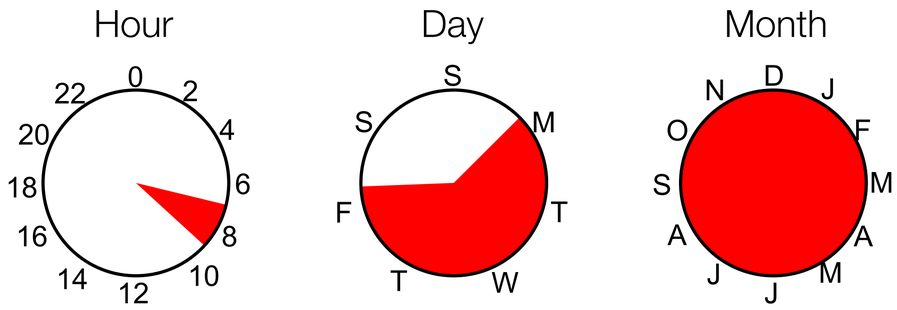
DCASE 2020 Task 5:
Urban Sound Tagging with Spatiotemporal Context

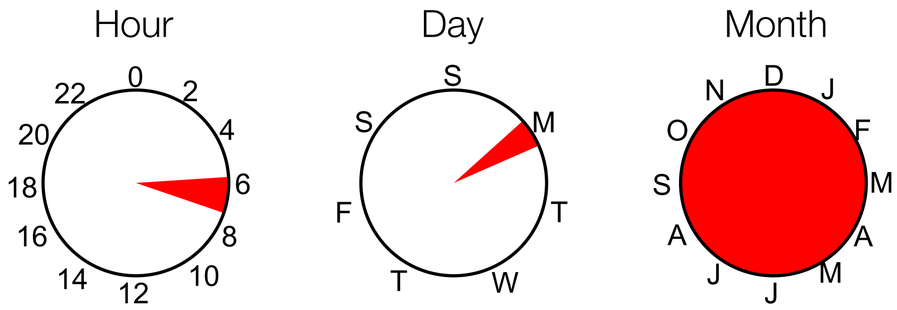
Mark Cartwright[1], Jason Cramer[1], Ana Elisa Mendez Mendez[1], Yu Wang[1],
Ho-Hsiang Wu[1], Vincent Lostanlen[2], Magdalena Fuentes[1], Justin Salamon[3], Juan Pablo Bello[1]

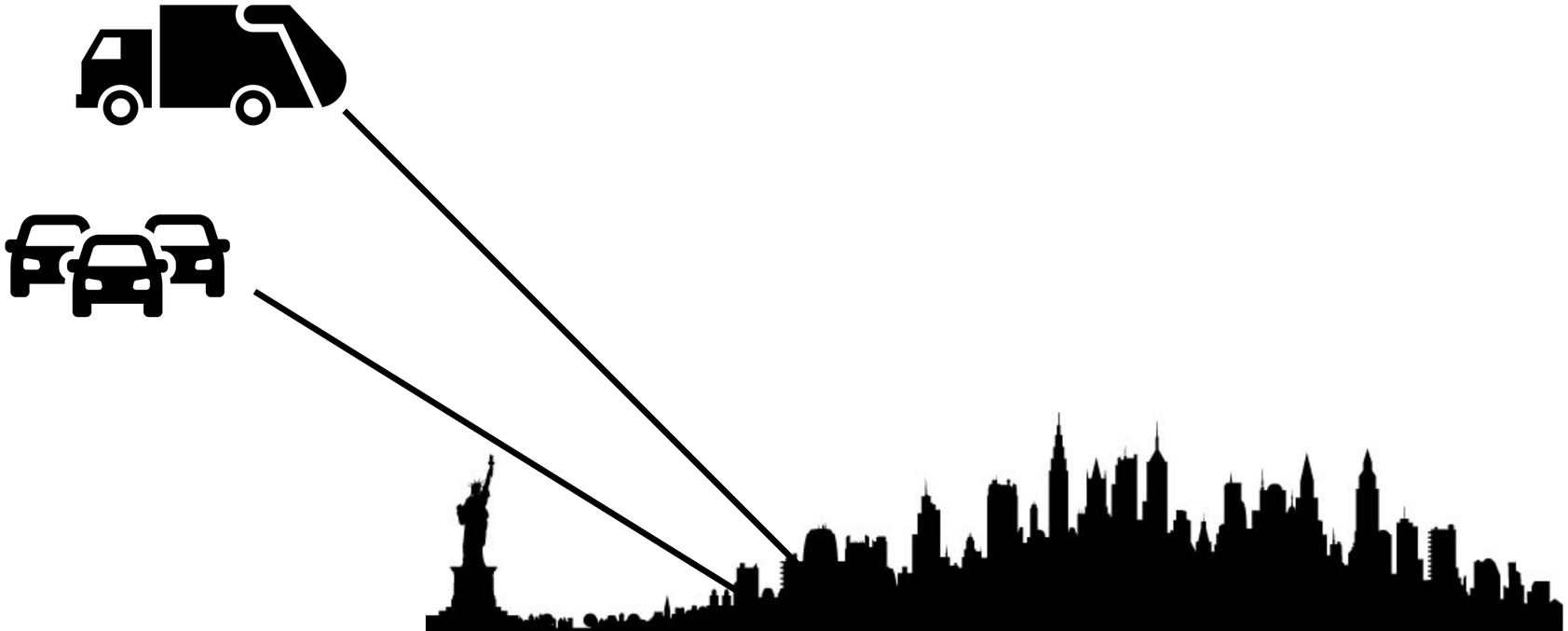
1. New York University Music and Audio Research Lab
2. CNRS
3. Adobe Research

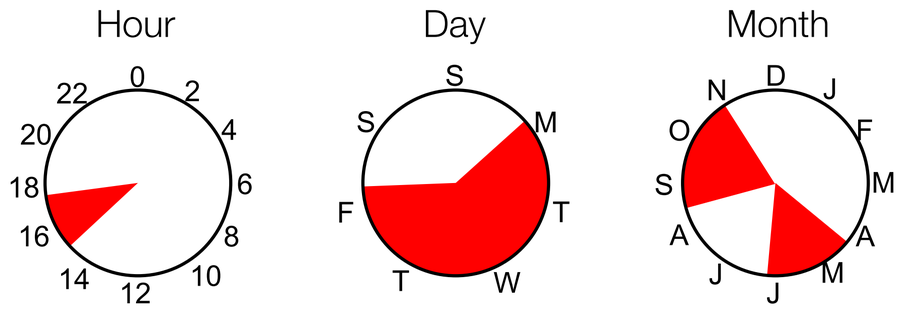
In many real-world machine-listening applications,
the recordings have context.

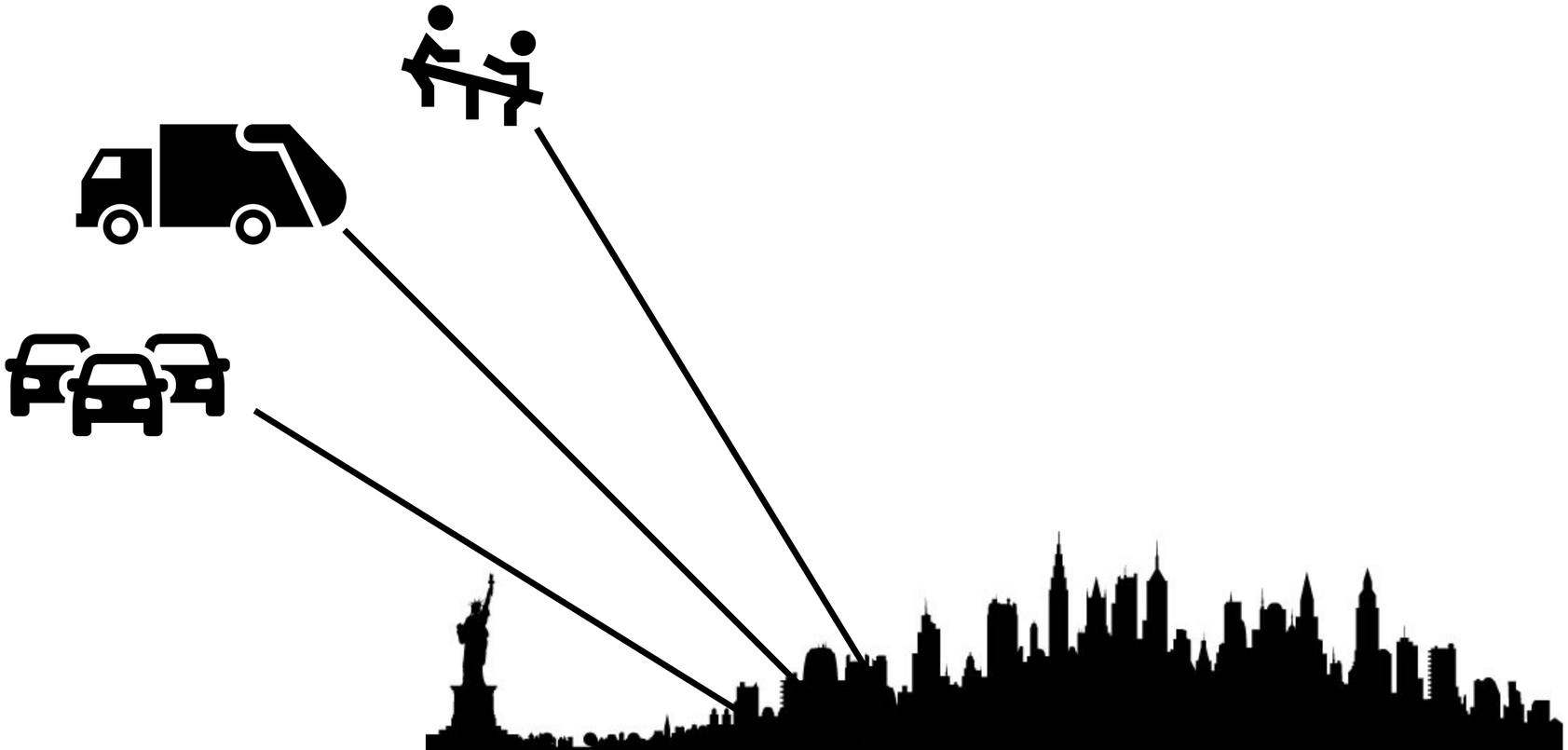

H&M
H&M

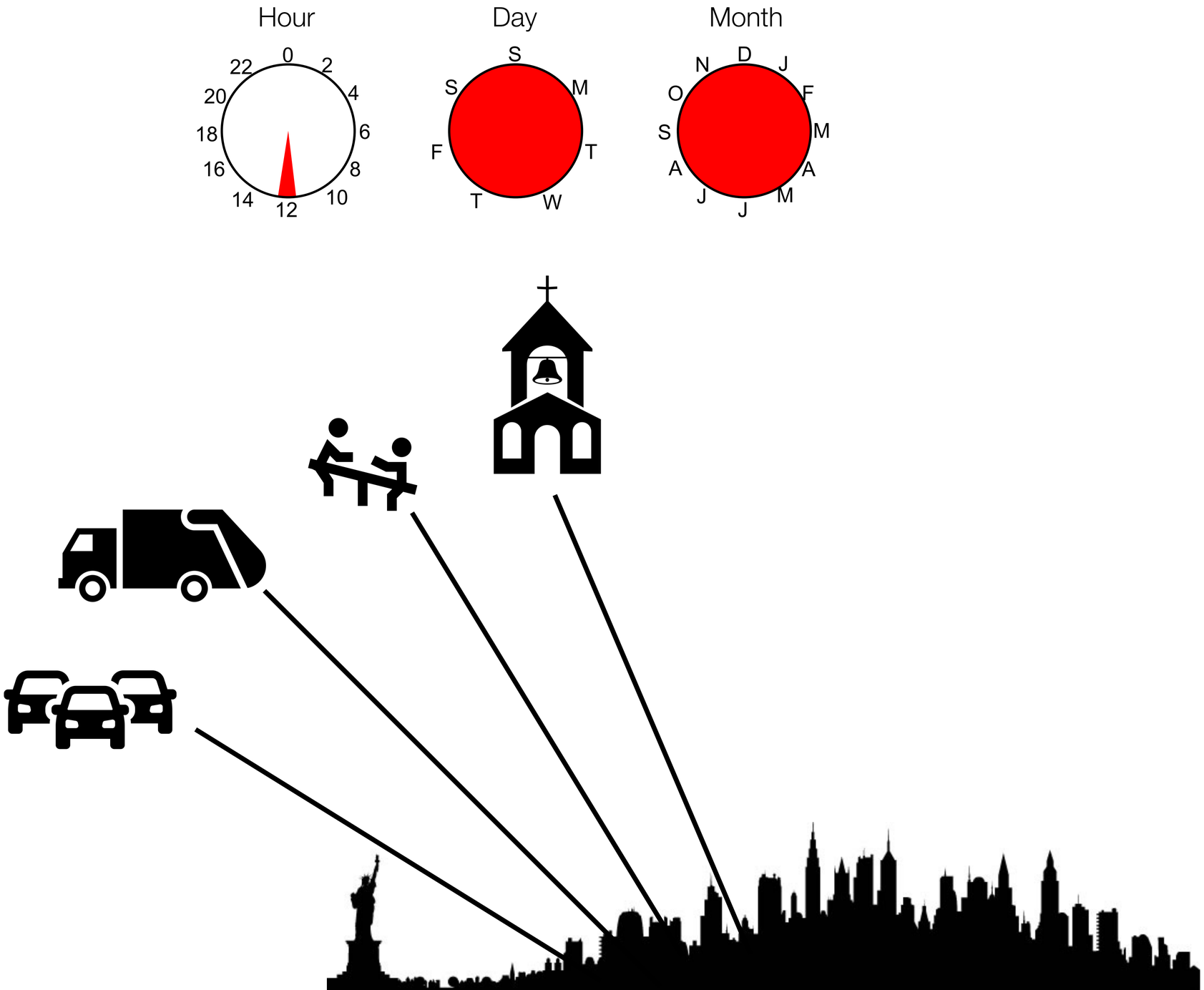

Hour
0
2
4
6
8
10
12
14
16
18
20
22
Day
S
M
T
W
T
F
S
Month
D
J
F
M
A
M
J
J
A
S
O
N

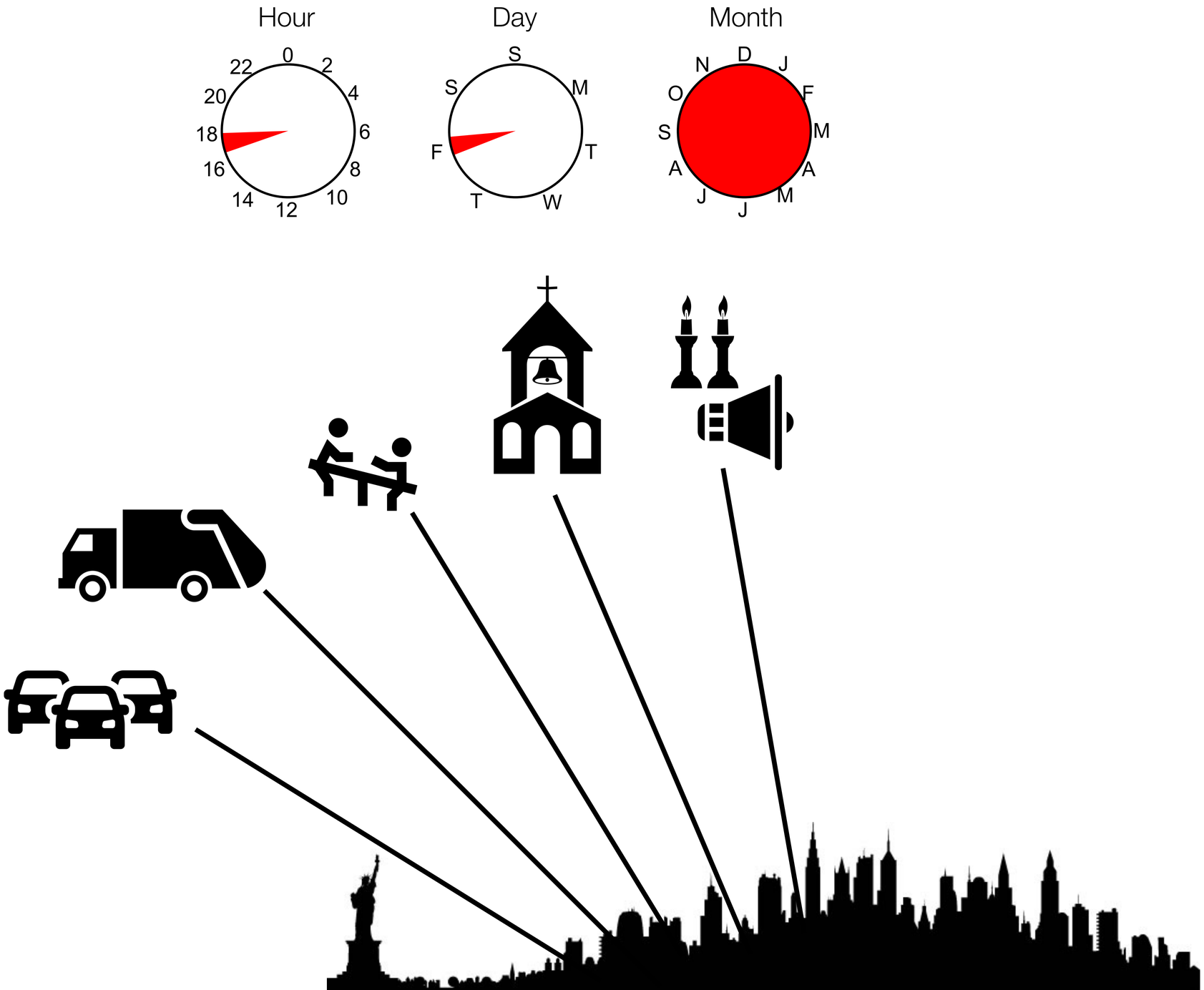

Hour
0
2
4
6
8
10
12
14
16
18
20
22
Day
S
M
T
W
T
F
S
Month
D
J
F
M
A
M
J
J
A
S
O
N

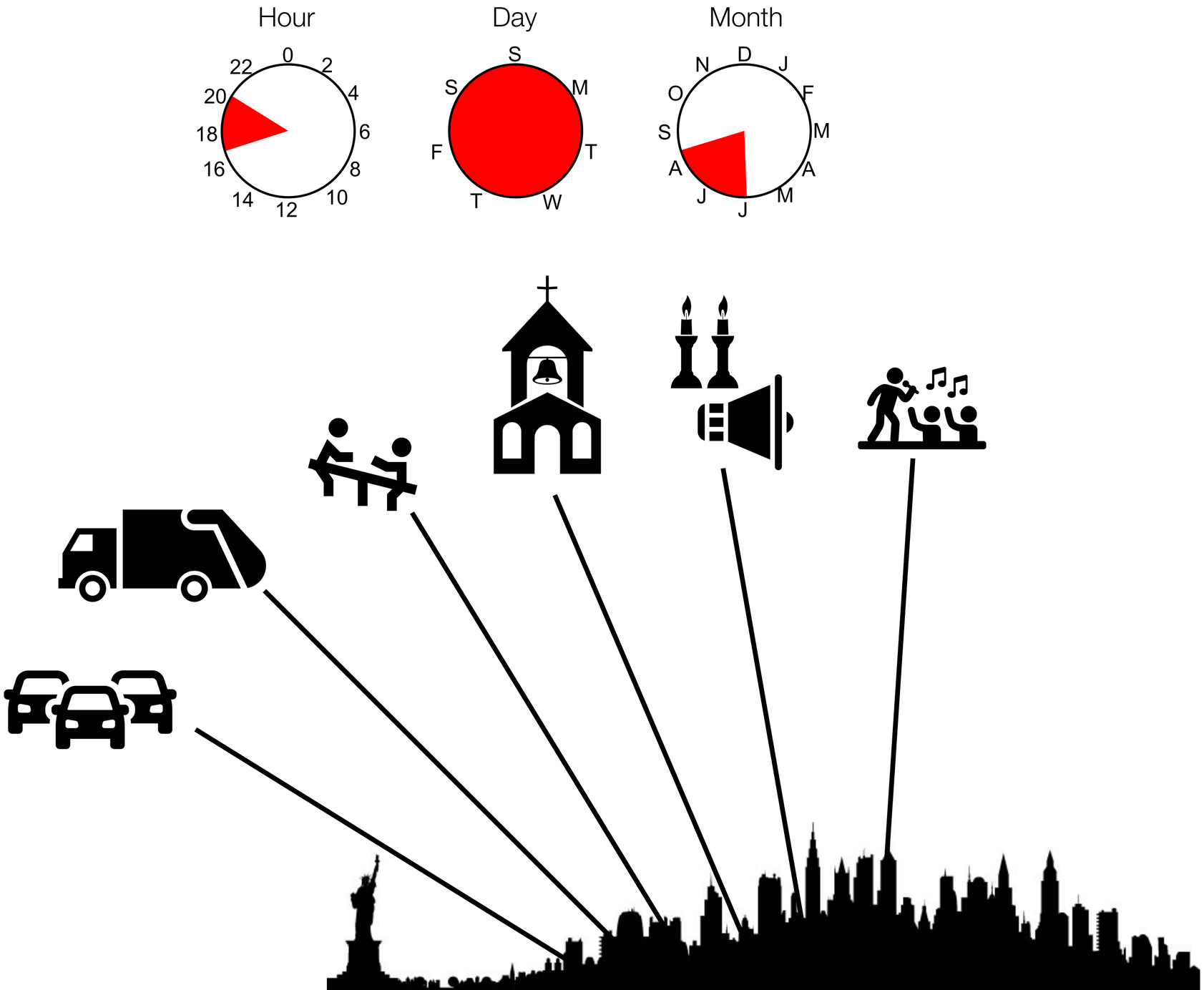


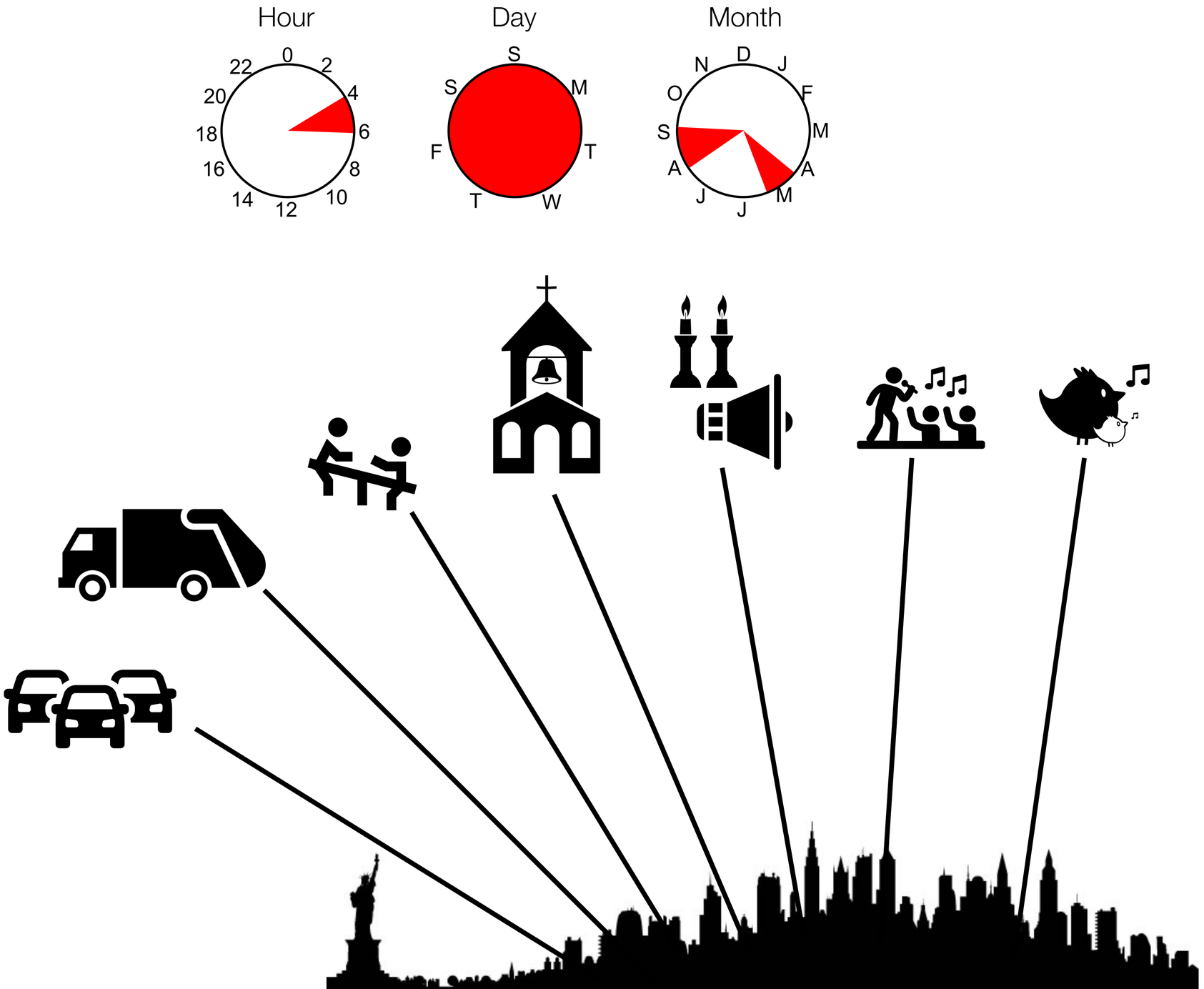

Hour
0
2
4
6
8
10
12
14
16
18
20
22
Day
S
M
T
W
T
F
S
Month
J
F
M
A
M
J
J
A
S
O
N
D



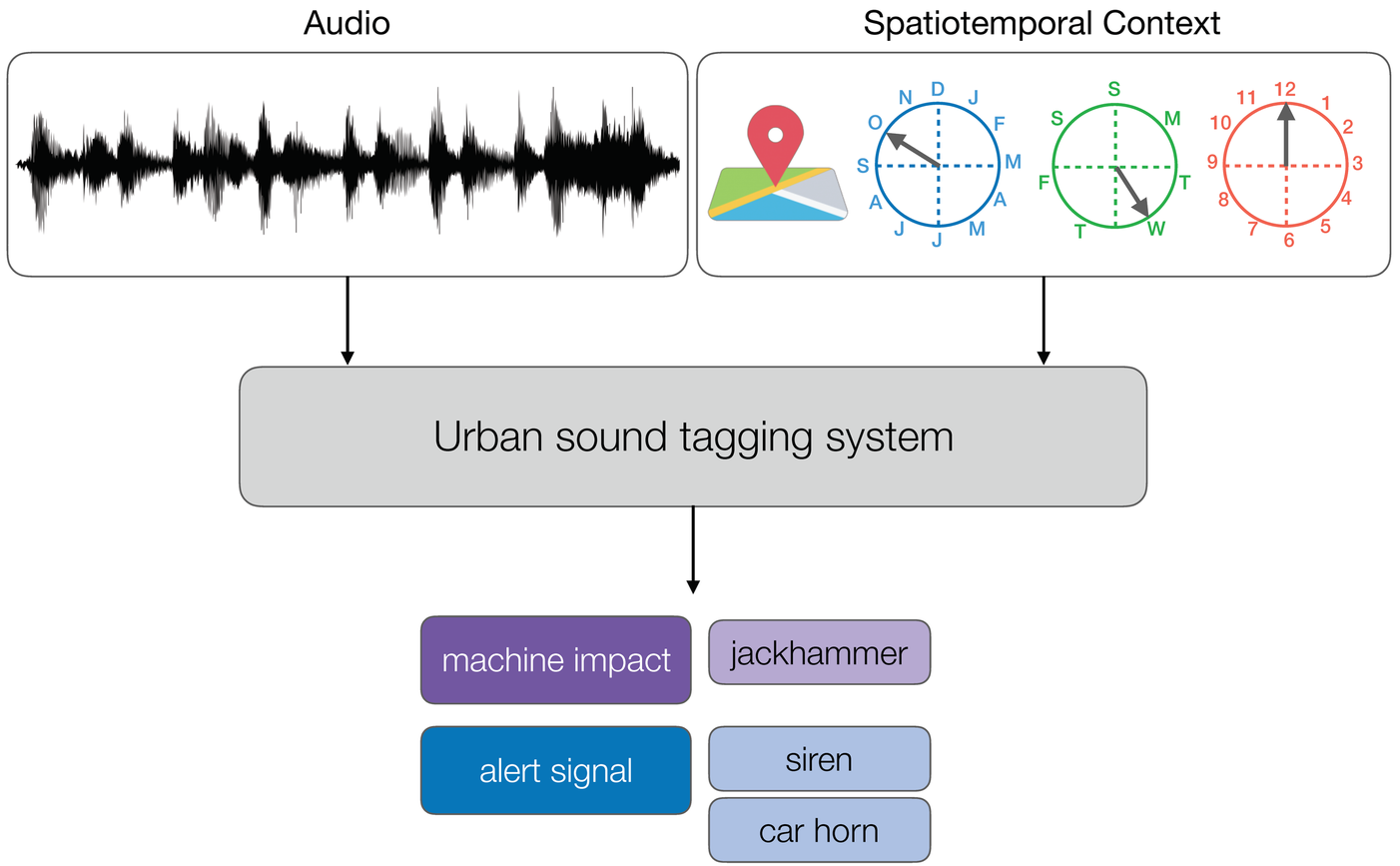

Hour
0
2
4
6
8
10
12
14
16
18
20
22
Day
S
M
T
W
T
F
S
Month
D
J
F
M
A
M
J
J
A
S
O
N


Hour
0
2
4
6
8
10
12
14
16
18
20
22
Day
S
M
T
W
T
F
S
Month
D
J
F
M
A
M
J
J
A
S
O
N


Hour
0
2
4
6
8
10
12
14
16
18
20
22
Day
S
M
T
W
T
F
S
Month
J
F
M
A
M
J
J
A
S
O
N
D


Hour
0
2
4
6
8
10
12
14
16
18
20
22
Day
S
M
T
W
T
F
S
Month
J
F
M
A
M
J
J
A
S
O
N
D

Can spatiotemporal metadata help in urban sound tagging?

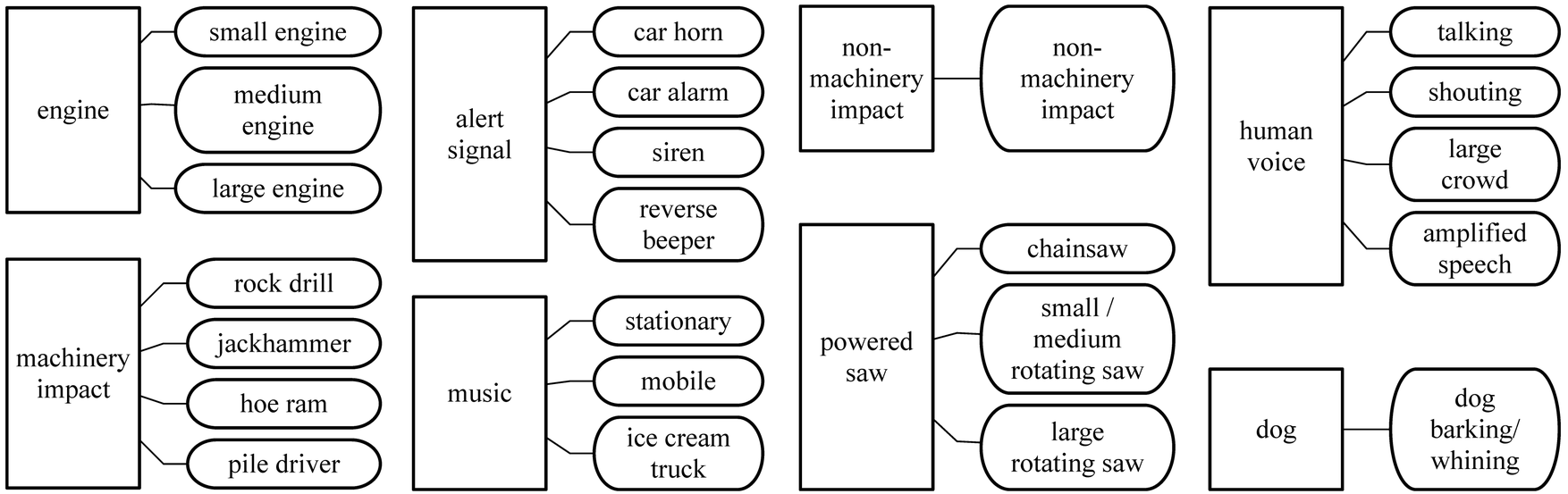


SONYC-UST-V2 dataset

- 18510 recordings (10 s) from 56 sensors in our Sounds of New York City (SONYC) sensor network (2016-2019)

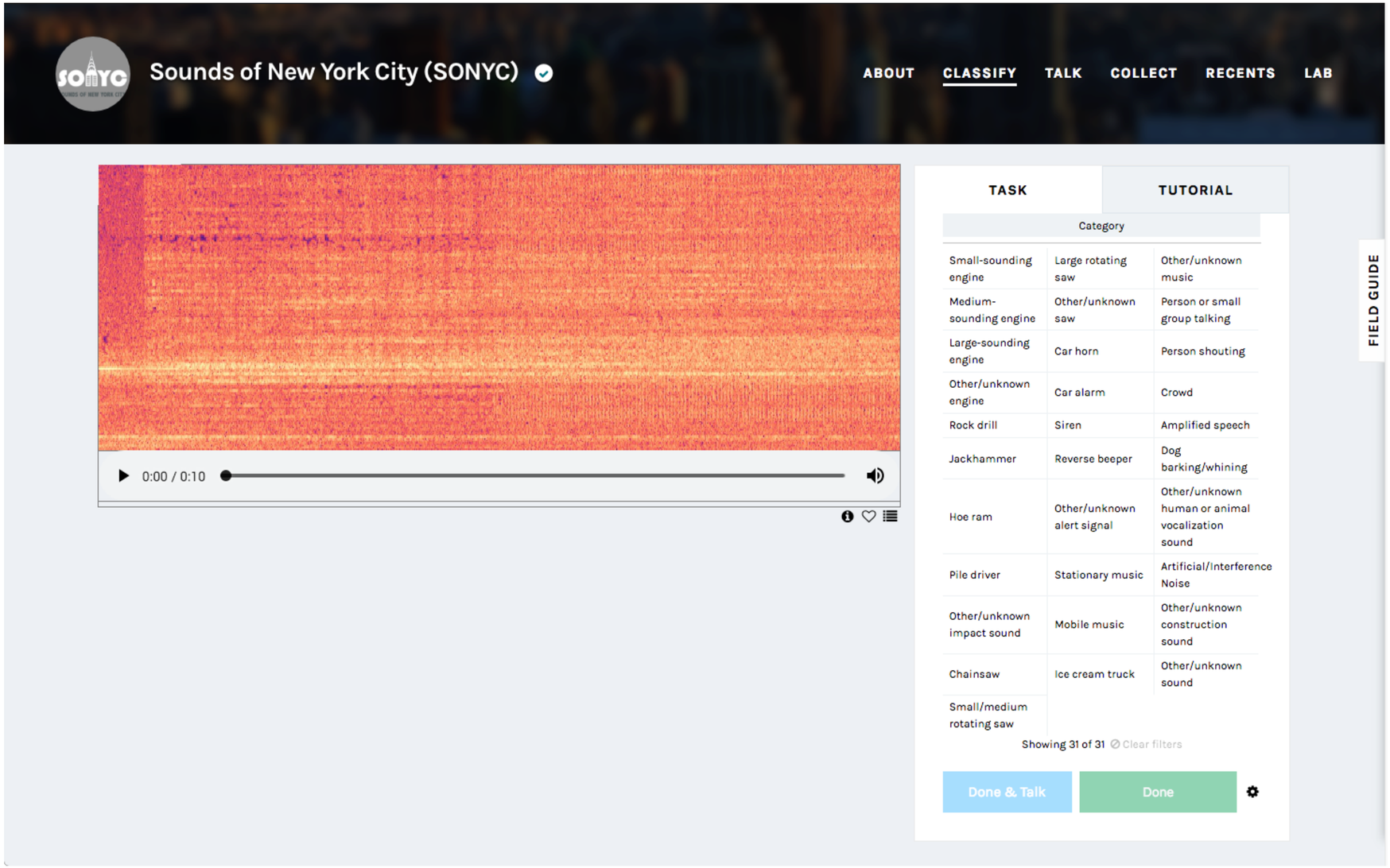
- Annotated with 23 fine-level urban sound tags from 8 coarse-level categories

SONYC UST Taxonomy



SONYC-UST-V2 dataset

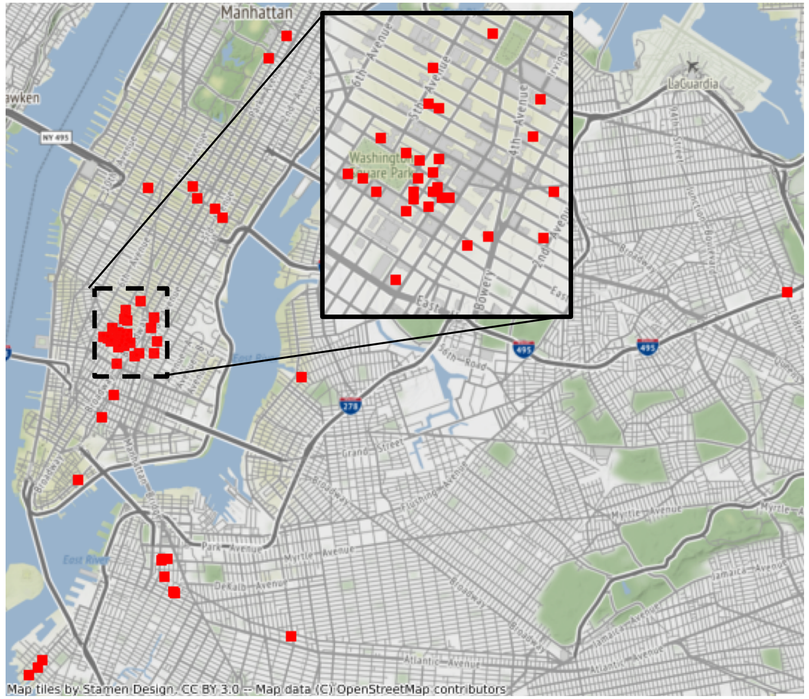
- 18510 recordings (10 s) from 56 sensors in our Sounds of New York City (SONYC) sensor network (2016-2019)
- Annotated with 23 fine-level urban sound tags from 8 coarse-level categories
- Each recording annotated by 3 volunteers on Zooniverse


Sounds of New York City (SONYC)
ABOUT
CLASSIFY
TALK
COLLECT
RECENTS
LAB
0:00 / 0:10
TASK
TUTORIAL
Category
Small-sounding engine
Large rotating saw
Other/unknown music
Medium-sounding engine
Other/unknown saw
Person or small group talking
Large-sounding engine
Car horn
Person shouting
Other/unknown engine
Car alarm
Crowd
Rock drill
Siren
Amplified speech
Jackhammer
Reverse beeper
Dog barking/whining
Hoe ram
Other/unknown alert signal
Other/unknown human or animal vocalization sound
Pile driver
Stationary music
Artificial/Interference Noise
Other/unknown impact sound
Mobile music
Other/unknown construction sound
Chainsaw
Ice cream truck
Other/unknown sound
Small/medium rotating saw
Showing 31 of 31
Clear filters
Done & Talk
Done
FIELD GUIDE

SONYC-UST-V2 dataset

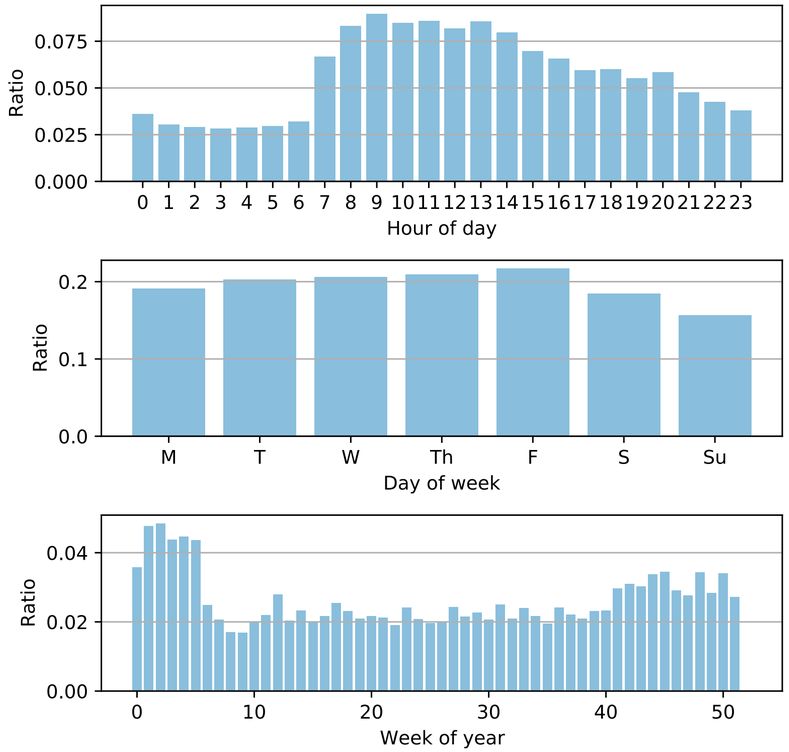
- 18510 recordings (10 s) from 56 sensors in our Sounds of New York City (SONYC) sensor network (2016-2019)
- Annotated with 23 fine-level urban sound tags from 8 coarse-level categories
- Each recording annotated by 3 volunteers on Zooniverse
- A subset of 1380 have “verified” annotations by the SONYC team
- **All recordings include spatiotemporal metadata**

SONYC-UST-V2 spatial distribution



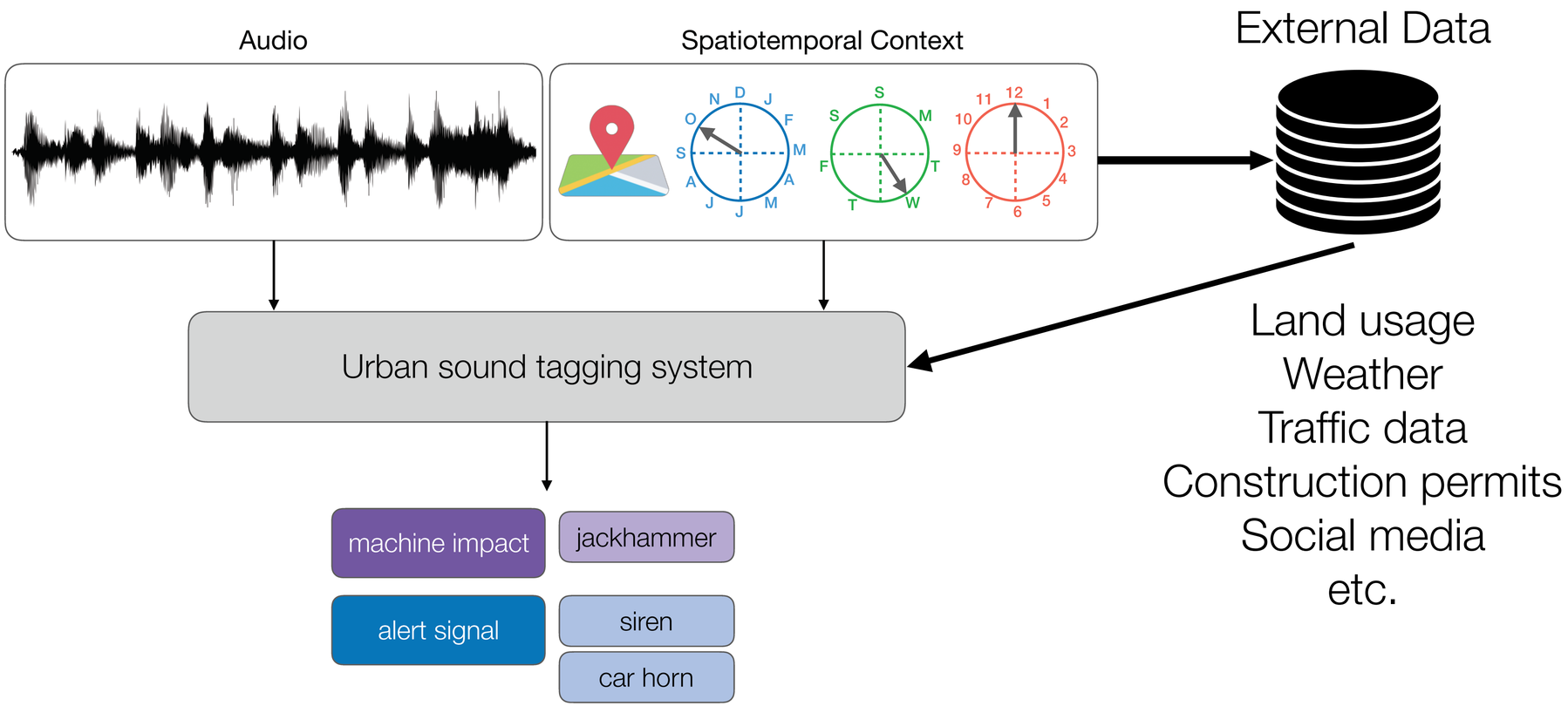
- Spatial location quantized to the city block level

SONYC-UST-V2 temporal distribution



- Temporal location quantized to the hour, expressed as:
  - hour of the day
  - day of the week
  - week of the year,
  - year

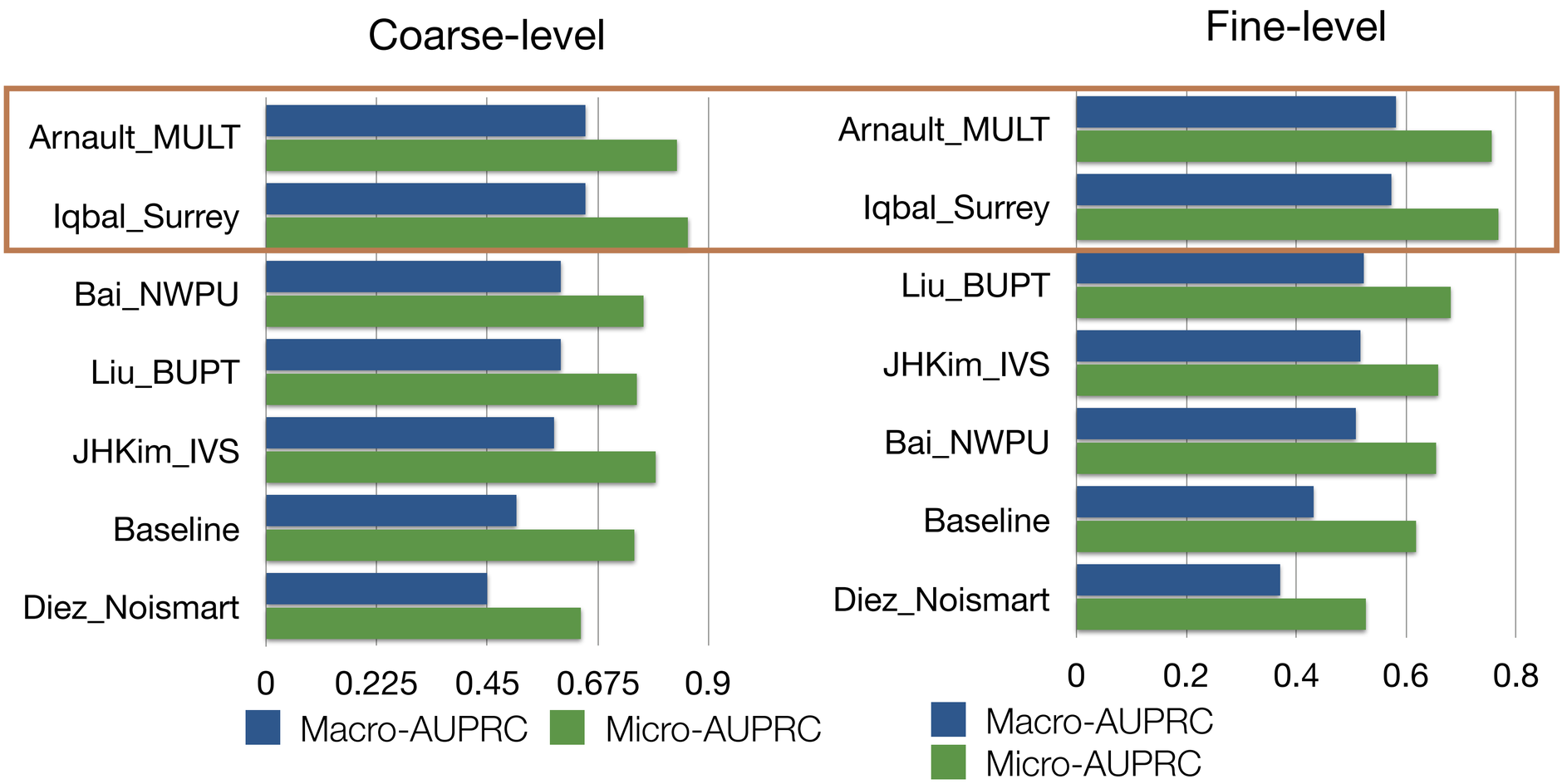
Encouraged the use of external data



Results

- 22 systems from 6 teams
- Most systems used a CNN/CRNN with spectrogram inputs
- Most systems increased the training data with augmentation techniques
  (e.g. mixup, scaling, shifting, masking)
- Some systems also used models pre-trained using external data
  (e.g. AudioSet, ImageNet)
- All the systems used temporal context (e.g., week of year, day of week, hour of day)
- Most systems used spatial context as well (e.g., latitude, longitude)
- **None** used the spatiotemporal context to query external data sources as input
  (e.g. weather, traffic, land usage, etc.)

Results


Coarse-level
Arnault_MULT
Iqbal_Surrey
Bai_NWPU
Liu_BUPT
JHKim_IVS
Baseline
Diez_Noismart
0
0.225
0.45
0.675
0.9
Macro-AUPRC
Micro-AUPRC
Fine-level
Arnault_MULT
Iqbal_Surrey
Liu_BUPT
JHKim_IVS
Bai_NWPU
Baseline
Diez_Noismart
0
0.2
0.4
0.6
0.8
Macro-AUPRC
Micro-AUPRC