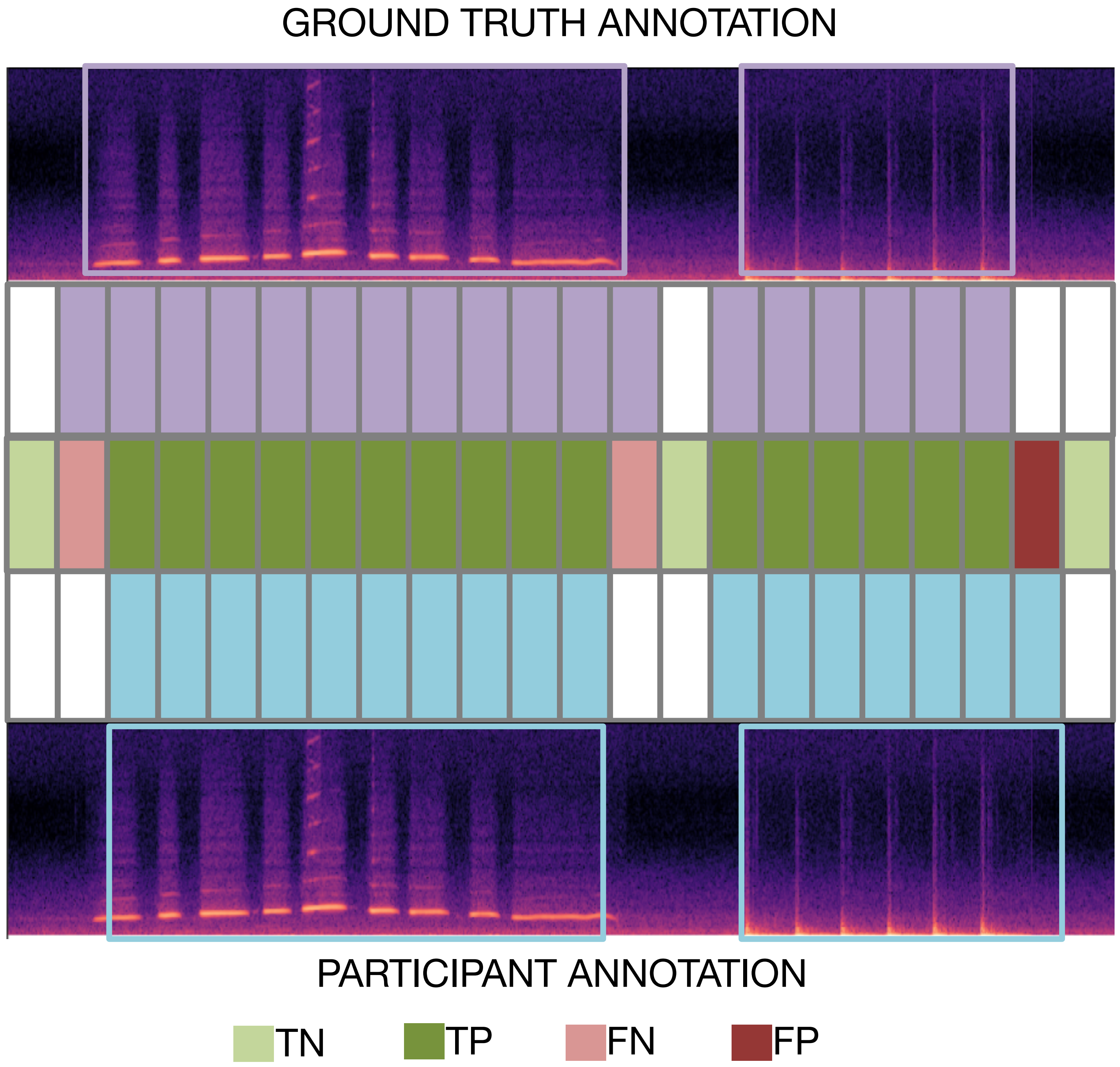
# Frame-based Evaluation

- Segment signal into 100ms frames.
- Round the annotations to the outer frame boundaries



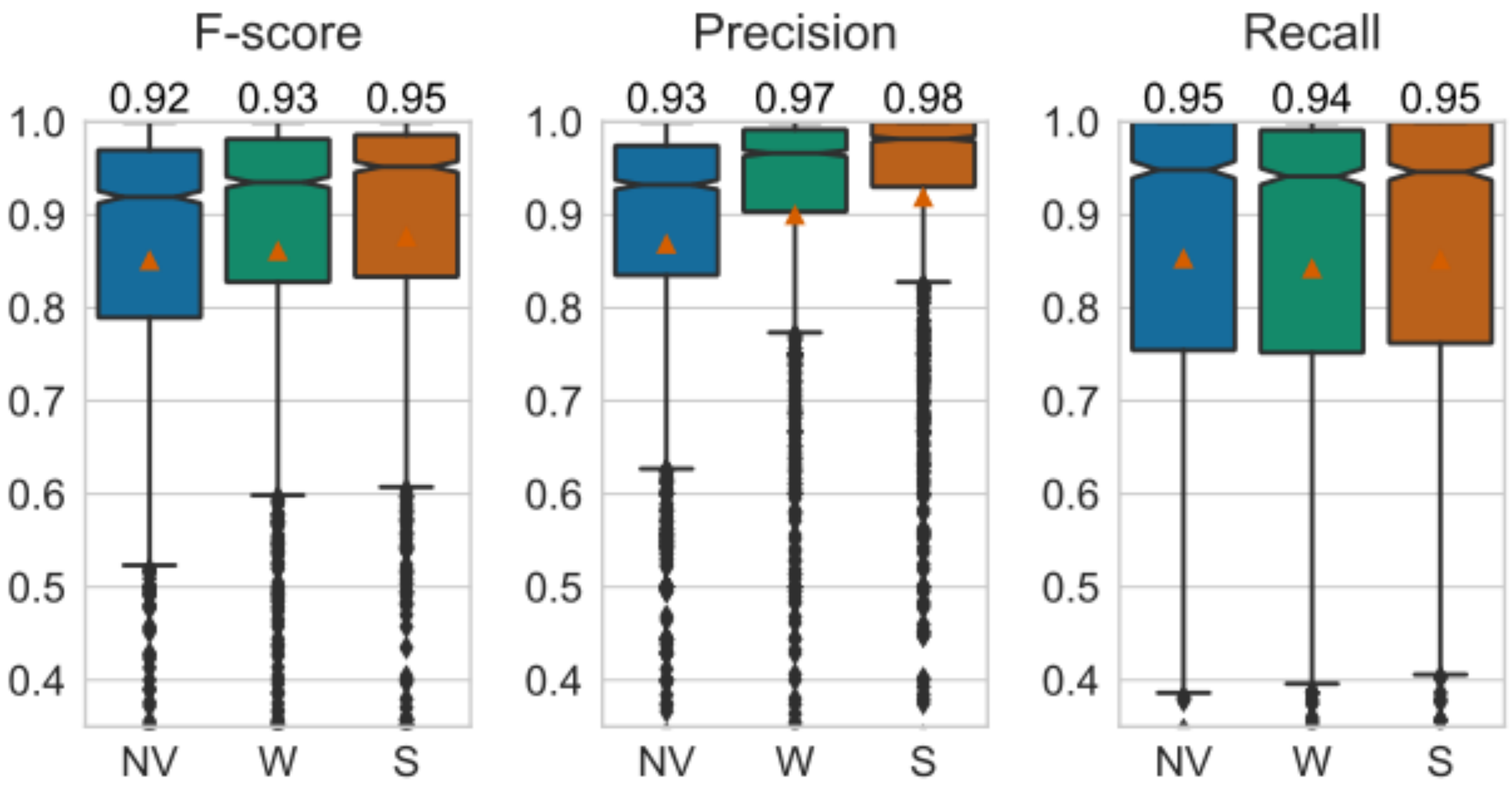
# Frame-based Evaluation

- Segment signal into 100ms frames.
- Round the annotations to the outer frame boundaries
- Count TP, FP, FN for each class and calculate precision, recall, F-score



# Results

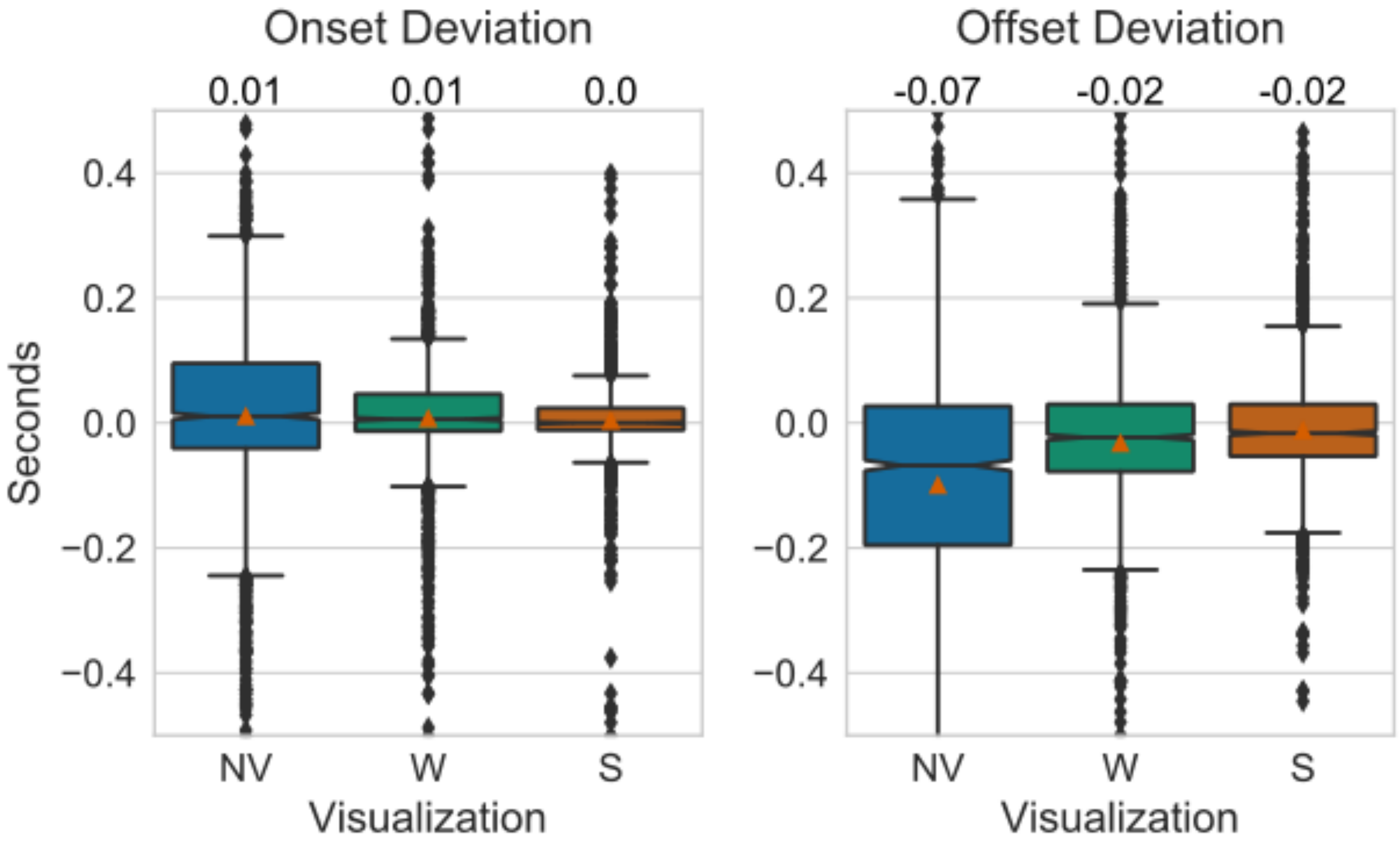
# Effect of Visualization on Quality of Annotations



NV = *no-vis*
W = *waveform*
S = *spectrogram*

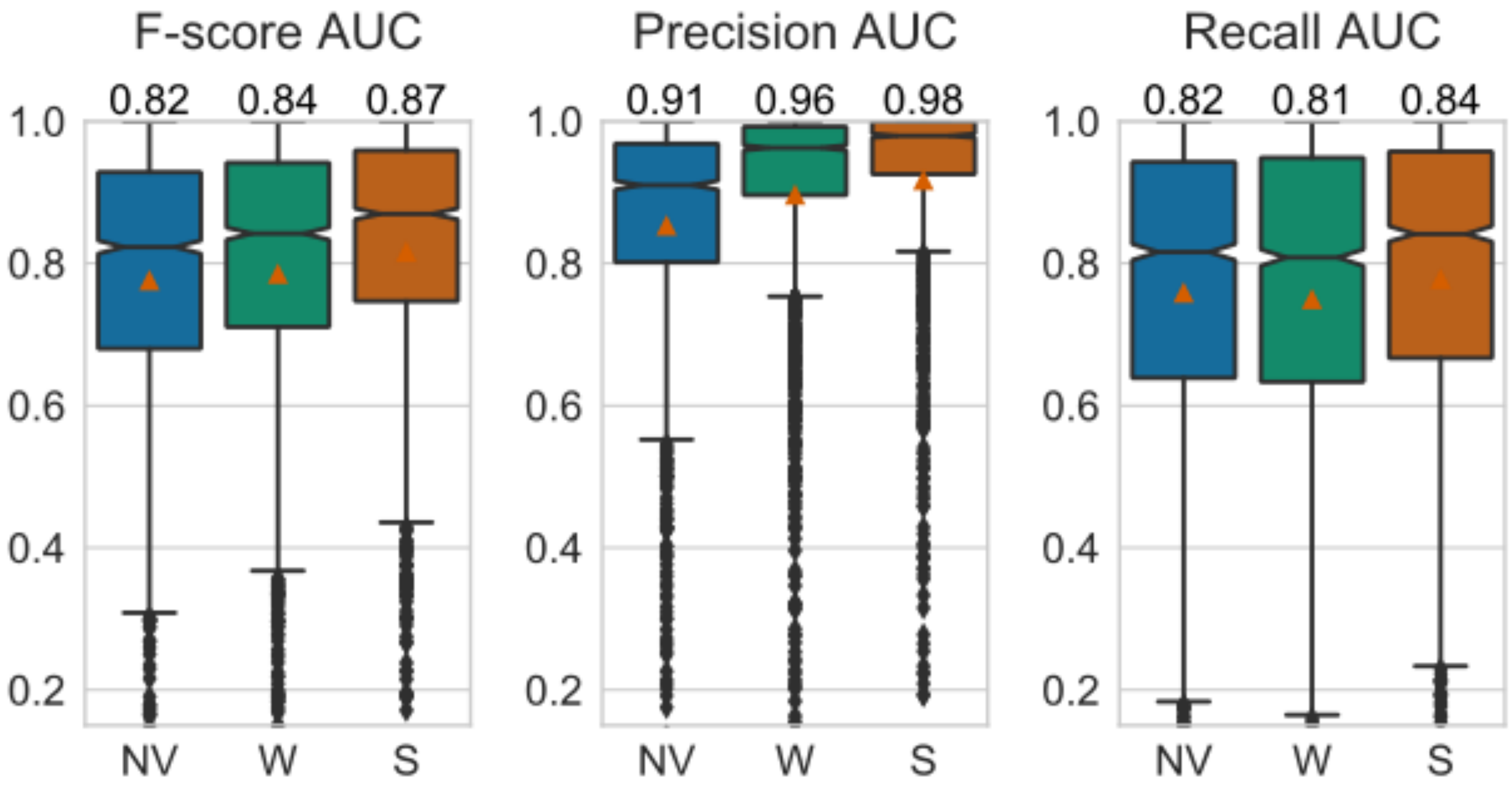
Spectrogram → higher-quality annotations

# Effect of Visualization on Quality of Annotations



NV = *no-vis*
W = *waveform*
S = *spectrogram*

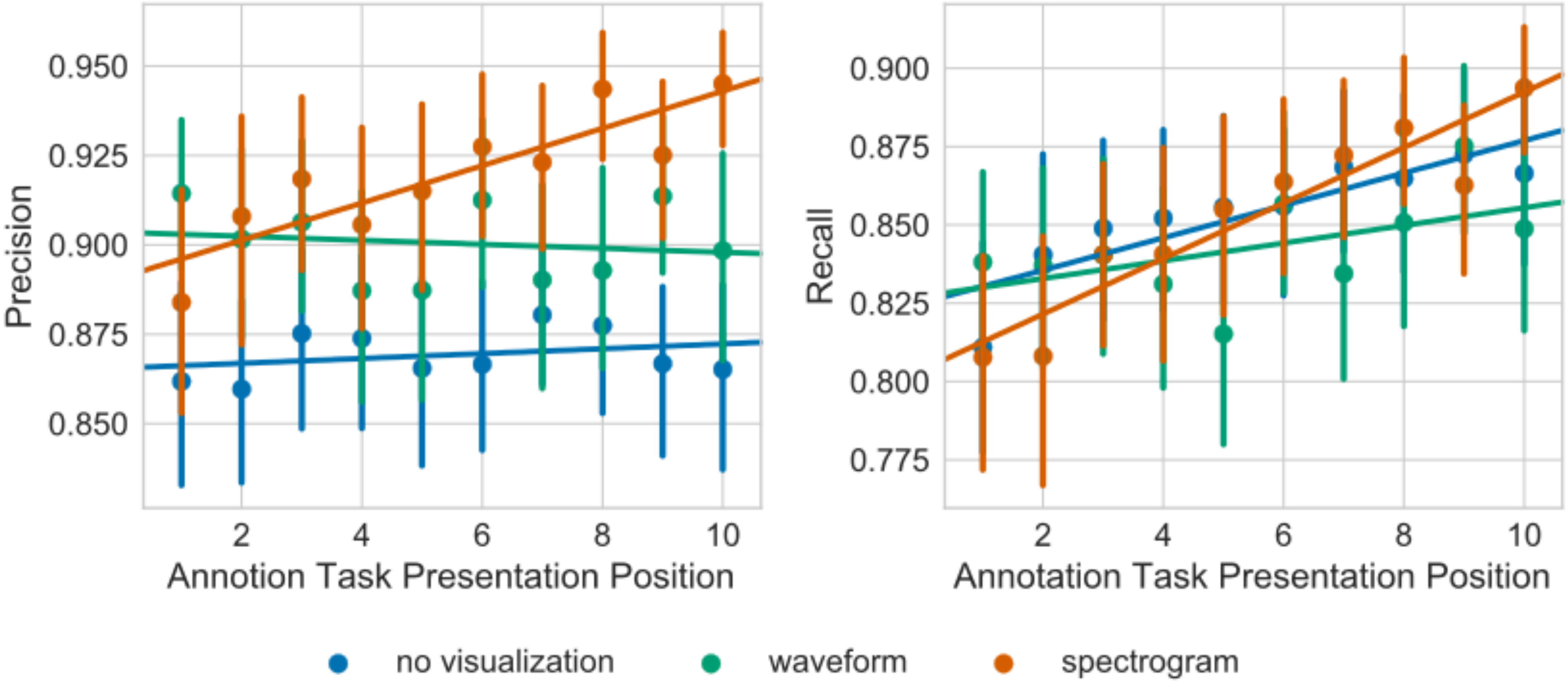
# Effect of Visualization on Quality and Speed of Annotations



NV = *no-vis*
W = *waveform*
S = *spectrogram*

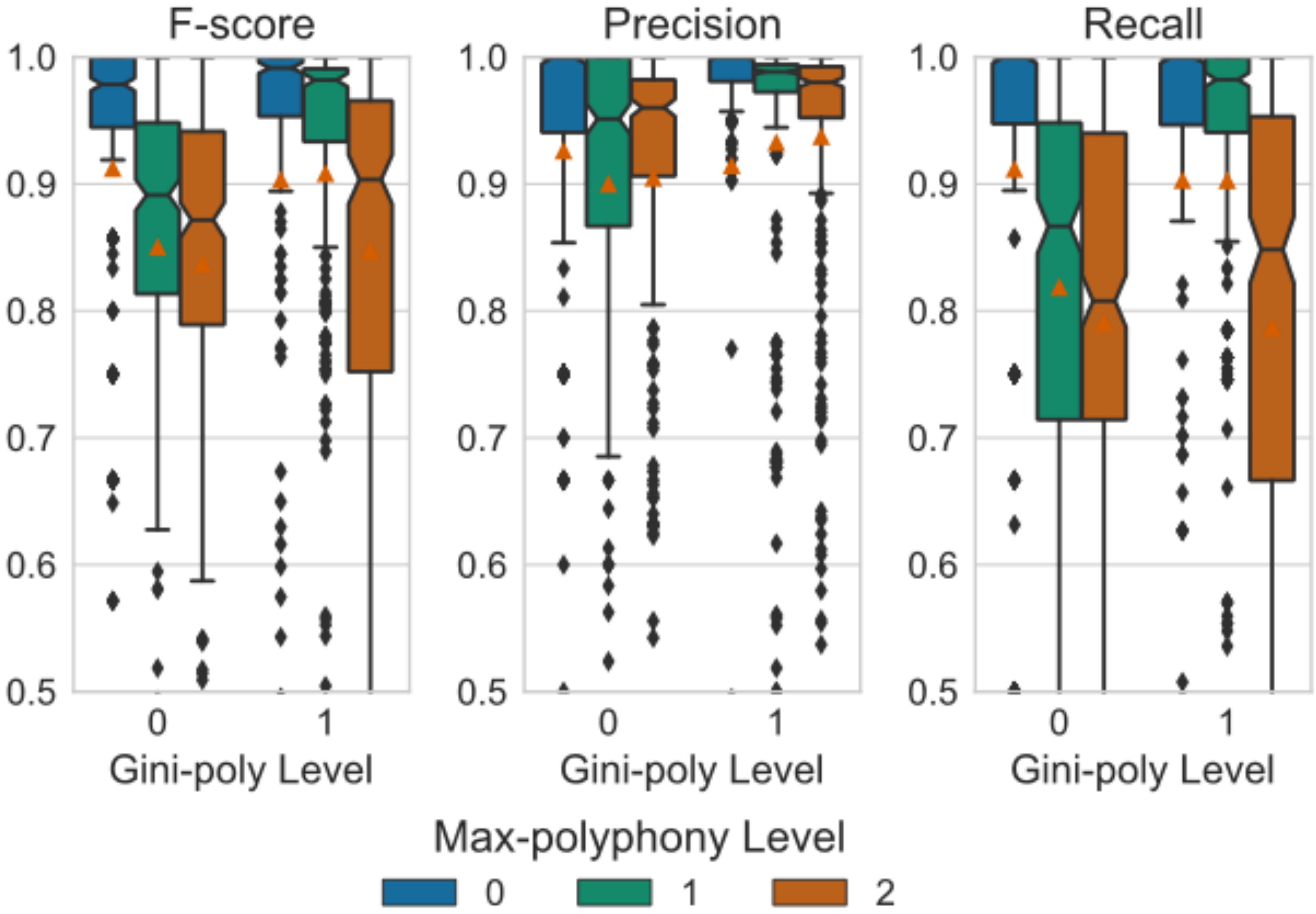
Spectrogram → higher-quality and faster annotations

# Effect of Visualization on Task Learning



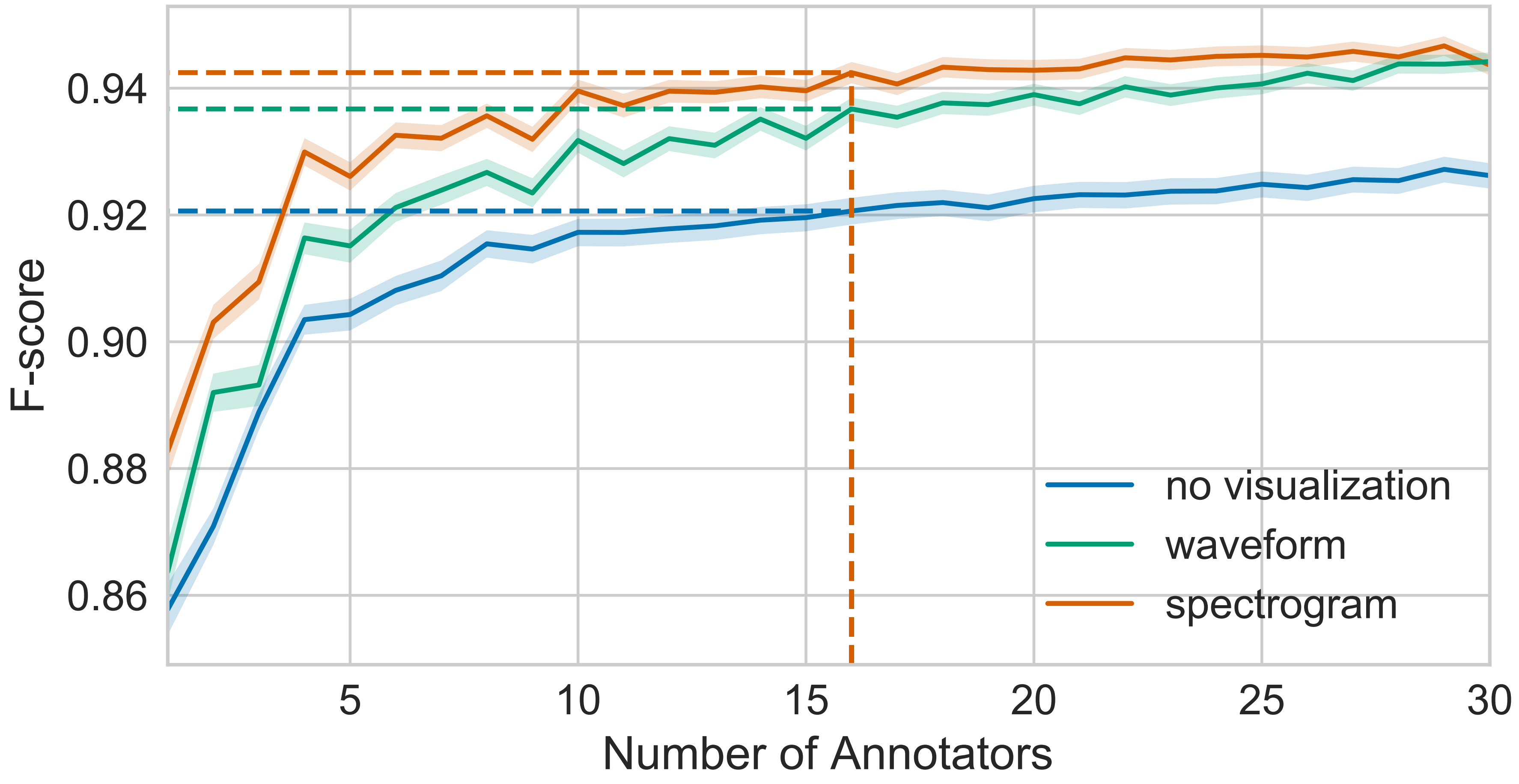
Expect even higher quality annotations after learning period

# Effect of Soundscape Complexity on Annotation Quality



Complex soundscapes → expect precise but incomplete annotations

# Effect of Number of Annotators on Aggregate Annotation Quality



16 annotators captured 90% of gain in annotation quality, but
5 annotators is reasonable choice with respect to cost/quality trade-off

# Takeaways

- Spectrogram → higher-quality and faster annotations
- Expect even higher quality annotations after learning period
- Complex soundscapes → expect precise but incomplete annotations
- 5 annotators is reasonable choice with respect to cost/quality trade-off

SONYC: wp.nyu.edu/sonyc

Audio Annotator: github.com/CrowdCurio/audio-annotator

Scaper: github.com/justinsalamon/scaper

CrowdCurio: crowdcurio.com

Data: https://doi.org/10.5281/zenodo.887924