Acoustic Intelligence in Machines

Anurag Kumar

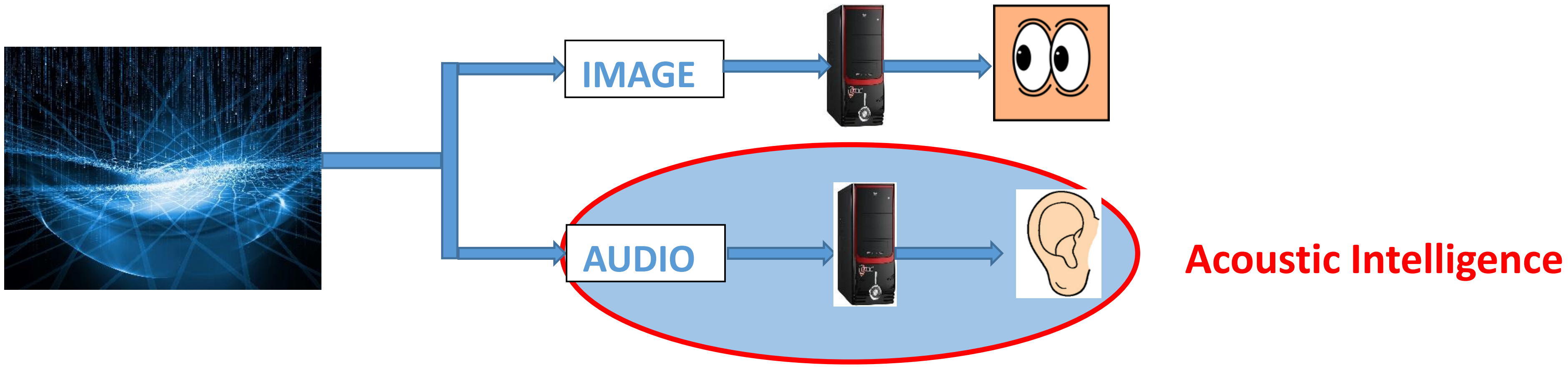
Language Technologies Institute, School of Computer Science
Carnegie Mellon University

If you are using/referencing any material from this presentation please cite my PhD thesis document

Machine Intelligence

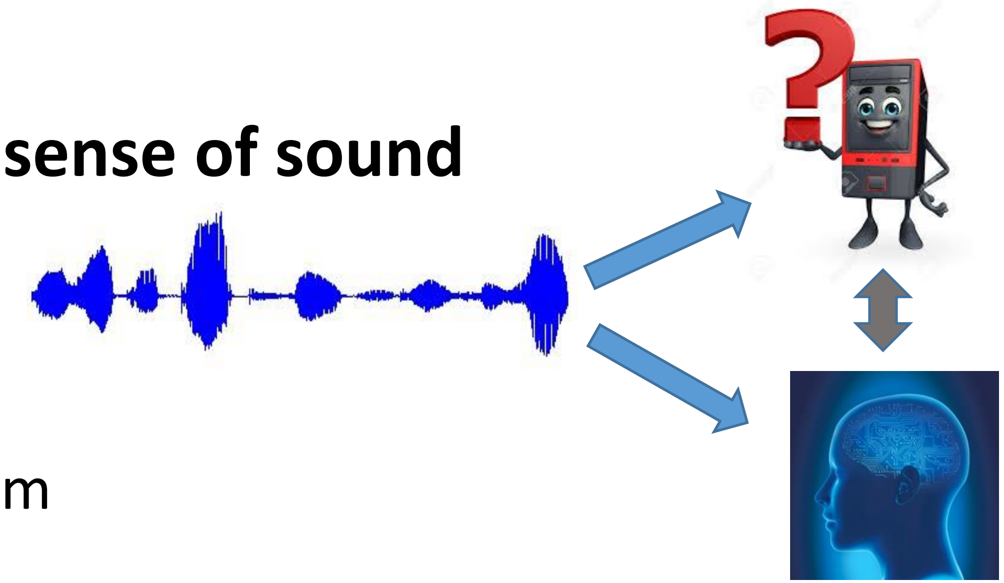
- Machine having capabilities to observe, understand, interpret and respond to the environment like humans do

  - Vision and Sound



Acoustic Intelligence

- **Machines should understand and make sense of sound**
  - know about various sounds
  - know or discover relationships between them
  - be able to recognize , categorize and index them



- Critical to a variety of applications

Acoustic Intelligence

Acoustic Intelligence

Acoustic Intelligence

Acoustic Intelligence

Acoustic Intelligence

Acoustic Intelligence - Problems

- Large number of Sounds
  - Completely Overlapping yet distinguishable by humans

- Unstructured
  - unlike speech from vocal cords

- No Language

Acoustic Intelligence

- Description, Interpretation, Saliency
  - Vision – You say what you see
  - Sound – *I heard car sound* – Means ???

- **Fundamental difference is that visual objects are formed from presence of *physical objects*, while sound objects result from *their actions***

Acoustic Intelligence in Machines

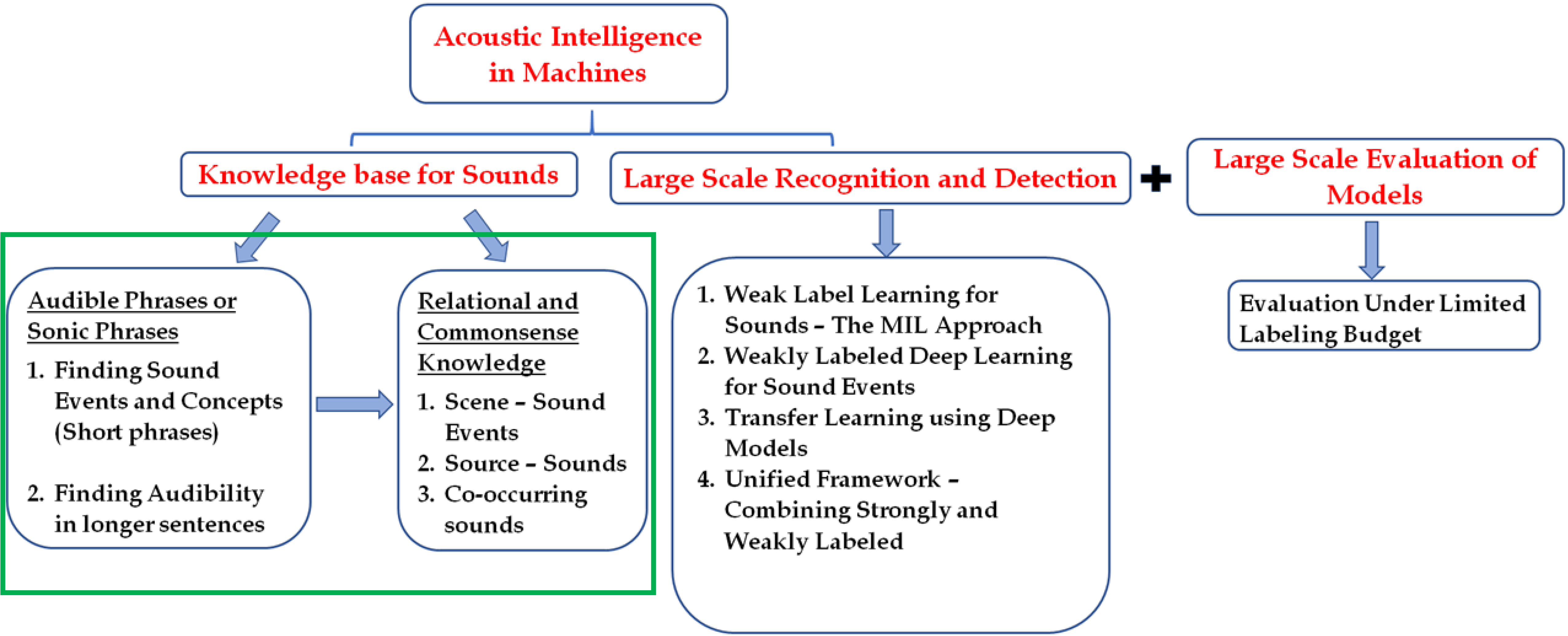
Knowledge of Sounds

Natural Language Understanding of Sounds

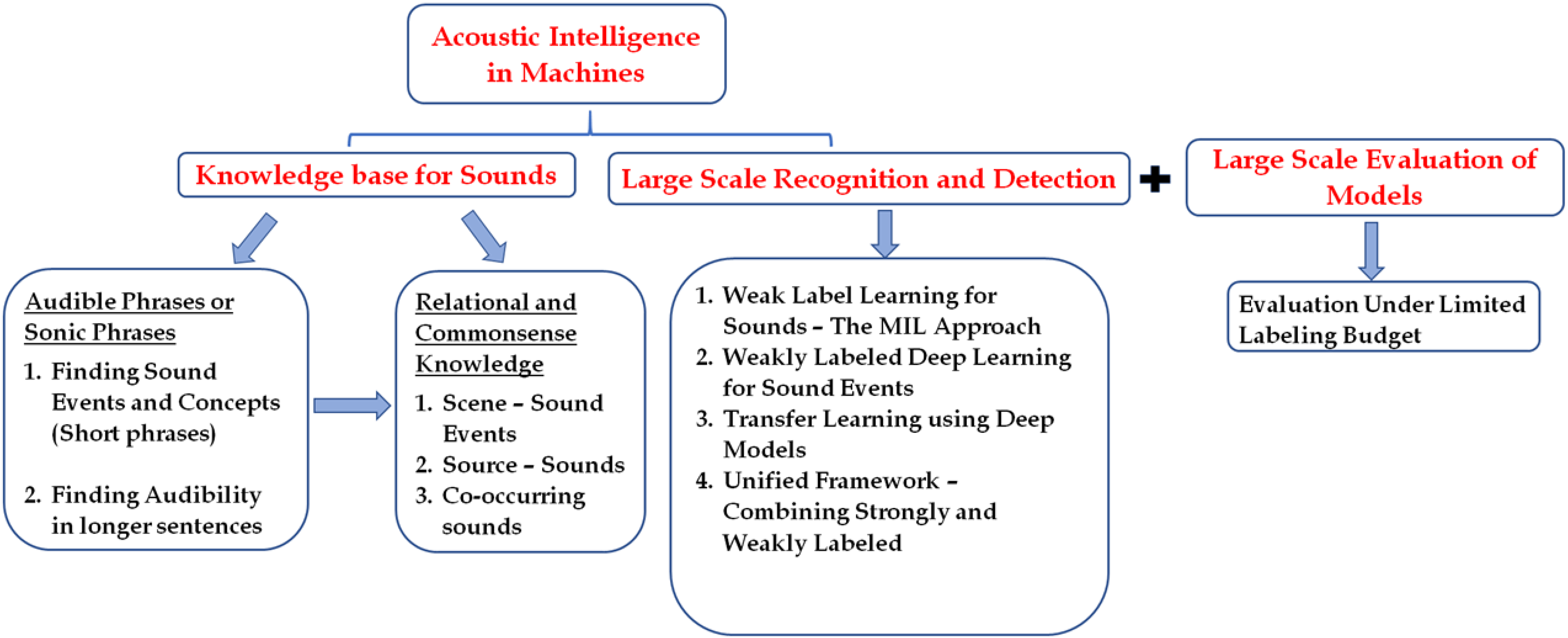
Linked and Facilitate Each Other

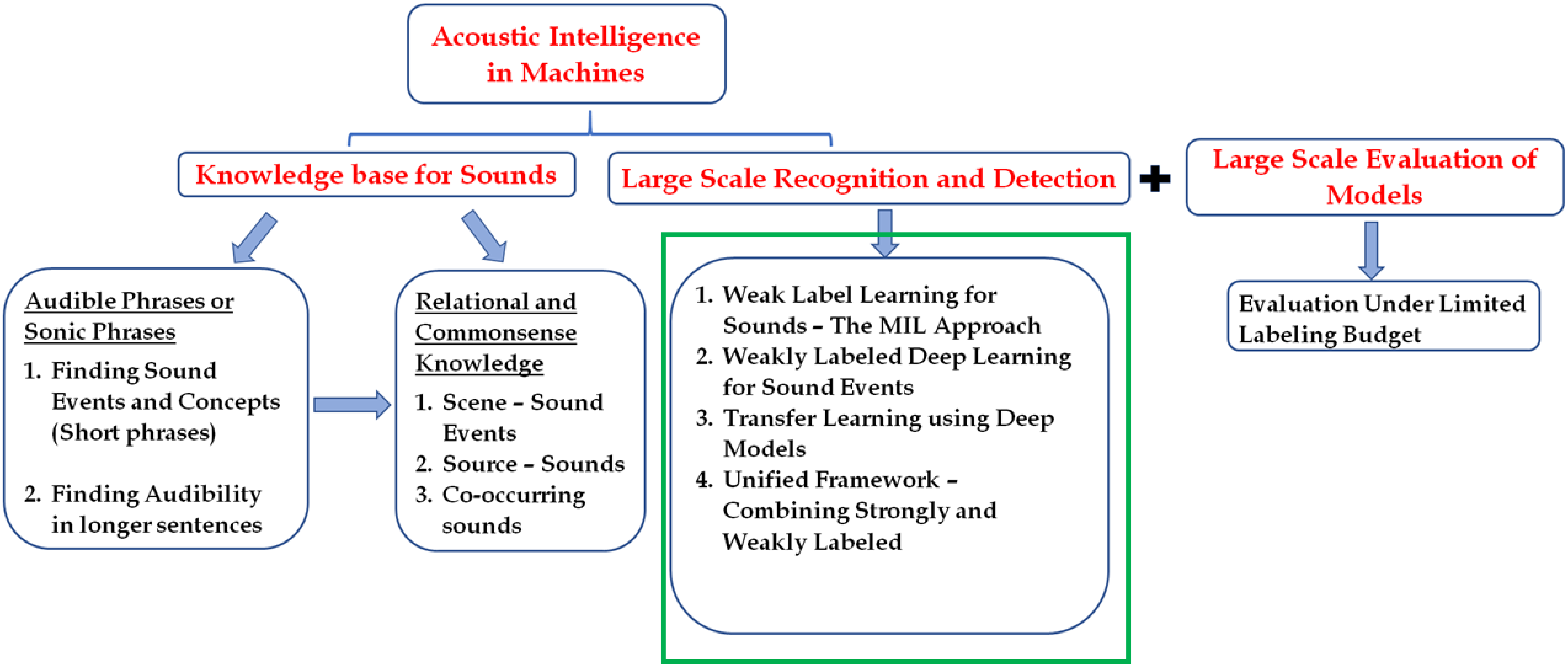
Recognition and Detection

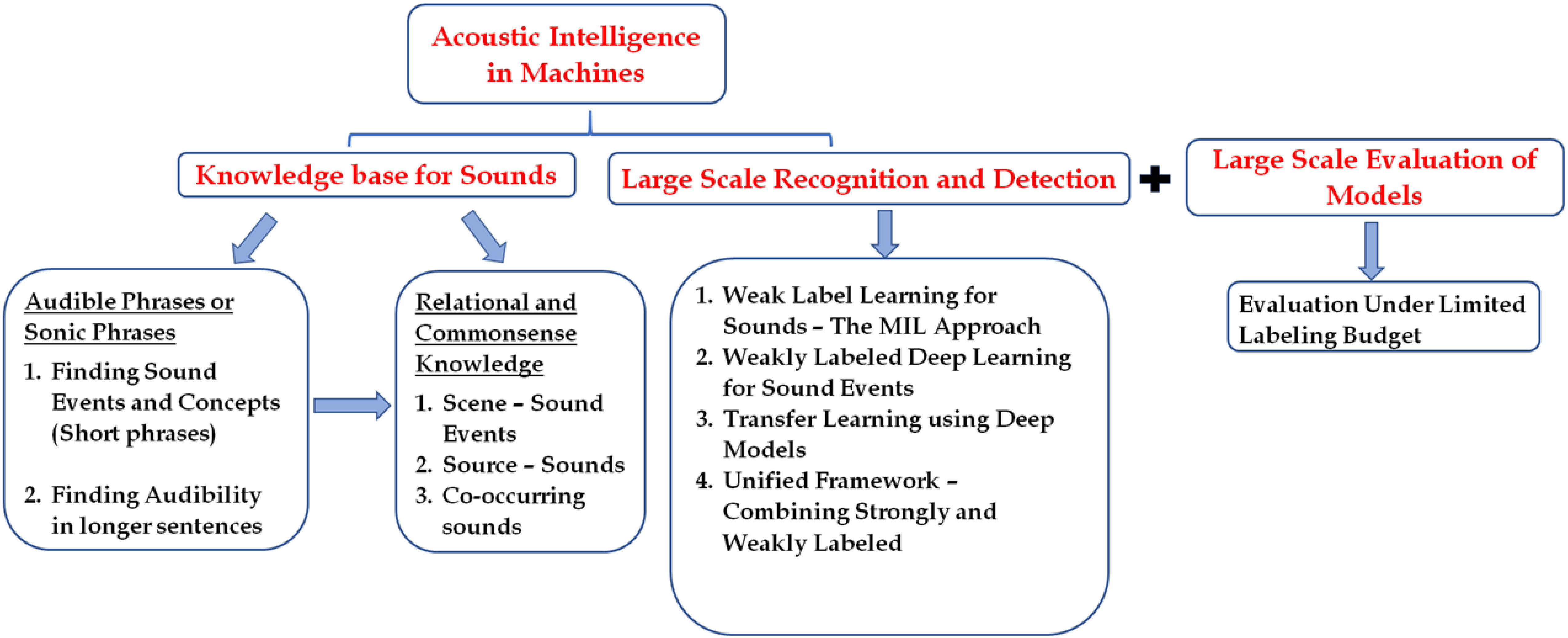
Large Scale Sound Event Detection

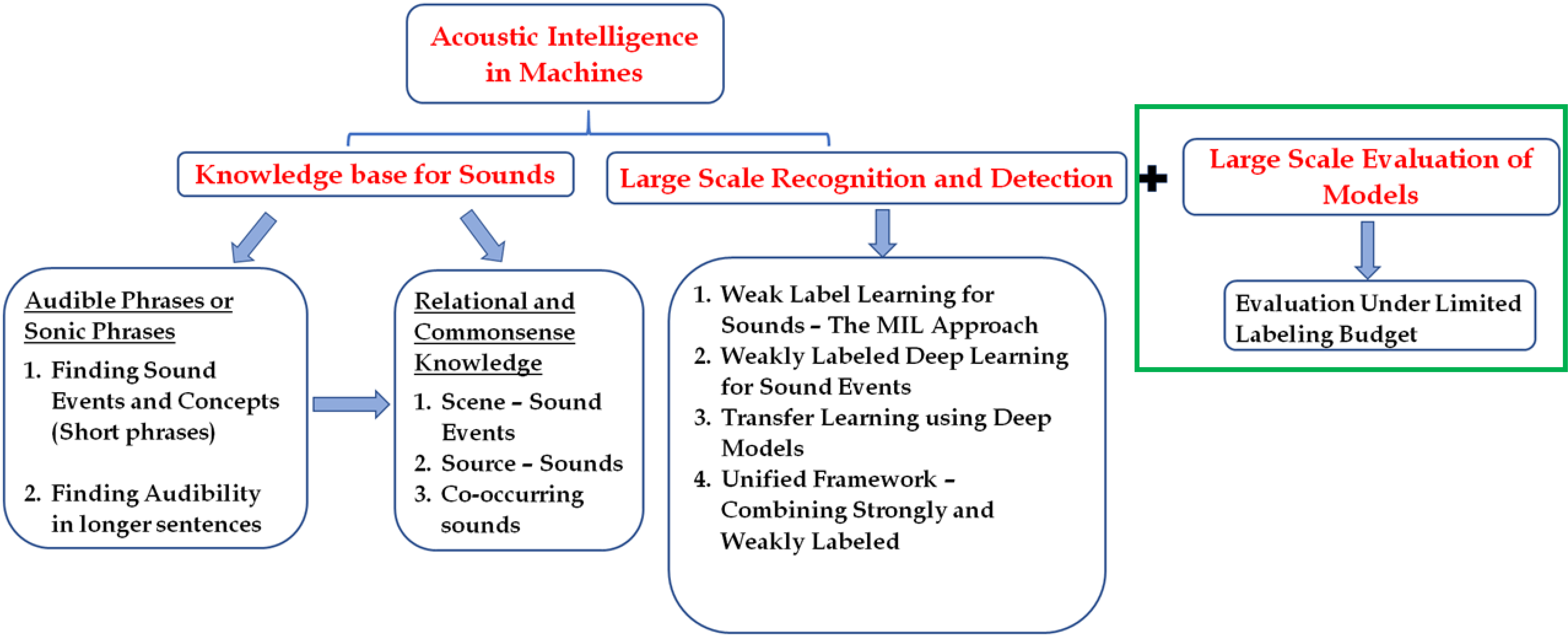


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget

Natural Language Understanding of Sounds

Cataloging, Understanding and Relating Sounds

Knowledge of Sounds

Natural Language Understanding of Sounds

Audible Phrases or Sonic Phrases

- Finding Sound Events and Concepts (Short phrases)

Relational and Commonsense Knowledge

- Scene – Sound Events

Knowledge of Sounds

- **Identifying "Audible Phrases" - a large list of sounds ?**
  - How humans describe or "name" sounds

- **Relationships and knowledge about sounds ?**
  - Commonsense knowledge and understanding
  - Source – Sound --- *Car* produces *honking, beeping, engine noise*
  - Scene – Sound --- *Children Laughing, Bird Chirping* can be found in *Park*
  - Co-occurrence relations --- *Laughing and Cheering* often occur together

Knowledge of Sounds

Knowledge of Sounds

How humans talk about sounds ? - Learn from text

Cataloging Sounds – "Audible Phrases"

- **Sounds are result of action on interaction between objects**
  - Same source different actions, Same action different sources

Cataloging Sounds – "Audible Phrases"

- **Sounds are result of action on interaction between objects**
  - Same source different actions, Same action different sources

- Different ways to express sounds – often composed of words which may have no relation to sound

  - *Music, Laughter, Screaming* are kind of "sound words" (onomatopoeia)

  - *Jackhammer, Garage door* not but are often used to denote sounds

Cataloging Sounds – "Audible Phrases"

Discover Potential Sound Concepts or Names and Then Filter

Cataloging Sounds – "Audible Phrases"

Discover Potential Sound Concepts or Names and Then Filter

Text Corpus → <sound of Y> → Potential Sound Names

<Sound of man yelling> , <sound of gunshots>

ClueWeb Corpus – 500 million webpages

Cataloging Sounds – "Audible Phrases"

Discover Potential Sound Concepts or Names and Then Filter

<Sound of man yelling> , <sound of gunshots>

Text Corpus → <sound of Y> → Potential Sound Names

ClueWeb Corpus – 500 million webpages

Unsupervised Filtering

Cataloging Sounds – "Audible Phrases"

Discover Potential Sound Concepts or Names and Then Filter

Text Corpus → <sound of Y> → Potential Sound Names

<Sound of man yelling> , <sound of gunshots>

ClueWeb Corpus – 500 million webpages

Unsupervised Filtering

Potential Sound Names → Parts of Speech Tag → Reduces to just 20 POS patterns → Select those Represent sound → Sound Concepts/ Names

<Sound of NN VBG> , <sound of NN(S)>

Kumar, Raj, Nakashole ICASSP 2017

Cataloging Sounds – "Audible Phrases"

- 6 Patterns which expresses sound

Cataloging Sounds – "Audible Phrases"

- 6 Patterns which expresses sound

<X> of (DT) VBG NN(S) → Honking cars, Crashing Waves, Laughing Children

Cataloging Sounds – "Audible Phrases"

- 6 Patterns which expresses sound

<X> of (DT) VBG NN(S)	→	Honking cars, Crashing Waves, Laughing Children
<X> of (DT) VBG	→	Yelling, Snoring, Knocking, Laughing

Cataloging Sounds – "Audible Phrases"

- 6 Patterns which expresses sound

<X> of (DT) VBG NN(S) →	Honking cars, Crashing Waves, Laughing Children
<X> of (DT) VBG →	Yelling, Snoring, Knocking, Laughing
<X> of (DT) NN(S) VBG →	Dog Barking, Birds Chirping, Metal hitting

Cataloging Sounds – "Audible Phrases"

- 6 Patterns which expresses sound

<X> of (DT) VBG NN(S) →	Honking cars, Crashing Waves, Laughing Children
<X> of (DT) VBG →	Yelling, Snoring, Knocking, Laughing
<X> of (DT) NN(S) VBG →	Dog Barking, Birds Chirping, Metal hitting
<X> of (DT) NN(S) →	Gunshots, Bell, Car, Hammer

Cataloging Sounds – “Audible Phrases”

- 6 Patterns which expresses sound

<X> of (DT) VBG NN(S)	Honking cars, Crashing Waves, Laughing Children
<X> of (DT) VBG	Yelling, Snoring, Knocking, Laughing
<X> of (DT) NN(S) VBG	Dog Barking, Birds Chirping, Metal hitting
<X> of (DT) NN(S)	Gunshots, Bell, Car, Hammer
<X> of (DT) NN NN(S)	Ocean Waves, Church Bells, Steel Drums

Cataloging Sounds – "Audible Phrases"

- 6 Patterns which expresses sound

<X> of (DT) VBG NN(S)	Honking cars, Crashing Waves, Laughing Children
<X> of (DT) VBG	Yelling, Snoring, Knocking, Laughing
<X> of (DT) NN(S) VBG	Dog Barking, Birds Chirping, Metal hitting
<X> of (DT) NN(S)	Gunshots, Bell, Car, Hammer
<X> of (DT) NN NN(S)	Ocean Waves, Church Bells, Steel Drums
<X> of (DT) JJ NN(S)	Live Music, Loud Crash, Heavy Machinery

Cataloging Sounds – "Audible Phrases"

- 6 Patterns which expresses sound

<X> of (DT) VBG NN(S)	Honking cars, Crashing Waves, Laughing Children
<X> of (DT) VBG	Yelling, Snoring, Knocking, Laughing
<X> of (DT) NN(S) VBG	Dog Barking, Birds Chirping, Metal hitting
<X> of (DT) NN(S)	Gunshots, Bell, Car, Hammer
<X> of (DT) NN NN(S)	Ocean Waves, Church Bells, Steel Drums
<X> of (DT) JJ NN(S)	Live Music, Loud Crash, Heavy Machinery

Total 116,729 sound concepts

Cataloging Sounds – "Audible Phrases"

- Manually inspect some frequent phrases
  - 100 most frequently occurring phrases for each pattern – 600 total
  - **Overall precision is ~77 %**

	Pattern	+ in 100 Most Freq.
P1	$<X>$ of (DT) VBG NN(S)	98
P2	$<X>$ of VBG	71
P3	$<X>$ of (DT) NN(S) VBG	91
P4	$<X>$ of (DT) NN(S)	59
P5	$<X>$ of (DT) NN NN(S)	93
P6	$<X>$ of (DT) JJ NN(S)	49

Cataloging Sounds – "Audible Phrases"

- Supervised Filtering
  - Classification Problem

- Representing phrases
  - Word Embeddings successful in syntactic and semantic similarity

  - Gives over 90% accuracy over 6000 phrases

Understanding and Relating Sounds

- DCASE 2016 Challenge – *Dishes, Object Banging*

- How would machine understand *Dishes* and *Object Banging* ?
  - Dishes clinking, Dishes Breaking, Washing Dishes, Running Water ?
  - What type of object ? Gavel, Iron, Glass

- The large list carries a lot of these information

Scene (Environment)– Sounds Relations

Scene (Environment)– Sounds Relations

- **What type of sounds can be found in an environment ?**
  - Commonsense knowledge
  - Useful for acoustic scene classification

Scene (Environment)– Sounds Relations

- **What type of sounds can be found in an environment ?**
  - Commonsense knowledge
  - Useful for acoustic scene classification

Children Laughing, Birds Chirping

Hammering, Drilling, Blasting

Scene (Environment)– Sounds Relations

- A relation classification task
  - Whether a sound and scene are related or not

- Sentences where a scene name and at least one of sound concept occur
  - The ***park*** was filled with the sound of ***Laughing***

Scene (Environment)– Sounds Relations

- Relate scene and sound concept through dependency paths
  - Shortest dependency paths good for relation classification
- Minimal Supervision for collecting labeled examples
  - Label most frequent dependency paths as positive or negative
- Train a classifier on the labeled examples

Scene (Environment)– Sounds Relations

Forest

Birds Singing, Breaking Twigs, Cooing

Bar

Piano Playing, Laughter, Clinking Glasses

Church

Church Bells, Singing, Applause

Scene (Environment)– Sounds Relations

Forest

Birds Singing, Breaking Twigs, Cooing

Bar

Piano Playing, Laughter, Clinking Glasses

Church

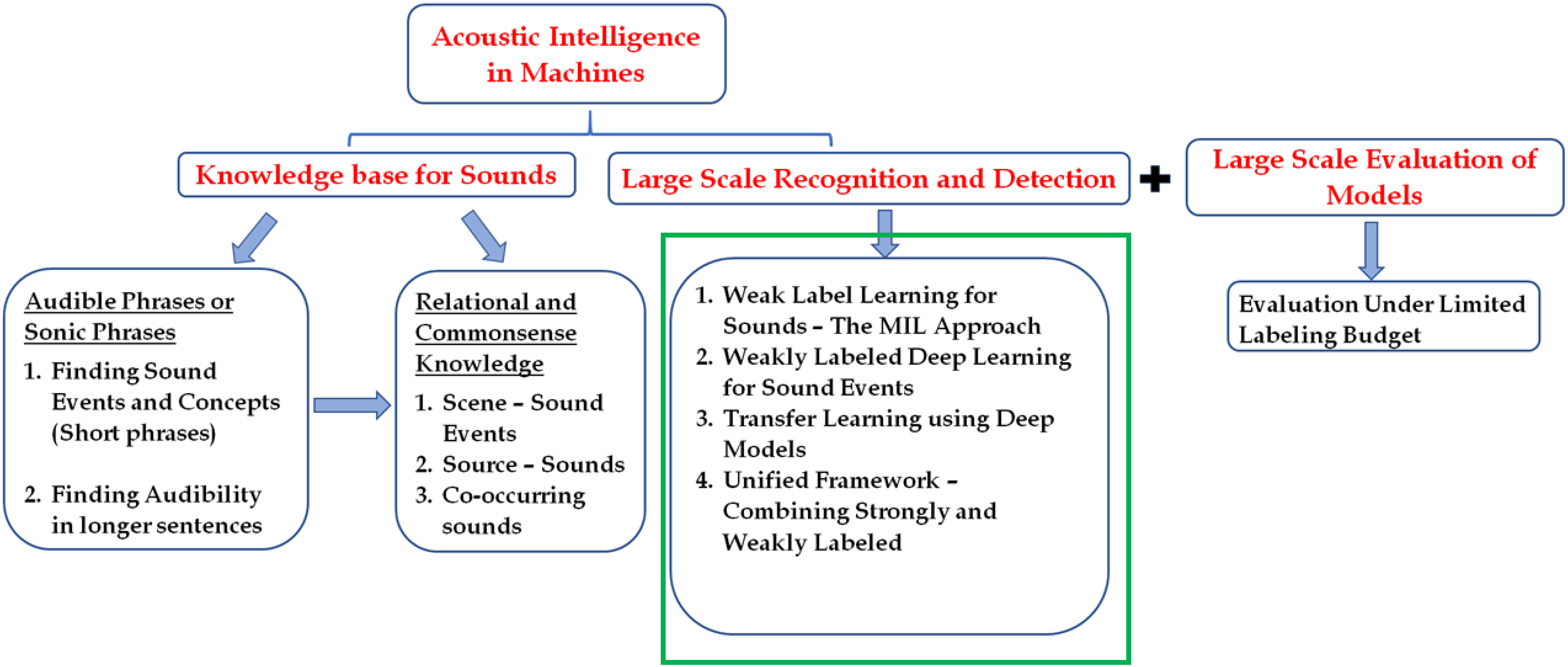
Church Bells, Singing, Applause

Some Unusual !!!

Farm – soldiers rampaging
Church – Rifle shots
Library – Chirping Birds

Summary so far

- Using text for creating a catalog and knowledge base of sounds
- Intricate and higher level information about sounds can be drawn
- Next Step – Recognition and Detection of Sound Events and Acoustic Scenes


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget

Recognizing and detecting sounds

Audio Event Detection (AED)

- **To recognize and detect sound events in audio (video) recordings**
- Learning on large scale
- Size of training and test for a given event
  - Really small !!!
- Limited Vocabulary
  - Which sounds to recognize ? --- We looked at previous part

Audio Event Detection - Scale

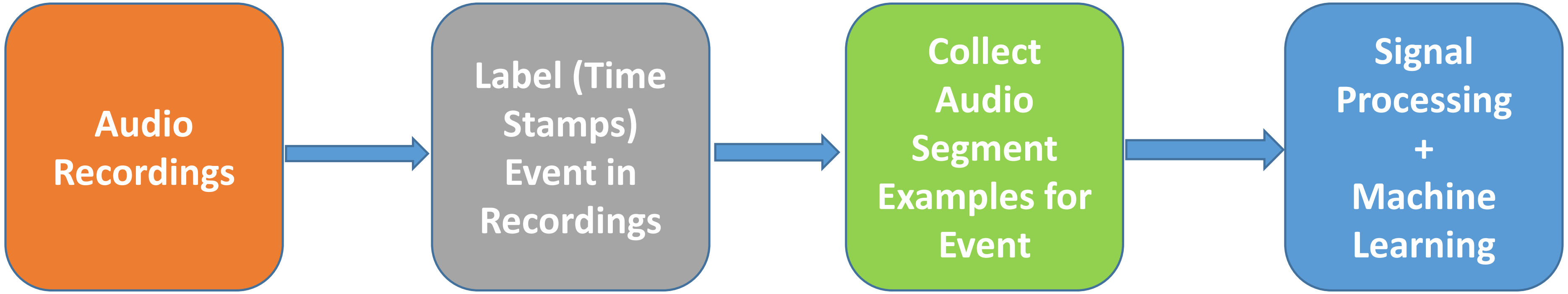
- A look at publicly available datasets

Dataset	# of Events	Audio Data
DCASE 2013	16	24 short (1 or 2 second) clips per class
DCASE 2016	18	Total audio data ~60 min
ESC-50	50	3.33 min per event
FBK-Irst	16	Total audio data ~1.7 hours
Urbansounds	10	~20 min per event (with repetitions)

ITC – Irst dataset – number of test samples as few as **3** and max of **12** for all events [widely used in several papers]

Audio Event Detection - Scale

- What's the bottleneck ?



General Framework for Audio Event Detection

Audio Event Detection - Scale

- What's the bottleneck ?

Bottleneck!! --- Strongly Labeled Data

Audio Recordings → **Label (Time Stamps) Event in Recordings** → **Collect Audio Segment Examples for Event** → **Signal Processing + Machine Learning**

General Framework for Audio Event Detection

Labeling data with time stamps – Biggest Problem

Audio Event Detection - Scale

- Strongly Labeled Data

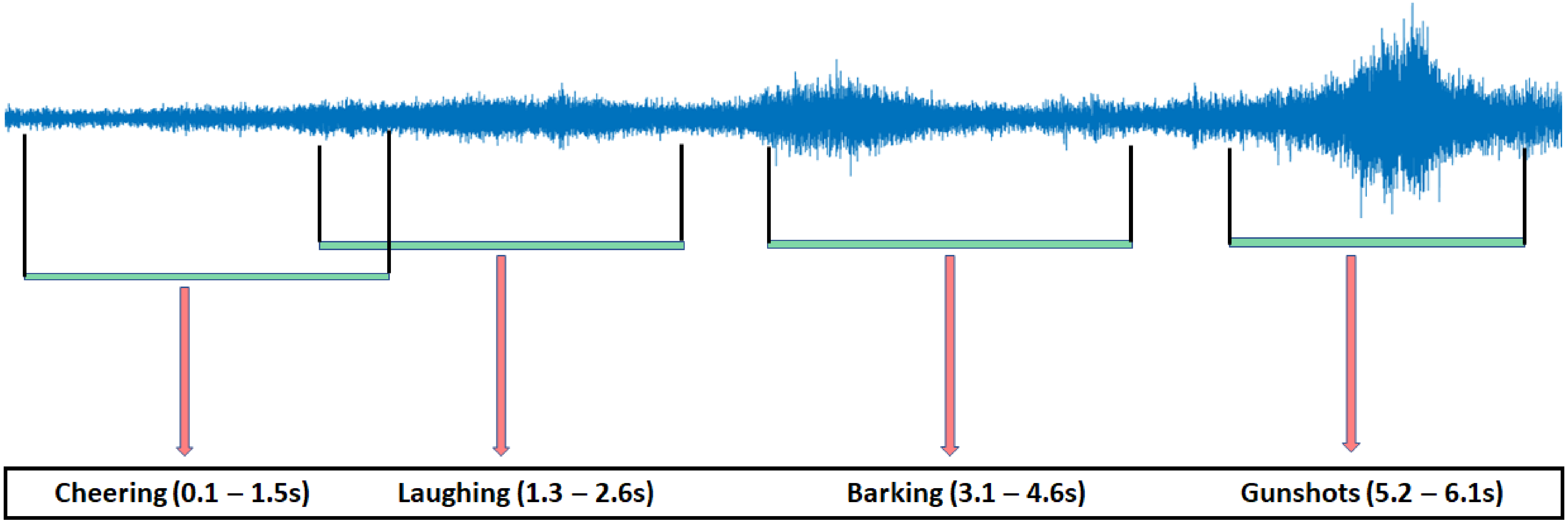
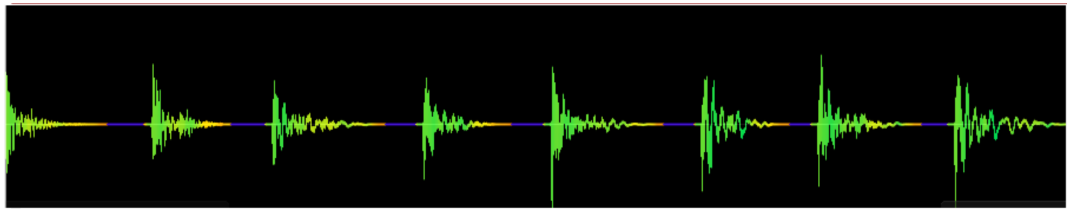


Illustration of Strong Labeling of Events

Audio Event Detection - Scale

- Strongly Labeled Data - **Time consuming and expensive**
  - Have to back and forth in audio to mark times
  - Overlapping events
  - Interpretation may create difficulties in marking times

How many beginnings and ends ?



Footsteps

AED – Moving to Weak Labels

Kumar and Raj, ACM Multimedia 2016

AED – Moving to Weak Labels

- A step down from strongly labeled data
    - Weaker form of supervised learning

AED – Moving to Weak Labels

- A step down from strongly labeled data
    - Weaker form of supervised learning
- **Weakly Labeled Data**

AED – Moving to Weak Labels

- A step down from strongly labeled data
  - Weaker form of supervised learning
- **Weakly Labeled Data**

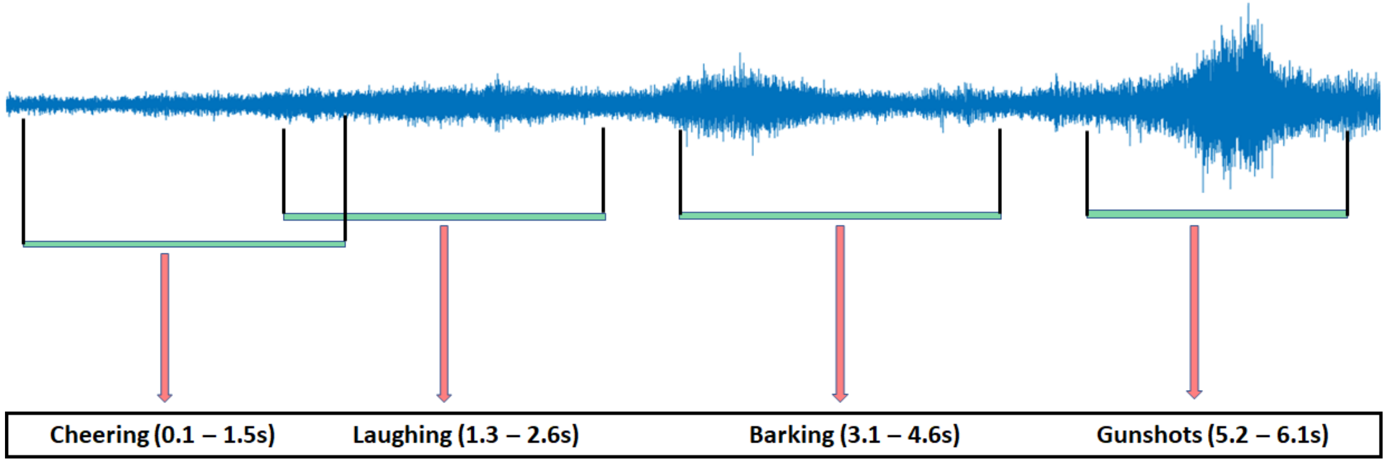


Illustration of Strong Labeling of Events

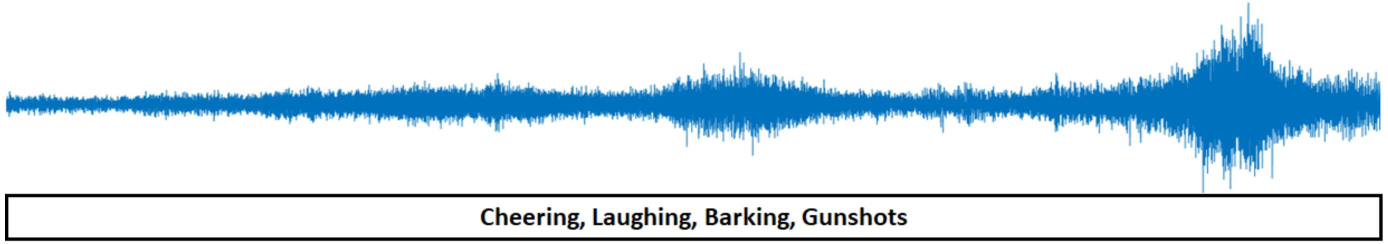


Illustration of Weak Labeling of Events

Kumar and Raj, ACM Multimedia 2016

AED – Moving to Weak Labels

- Weakly labeled data
  - Much easier to label
- Possible to use the massive amount of data on web
  - Possibly without any manual labeling effort
- Example – Audioset Large Scale Weakly Labeled Dataset

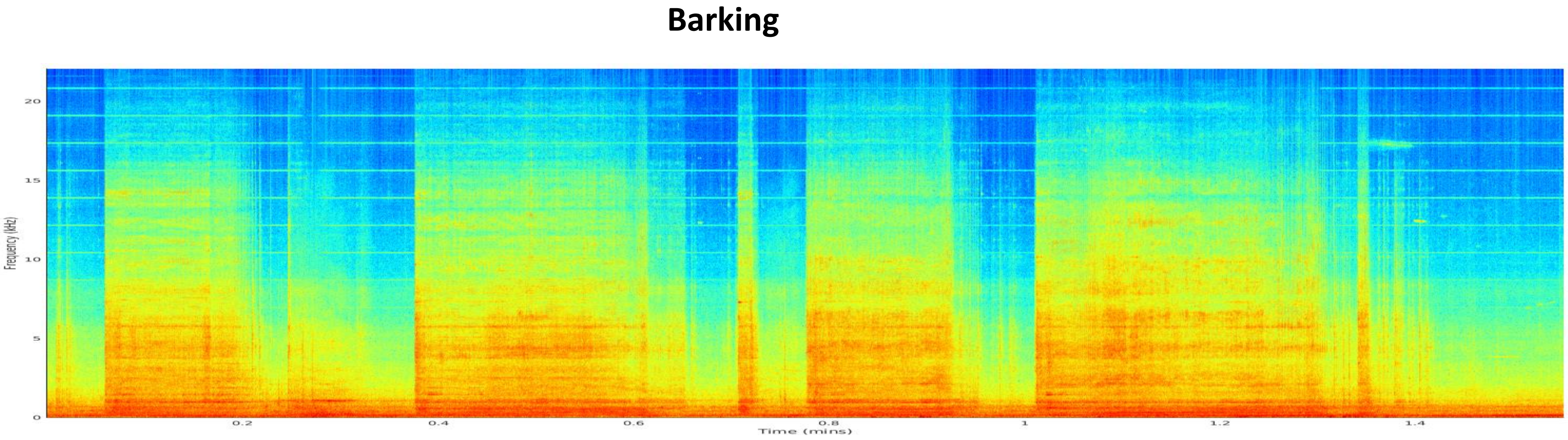
A bit of Nomenclature!!

- **Frame** – small 20, 30 or so milisecond STFT window
  - Correspond to 1 single STFT frame
  - Or say 1 frame of Spectrogram, MFCC, Logmel, CQT

- **Recording** – Full audio recording
  - from a few seconds to several minutes
  - Represented by several frames – from a few to possibly thousands

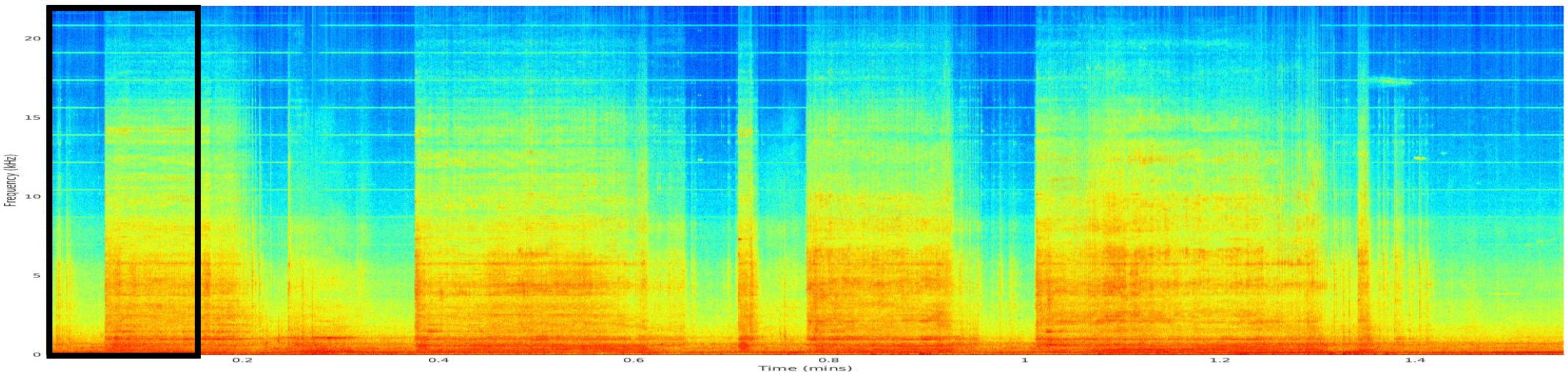
- **Segments** – Small chunks or "segments" of Recording
  - Usually 0.5, 1 or 1.5 seconds
  - Represented by several frames – from a few to may be up to hundred or so

AED Using Weak Labels



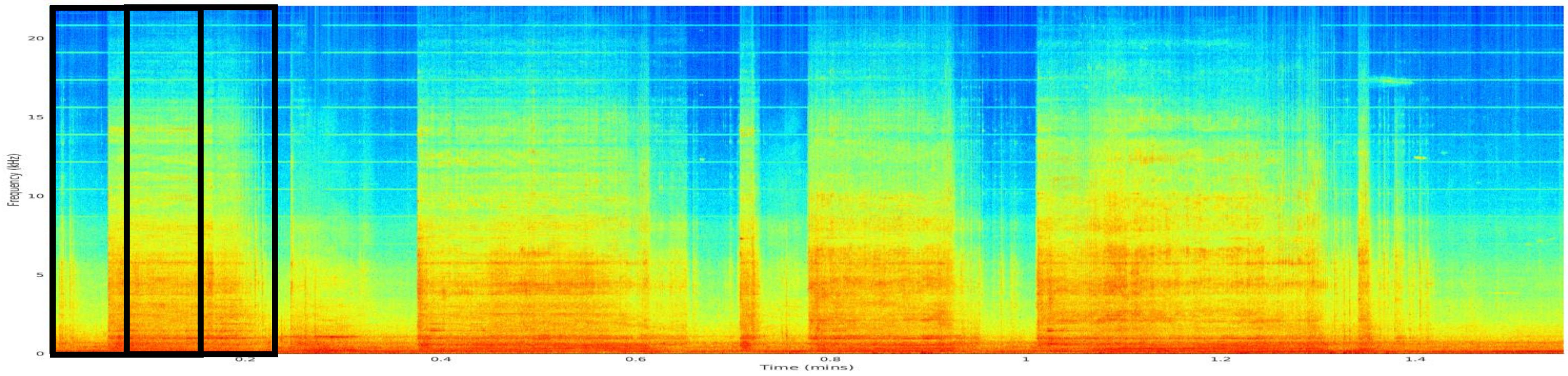
AED Using Weak Labels

Barking

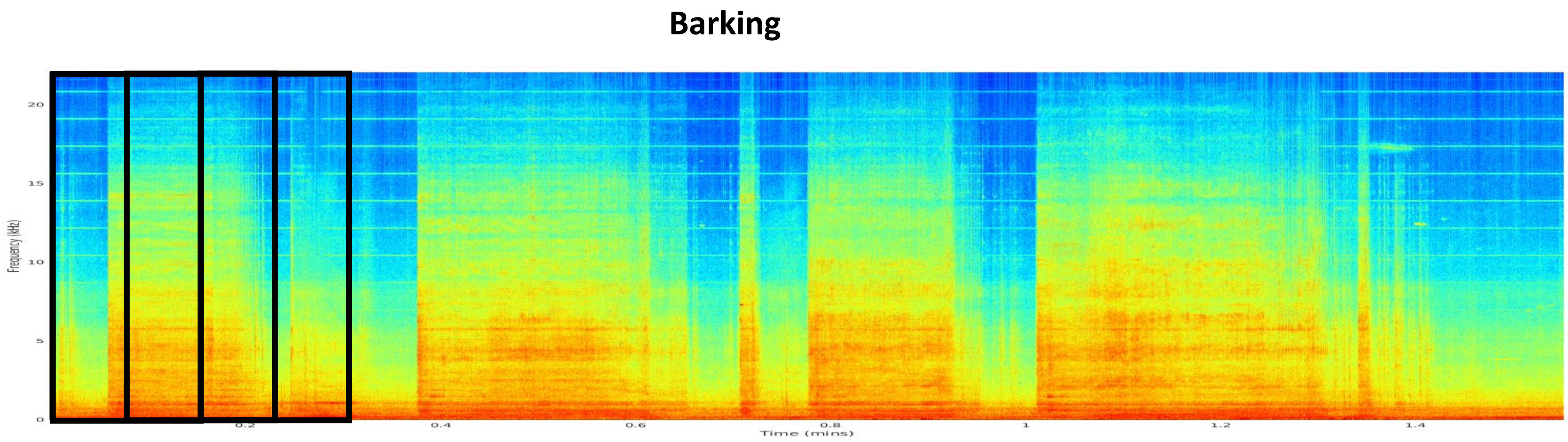


AED Using Weak Labels

Barking

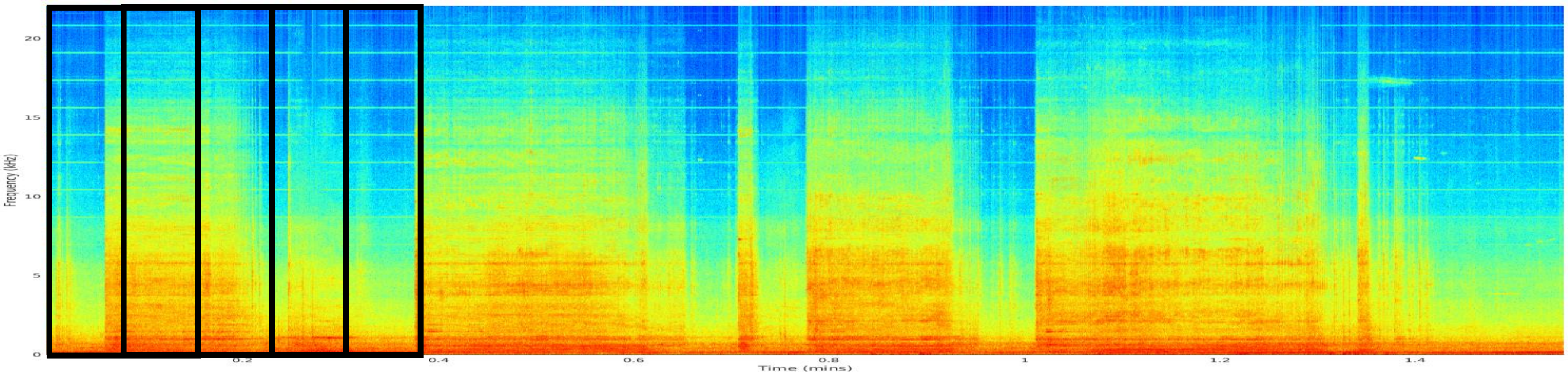


AED Using Weak Labels



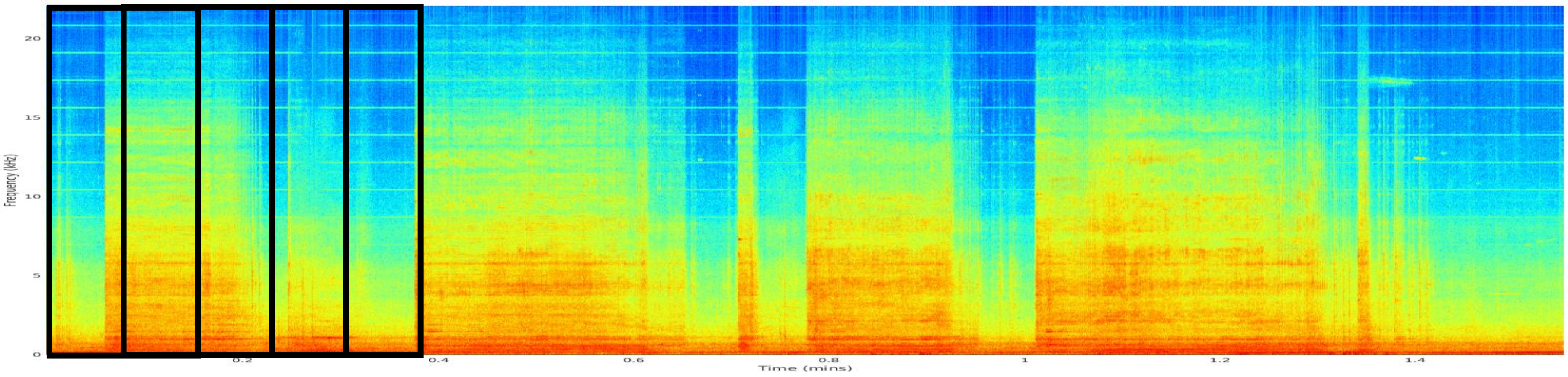
AED Using Weak Labels

Barking



AED Using Weak Labels

Barking

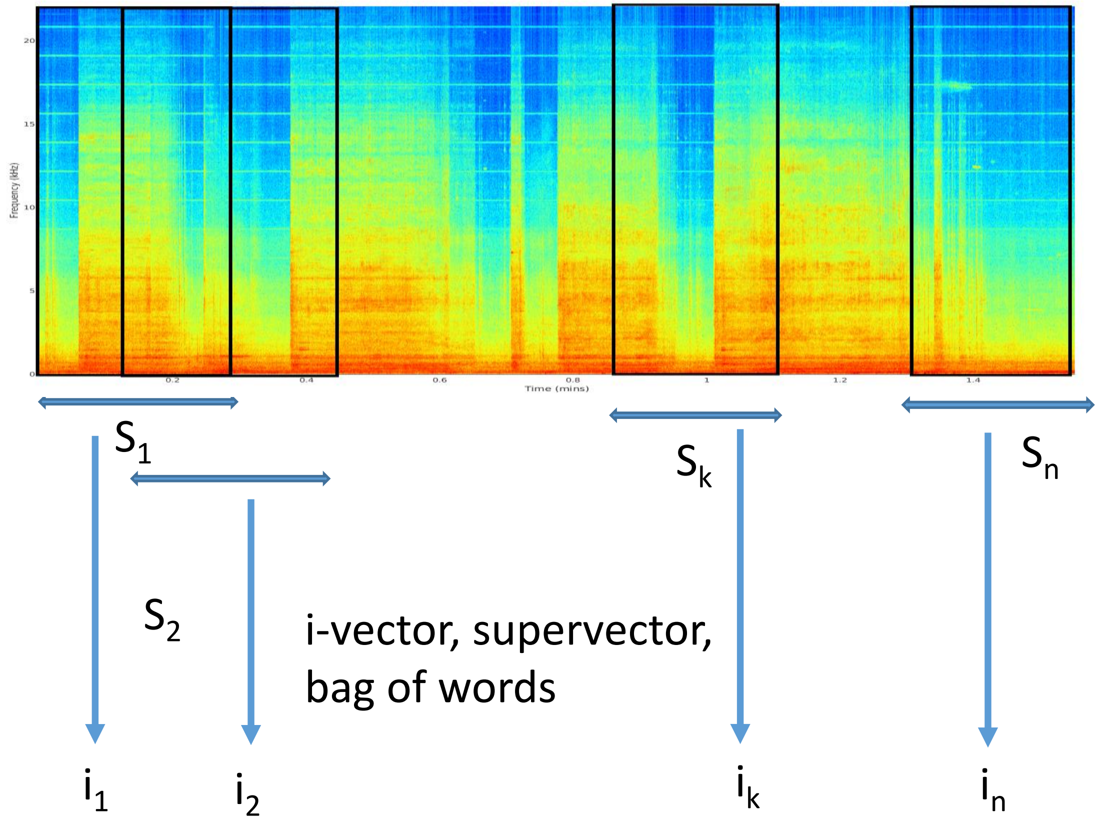


Look through the recording for where the event might have occurred. Work with that!

A general algorithmic framework for doing this

AED using Weak Labels - Framework

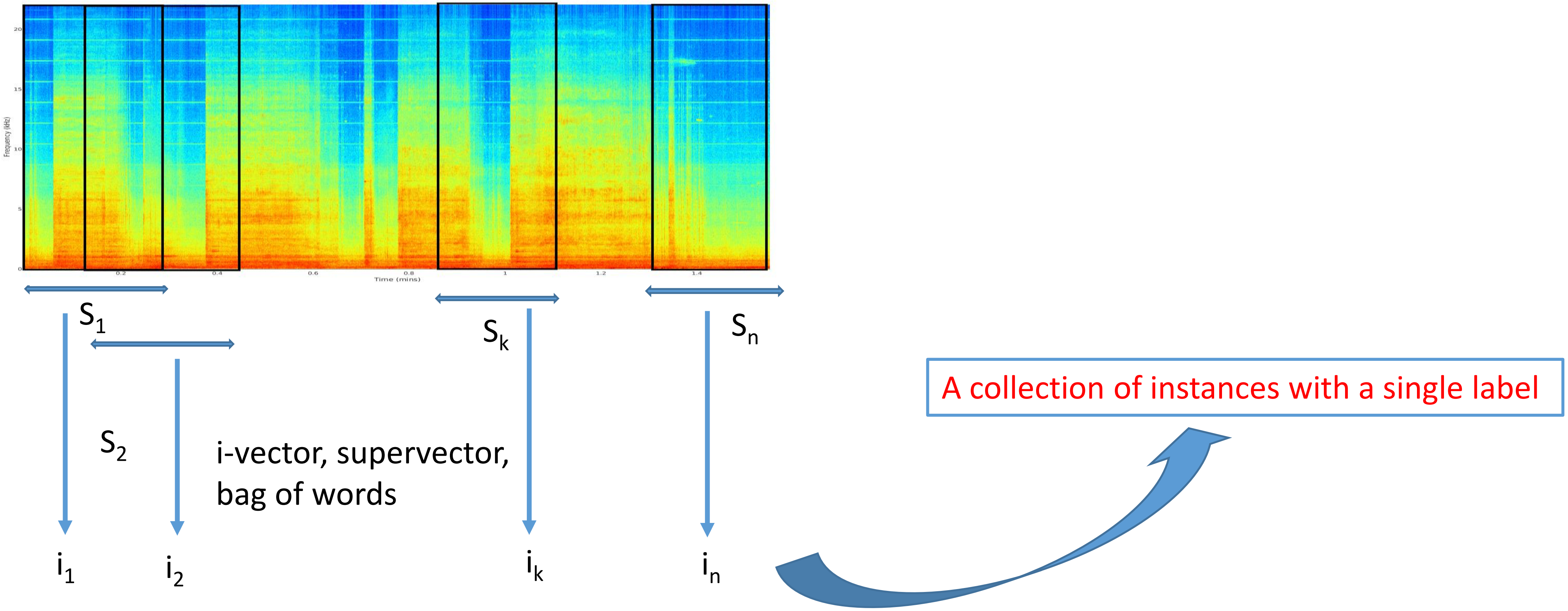
An algorithmic framework (for each event)



Kumar and Raj, ACM Multimedia 2016

AED using Weak Labels - Framework

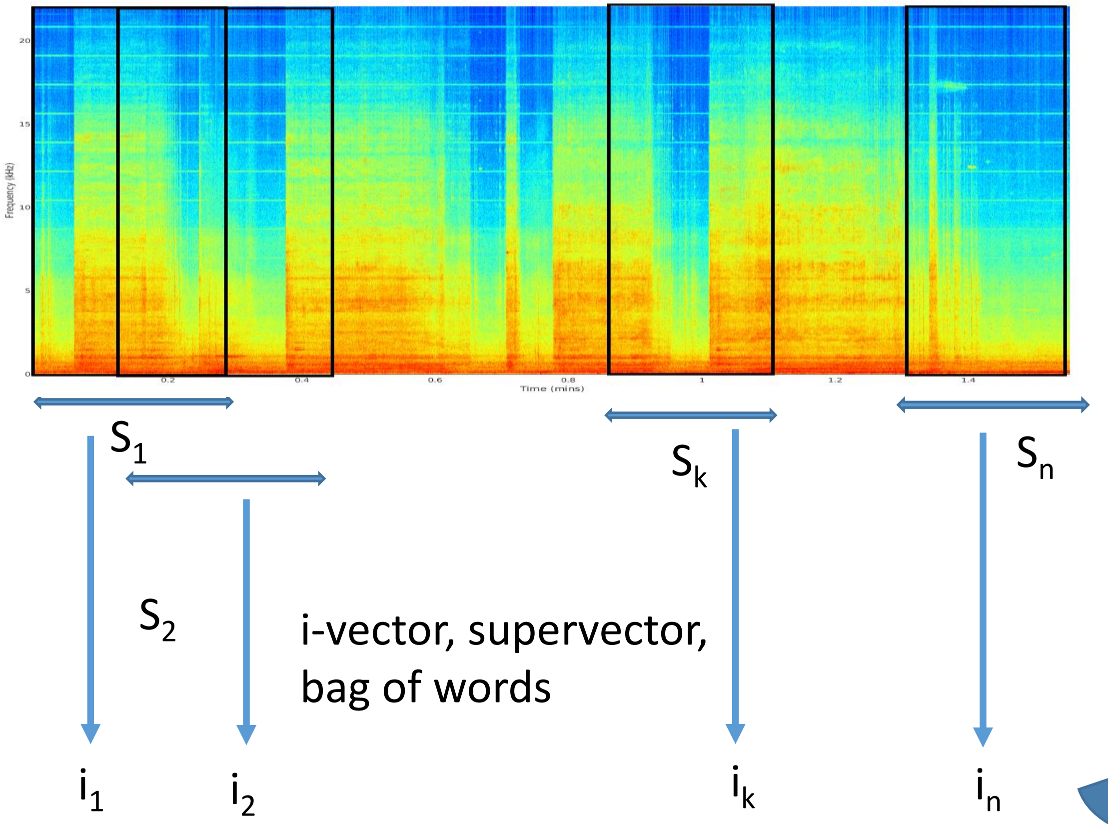
An algorithmic framework (for each event)



Kumar and Raj, ACM Multimedia 2016

AED using Weak Labels - Framework

An algorithmic framework (for each event)

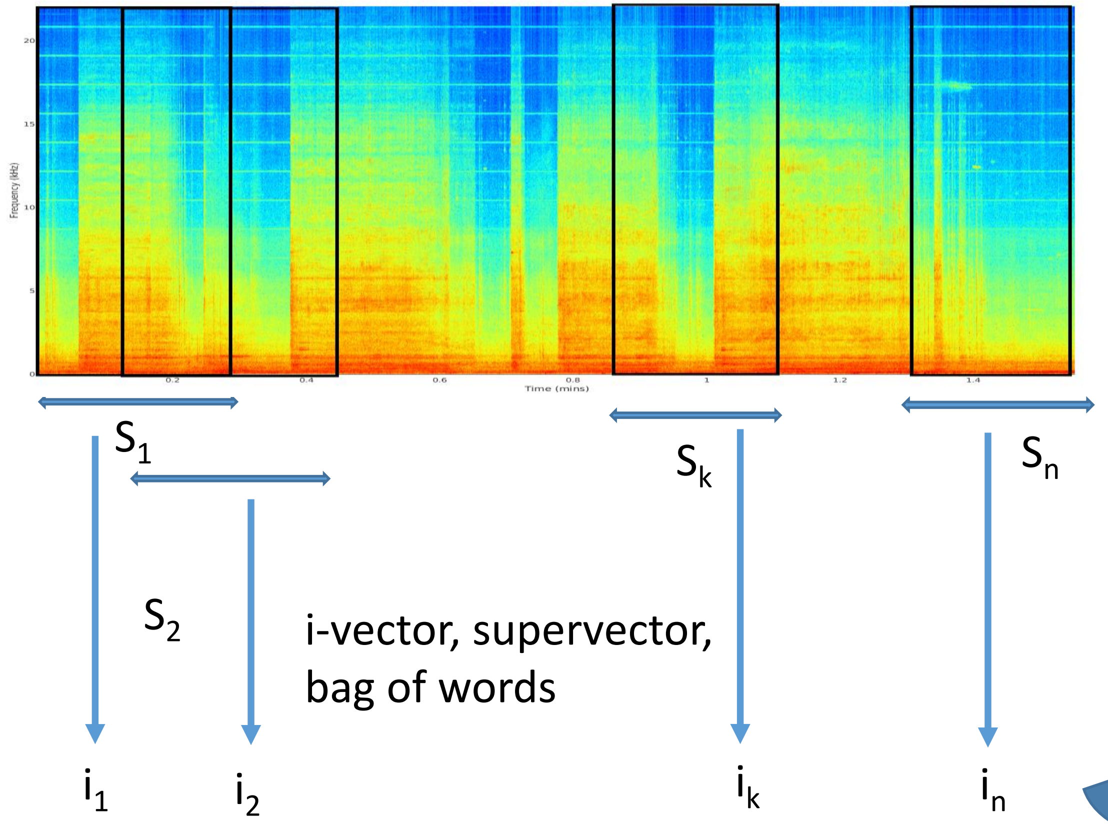


At least one of these segments contain event and is a good representation of it

A collection of instances with a single label

AED using Weak Labels - Framework

An algorithmic framework (for each event)



General Framework

Multiple Instance Learning

At least one of these segments contain event and is a good representation of it

A collection of instances with a single label

AED with weak labels – Framework

- **Multiple Instance Learning – Weak form of supervised learning**

- Labels are available for a group of instances called *Bags*

- Negative bag – All Instances in are negative

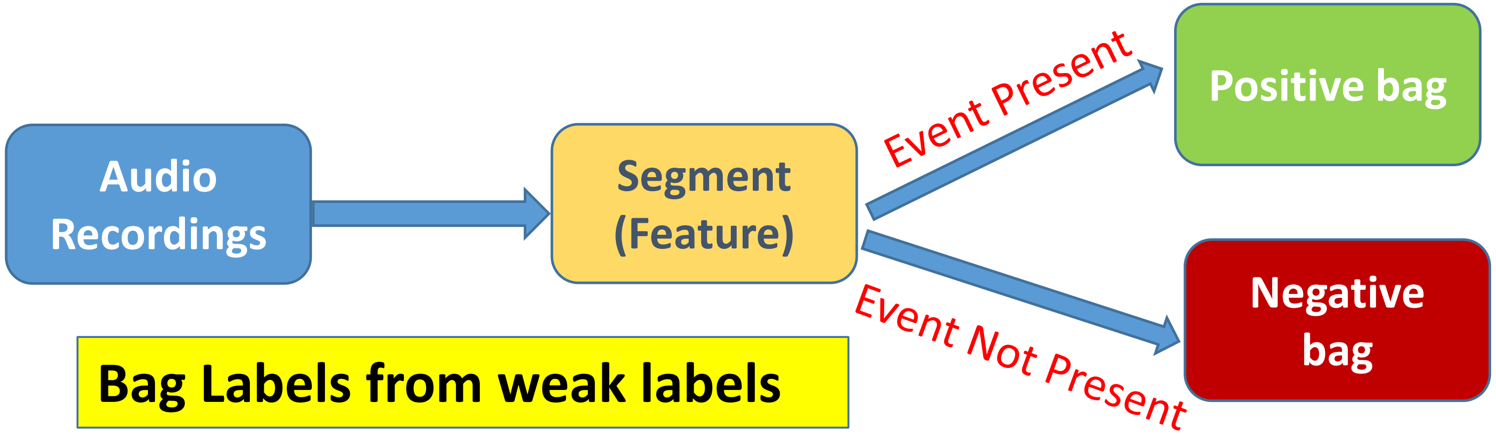
- Positive bag -- **At least one instance is positive**

AED with weak labels - MIL

- **Segment audio recordings to form bags with labels (for each event)**

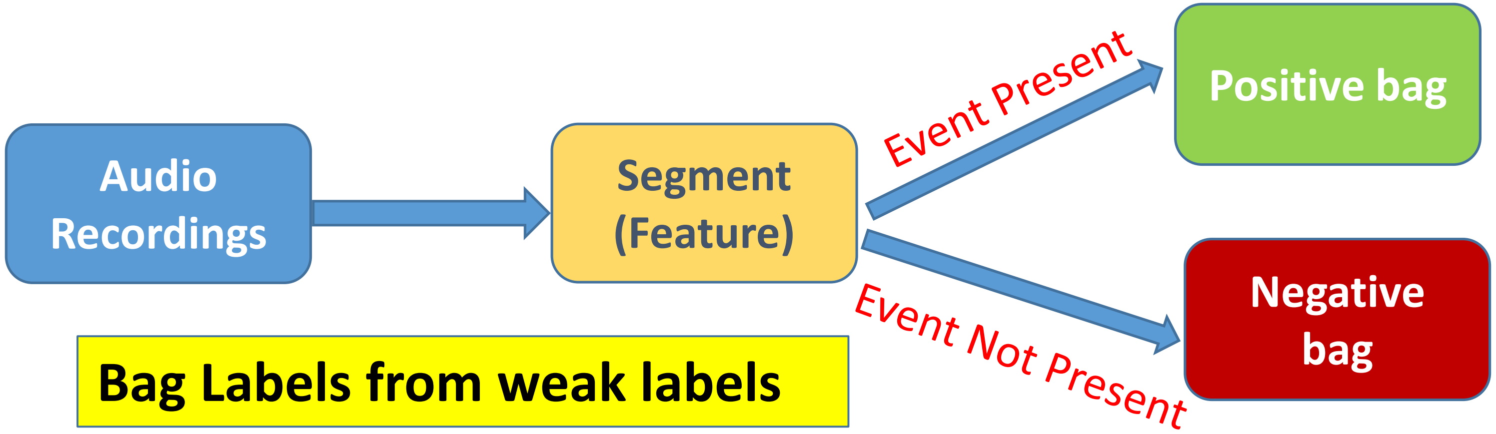
AED with weak labels - MIL

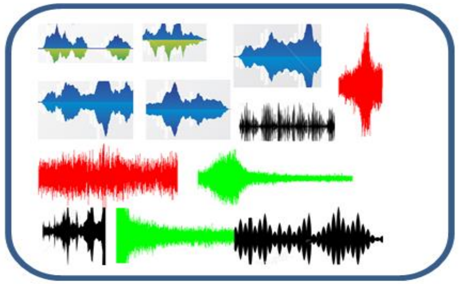
- **Segment audio recordings to form bags with labels (for each event)**



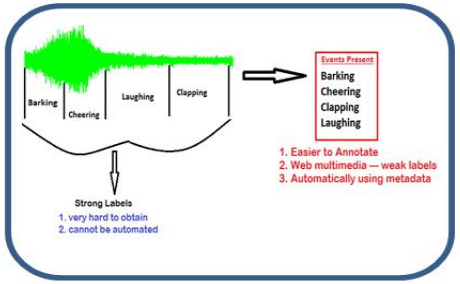
AED with weak labels - MIL

- **Segment audio recordings to form bags with labels (for each event)**



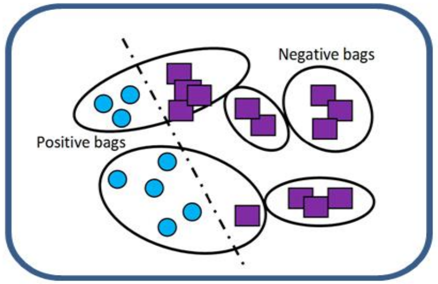


Audio Recordings



Weak Labels

Segment to Form Bags

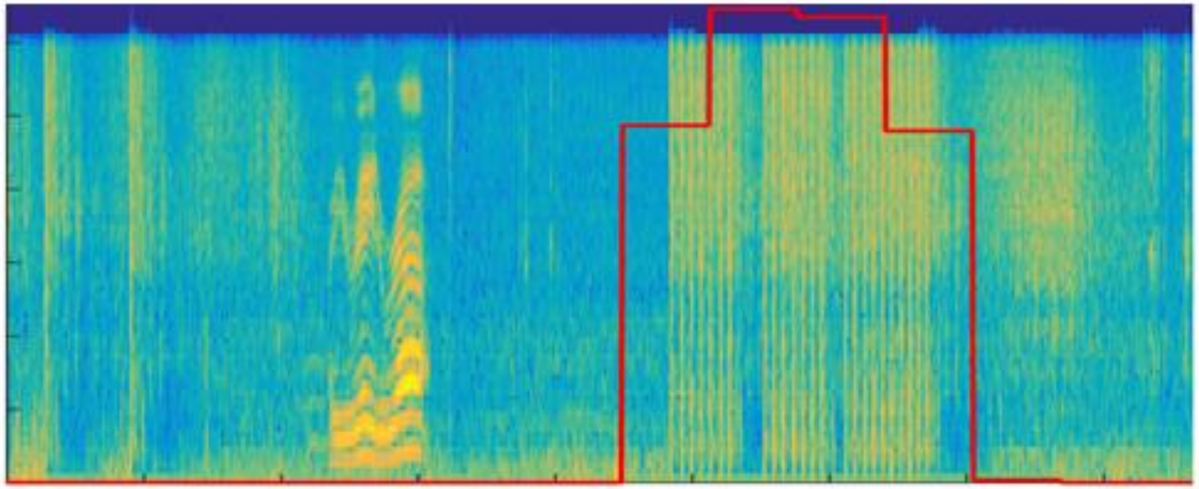


Multiple Instance Learning

AED using weakly labeled data - Framework

AED with weak labels - MIL

- **Temporal Localization**
  - **Where the event occurred in the recording**



Machine Gun, Gunfire Sound

- After training prediction can be done on each segment of recording

Kumar and Raj, ACM Multimedia 2016

AED with weak labels - MIL Methods

- **miSVM -** Impose the "*at least one positive instance in positive bag*" constraints

$$\sum_{j=1}^{n_i} \frac{y_{ij}+1}{2} \geq 1 \ \forall \ i \ \text{s.t} \ Y_i = 1, \ ; \ y_{ij} = -1 \ \forall \ i \ \text{s.t} \ Y_i = -1$$

- **MISVM** - Define margin with respect to a "witness instance"
  - Witness Instance – Maximal output instance

- **NN – MIL** – Error with respect to maximal output instance
  - Back-propagation Training

Kumar and Raj, ACM Multimedia 2016

AED with weak labels - Analysis

- Manual weak labels on TRECVID dataset
  - ~20 hours of data
- Temporal Localization of Events
  - **Mean AUC around 0.66**
  - Segment Size
- **On the right track !!!**

Area Under ROC Curves

Event	AUC miSVM	AUC NN-MIL
Cheering	0.668	0.759
Children Voices	0.730	0.767
Clanking	0.859	0.764
Clapping	0.680	0.781
Drums	0.639	0.601
Engine Noise	0.642	0.698
Hammering	0.660	0.603
Laughing	0.685	0.632
Marching Band	0.745	0.618
Scraping	0.744	0.785
Mean	**0.704**	**0.701**

Recording level prediction results

Kumar and Raj, ACM Multimedia 2016

AED with weak labels – Analysis

AED with weak labels – Analysis

- Weakly labeled shows a way to get data on large scale

AED with weak labels – Analysis

- Weakly labeled shows a way to get data on large scale

- What about the scalability of the MIL methods

AED with weak labels – Analysis

- Weakly labeled shows a way to get data on large scale
- What about the scalability of the MIL methods
- **Most of the MIL algorithms suffers from scalability issues**

AED with weak labels – Analysis

- Weakly labeled shows a way to get data on large scale
- What about the scalability of the MIL methods
- **Most of the MIL algorithms suffers from scalability issues**
  - Complexity of hypothesis space in bag representation is large, harder to learn

AED with weak labels – MIL Scalability

AED with weak labels – MIL Scalability

- **Embed each bag into a vector**
  - Capture non-redundant information from instances in bag into a single vector

AED with weak labels – MIL Scalability

- Embed each bag into a vector
  - Capture non-redundant information from instances in bag into a single vector
- MIL now essentially becomes supervised learning
- Use any efficient, scalable supervised learning method

AED with weak labels – MIL - Scalability

- Two ways to encode bags

- miFV
  - Fisher Vectors (FV) for encoding bags

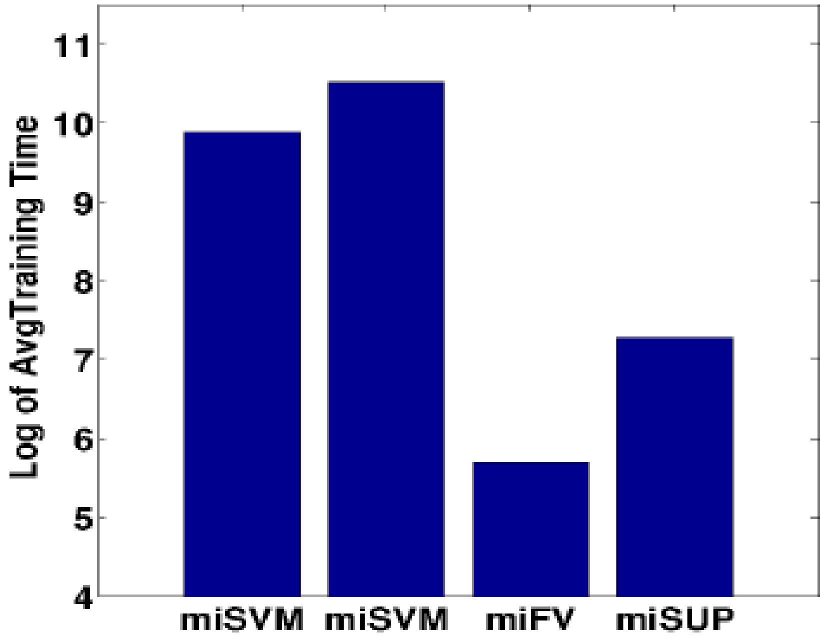
- miSUP
  - Use maximum-a-posteriori to adapt GMM parameters to a given bag

Kumar and Raj, IEEE ICME 2016

AED with weak labels – MIL - Scalability

- Scalable vs Non-scalable MIL methods
  - **12 – 15% improvement in MAP** (mean average precision)

- Avg. Training Time Comparison
  - 20 to 100 times faster



- Temporal Localization is a concern

Kumar and Raj, IEEE ICME 2016

AED using weak labels

- AED with weak labels

  - **First work on audio or sound event using Weak Labels [ACM Multimedia'16]**

  - Audioset (ICASSP 2017): A large scale weakly labeled dataset for sounds

  - Weak Label based learning for sounds is now part of **annual IEEE Sound Events and Scenes Challenge (2017, 2018)**

  - A large body of works have followed on this idea of learning from weak labels

Audio Event Detection So Far

Audio Event Detection So Far

Strongly Labeled

Small scale

Audio Event Detection So Far

Strongly Labeled → **Weakly Labeled**

Small scale → **Large Scale**

Audio Event Detection So Far

Strongly Labeled → **Weakly Labeled**

Small scale → **Large Scale**

↓

Deep Learning Methods

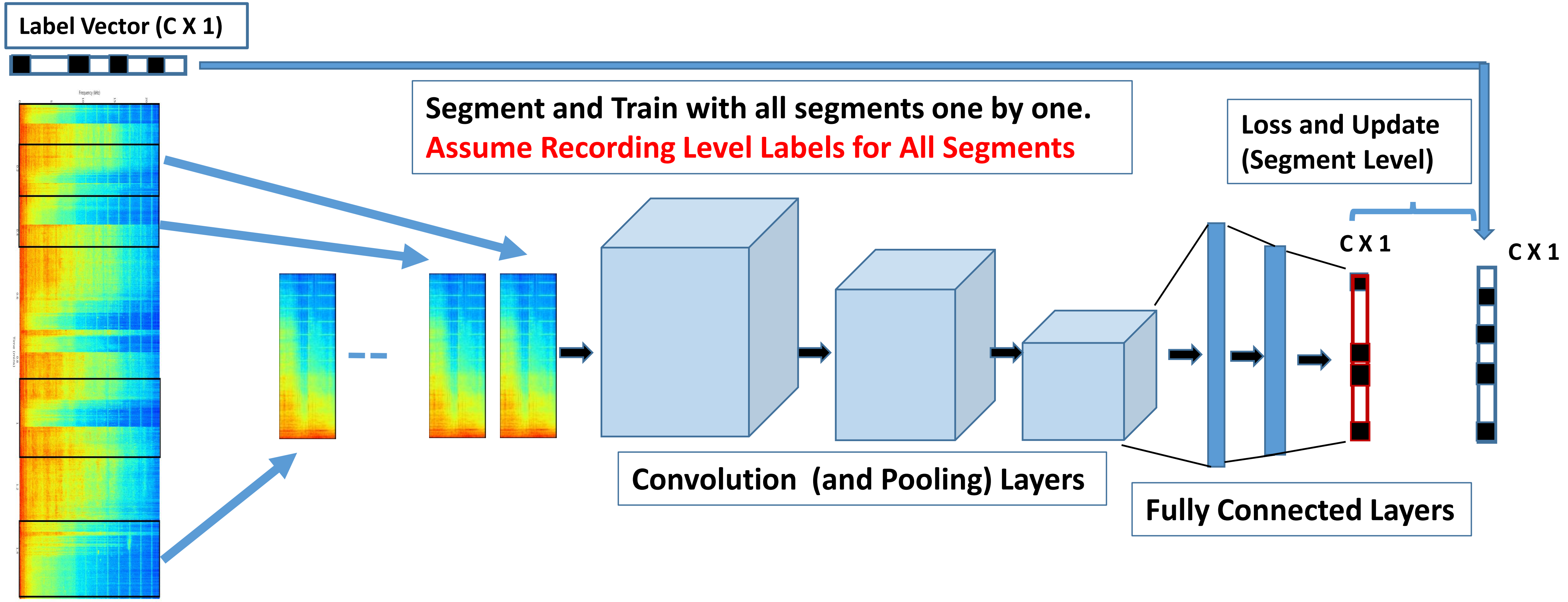
Deep Learning for AED using Weak Labels

Deep Learning for AED using Weak Labels

- **Simplest - Strong Label Assumption Training (SLAT) – Segment and assume events are present**

Deep Learning for AED using Weak Labels

- **Simplest - Strong Label Assumption Training (SLAT) – Segment and assume events are present**



Deep Learning for AED using Weak Labels

- **Can we do better ?**

- **What would we like to be able to do**

  - Learning process should treat weak labels as weak

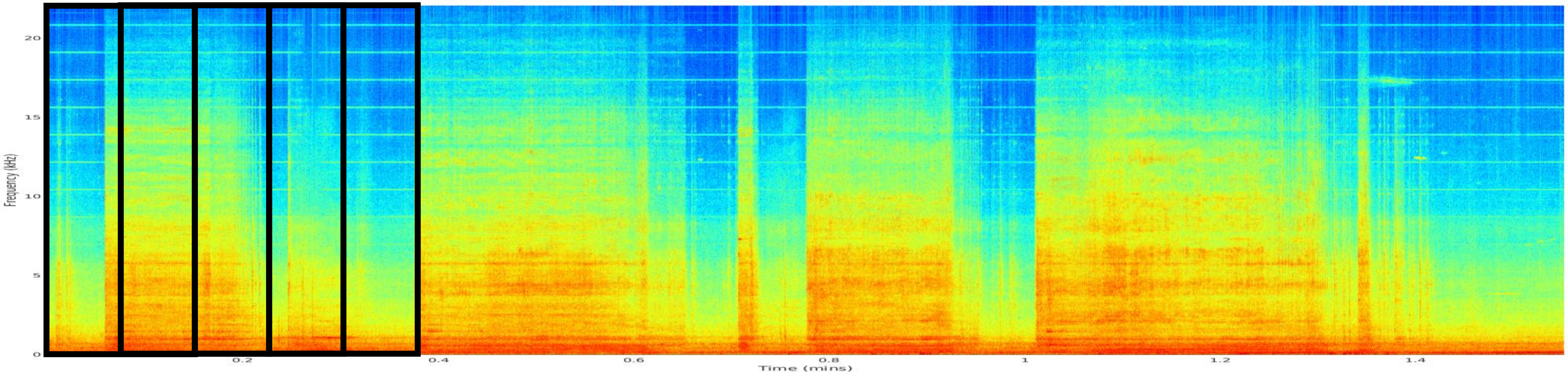
  - Be able to handle recordings of variable length

  - Let the network do the scan and segmentation instead of having a preprocessing

Again the bottom up approach - From segment level posteriors to recording level posteriors

Deep Learning for AED using Weak Labels

Barking, Laughing



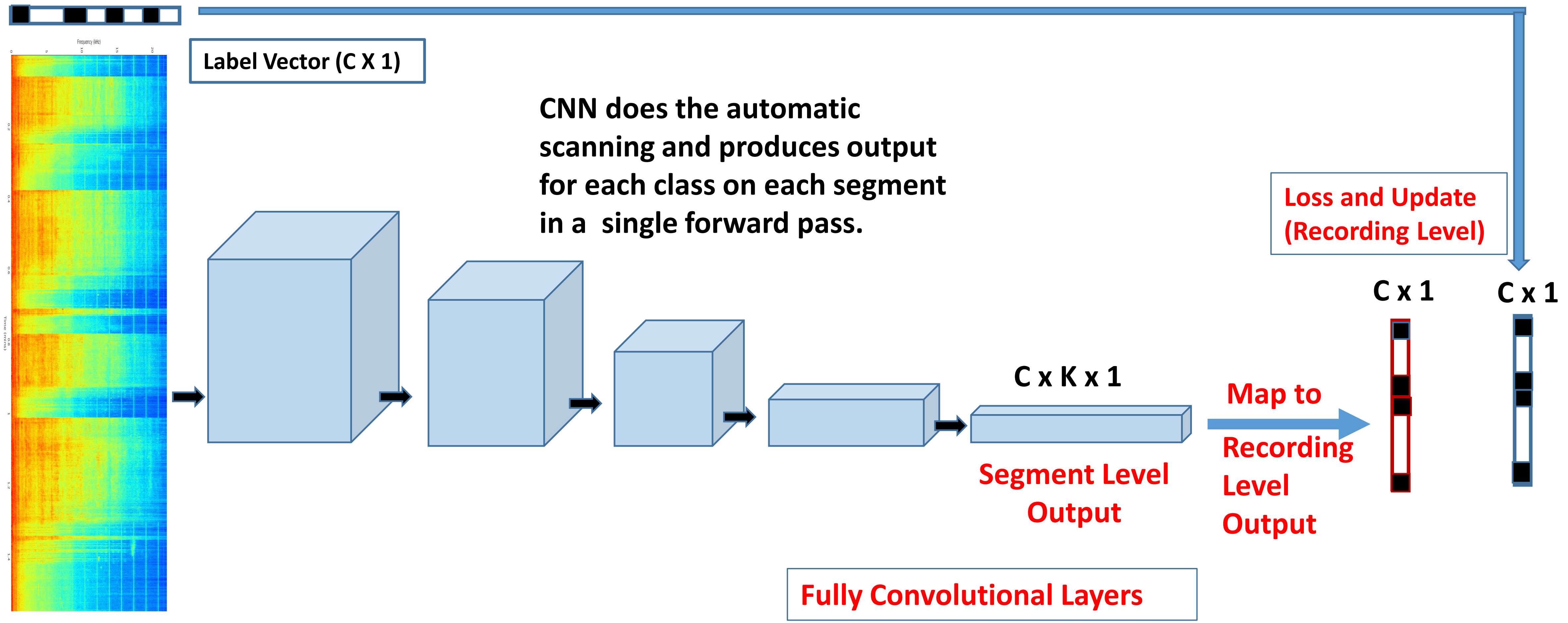
CNN to scan and produce outputs for all segments in one forward pass of whole recording

C X 1 **C X 1** **C X 1** **C X 1**

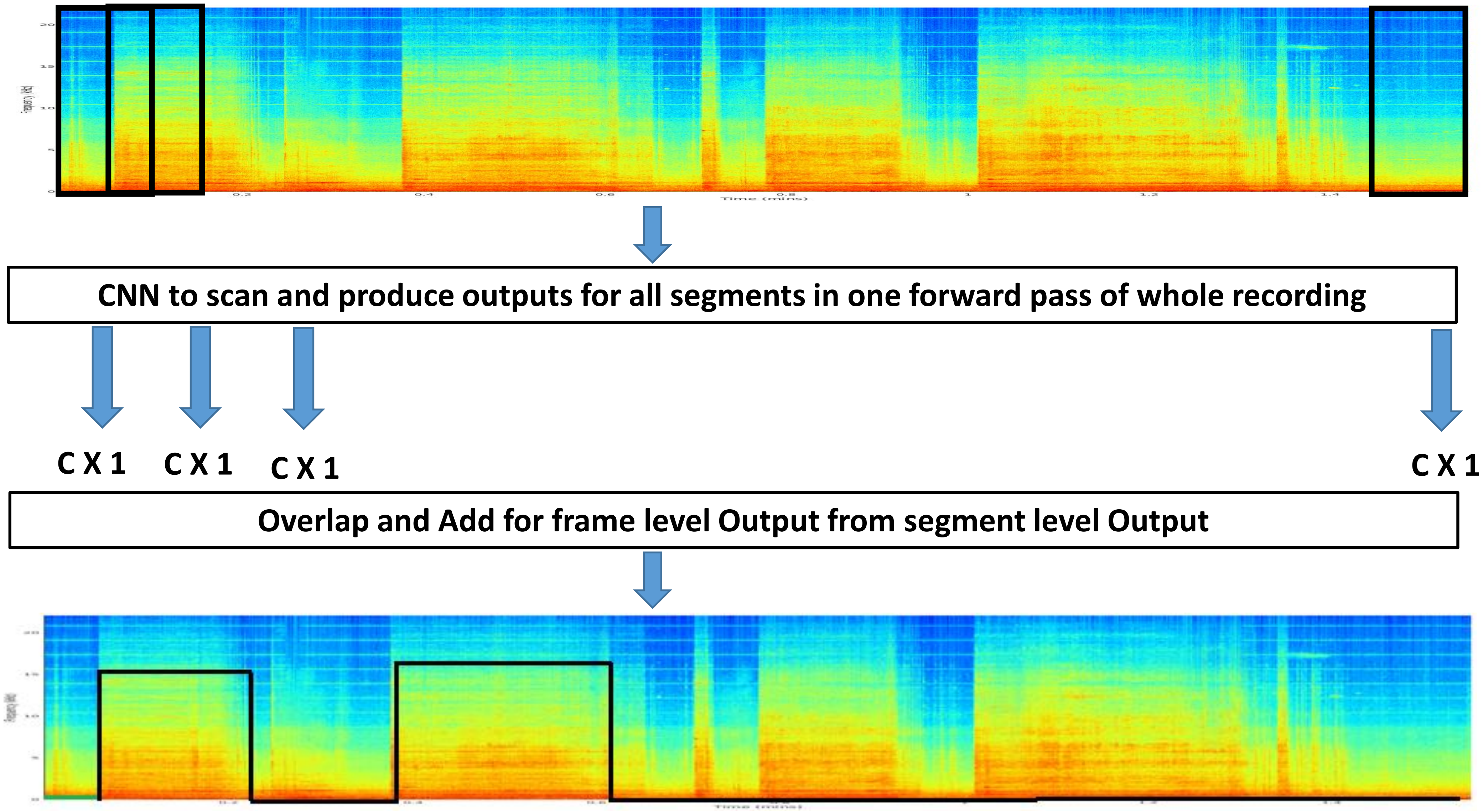
C – Number of classes

Deep Learning for AED using Weak Labels

- **Weak Label Training (WLAT) – Map segment level posteriors to recording level posterior**



Temporal Localization



Deep Learning for AED using Weak Labels

- Can use a variety of methods to map segment level outputs to recording level outputs
  - Simple linear functions – weighted combinations
    - Max – sparse one hot vector
    - Avg – dense
  - Learnable weights -- Attention like
  - A recurrent architecture – using LSTM

Results

- Urbansounds
  - 10 sound events, 27 hours of audio
  - Weak labels
  - Duration – A few seconds to up to several minutes

Comparison of WLAT and SLAT

7.6% improvement in MAP

Event Name	AP		AUC	
	$\mathcal{N}_{slat}$	$\mathcal{N}_{W}$	$\mathcal{N}_{slat}$	$\mathcal{N}_{W}$
Air Conditioner	0.507	0.477	0.807	0.817
Car Horn	0.693	0.834	0.884	0.957
Children Playing	0.774	0.879	0.951	0.978
Dog Bark	0.859	0.918	0.911	0.944
Drilling	0.669	0.622	0.931	0.922
Engine Idling	0.444	0.540	0.795	0.871
Gunshot	0.832	0.929	0.983	0.990
Jackhammer	0.685	0.703	0.940	0.939
Siren	0.703	0.694	0.902	0.954
Street Music	0.800	0.907	0.949	0.978
Mean	**0.697**	**0.750**	**0.905**	**0.935**

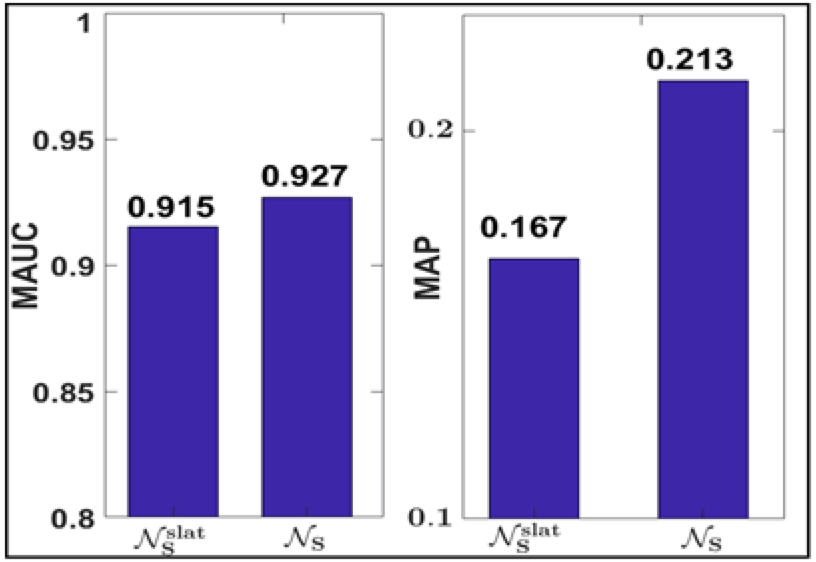
Kumar and Raj, under review IEEE Trans. NNLS

Results – Large Vocabulary

- Audioset
  - 527 sound events
  - Balanced Train Set - ~ 22, 000 mostly 10 second weakly labeled recordings
  - Evaluation set - ~20,000 audio recordings

Results – Large Vocabulary

- Audioset
  - 527 sound events
  - Balanced Train Set - ~ 22, 000 mostly 10 second weakly labeled recordings
  - Evaluation set - ~20,000 audio recordings

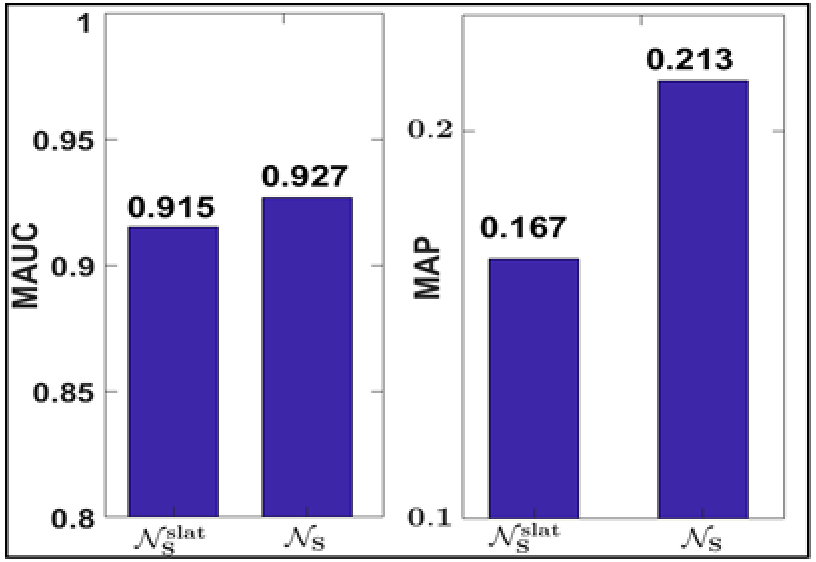


Performance Comparison

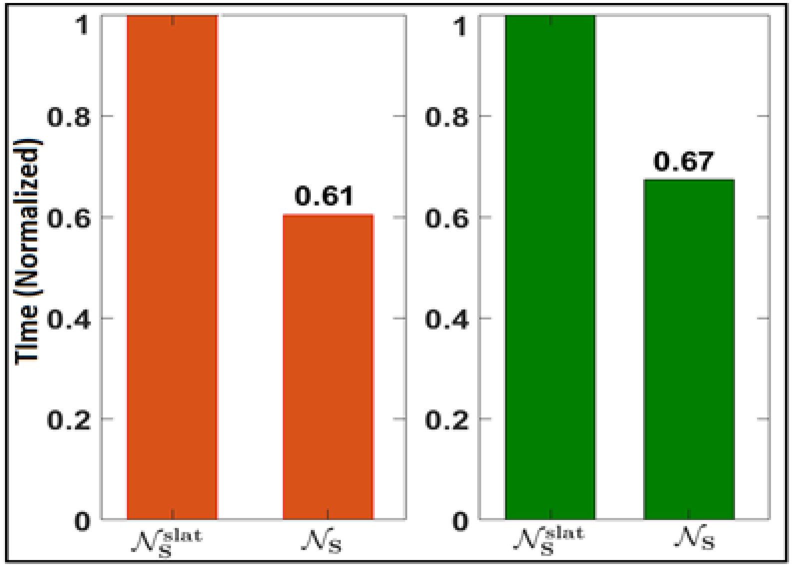
Kumar, Khadkevich and Fügen ICASSP 2018

Results – Large Vocabulary

- Audioset
  - 527 sound events
  - Balanced Train Set - ~ 22, 000 mostly 10 second weakly labeled recordings
  - Evaluation set - ~20,000 audio recordings



Performance Comparison



Relative Training (O) and Inference (G) Time

Results – Large Vocabulary

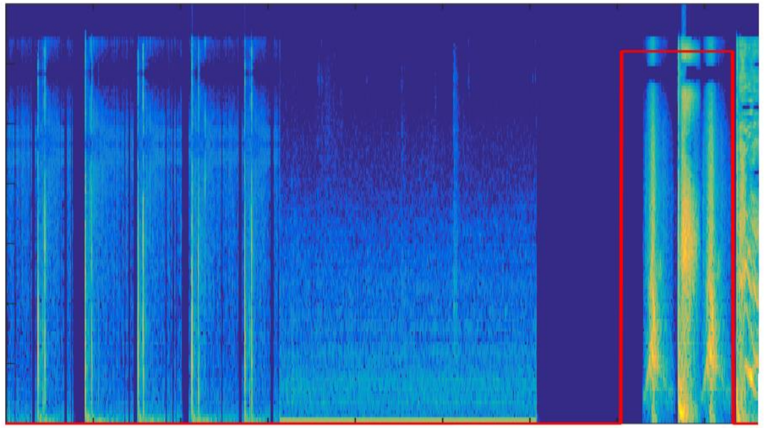
- Audioset
  - Comparison for 10 worst and best performing classes

1. **For 10 "worst" classes – performance almost doubles**
2. **For 10 "best" classes 8.5 % relative improvement**

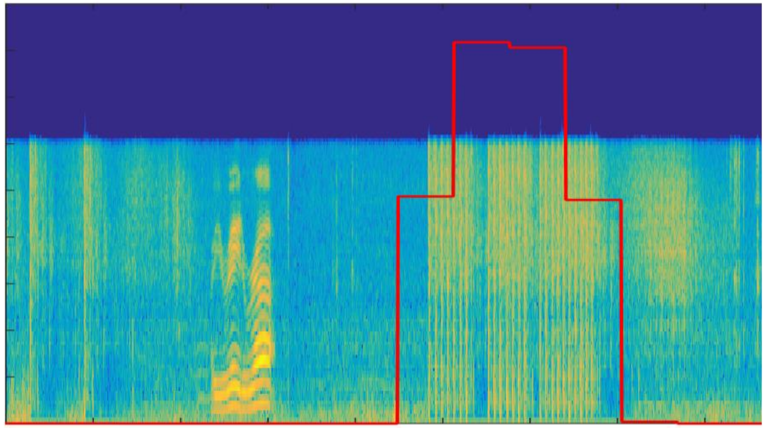
Lowest 10			Highest 10		
Event	$\mathcal{N}_S^{slat}$	$\mathcal{N}_S$	Event	$\mathcal{N}_S^{slat}$	$\mathcal{N}_S$
Scrape	0.0058	0.0092	Music	0.728	0.749
Crackle	0.0078	0.0097	Siren (Civil Defense)	0.671	0.641
Man Speaking	0.0080	0.0202	Bagpipes	0.646	0.786
Mouse	0.0092	0.0368	Speech	0.631	0.661
Buzz	0.0095	0.0077	Purr (Cats)	0.575	0.600
Squish	0.0102	0.0122	BattleCry	0.575	0.651
Gurgling	0.0111	0.0125	Heartbeat	0.559	0.569
Door	0.0115	0.0685	Harpsichord	0.544	0.630
Noise	0.0116	0.0107	Ringing (Campanology)	0.538	0.690
Zipper	0.0121	0.0161	Timpani	0.538	0.528
Mean	**0.0097**	**0.0203**	**Mean**	**0.600**	**0.651**

Deep Learning for AED using Weak Labels

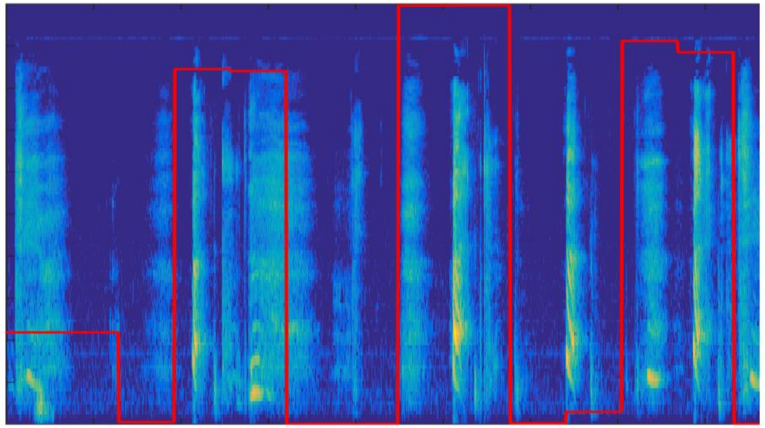
- **Audioset – Examples of Event Localization**



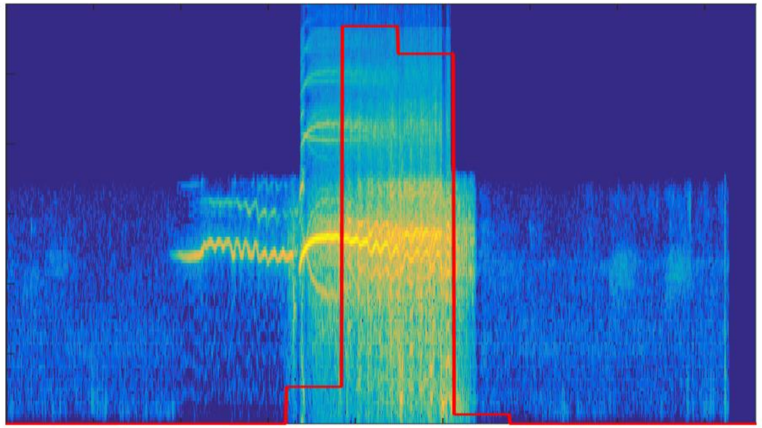
Whoosh-Swoosh-Swish Sound



Machine Gun, Gunfire Sound



Sneeze



Horse (Neigh-Whinny) Sound

Kumar, Khadkevich and Fügen ICASSP 2018

Deep Learning for AED using Weak Labels

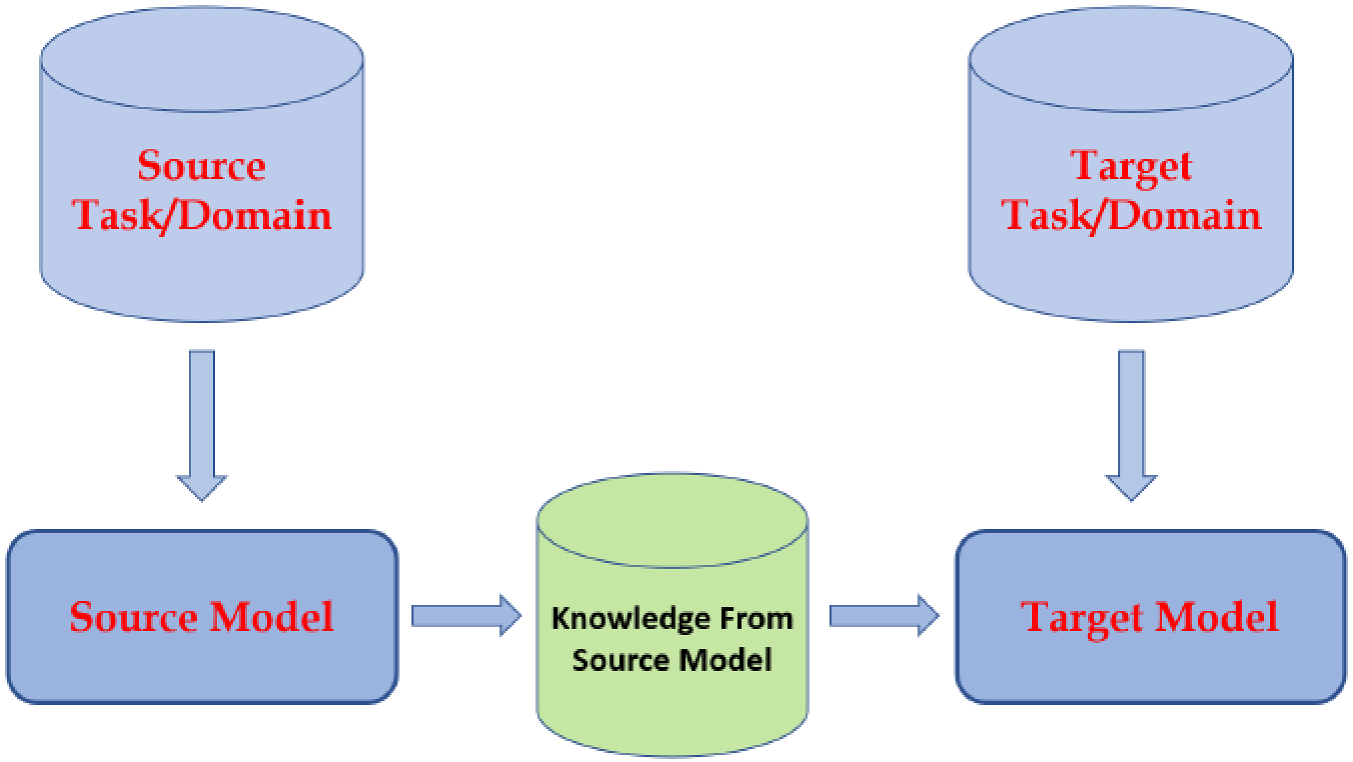
Methods	MAP	MAUC
ResNet-Attention [Xu et al., 2017]	22.0	93.5
ResNet-SPDA [Zhang et al., 2016]	21.9	93.6
M&mnet [Chou et al., 2018]	22.6	93.8
M&mnet (Multiscale) [Chou et al., 2018]	23.2	94.0
WLAT	**22.8**	**93.5**
WLAT (Attention)	**23.1**	**93.1**

Comparison of latest state of the art on Audioset

Deep Learning for AED using Weak Labels

- Large scale learning
- What are we learning
- Can we use these learned knowledge in other tasks

How Useful ? – Transfer Learning

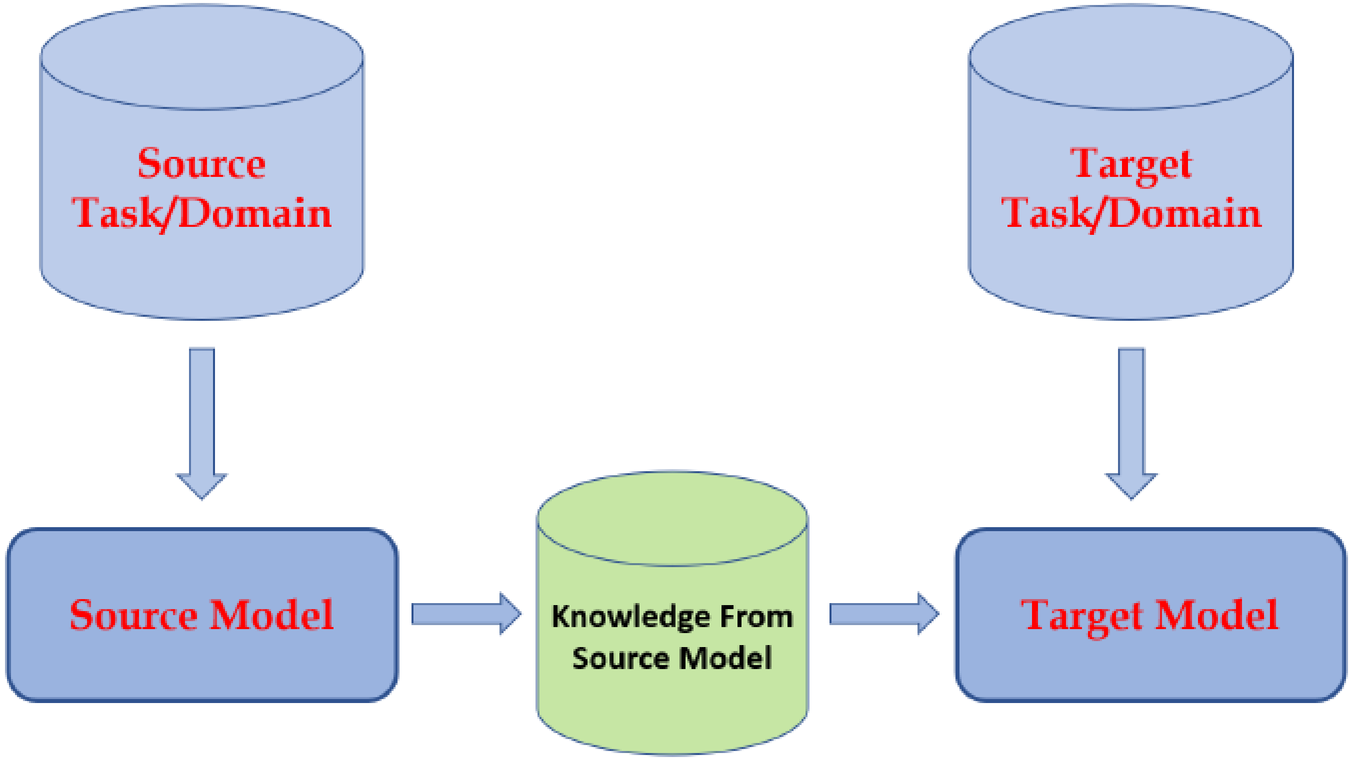


Transfer Learning Basics

Using these large scale models of sounds in other tasks

Kumar et. al. ICASSP 2018

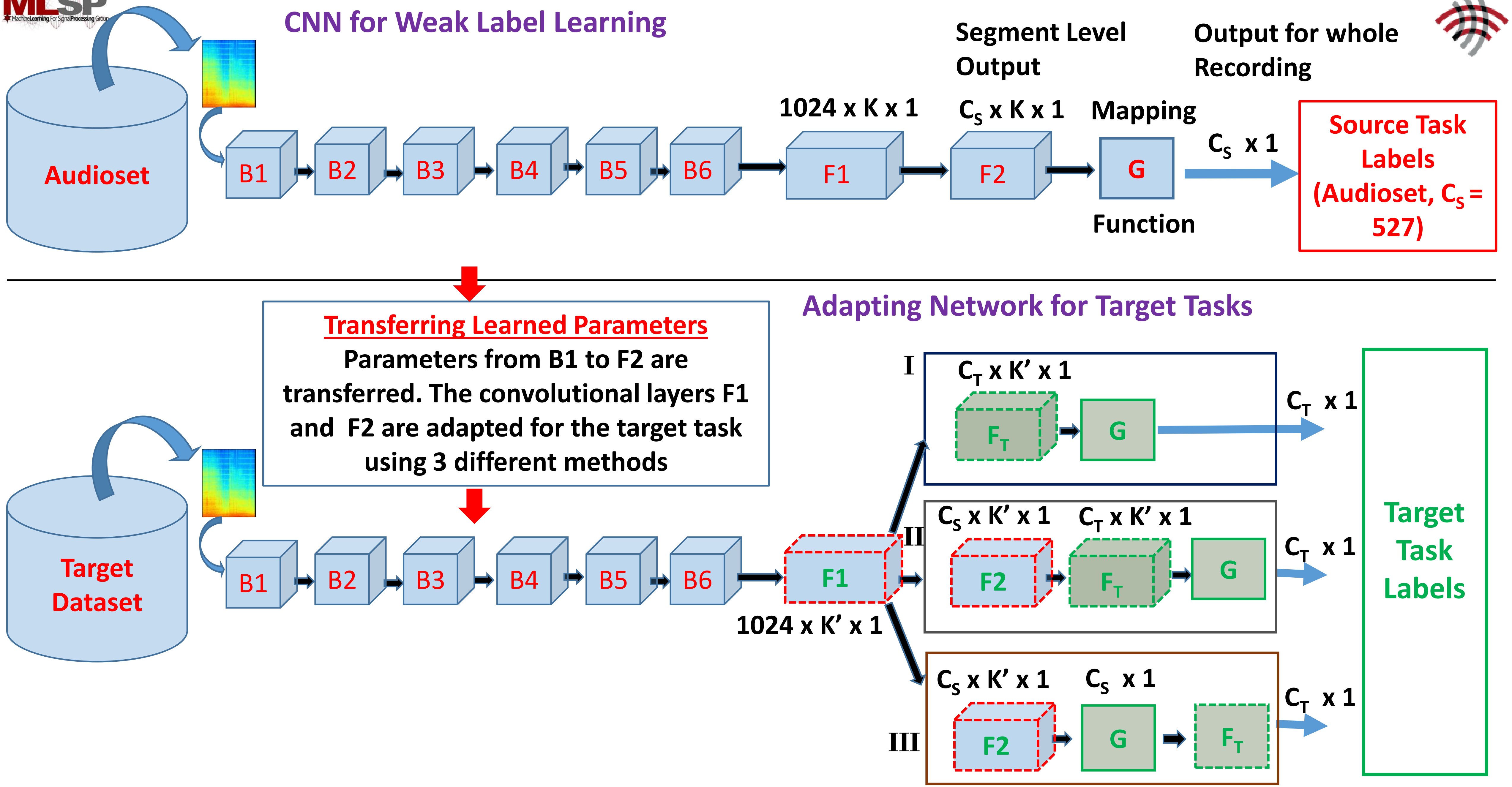
How Useful ? – Transfer Learning



Transfer Learning Basics

Using these large scale models of sounds in other tasks

Kumar et. al. ICASSP 2018


CNN for Weak Label Learning
Audioset
B1
B2
B3
B4
B5
B6
1024 x K x 1
F1
Segment Level Output
C_S x K x 1
F2
Mapping
G
Function
Output for whole Recording
C_S x 1
Source Task Labels (Audioset, C_S = 527)
Transferring Learned Parameters
Parameters from B1 to F2 are transferred. The convolutional layers F1 and F2 are adapted for the target task using 3 different methods
Adapting Network for Target Tasks
Target Dataset
B1
B2
B3
B4
B5
B6
F1
1024 x K' x 1
I
C_T x K' x 1
F_T
G
C_T x 1
II
C_S x K' x 1
C_T x K' x 1
F2
F_T
G
C_T x 1
III
C_S x K' x 1
C_S x 1
F2
G
F_T
C_T x 1
Target Task Labels

How Useful ? – Transfer Learning

Kumar et. al. ICASSP 2018

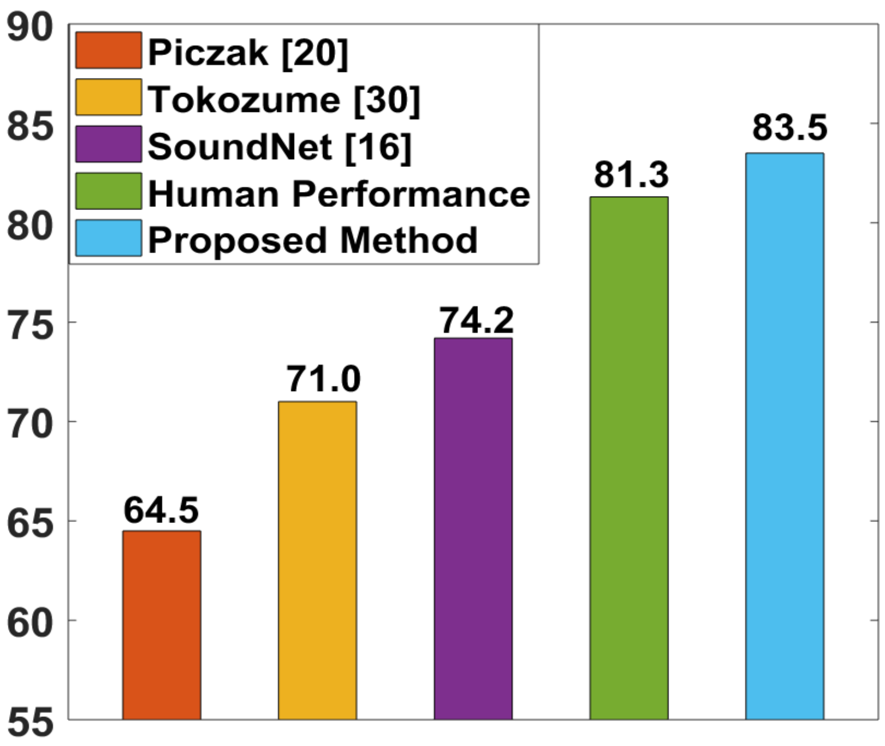
How Useful ? – Transfer Learning

- **Learning Representations**
  - Sets state of art performance on ESC-50 sound events dataset
  - 50 sound events
  - Total 2.7 hours of data

Kumar et. al. ICASSP 2018

How Useful ? – Transfer Learning

- **Learning Representations**
  - Sets state of art performance on ESC-50 sound events dataset
  - 50 sound events
  - Total 2.7 hours of data



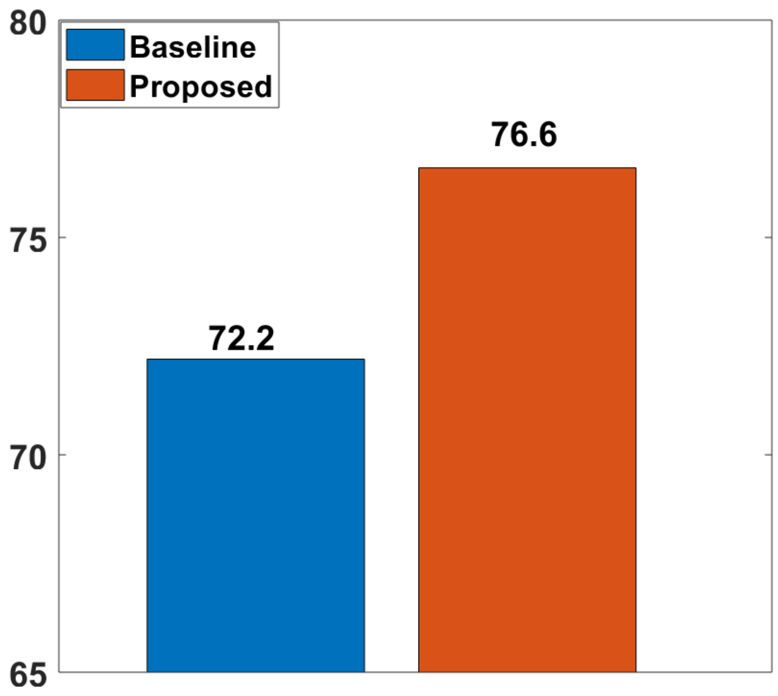
Kumar et. al. ICASSP 2018

Transfer Learning using Weak Labels

- Acoustic Scenes
  - Classification on DCASE 16 task

- Understanding Acoustic Scenes through sound events

  - Establishing relationship - Which events are most active for inputs of a given scene



Kumar et. al. ICASSP 2018

Transfer Learning using Weak Labels

- Established Relations through events most active in a given scene

Scene	Frequent Highly Activated Sound Events
Cafe	Speech, Chuckle-Chortle, Snicker, Dishes, Television
City Center	Applause, Siren, Emergency Vehicle, Ambulance
Forest Path	Stream, Boat Water Vehicle, Squish, Clatter, Noise, Pour
Home	Speech, Finger Snapping, Scratch, Dishes, Baby Cry, Cutlery
Beach	Pour, Stream, Applause, Splash - Splatter, Gush
Park	Bird Song, Crow, Stream, Wind Noise, Stream

Kumar et. al. ICASSP 2018

AED using Weak Labels - A closer Look

- Analyzing Weak Labels
- Density of labels or *weakness* of labels
- Label Noise - No manual labeling

Shah*, Kumar* et. al. [under review]

Unified Framework – Strong + Weak Labels

- **To leverage labeled data in both strong and weak form**
  - **(Wea)kly and Strongly Labeled learning (WEASL)**

- Can Address problems previously mentioned

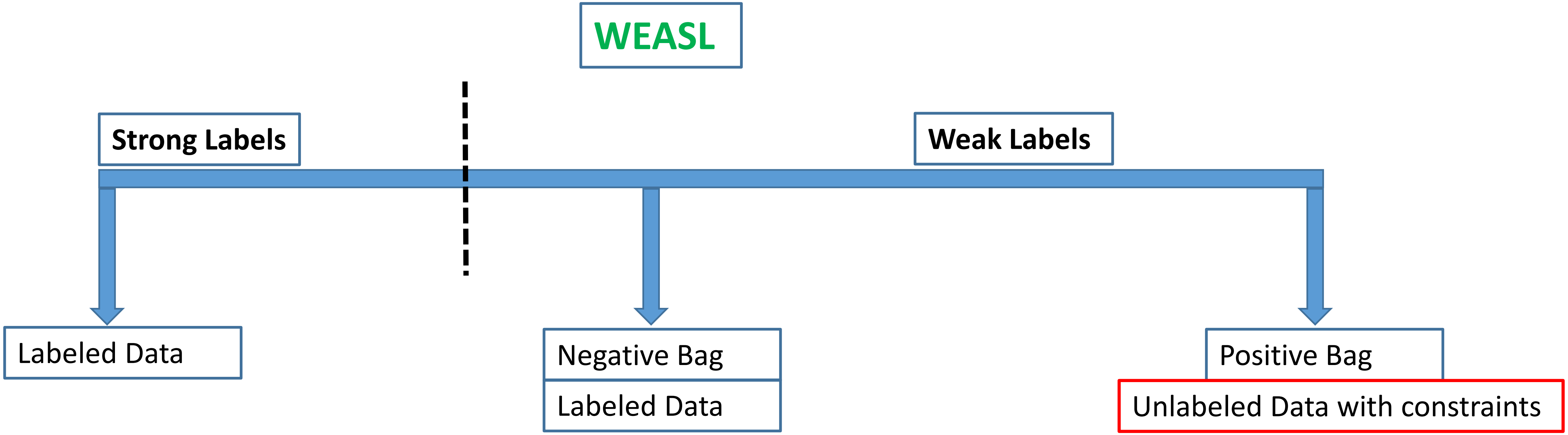
- When strongly labeled available, even in small amount

  - A unified approach is desirable

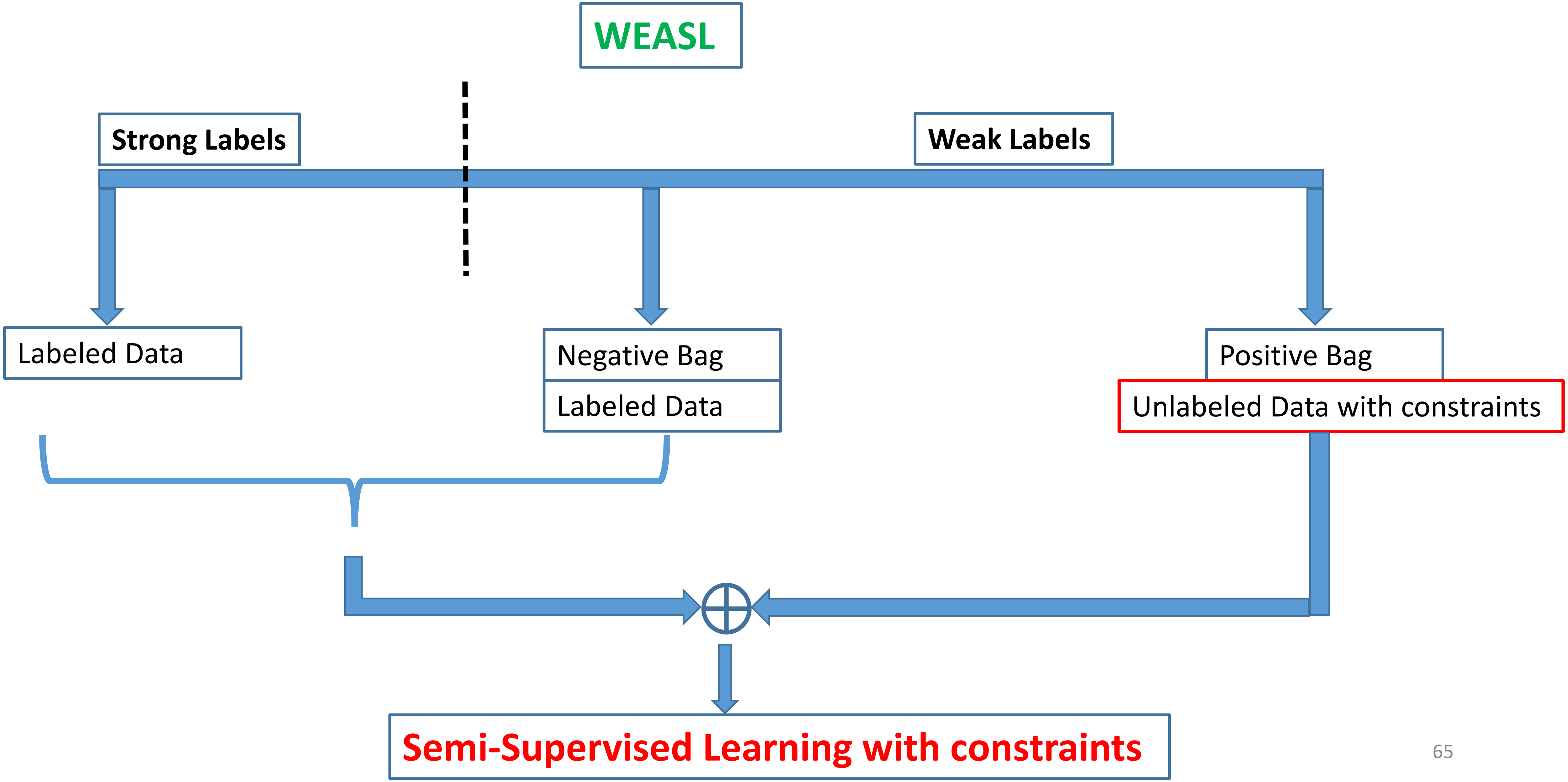
Kumar and Raj, IJCNN 2017

Unified Framework

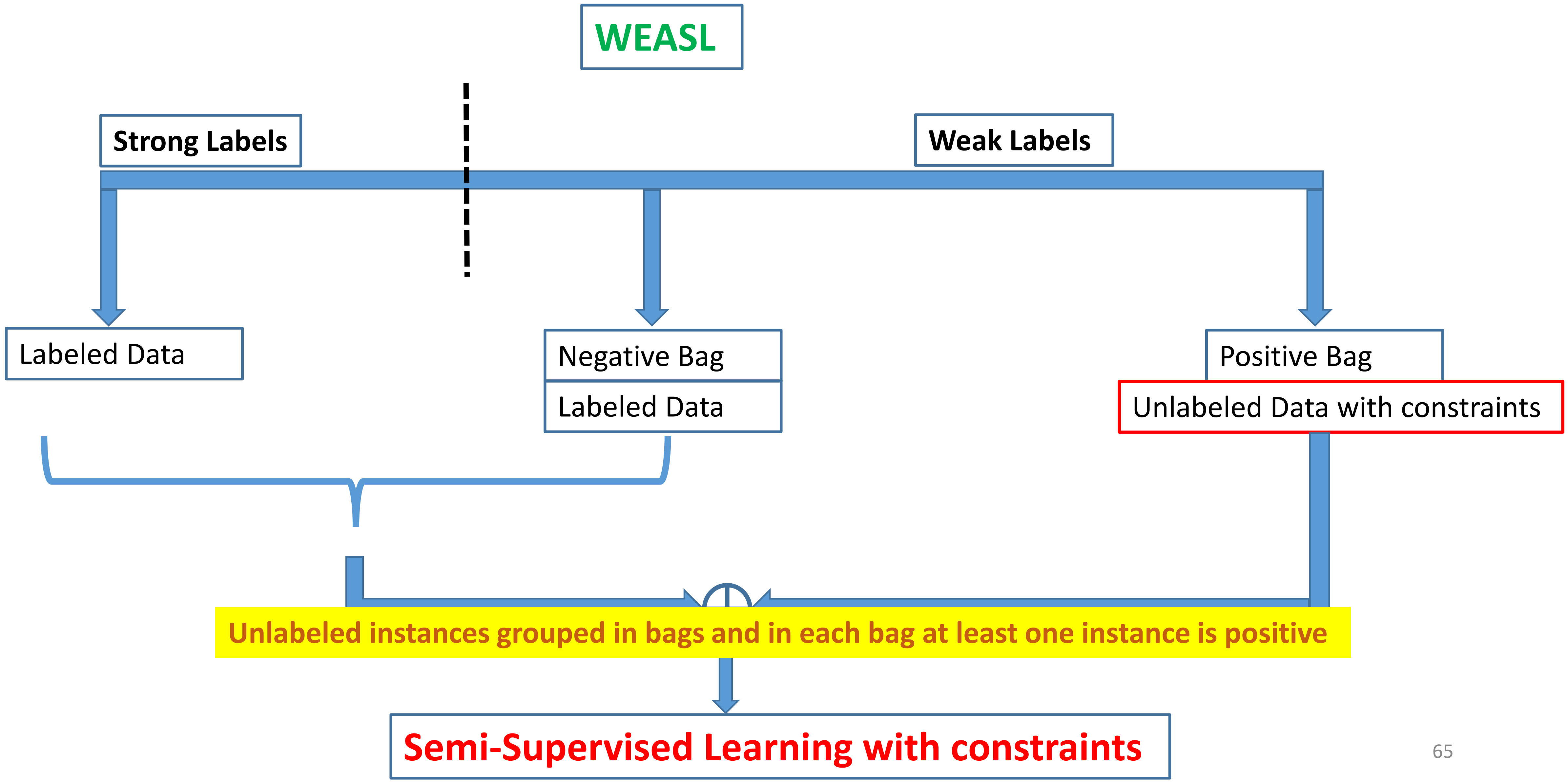
Unified Framework



Unified Framework



Unified Framework



Unified Framework – Graph Based Approach

- **Graph-WEASL -- Using Graph based semi-supervised learning**

- Manifold regularization on graphs

$$\min_{\mathbf{f}} \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - f(\mathbf{x_i}))^{\mathbf{2}} + \lambda_{\mathbf{1}} ||f||_{\mathcal{H}}^{\mathbf{2}} + \lambda_{\mathbf{2}} ||f||_{\mathbf{I}}^{\mathbf{2}}$$

Kumar and Raj, IJCNN 2017

Unified Framework – Graph Based Approach

- **Graph-WEASL -- Using Graph based semi-supervised learning**

- Manifold regularization on graphs

$$\min_{\mathbf{f}} \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} (\mathbf{y_i} - f(\mathbf{x_i}))^{\mathbf{2}} + \lambda_{\mathbf{1}} ||f||_{\mathcal{H}}^{\mathbf{2}} + \lambda_{\mathbf{2}} ||f||_{\mathbf{I}}^{\mathbf{2}}$$

- Add weak label loss

Kumar and Raj, IJCNN 2017

Unified Framework – Graph Based Approach

- **Graph-WEASL -- Using Graph based semi-supervised learning**

- Manifold regularization on graphs

$$\min_{\mathbf{f}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_1||f||^2_{\mathcal{H}} + \lambda_2||f||^2_{I}$$

- Add weak label loss

$$\min_{\mathbf{f}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_1||f||^2_{\mathcal{H}} + \lambda_2||f||^2_{I}$$

$$+\frac{\lambda_3}{T}\sum_{t=1}^{T}(1 - \max_{j=p_t,\ldots,q_t} f(x_j))^2$$

Unlike Previous Case it is non-convex

Unified Framework – Graph Based Approach

- Solved through Convex-Concave Procedure (CCCP)

$$\min_{\alpha,\xi} \ (Y - JK\alpha)^T (Y - JK\alpha) + \lambda_1 \alpha^T K\alpha$$

$$+ \lambda_2 \frac{1}{N^2} \alpha^T KLK\alpha + \lambda_3 \sum_{t=1}^{T} \xi_t^2$$

$$s.t$$

$$1 - (\max_{j=p_t \dots q_t} K_j' \alpha^{(k)} + \sum_{j=p_t}^{q_t} \delta_{tj}^{(k)} K_j' (\alpha - \alpha^{(k)})) \leq \xi_t$$

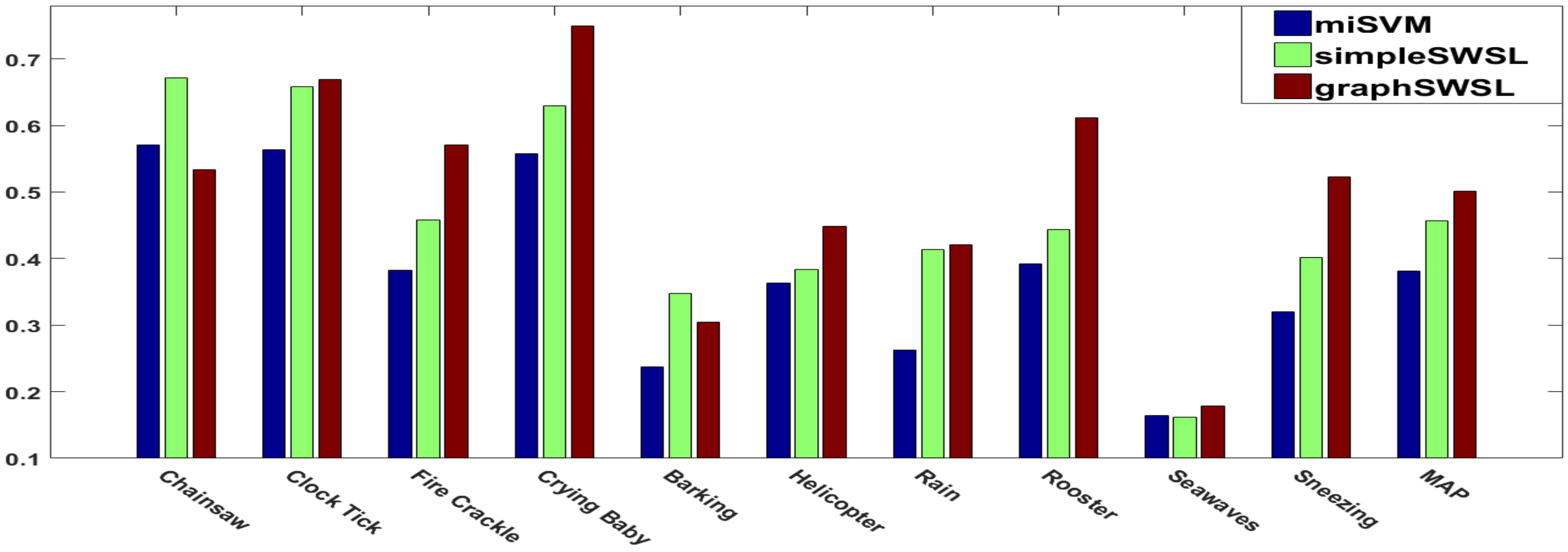
$$t = 1, ..., T$$

$$\xi_t \geq 0, \ t = 1, ..., T$$

Kumar and Raj, IJCNN 2017

Unified Framework – Graph Based Approach

- Weakly Labeled Data – Youtube
- Strongly labeled data – ESC-10 dataset
  - Approx. 1/13 of weakly labeled data



MAP Improvement

Simple-WEASL +7%
Graph-WEASL +12%

Kumar and Raj, IJCNN 2017

Summary So Far

Acoustic Intelligence in Machines

Knowledge base for Sounds

<u>Audible Phrases or Sonic Phrases</u>

1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences

<u>Relational and Commonsense Knowledge</u>

1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds

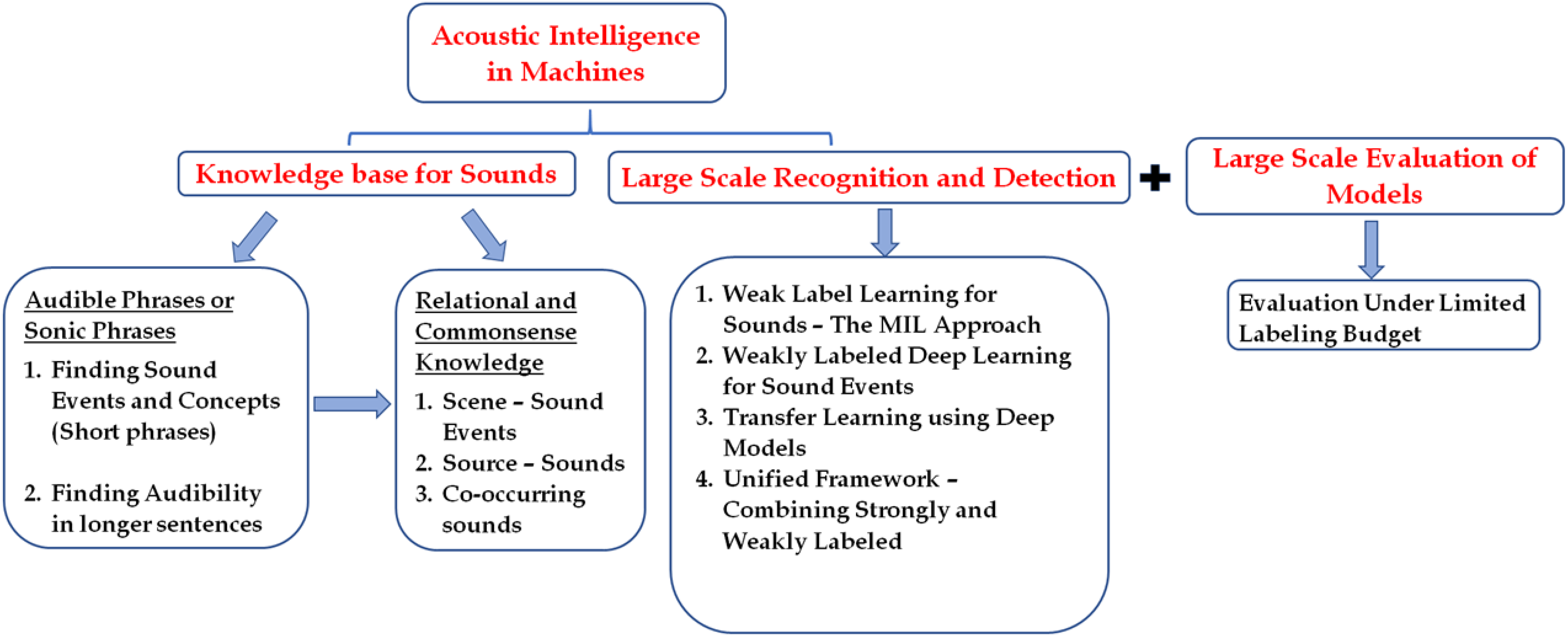
Large Scale Recognition and Detection

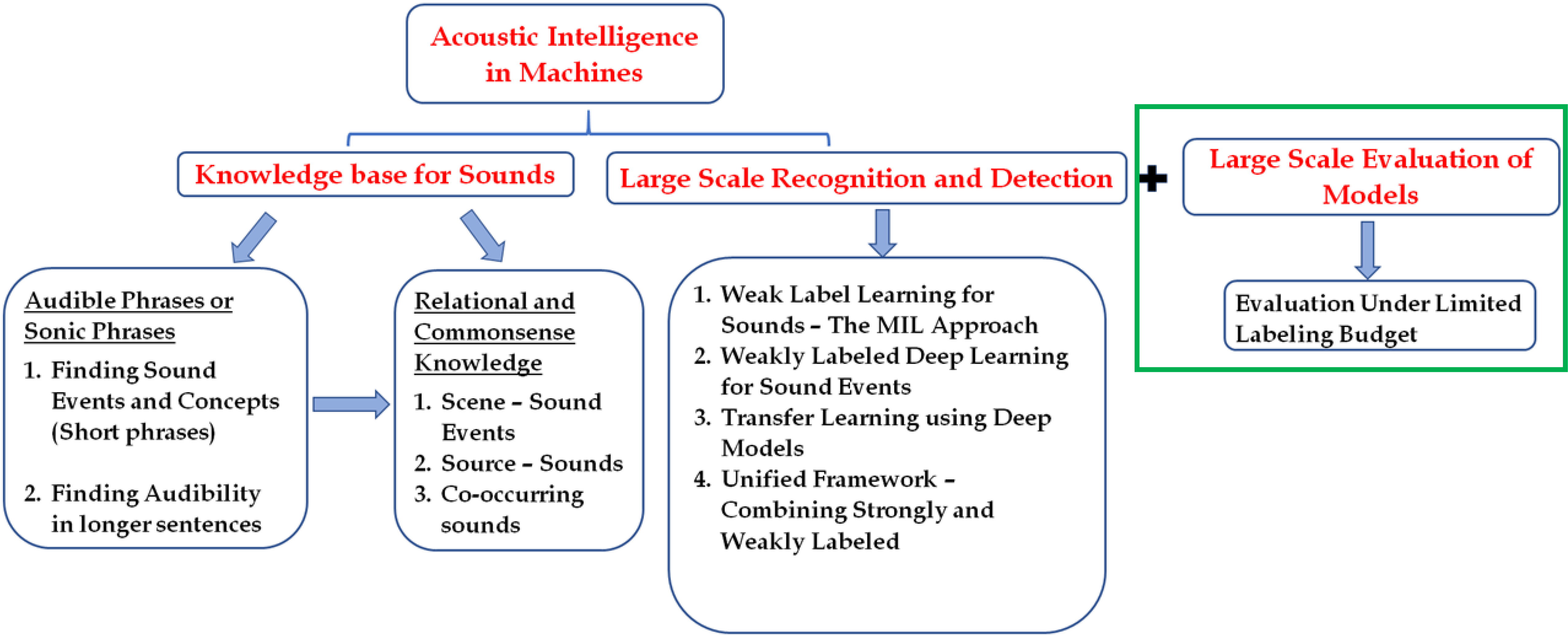
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled

\+

Large Scale Evaluation of Models

Evaluation Under Limited Labeling Budget


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget


Acoustic Intelligence in Machines
Knowledge base for Sounds
Large Scale Recognition and Detection
Large Scale Evaluation of Models
Audible Phrases or Sonic Phrases
1. Finding Sound Events and Concepts (Short phrases)
2. Finding Audibility in longer sentences
Relational and Commonsense Knowledge
1. Scene – Sound Events
2. Source – Sounds
3. Co-occurring sounds
1. Weak Label Learning for Sounds – The MIL Approach
2. Weakly Labeled Deep Learning for Sound Events
3. Transfer Learning using Deep Models
4. Unified Framework – Combining Strongly and Weakly Labeled
Evaluation Under Limited Labeling Budget

Evaluation Under Limited Labeling Budget

Evaluation of Models

- Large scale learning – train and test on large data
  - Fixed labeling budget – where to spend budget ?

- Large scale testing
  - A audio/multimedia event detection system testing on Youtube
  - A text categorization, semantic content analysis system classifying webpages
  - For audio event detection [ICASSP 2018]

- **Can label only a small number of test samples**

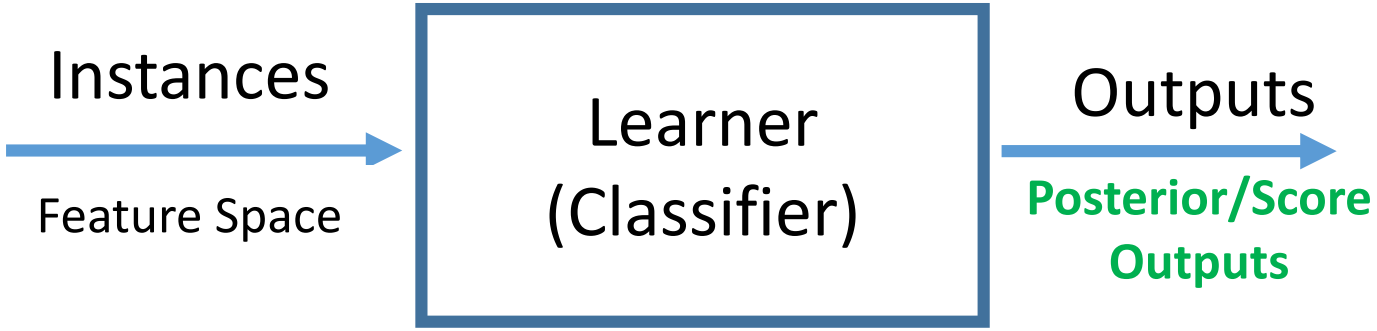
Badlani, etc. ICASSP 2018

Evaluation of Models

- **How to precisely estimate performance using as little labeling resource as possible ?**
  - For a fixed labeling budget

Evaluation of Models

- **How to precisely estimate performance using as little labeling resource as possible ?**
  - For a fixed labeling budget



Evaluation of Models

- **How to precisely estimate performance using as little labeling resource as possible ?**
  - For a fixed labeling budget

Evaluation of Models

- **How to precisely estimate performance using as little labeling resource as possible ?**
  - For a fixed labeling budget

Evaluation of Models

- **How to precisely estimate performance using as little labeling resource as possible ?**
  - For a fixed labeling budget

- Select instances for labeling for accuracy estimation

Accuracy Estimation: Random Sampling Estimate

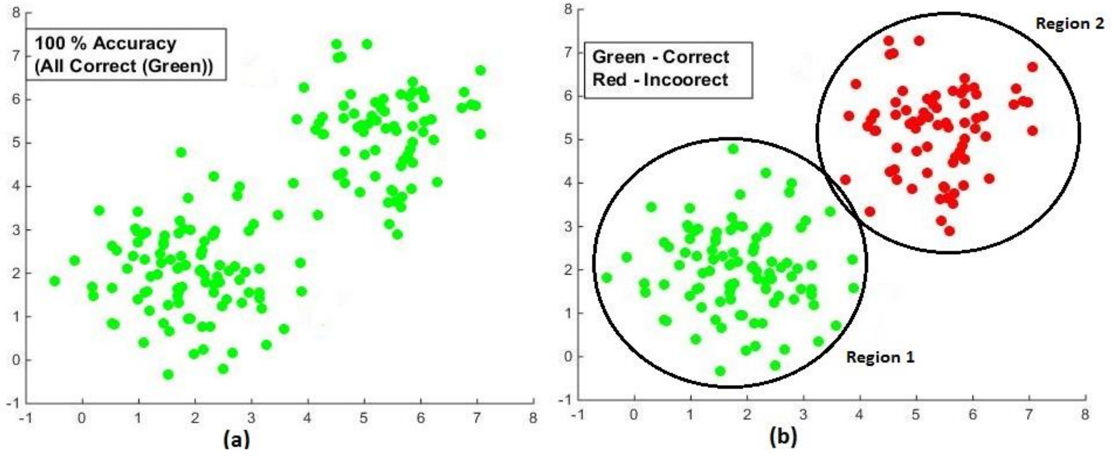
- Random Sampling – Naïve Solution
- Ignoring what the classifier is doing
- Ignoring how the instances are distributed
- **Inefficient – High Variance**
  - Estimated accuracy can be far off from true accuracy

Accuracy Estimation: Limited Budget

- Reformulate the problem – How many samples to estimate accuracy ?

Accuracy Estimation: Limited Budget

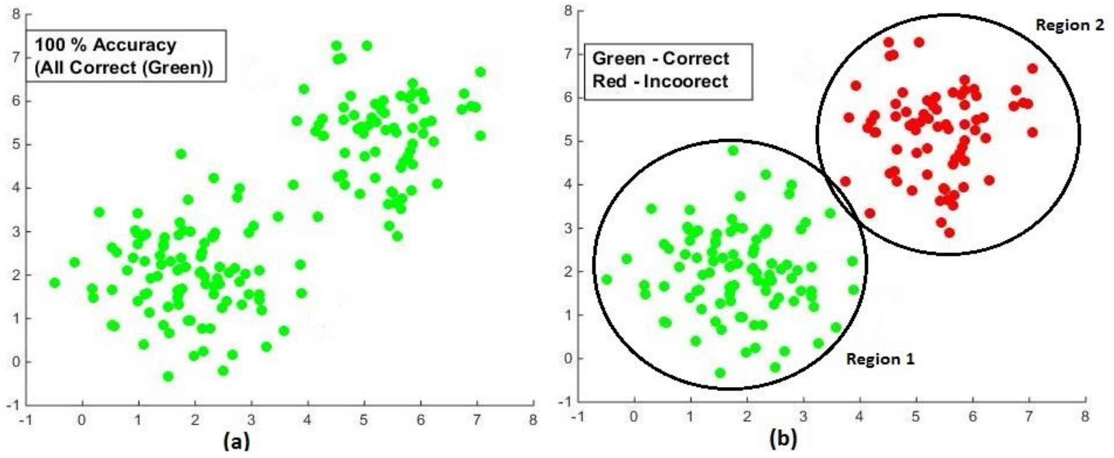
- Reformulate the problem – How many samples to estimate accuracy ?



Accuracy Estimation: Limited Budget

- Reformulate the problem – How many samples to estimate accuracy ?

- **Cases**
  - Case (a) – Label just one sample
  - Case (b) – One sample from each reason
  - Accuracy = $(1*N_1 + 0*N_2)/N$

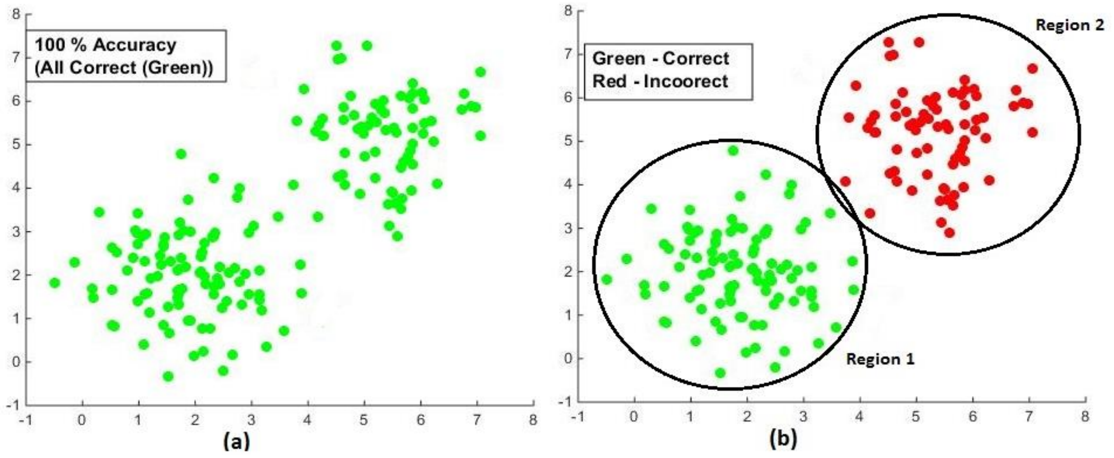


Kumar and Raj, PAKDD 2018

Accuracy Estimation: Limited Budget

- Reformulate the problem – How many samples to estimate accuracy ?

- **Cases**
  - Case (a) – Label just one sample
  - Case (b) – One sample from each reason
  - Accuracy = $(1*N_1 + 0*N_2)/N$



- Homogeneity within a region
  - Low variance

Kumar and Raj, PAKDD 2018

Accuracy Estimation: Stratified Estimate

- Main Idea !!
  - Group data (Stratify) such that each group is as homogeneous as possible
  - Sample more from groups which are less homogeneous

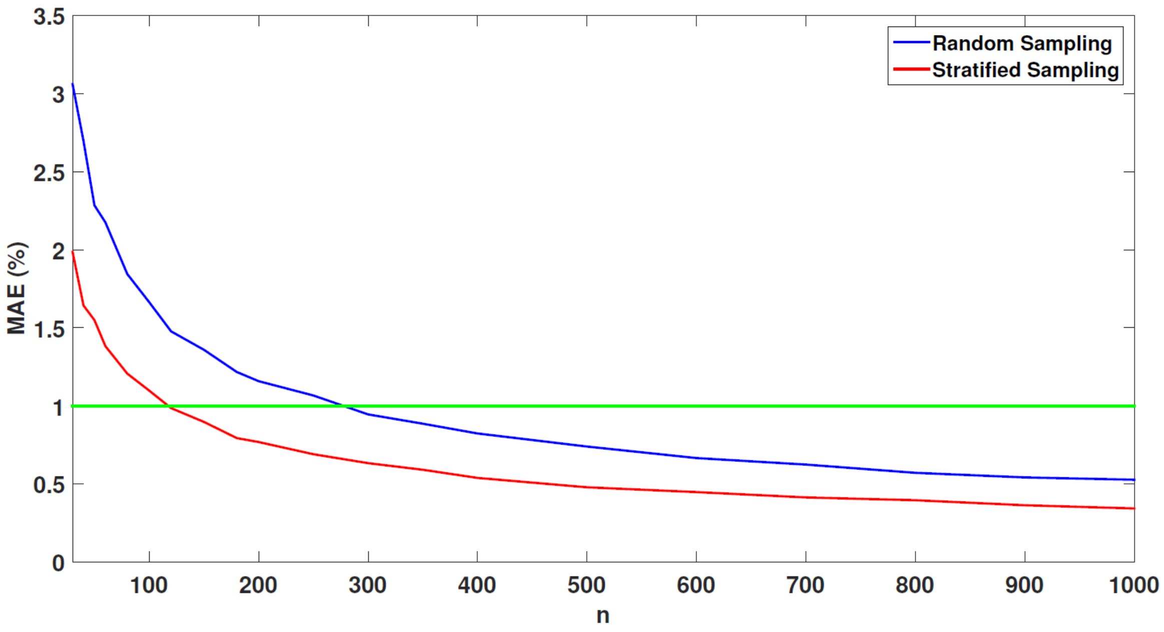
- **Stratification and Allocation Methods**

- **Significance - Reduces the variance of the estimator**
  - Optimal Allocation (minimum variance)

Accuracy Estimation

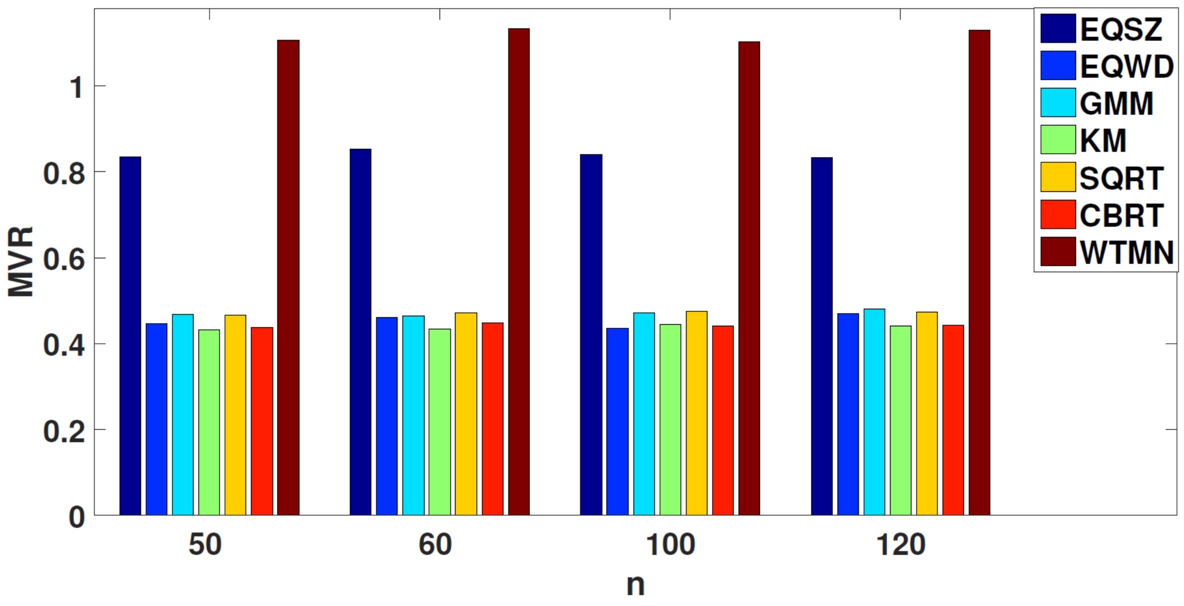
rcv1 dataset – 0.7 million test instances

58% reduction in labeling budget !!



Estimation Error vs Labeling Budget

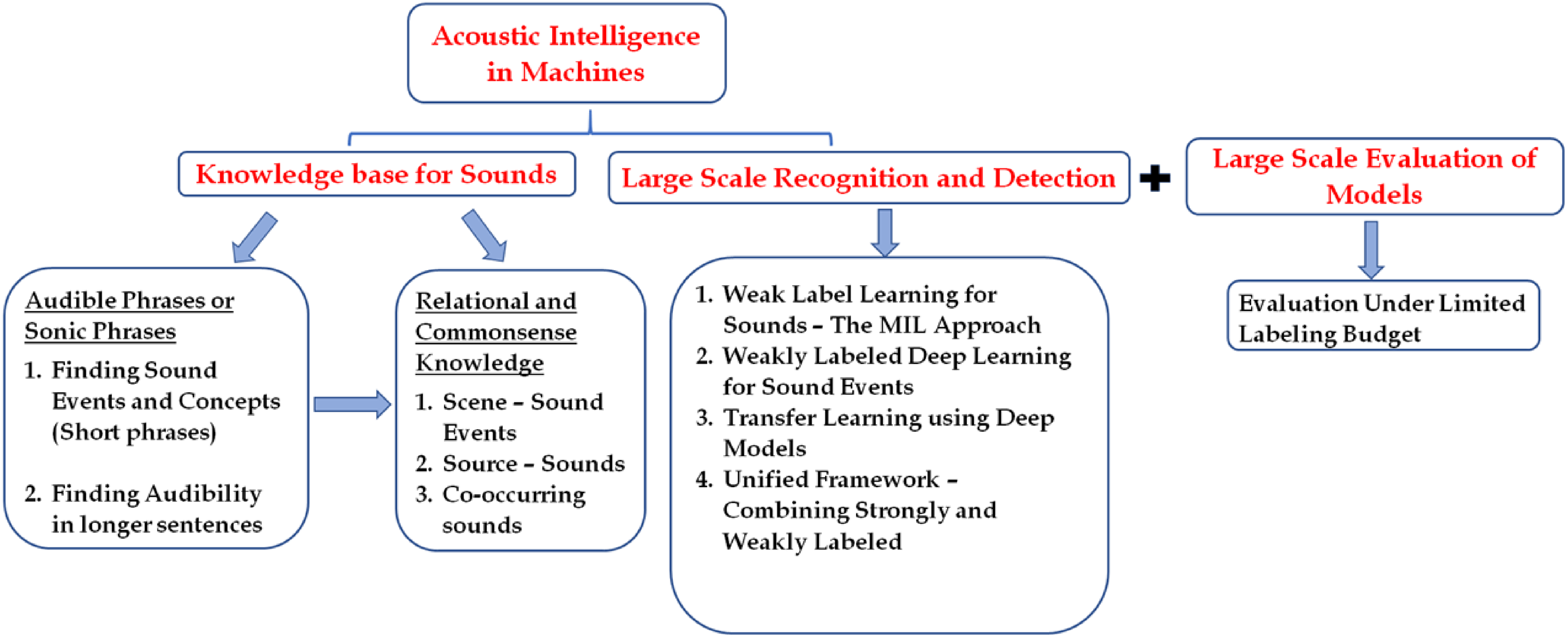
Upto 60 % reduction in variance



Mean Variance Ratio at different labeling budget

Kumar and Raj, PAKDD 2018

Summary and More!



Summary and More!

- Applications –
  - Query by example retrieval [ICASSP 2018]
  - Geotagging [Interspeech 2017]
  - Never Ending Learning of Sounds

Interspeech 2017 – Kumar et. al. , ICASSP 2018 – Manocha et. al.

Summary and More!

- Knowledge of sounds
  - Other relations
- Learning from weakly labeled without manual labeling
- Linking the two sub-problems
- Multimodal understanding
  - Incorporating visual understanding
  - Relating to visual objects


thank you for attention
any
questions ???????
quickmeme.com