Deep Learning

“Master Class”

17 December 2019

Univ.-Prof. Dr. habil. Björn W. Schuller

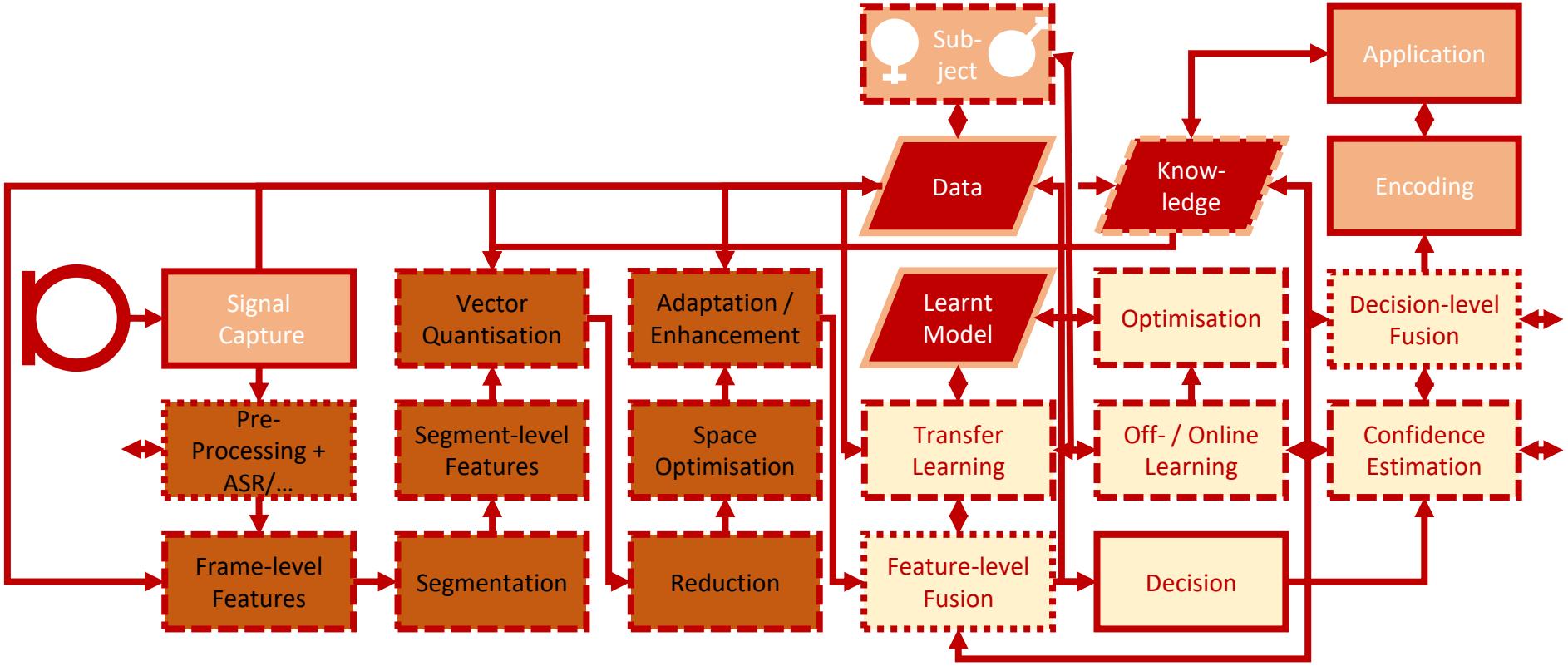
Faculty of Applied Computer Science
Faculty of Medicine

Universität Augsburg
Embedded Intelligence for Health Care and Wellbeing

End-to-End.

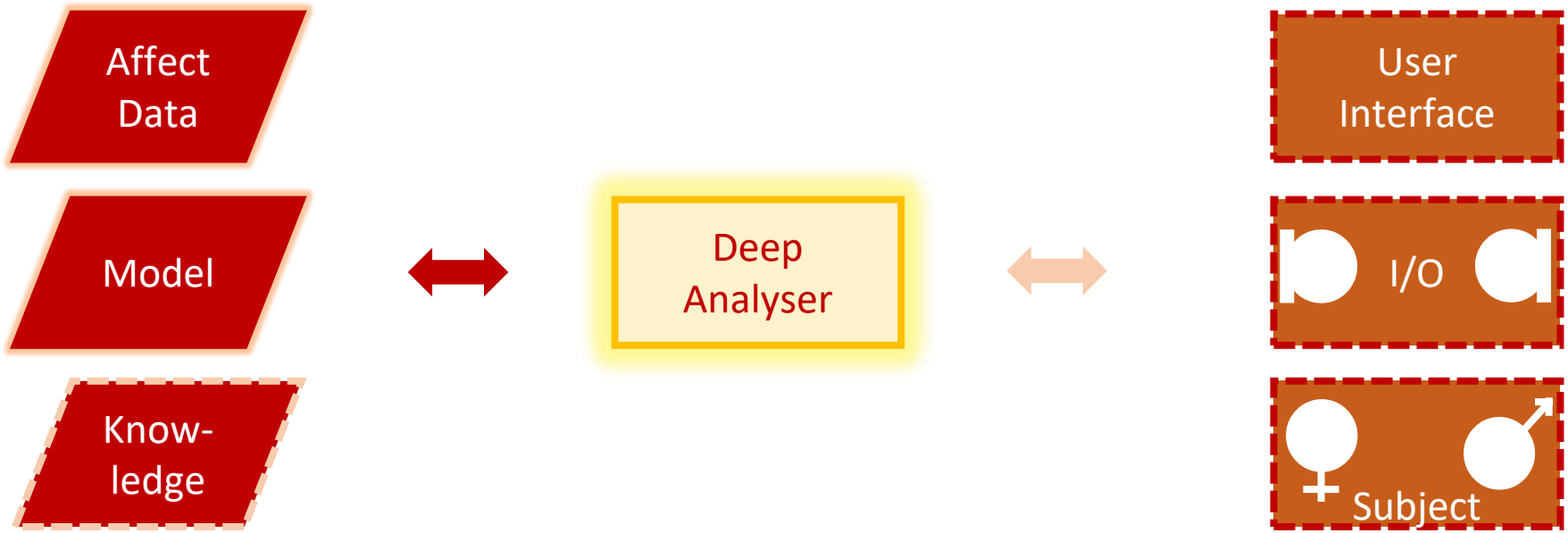
Pattern Recognition 1.0?

- **The "Traditional" Engine**



Pattern Recognition 2.0?

- **The "Modern" Engine?**



1: Preprocessing.

- **openBliSSART**

  Non-negative

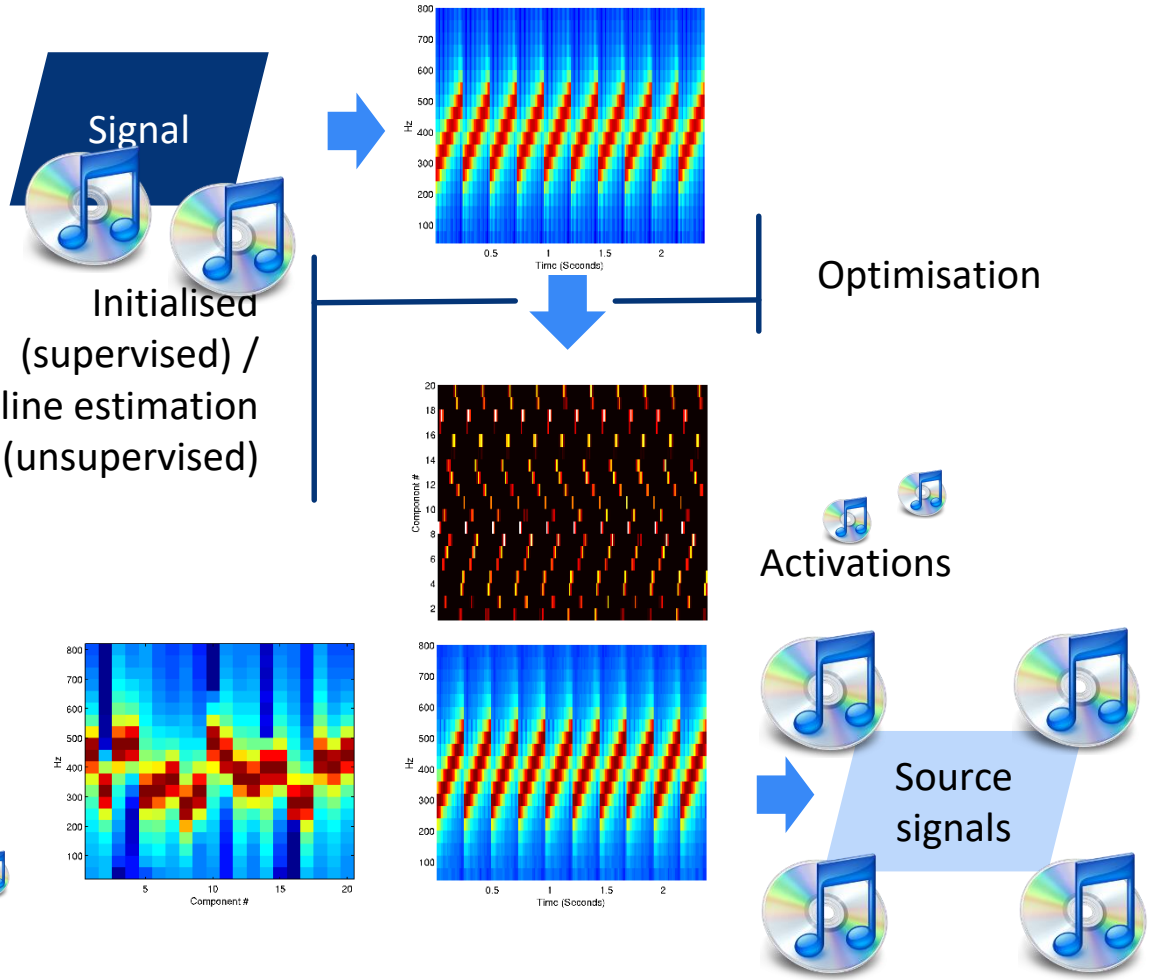
  Matrix Factorisation

  Iterative

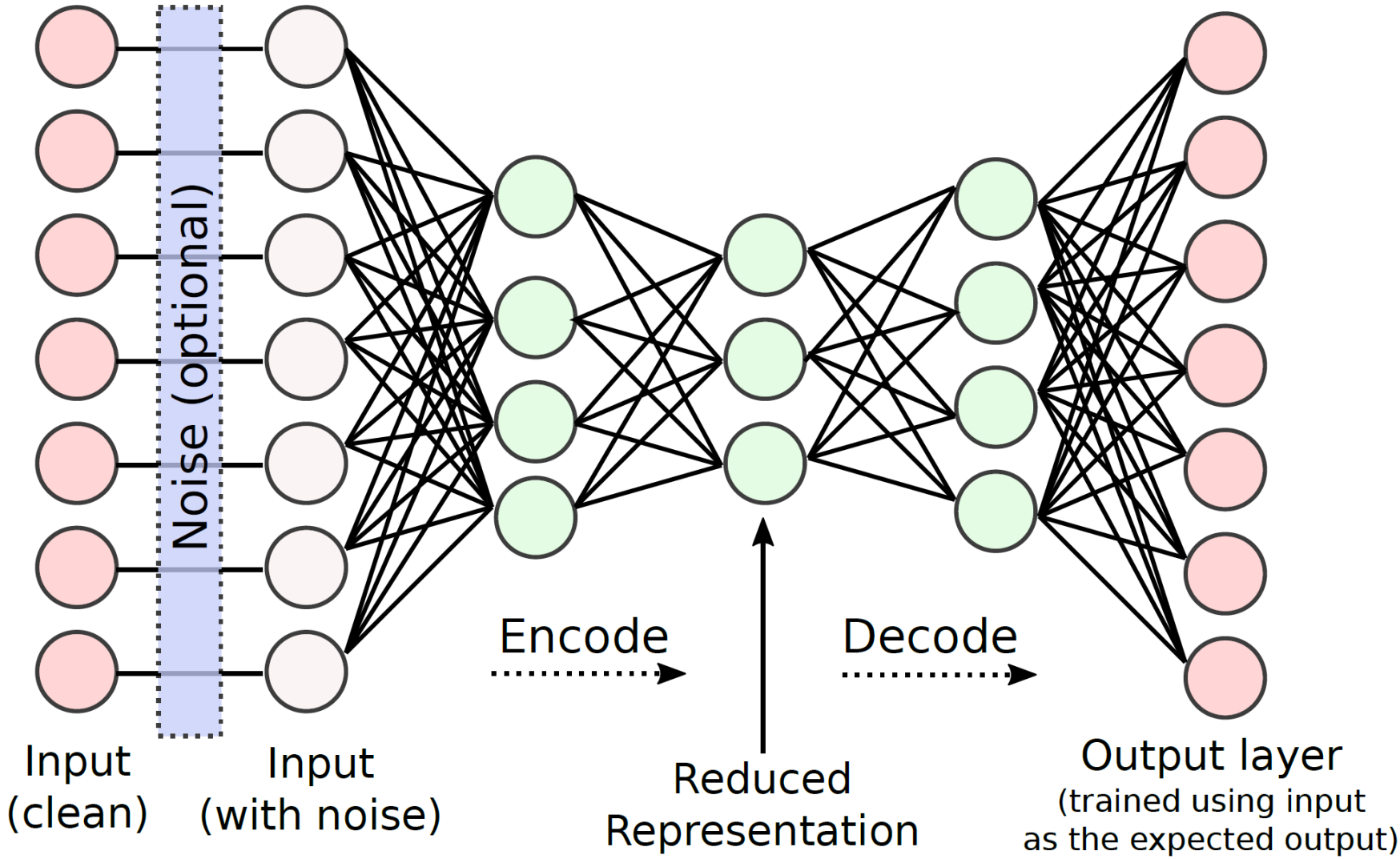
  Optimisation

  Find (local) min of

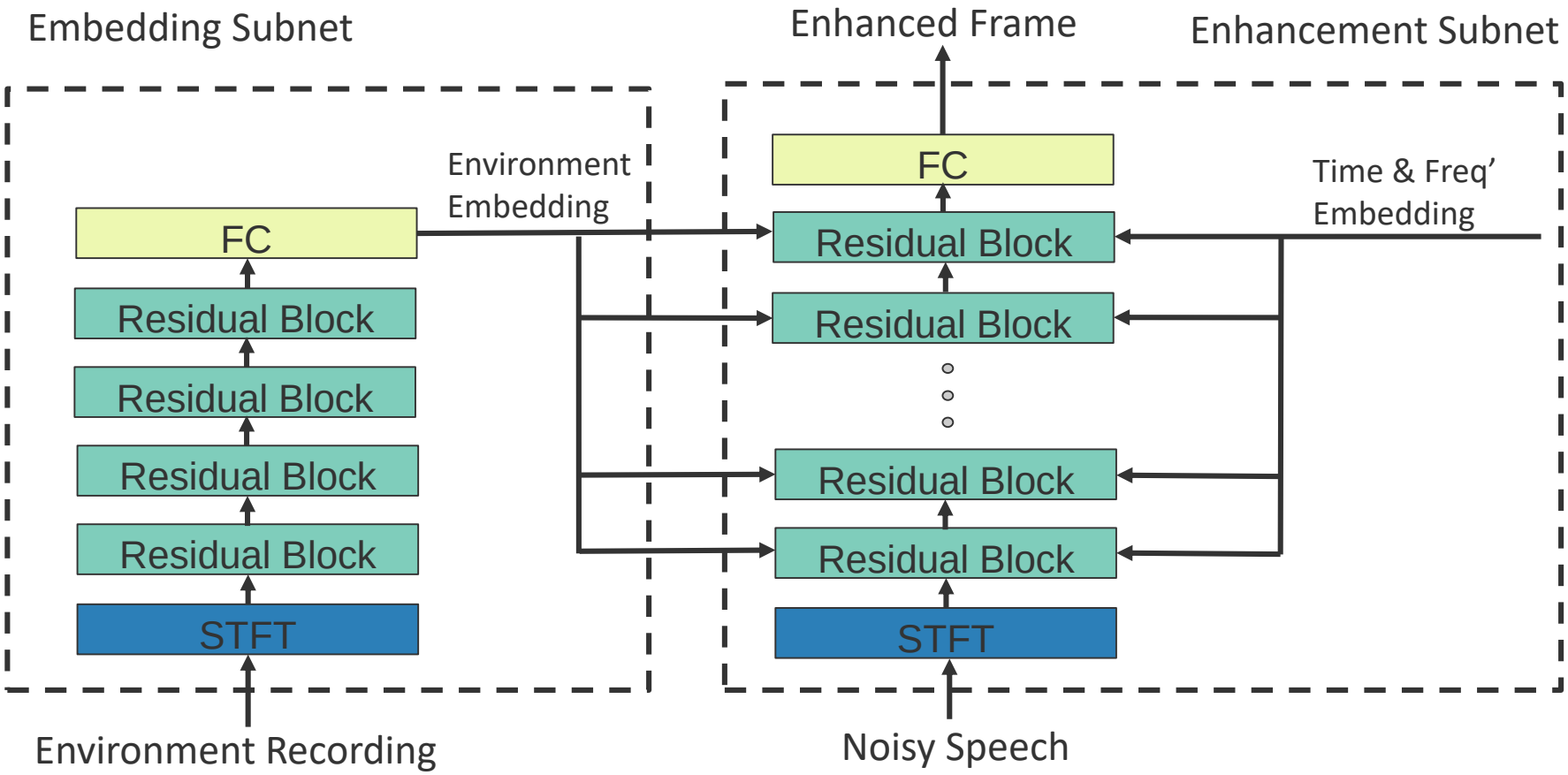
  $E(\mathbf{Z}.\mathbf{H}) = D(\boldsymbol{X}.\mathbf{ZH})$



1: Preprocessing.

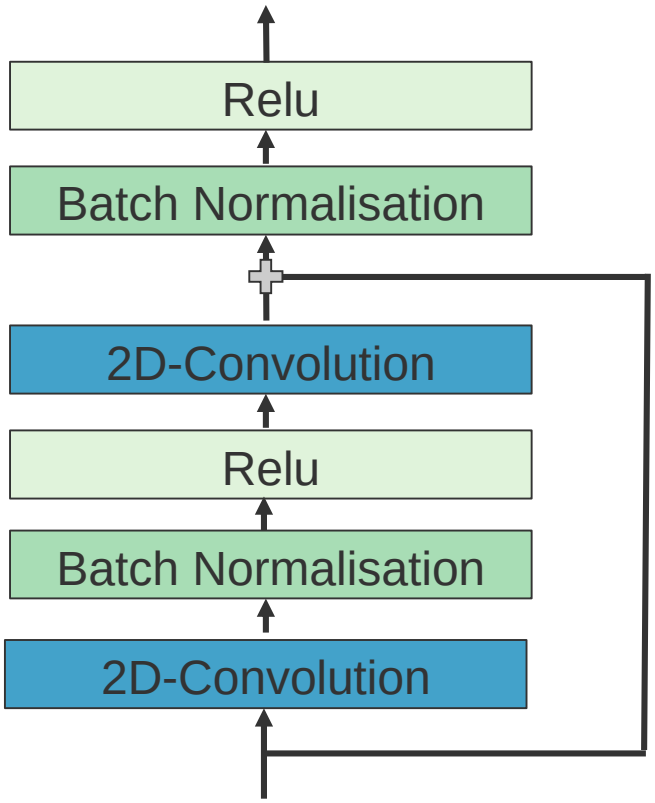


1: Preprocessing



1: Preprocessing

A Residual Block:



	WER	PESQ	SegSNR	LSD
Clean Speech	4.21	–	–	–
Noisy Speech	34.04	2.59	7.02	0.94
Log-MMSE	35.38	2.66	7.12	0.91
Noise Aware	25.30	2.96	11.01	0.54
w/o Embed.	16.78	3.25	11.71	0.48
w/ Embed.	15.46	3.30	12.99	0.45

25 dB

20 dB

15 dB

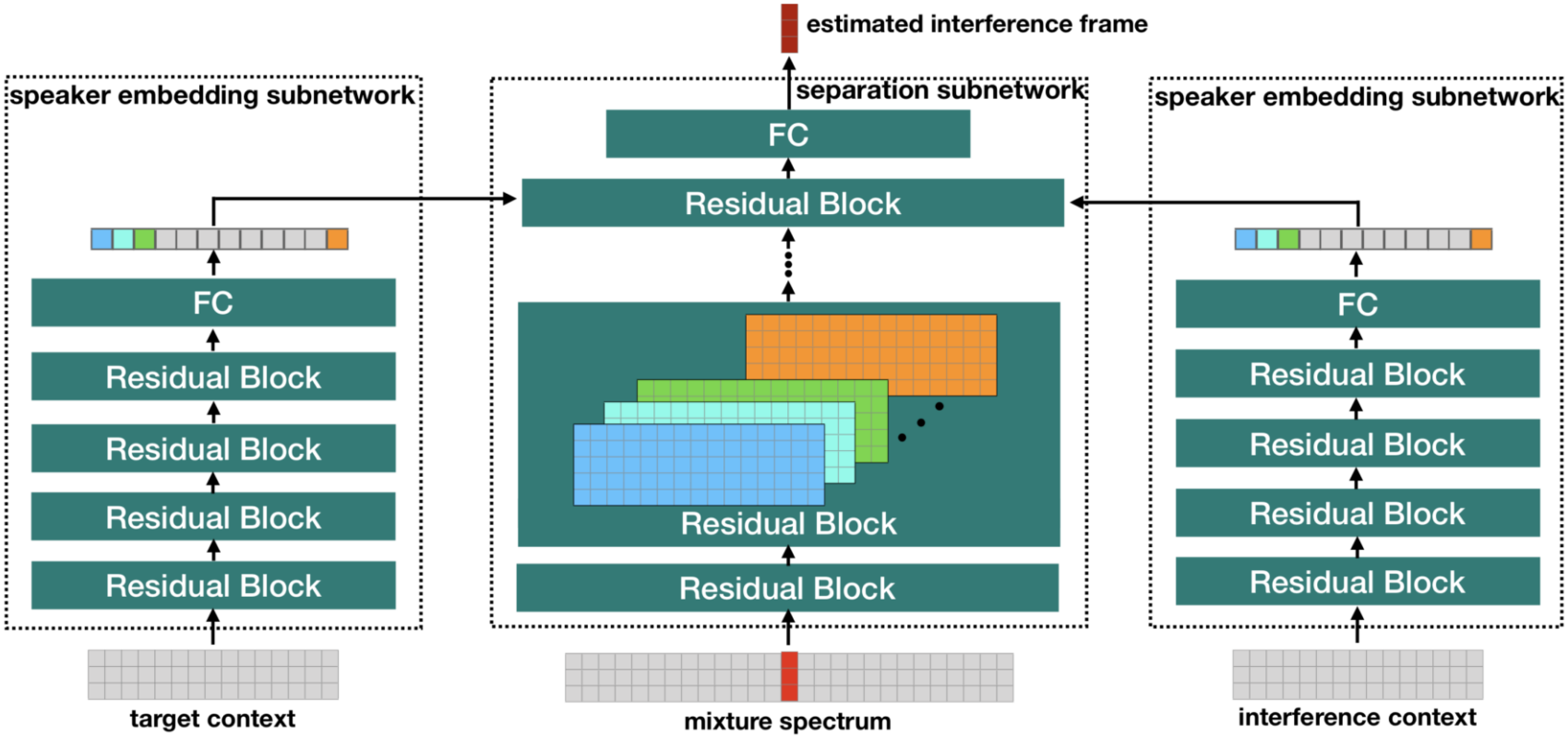
10 dB

5 dB

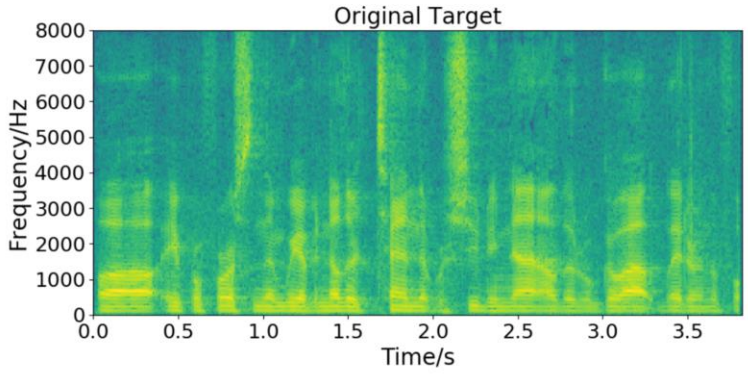
0 dB

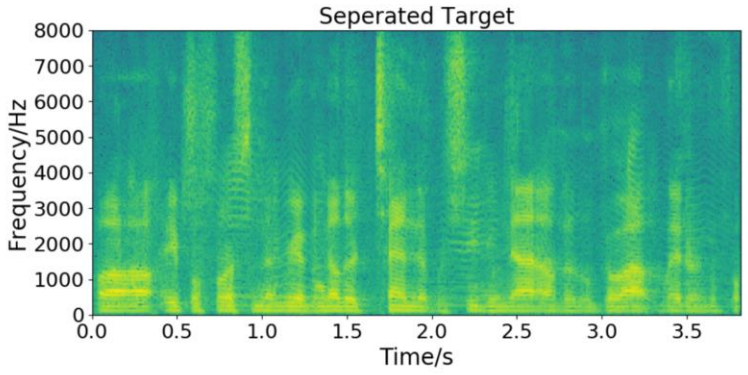
1: Preprocessing

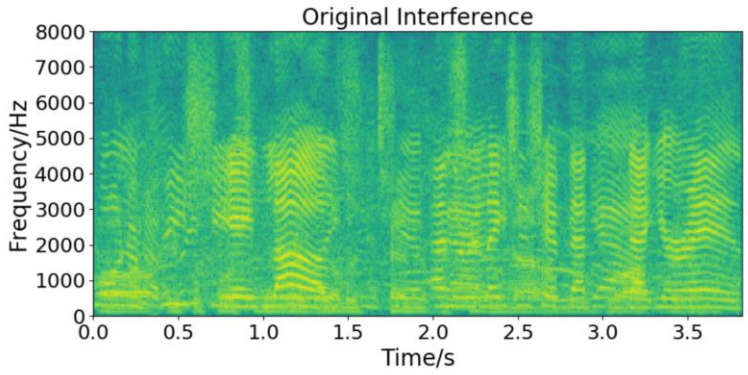
	SDR	SAR	SIR
DC	0.84	**2.09**	6.58
DaNet	1.81	3.29	10.41
Proposed	4.79	8.44	7.11

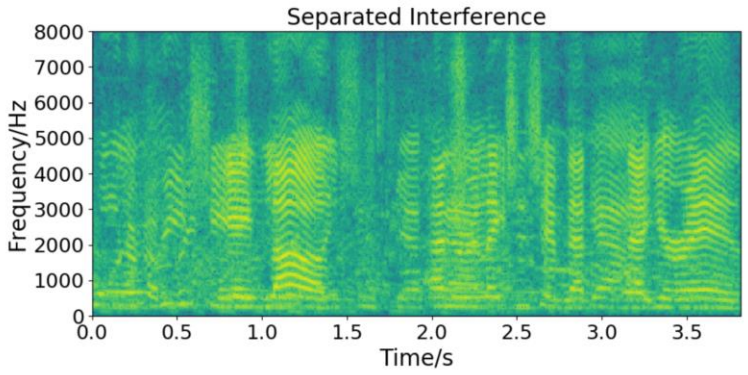


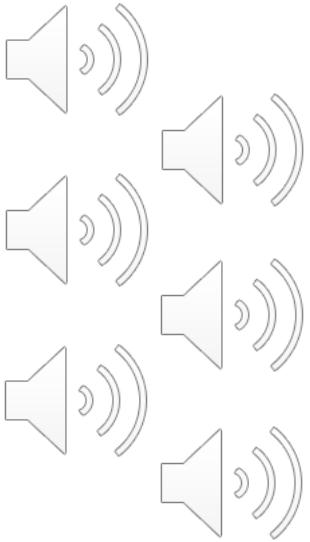
1: Preprocessing.



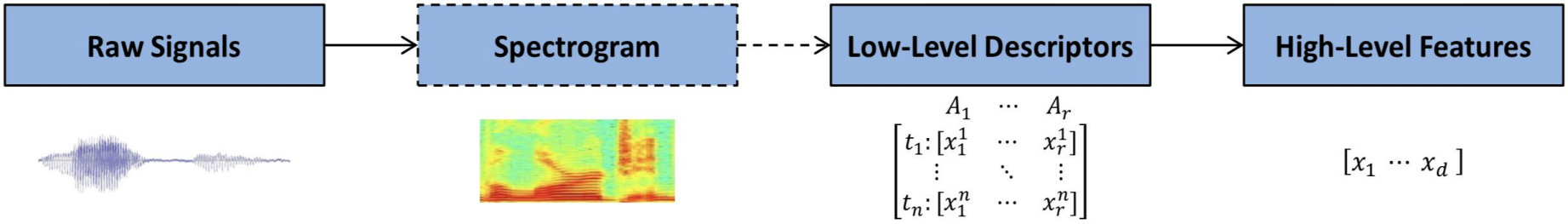








2: Feature Extraction.



Format: pcm, 8/16/44.1 kHz	**Format**: (log) power, (log) Mel	**Format**: (frame-level acoustic features) MFCC, eGeMAPS, ComParE13, etc.	**Format**: (segment-level statistic features) functional, BoAW, etc.
Pros: complete information	**Pros**: largely reduced temporal dimension; information is largely remained	**Pros**: easy explainable	**Pros**: easy explainable; greatly reduced feature size
Cons: instantaneous temporal change; high computational complexity	**Cons**: somewhat high computational complexity	**Cons**: high requirement of experts and experience	**Cons**: : high requirement of experts; severe information loss
Typical **models**: end-to-end, etc.	Typical **models**: CNN, gridGRU, etc.	Typical **models**: RNN, HMM, etc.	Typical **models**: SVM, etc.

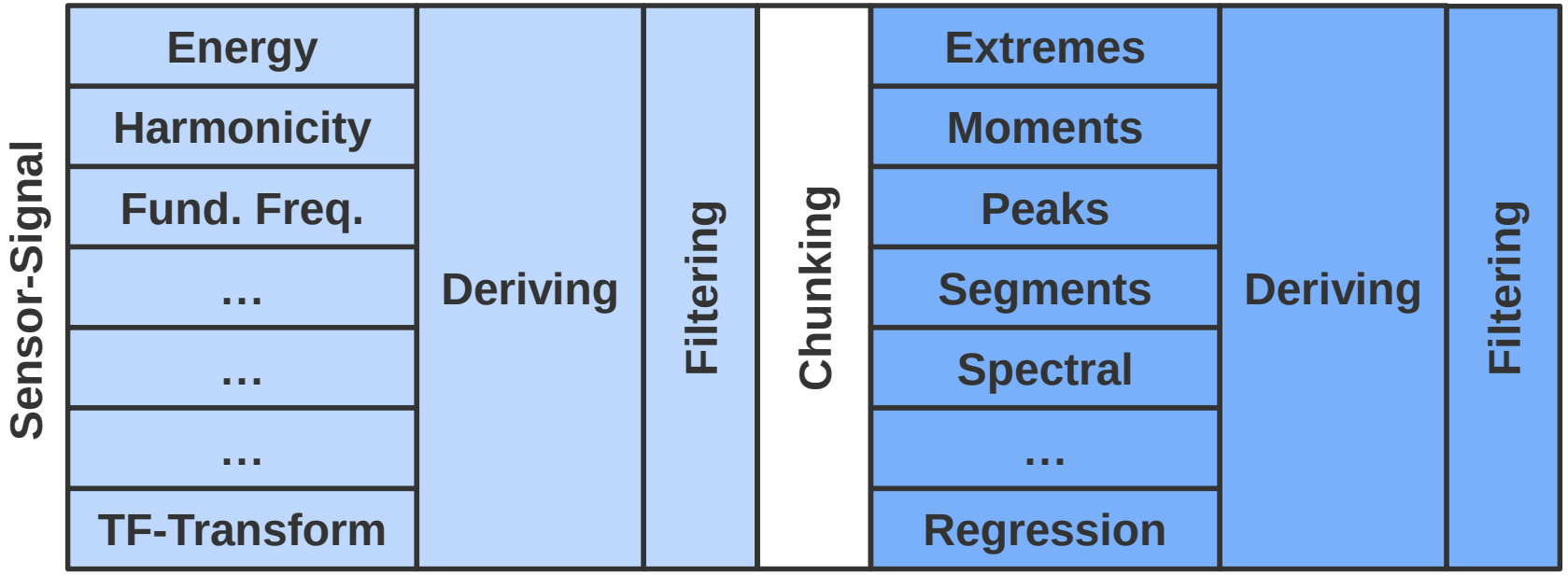
"Towards Automated Musical Behaviour Analysis for Infant Development: A Comprehensive Study", submitted.

2: Feature Extraction.

- **Brute-force**

  High-Dim. Space → Basis for selection

  Online update



2: Feature Extraction.

openSMILE:)

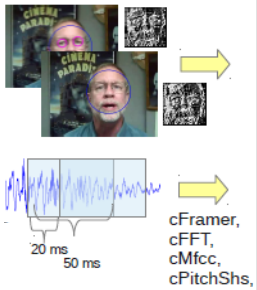
- **On-device Feature Extraction**

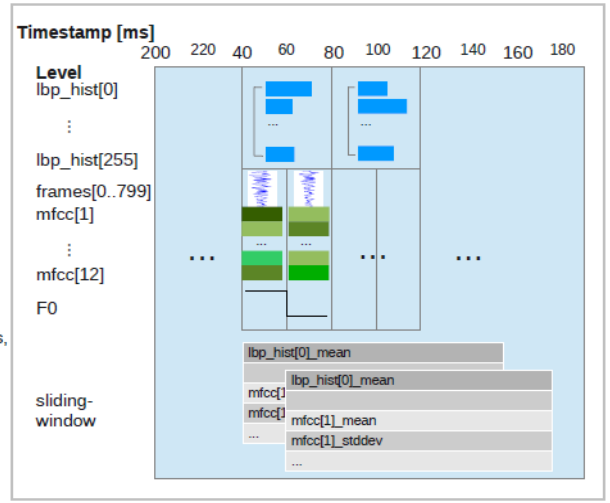
  Fast computation

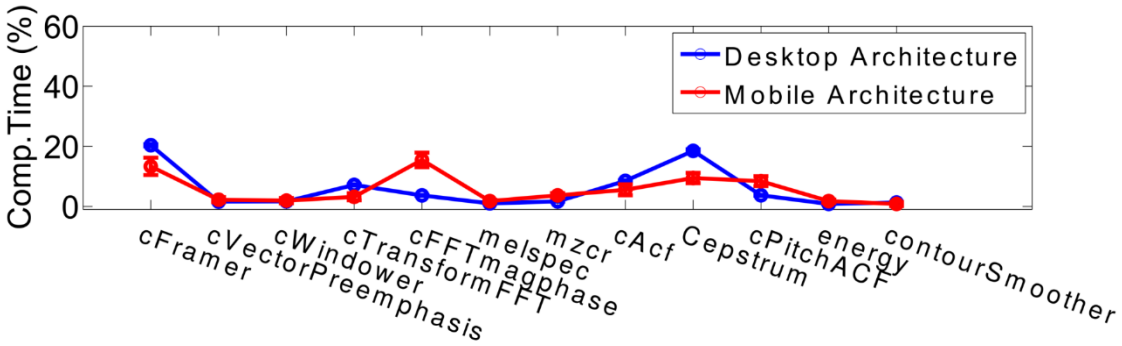
  Cross-signal

  Energy/speed-aware selection

RTF (#feat)	Intel i7	HTC OneM9	Galaxy S3
.4k	.01	.06	.43
6.4k	.04	.23	.63







2: Feature Extraction.

openSMILE:) by audEERING™

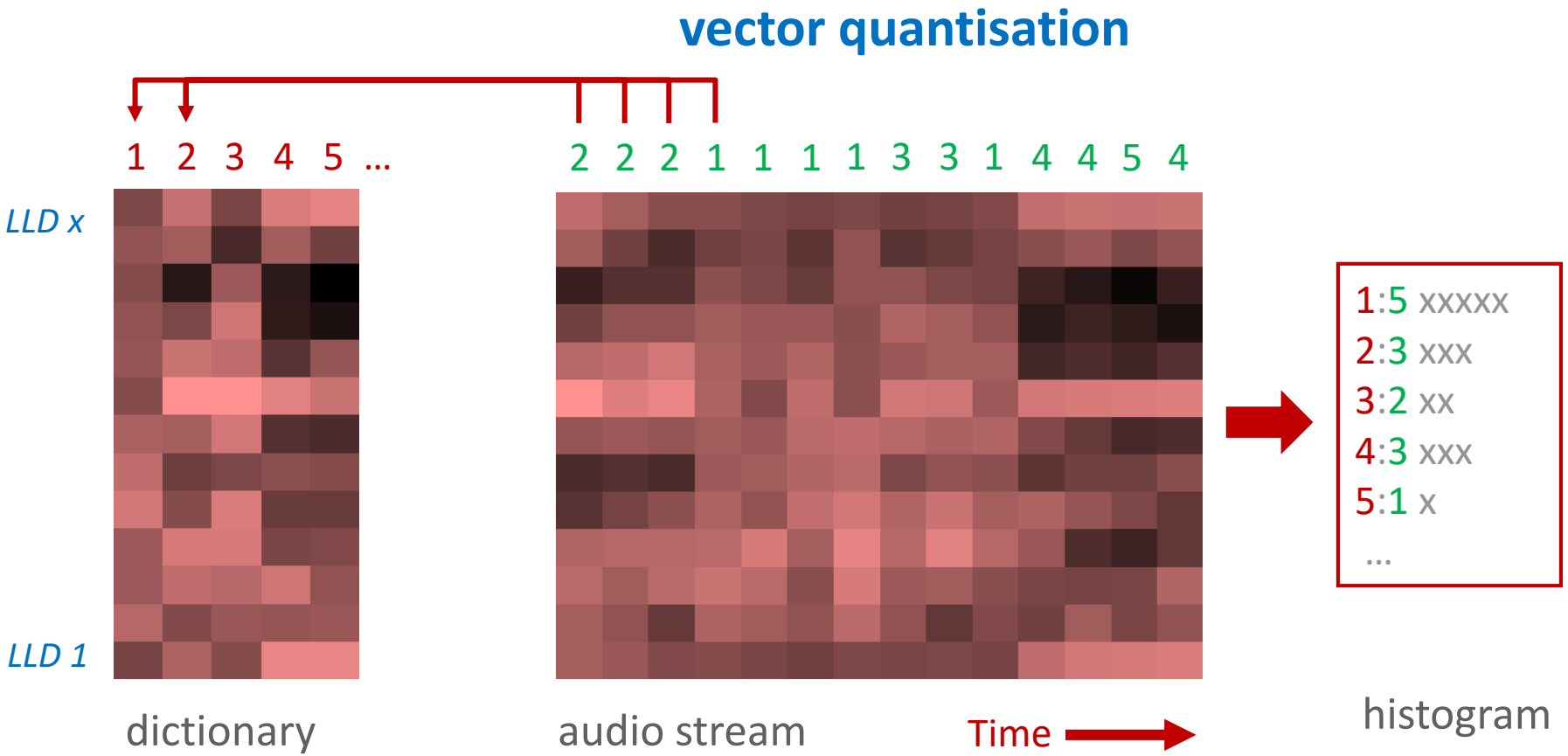
- **Traditional Features… Minimalistic: GeMAPS**

4 energy related LLD	Group
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0-8 k Hz)	spectral
MFCC 1–14 cepstral	cepstral
Spectral energy 250–650 Hz, 1 k–4 k Hz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice quality
log. HNR, Jitter (local & DDP), Shimmer (local)	voice quality

1 energy related LLD	Group
Sum of auditory spectrum (loudness)	Prosodic
25 spectral LLD	**Group**
α ratio (50–1 000 Hz / 1-5 k Hz)	Spectral
Energy slope (0–500 Hz, 0.5–1.5 k Hz)	Spectral
Hammarberg index	Spectral
MFCC 1–4	Cepstral
Spectral Flux	Spectral
6 voicing related LLD	**Group**
F0 (Linear & semi-tone)	Prosodic
Formants 1, 2, (freq., bandwidth, ampl.)	Voice Quality
Harmonic difference H1–H2, H1–A3	Voice Quality
log. HNR, Jitter (local), Shimmer (local)	Voice Quality

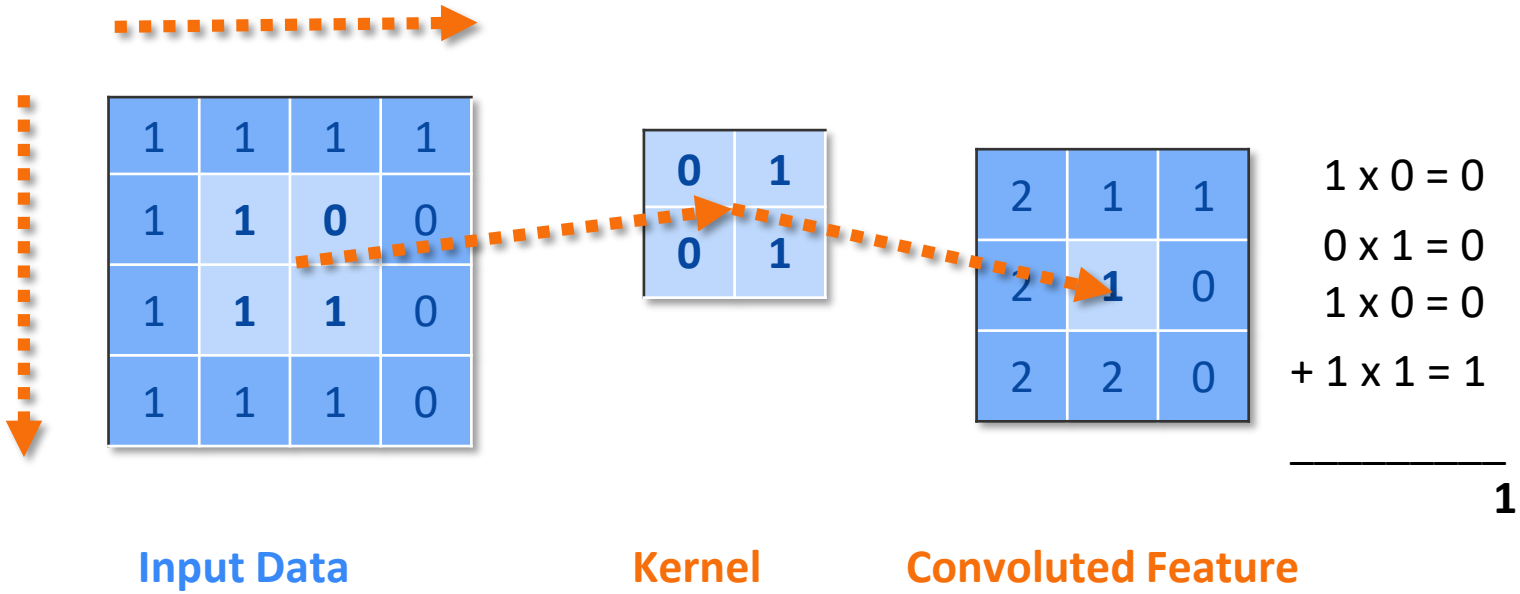
2: Feature Extraction.

openXBOW –|)→



2: Feature Extraction: Enter CNNs.

- **Convolution: Multiplication of Kernel with Pixel Values**

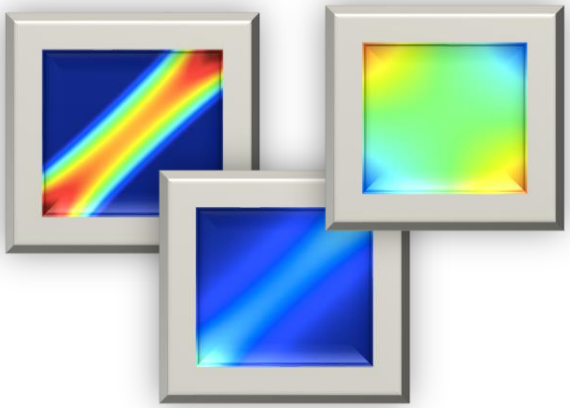


Kernel K. Window X_A of Input X of same size

→ Strength of feature: Sum of element-wise (!) Product

(not matrix multiplication!): $\sum_i X_A(i) \cdot K(i)$

2: Feature Extraction: Enter CNNs.



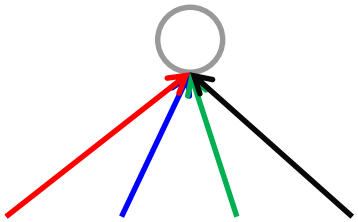
Convolution of a single window of the input:
Neural Net with a single output neuron:

Kernel → weights of the neural connection

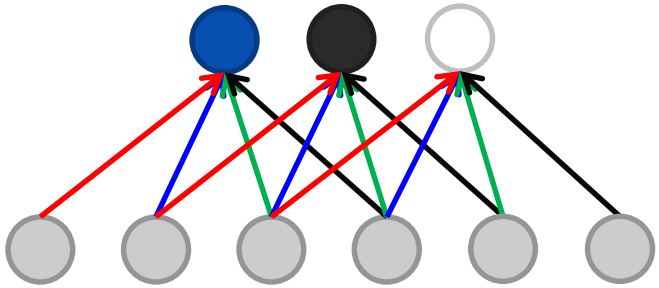
Learning weights of the neuron: learn the Kernel

Kernel is shifted over input:

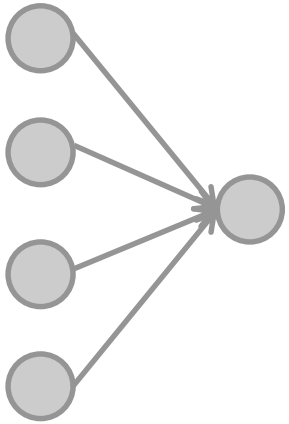
Weights remain. but the output changes



Kernel as neuronal weights



Triple application of Kernel. shifted by 1



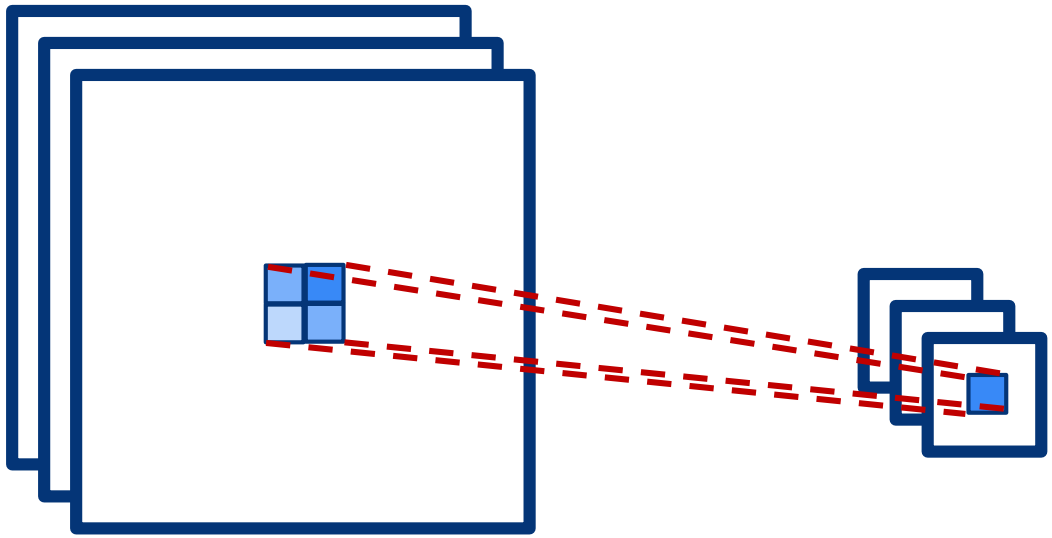
2: Feature Extraction: Enter CNNs.

- **Pooling**

  Reduces size of feature maps. e.g.. Max Pooling:

  1 neuron in max pooling layer forwards max activation of several previous → forward only most important information

  → reduce number neurons



2: Feature Extraction: Enter CNNs.

- **Normalisation Layers**

  → ensure normalisation of input also for higher layers

- **Batch Normalisation**

  input of each neuron normalised over “batch” (such as 50 instances)

  allows for higher learning rates. reduces overfitting

  only in forward networks

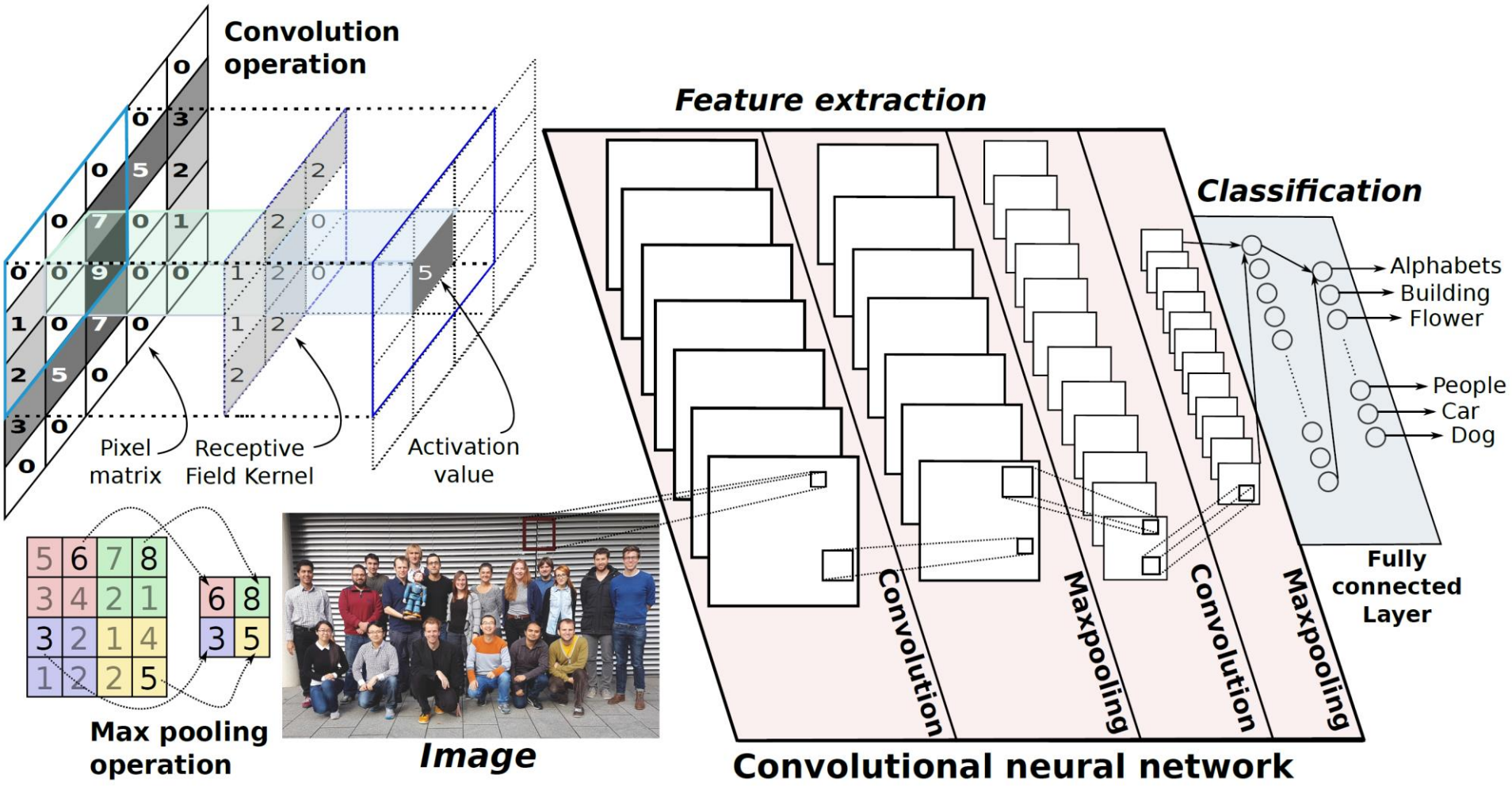
m: batch size. a_i: activation of neuron in step i of the batch ($1 \leq i \leq m$)

batch mean: $\mu_B = \frac{1}{m}\sum_{i=1}^{m} a_i$

batch variance: $\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(a_i - \mu_B)^2$

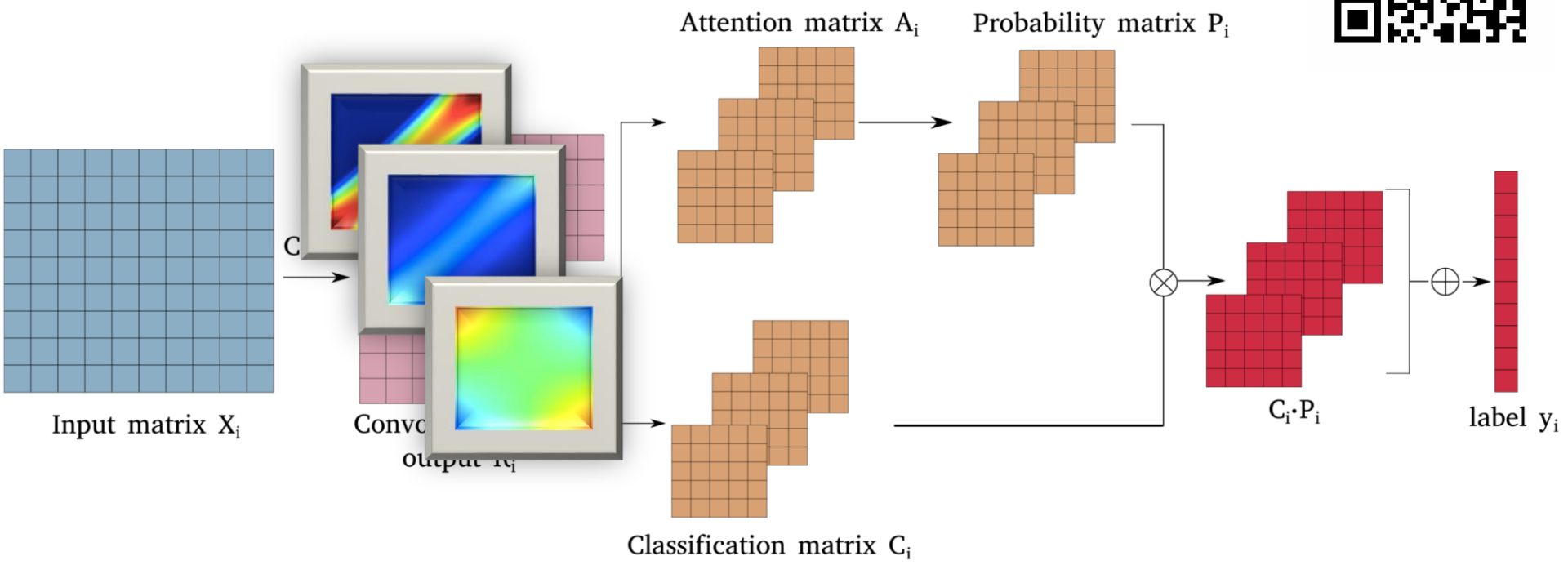
normalised activation: $\hat{a}_i = \frac{a_i - \mu_B}{\sigma_B}$

2: Feature Extraction.

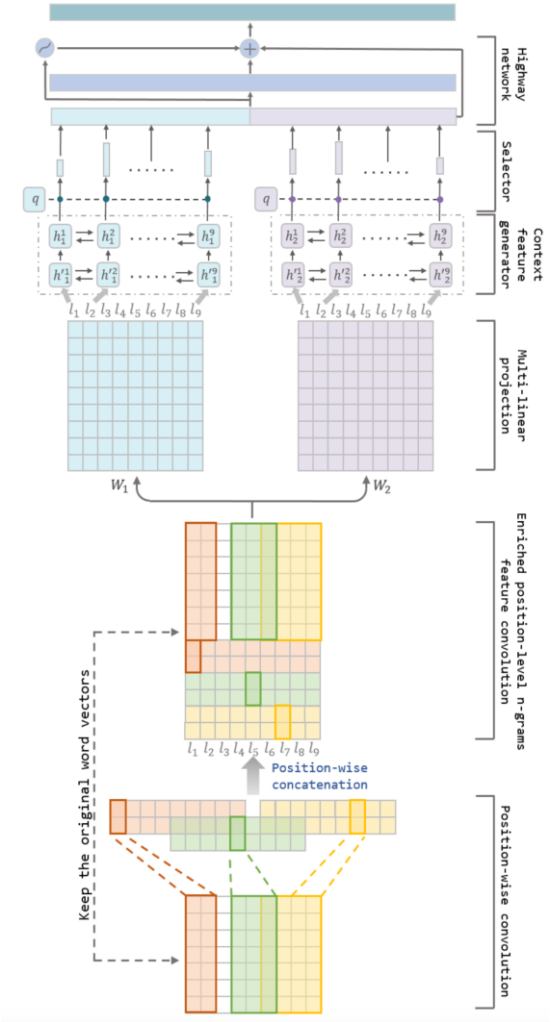


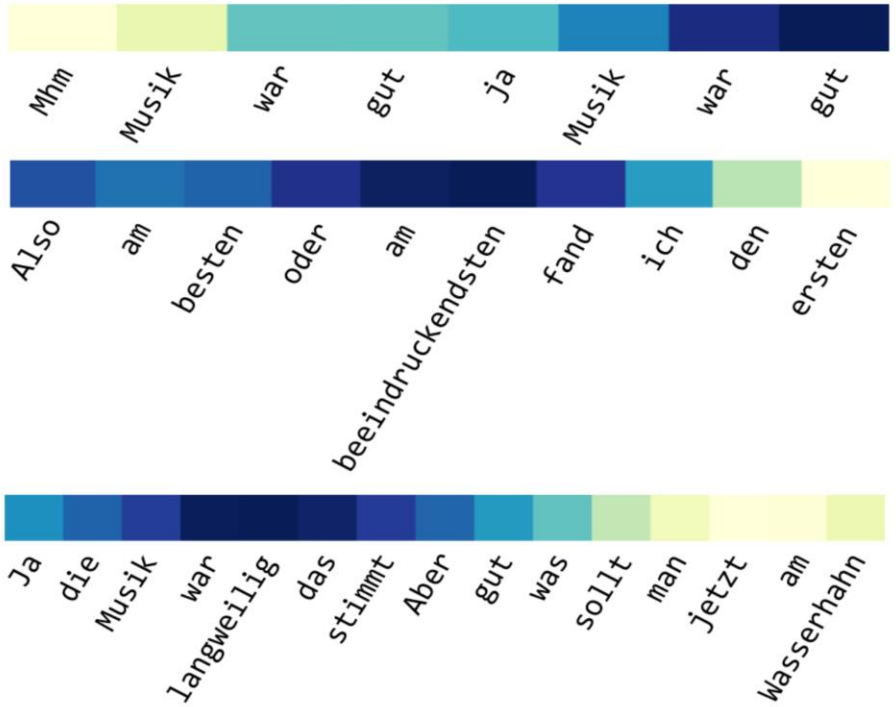
2: Feature Extraction.

- **Self-Learning Representation: CNNs & Attention**

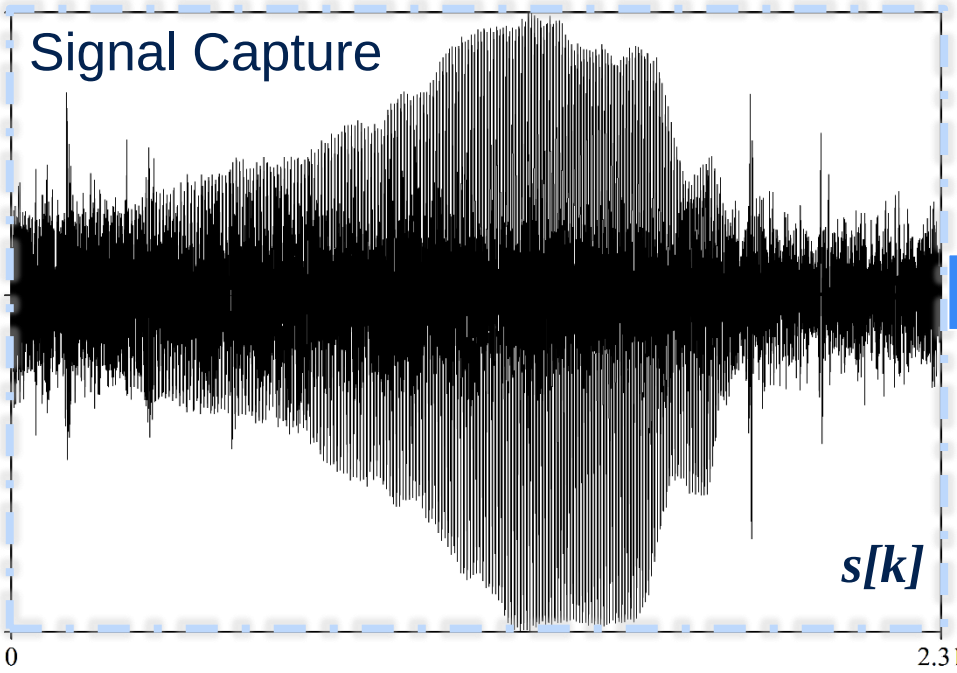


2: Feature Extraction.





End-to-End.



"Big Data vs Little Labels"

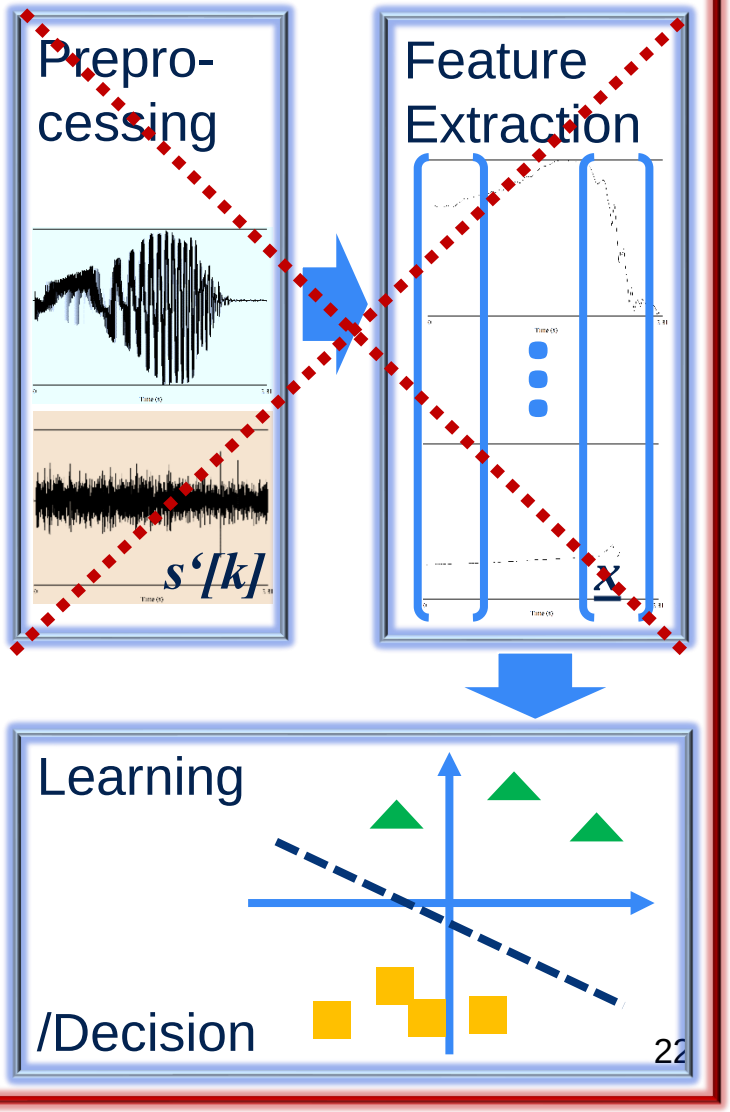
Efficient Weakly Supervised

<label>

(e.g.. "burst")

y

Learning at Signal's Edge



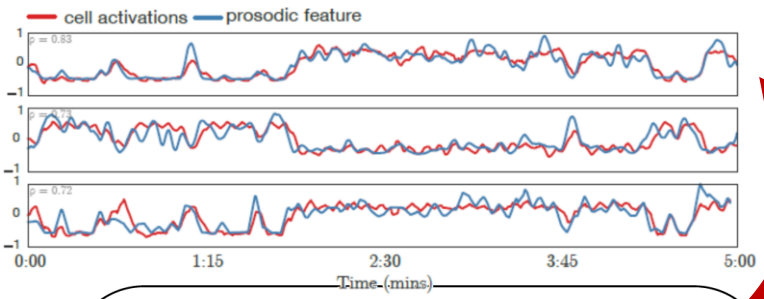
End-to-End.

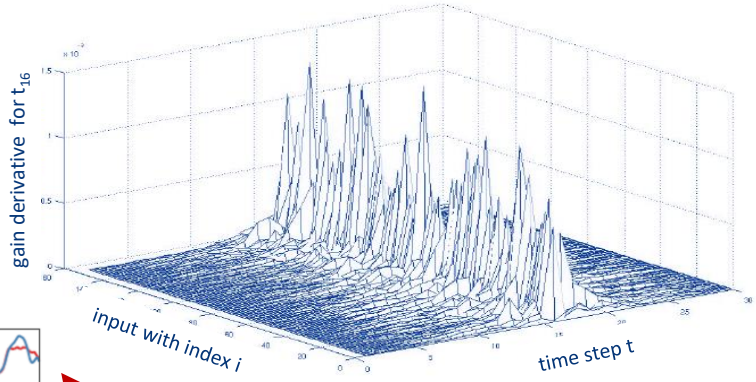
- **CNN + LSTM RNN**

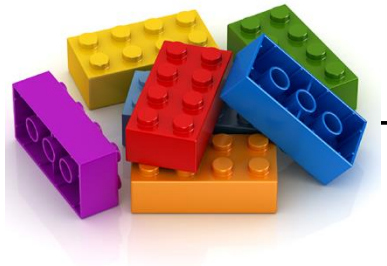
energy range (.77)

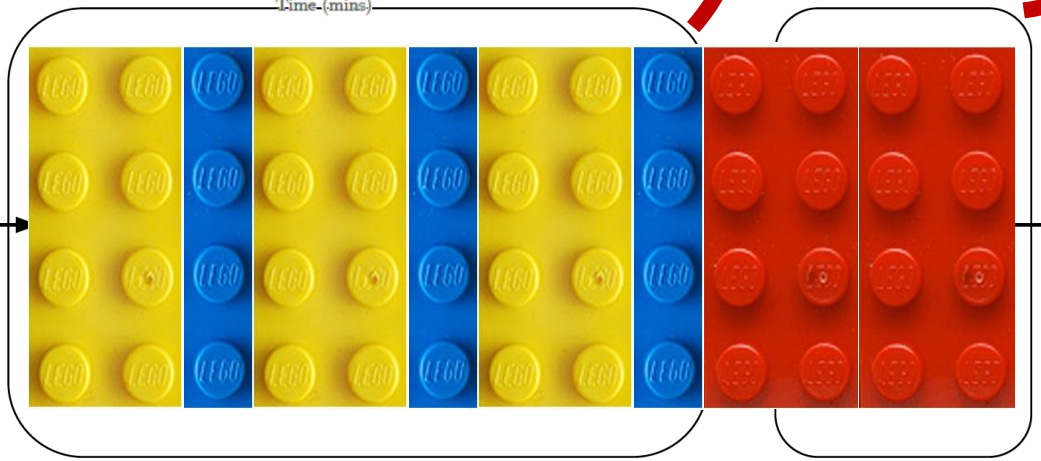
loudness (.73)

F0 mean (.71)









Arousal

Valence

Input raw waveform at 16kHz

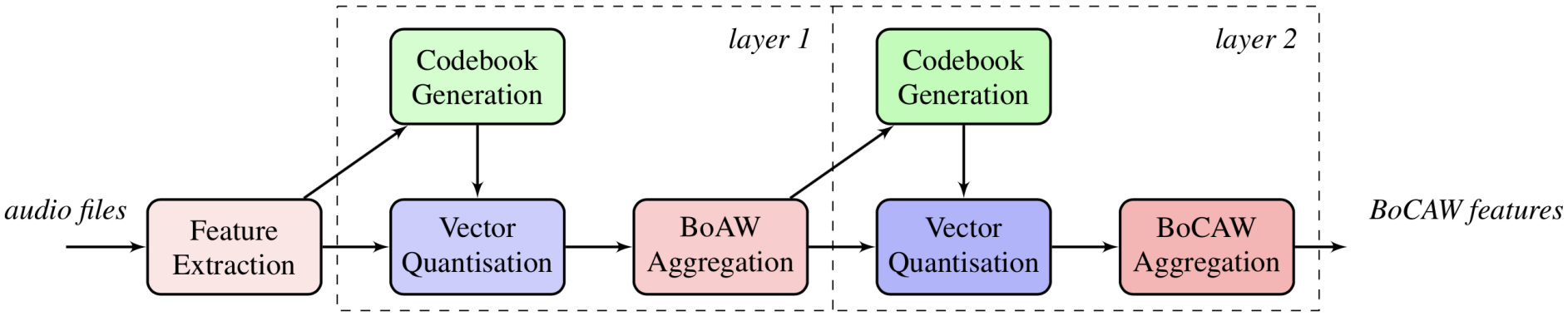
Convolution layers

Recurrent layers

Output label at 25Hz

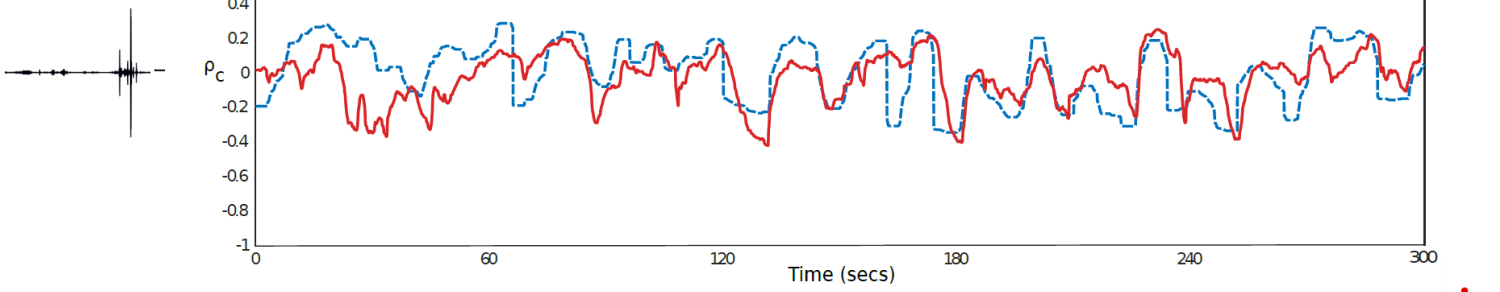
Performance?

CCC AVEC 16-18 RECOLA	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261
BoCAW	.750	.465
e2e (2018)	.787	.440

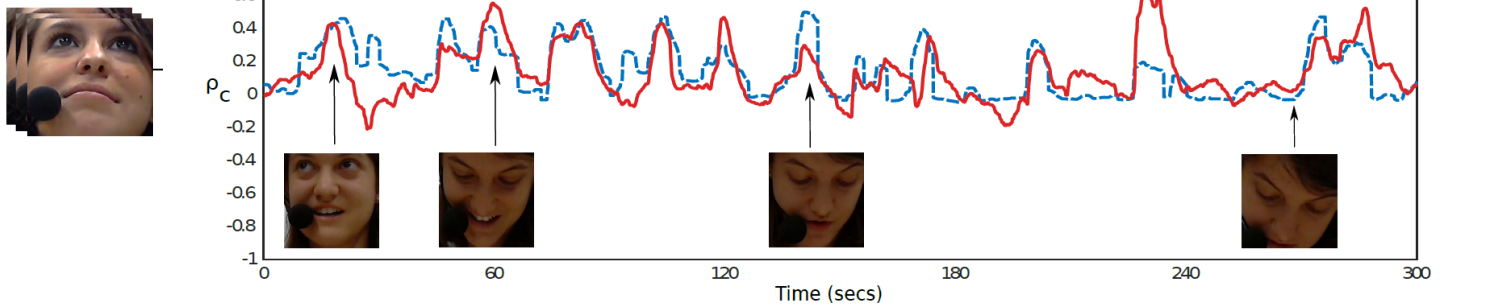


Performance?

CCC Recola	Arousal	Valence
A/V end-to-end	.770	.612

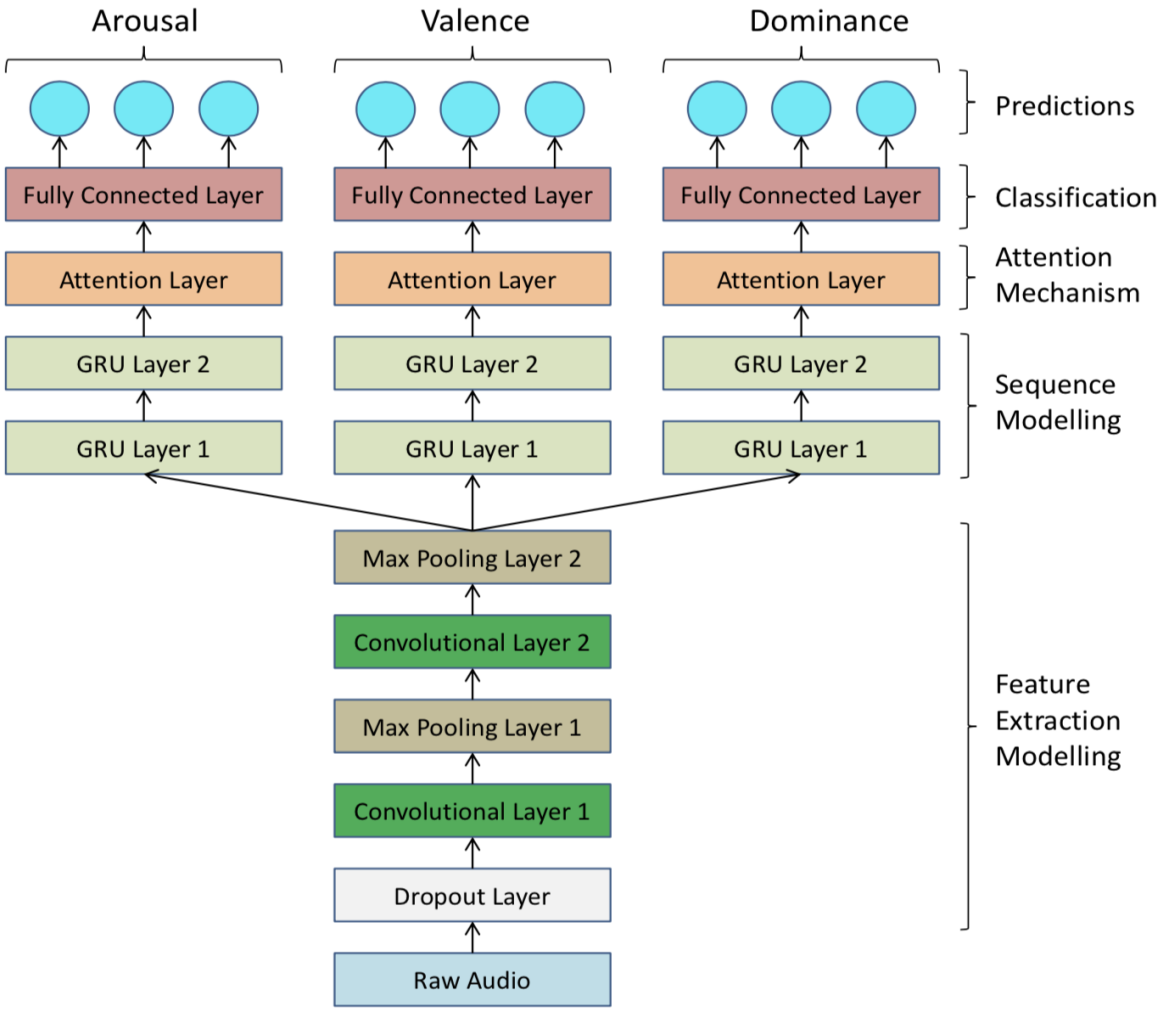


(a) Arousal

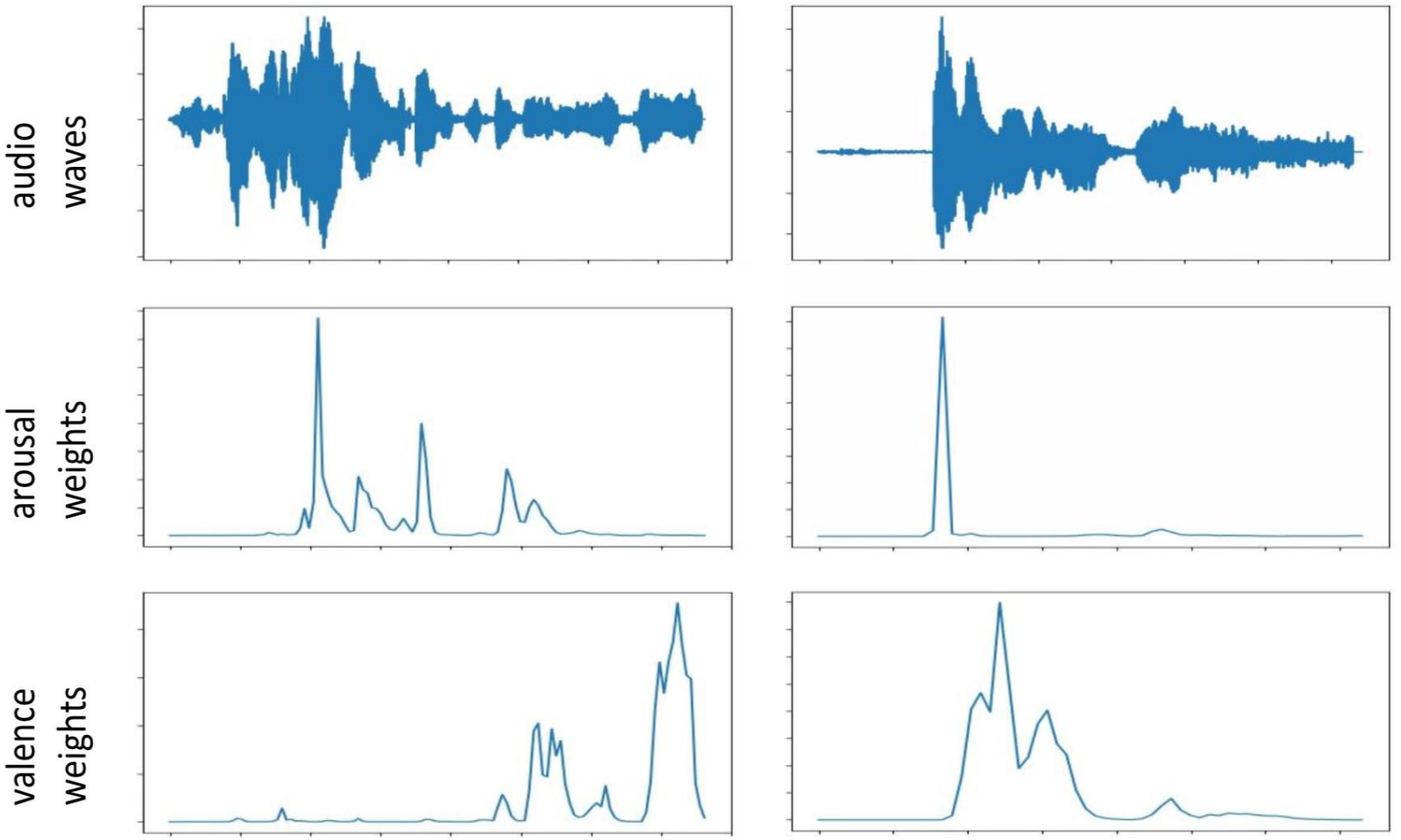


(b) Valence

Decision Making w/ Attention.

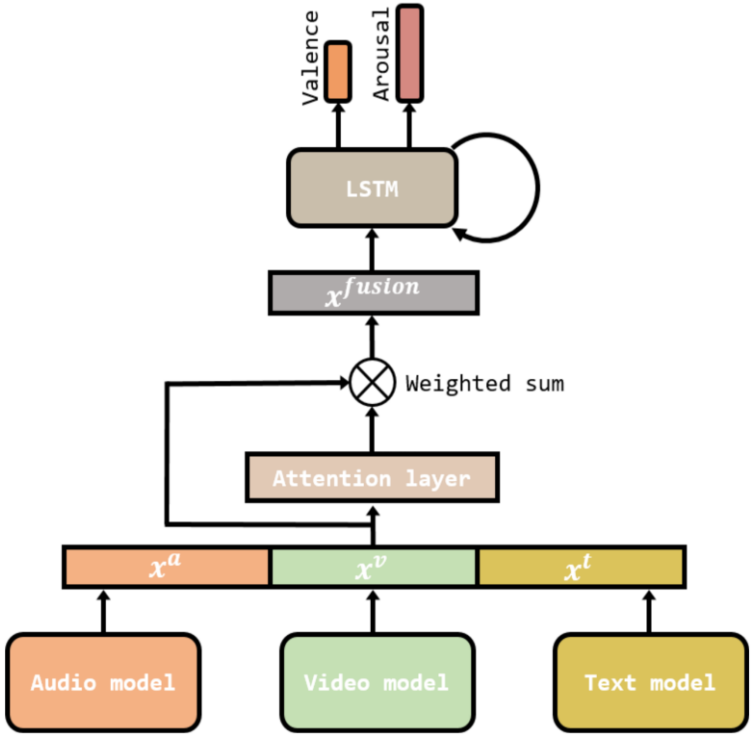


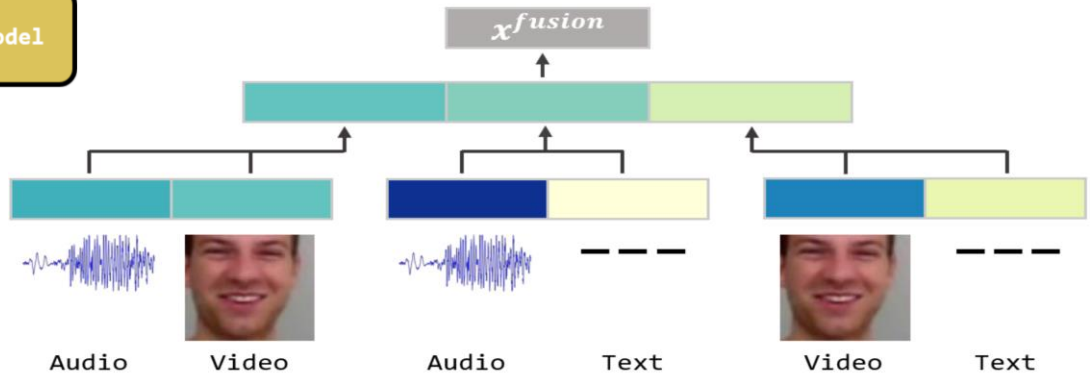
Decision + Attention.



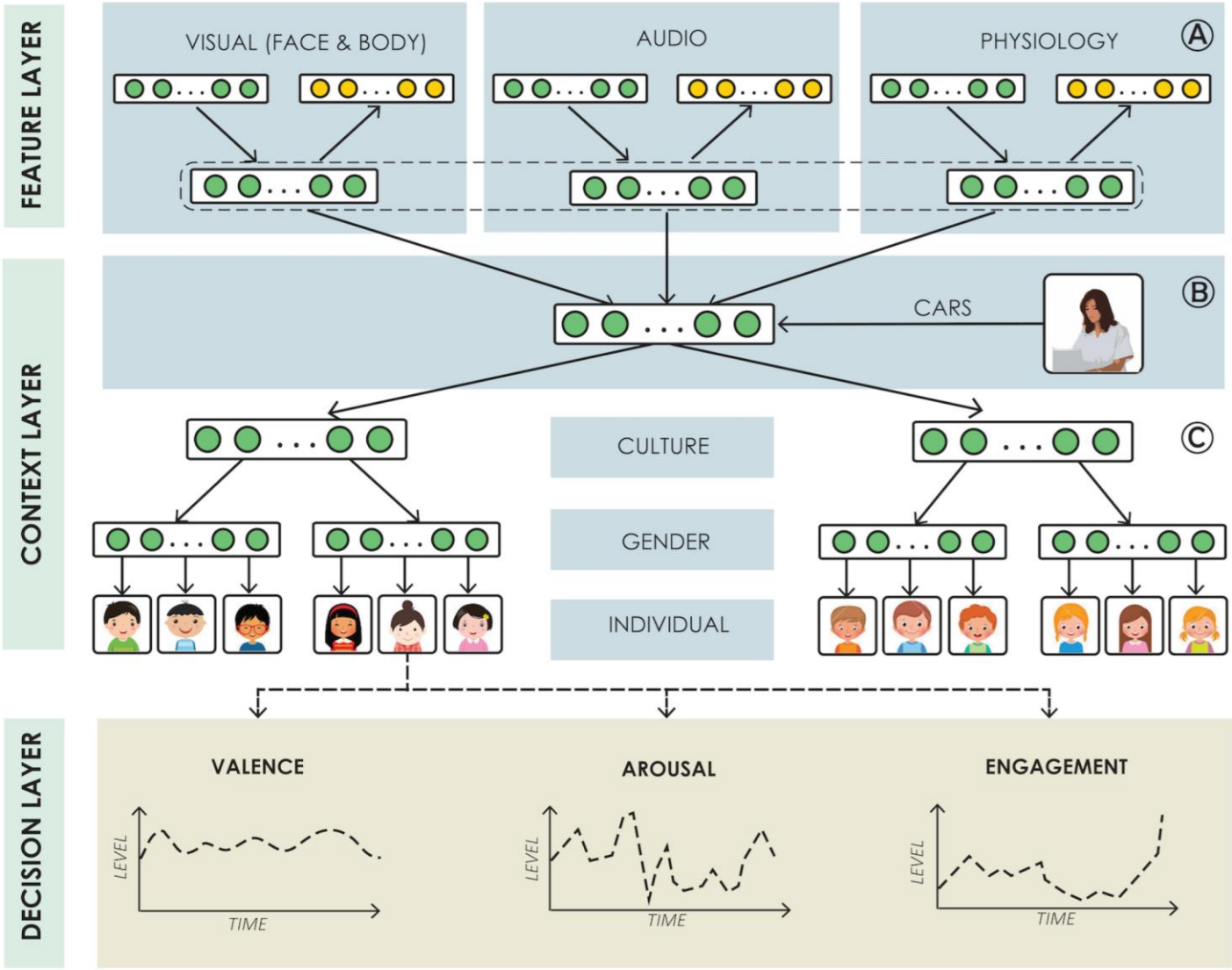
Decision + Attention.

CCC AVEC 2017 SEWA	Arousal	Valence
Audio	.456	.438
Video	.603	.673
Text	.508	.554
Fusion	.664	.735



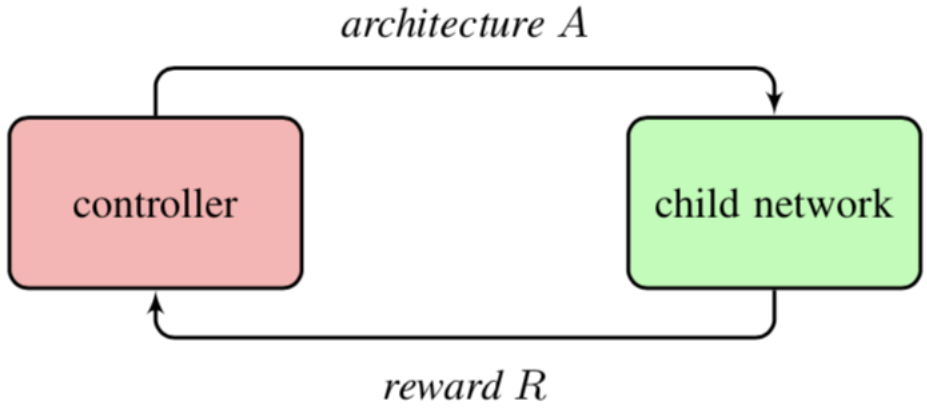


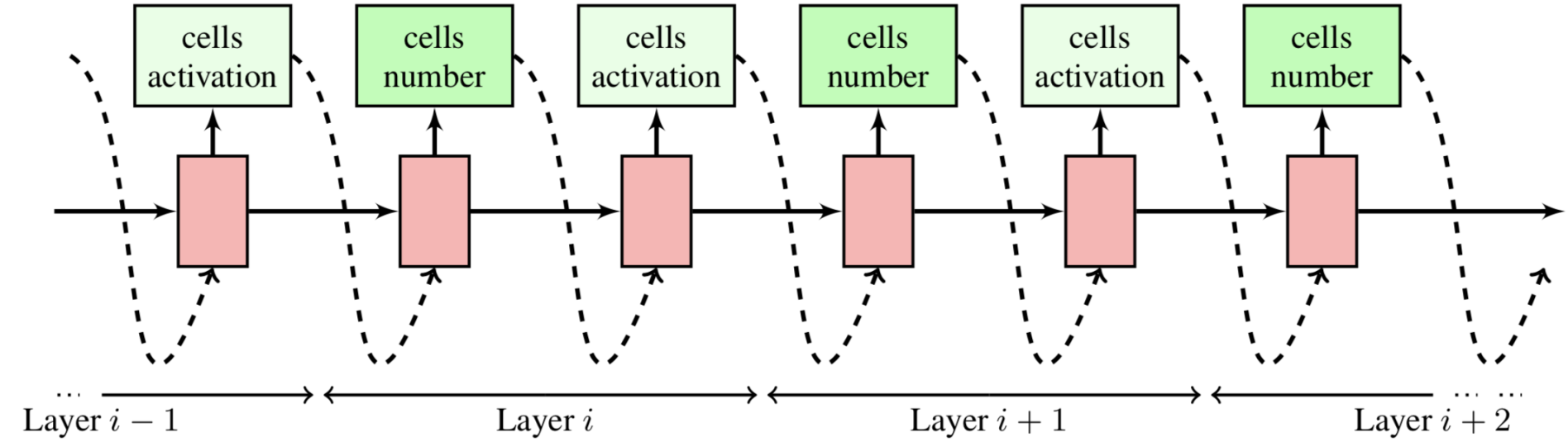
Shape?


FEATURE LAYER
VISUAL (FACE & BODY)
AUDIO
PHYSIOLOGY
A
CONTEXT LAYER
CARS
B
CULTURE
GENDER
INDIVIDUAL
C
DECISION LAYER
VALENCE
AROUSAL
ENGAGEMENT
LEVEL
TIME
LEVEL
TIME
LEVEL
TIME

AutoML.

ComParE 18 Crying	%UA
Seq2Seq	62.1
End2End	63.5
BoAW	67.7
Evolved GRU-RNN	70.1

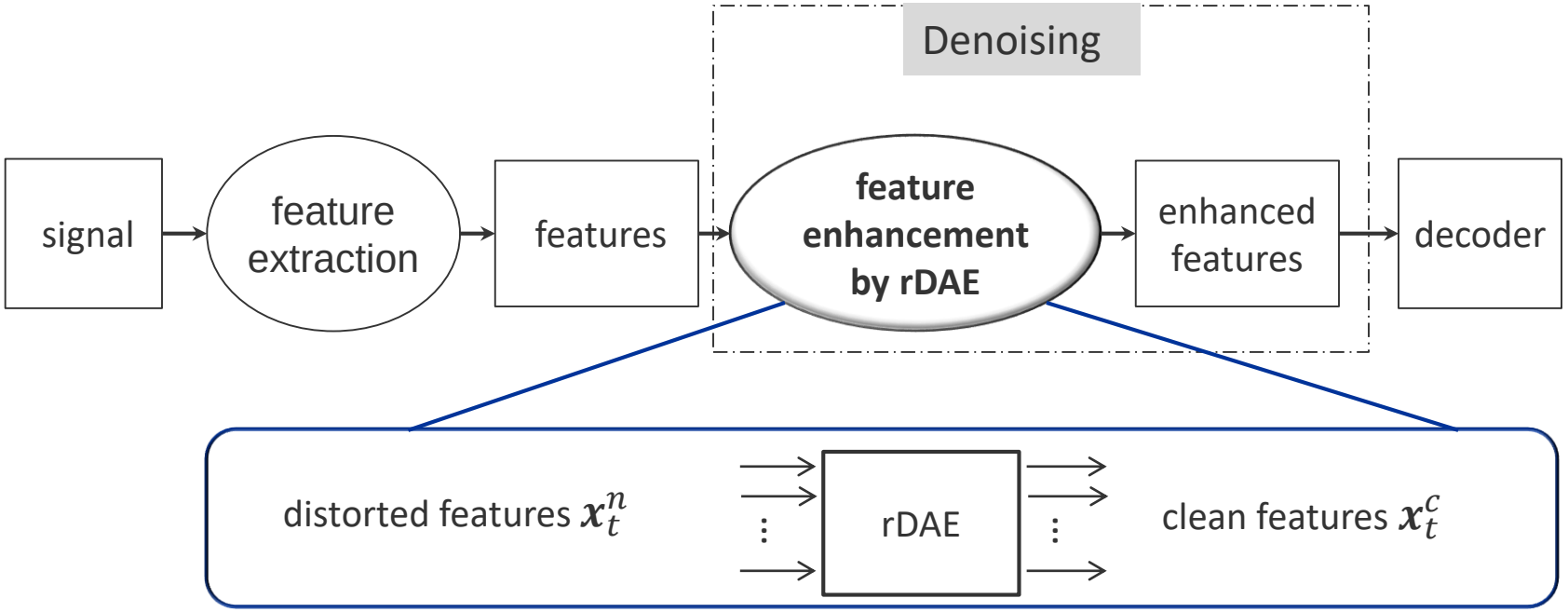




End-to-End: Some More?

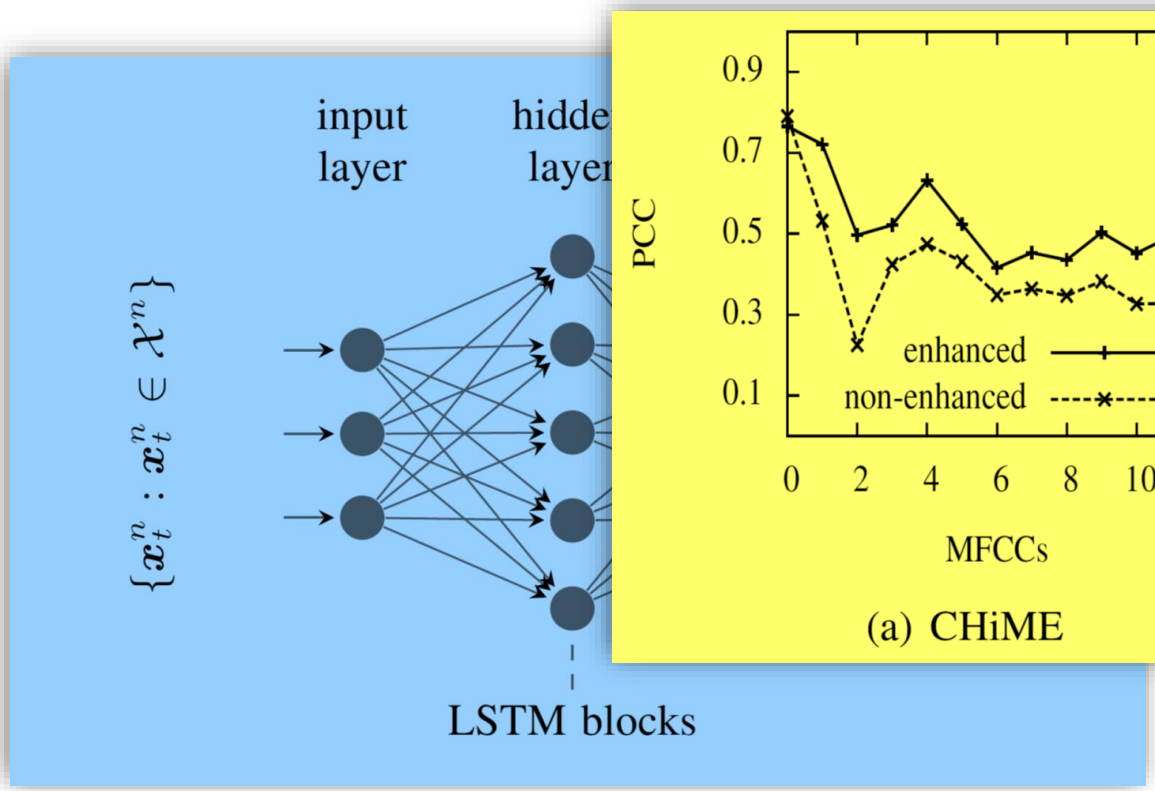
- **Feature Enhancement**

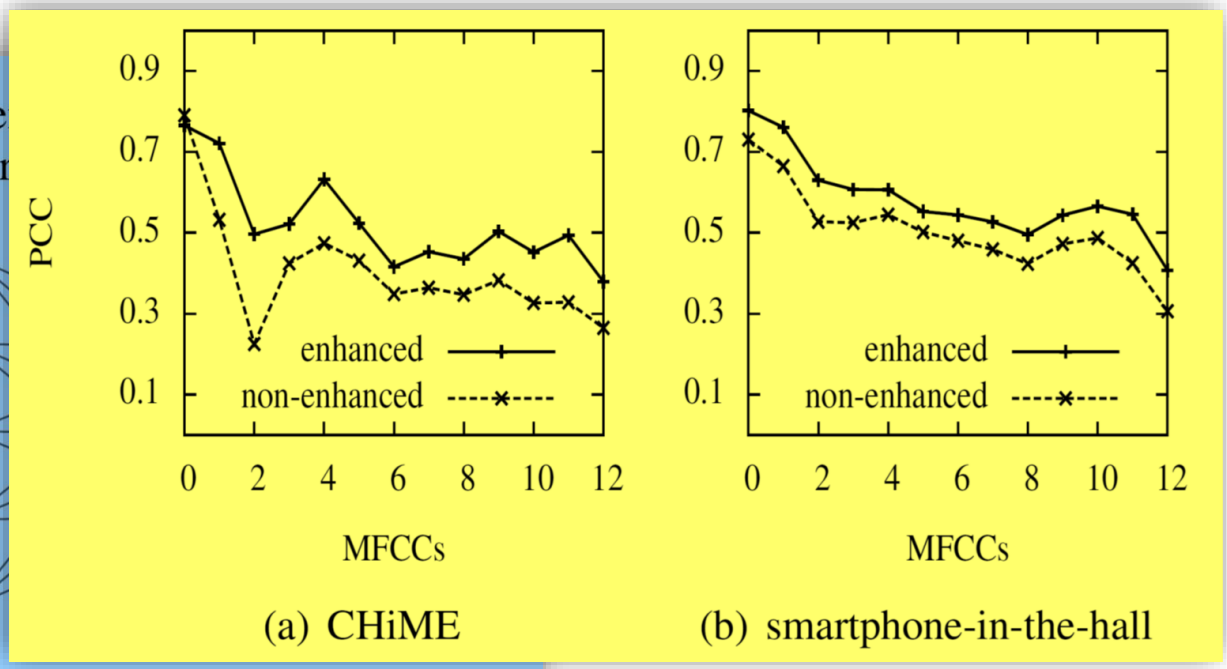
  Recurrent Denoising Autoencoder



End-to-End: Some More?

- **Autoencoding**

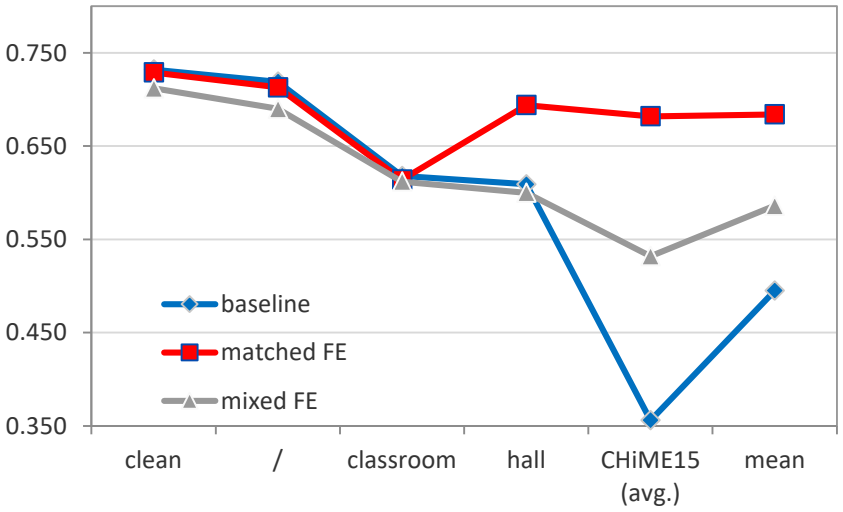




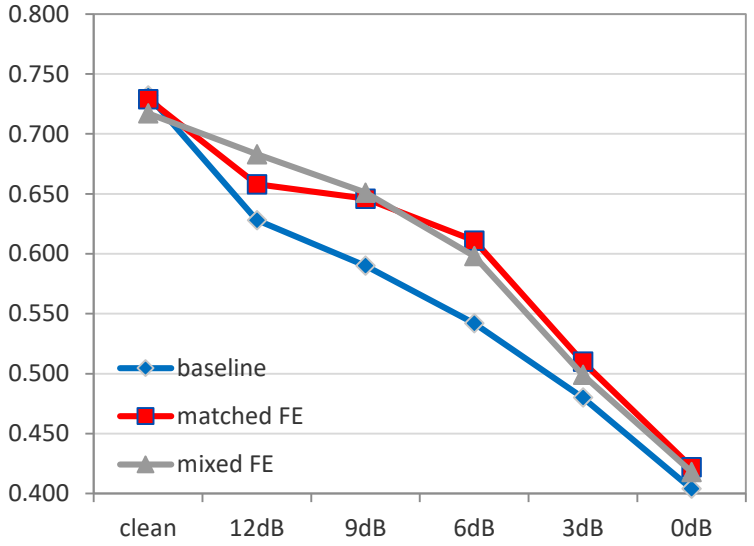
End-to-End: Some More?

- **Feature Enhancement: Arousal**

Smartphone



CHiME

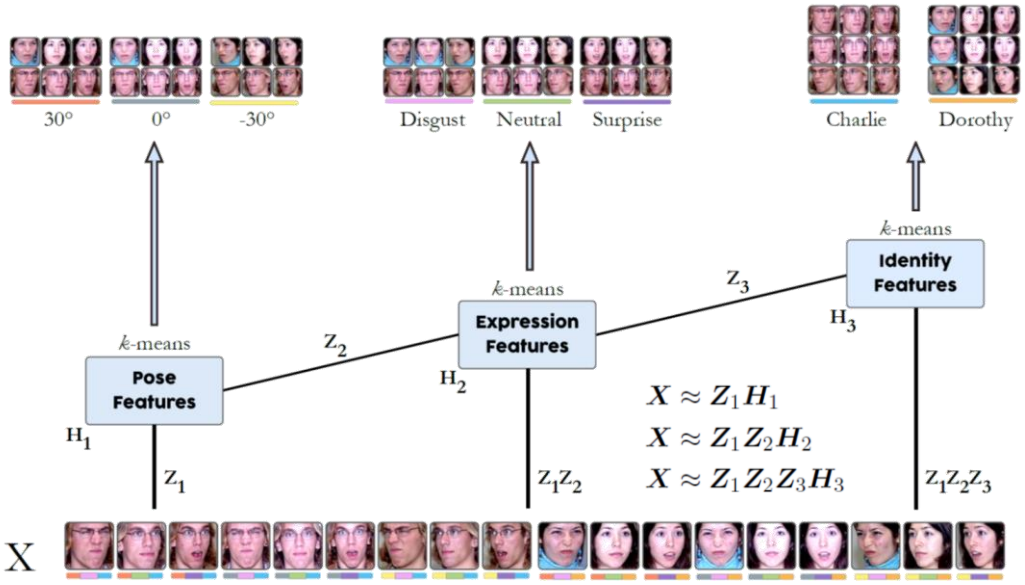


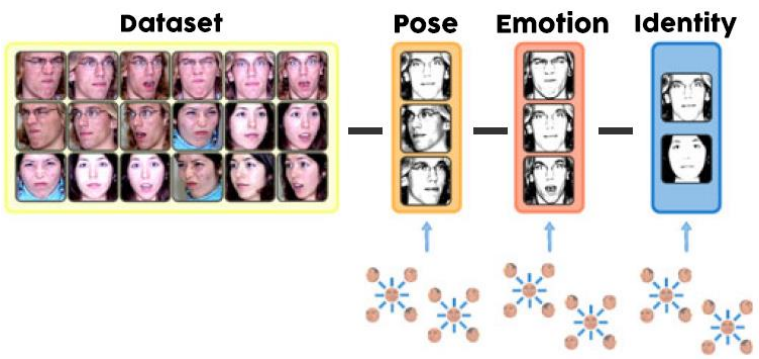
CCC	base	matched	mixed
average	.495	.684	.586

End-to-End: Some More?

- **Deep Clustering**
  - Learn latent attribute hierarchy
  - Better representation of attribute
    - w/ lowest variability

Emotion	%UA
Baseline	61
DSNMF	83





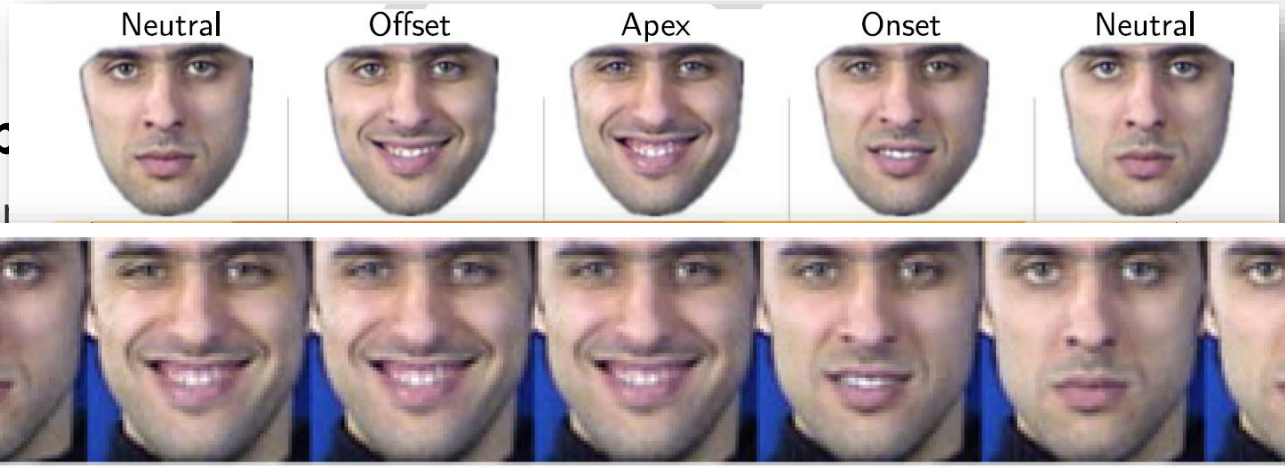
$$C_{\text{deep}} = \frac{1}{2}\|X - Z_1 Z_2 \cdots Z_m H_m\|_F^2$$

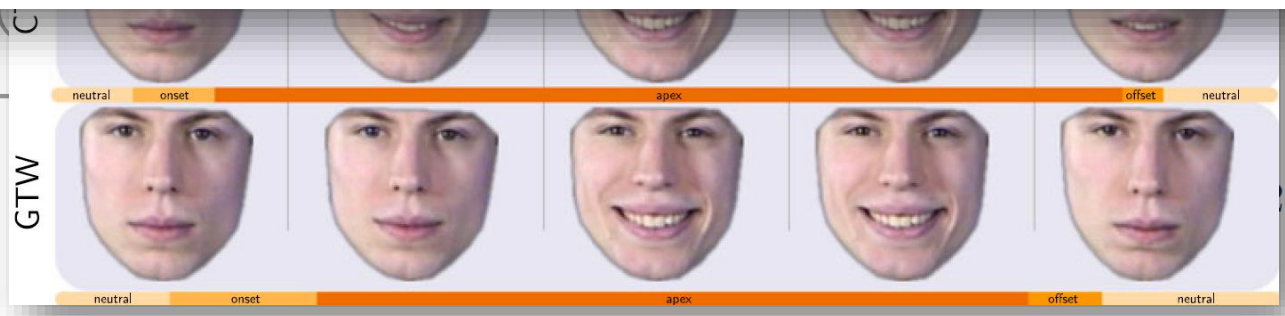
End-to-End: Some More?

"Deep Canonical Time Warping for simultaneous alignment and representation learning of sequences".

- **Deep**

Maxi

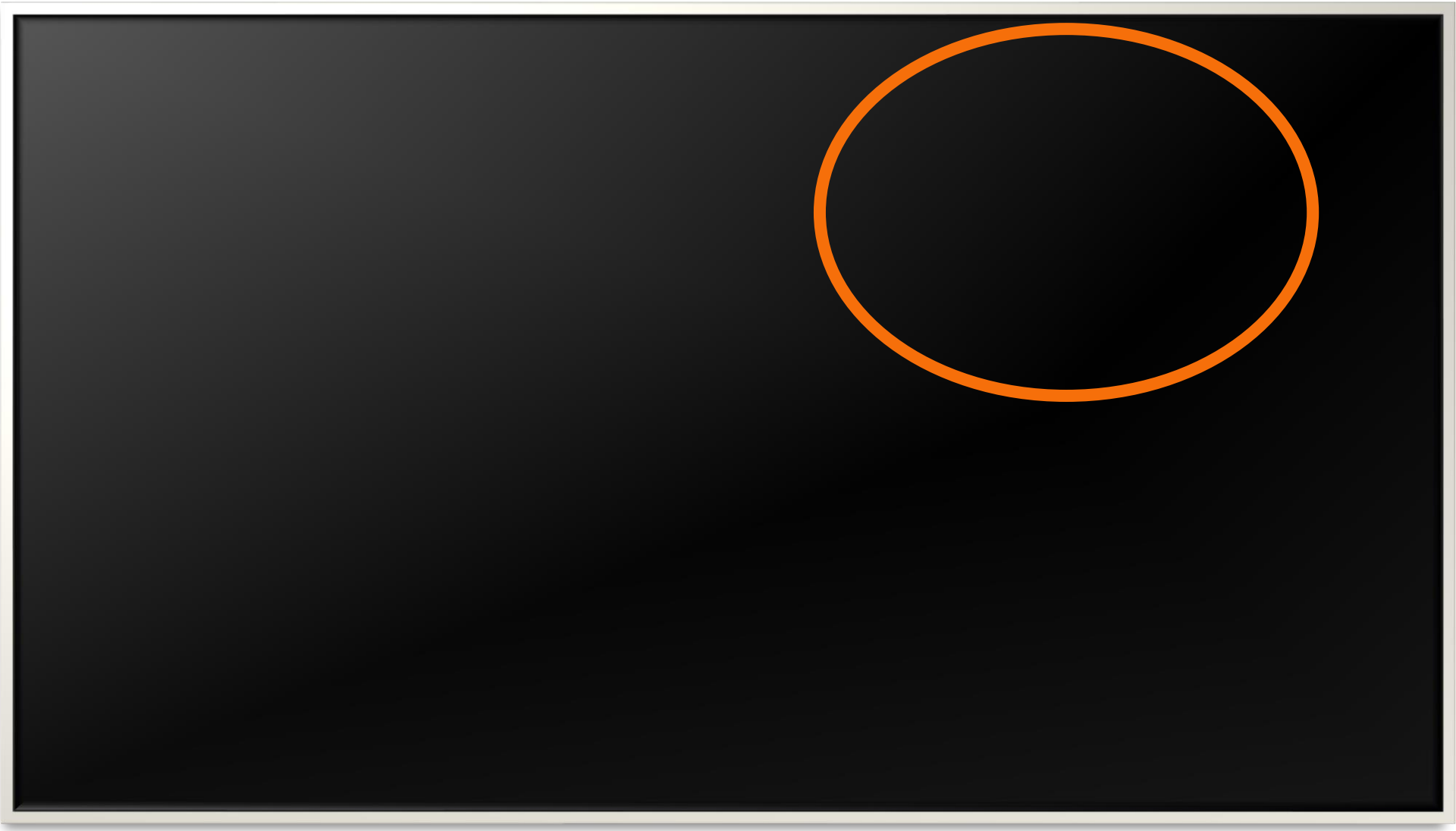




DCTW

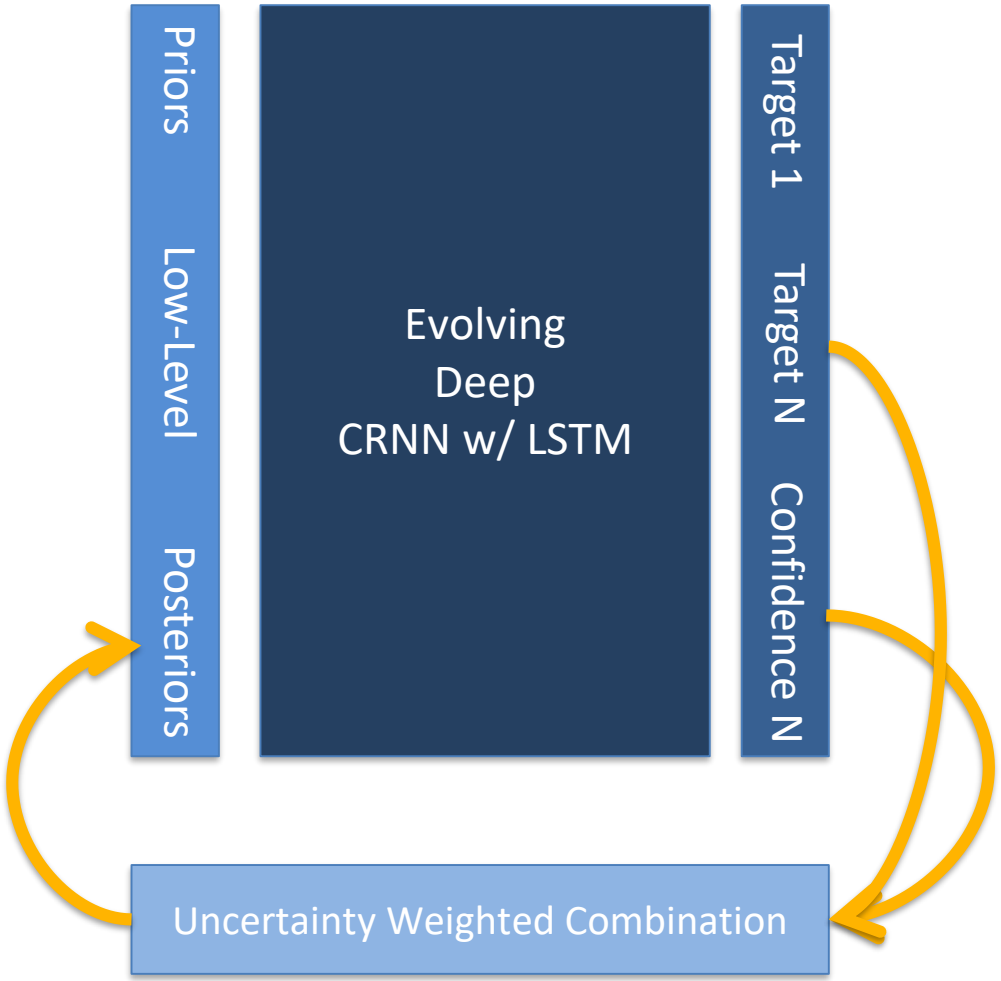
$\|\mathbf{X}_1\mathbf{W}_1 - \mathbf{X}_2\mathbf{W}_2\|_F^2 \quad \text{s.t.} \quad \mathbf{W}_1 = \mathbf{A}(\mathbf{X}_1; \theta_1)\mathbf{\Delta}_1, \mathbf{W}_2 = \mathbf{A}(\mathbf{X}_2; \theta_2)\mathbf{\Delta}_2$

End-to-End: Some More?



End-to-End: Some More?

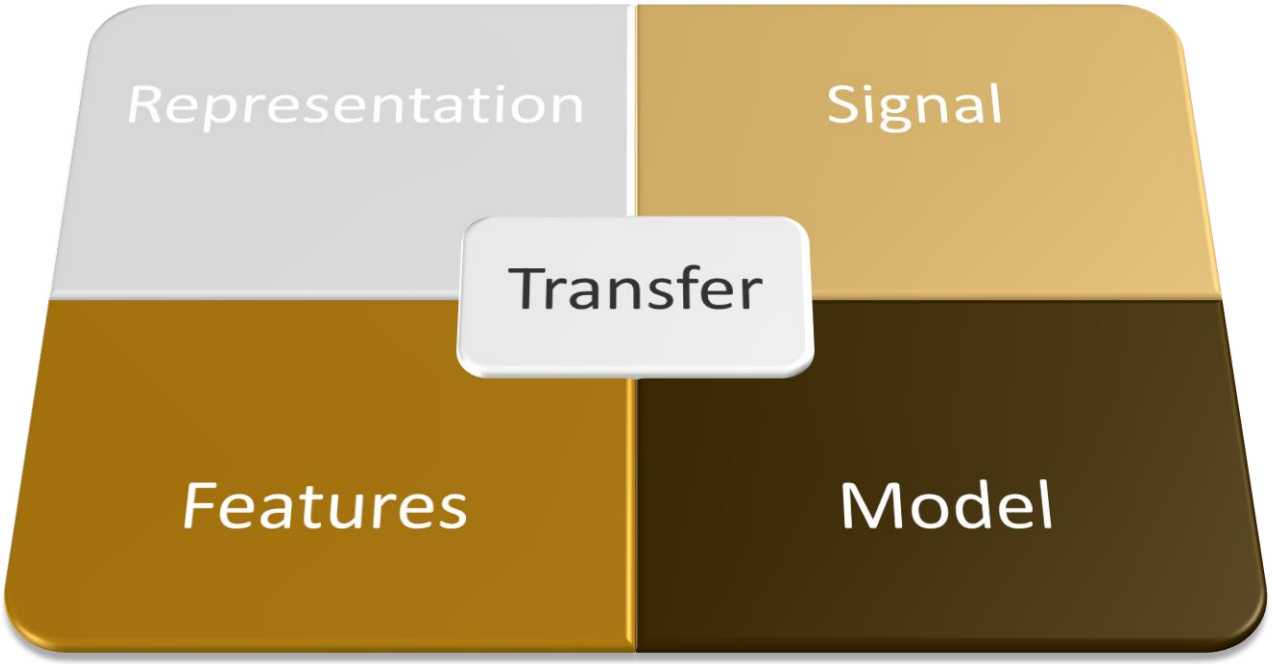
- **Seamless Holism**
- **Horizontal:**
  - Signal Enhancement
  - Feature Extraction
  - Feature Enhancement
  - Feature Transfer
  - Feature Alignment
  - Feature Selection (Bottleneck)
  - Classification / Regression
  - Language Modelling
- **Vertical:**
  - Multitarget
  - w/ Confidences (e.g.. agreement)



Transfer Learning.

Transfer Learning.

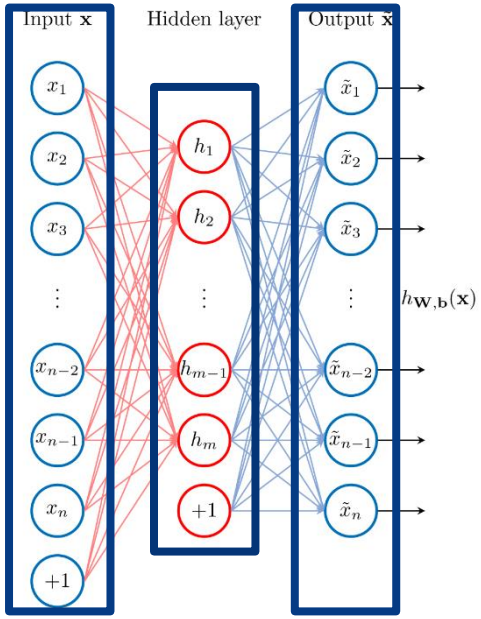
- **>2 Decades**
- **Reuse Data!**
- **Human: Great @it!**
  Know French → Understand some Spanish, Portuguese, etc.

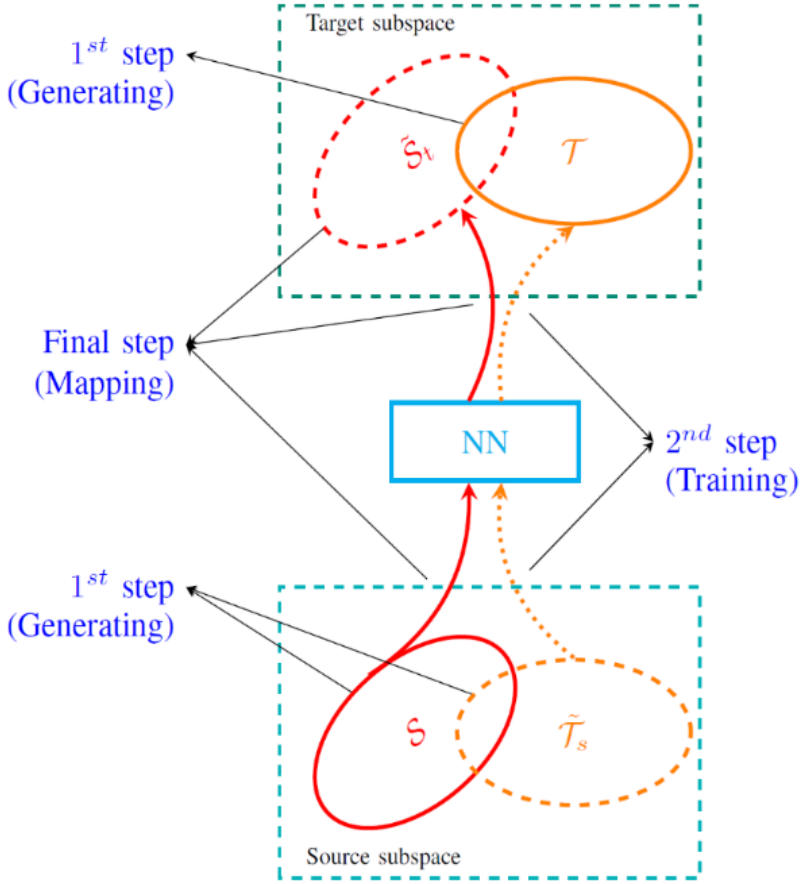


Shared Hidden Layers?

% UA	Target	w/o	DAE	DAE-NN
ComParE: EC	60.4	56.3	59.2	64.2

:
=





Audio = Audio ?

- **Cross-Audio**

  Speech
  Spontaneous: VAM
  Enacted: GEMEP

  Music
  NTWICM

  Sound
  EmoFindSound

  Task-adapted features

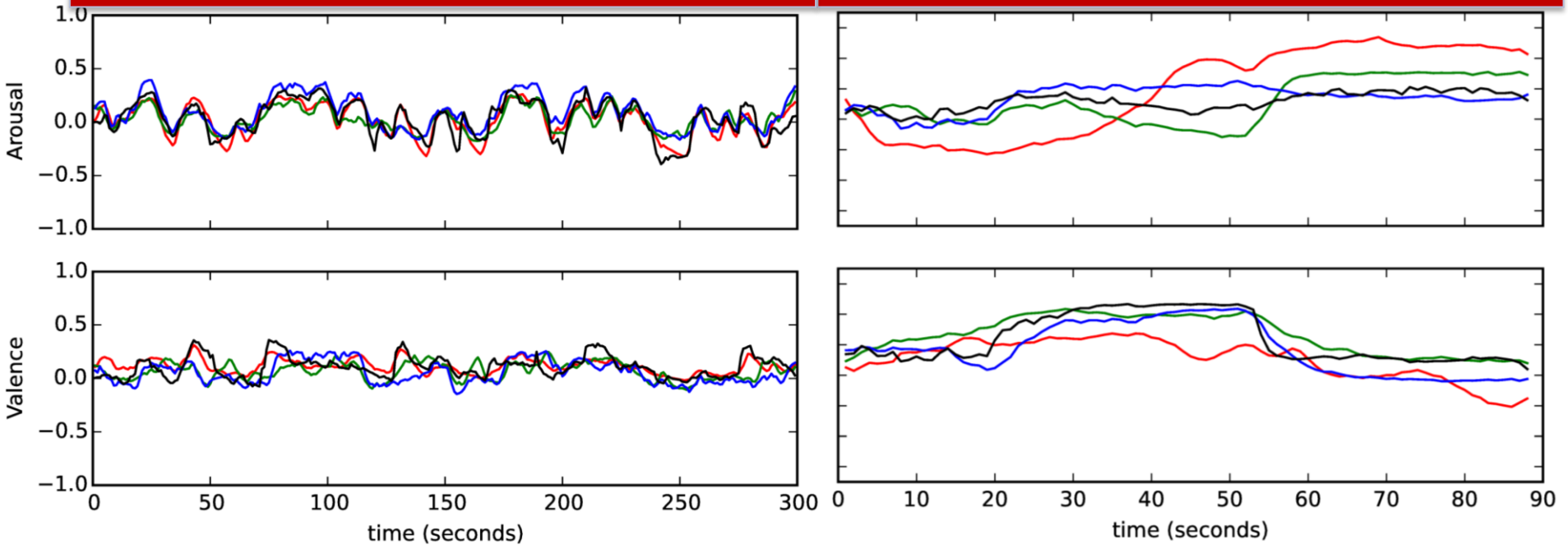
r	Test on				Mean
Train on	Sound	Music	Speech		
			Sp.	En.	
Arousal					
Sound	0.59**	0.46**	0.76**	0.79**	0.65
Music	0.46**	0.67**	0.73**	0.75**	0.65
Speech/Sp.	0.54**	0.47**	0.83**	0.78**	0.66
Speech/En.	0.56**	0.46**	0.77**	0.85**	0.66
Mean	0.54	0.52	0.77	0.79	0.65
Valence					
Sound	0.51**	0.36**	0.27**	0.48**	0.41
Music	0.40**	0.82**	0.33**	0.52**	0.52
Speech/Sp.	0.30**	0.45**	0.44**	0.26°	0.36
Speech/En.	0.45**	0.60**	0.36**	0.50**	0.48
Mean	0.41	0.56	0.35	0.44	0.44

Audio = Audio ?

- **Cross-Domain & Transfer Learning**

CCC Speech	Arousal	Valence
S2S	.749	.332
M2S	.545	.164
M2S-TL	.567	.181

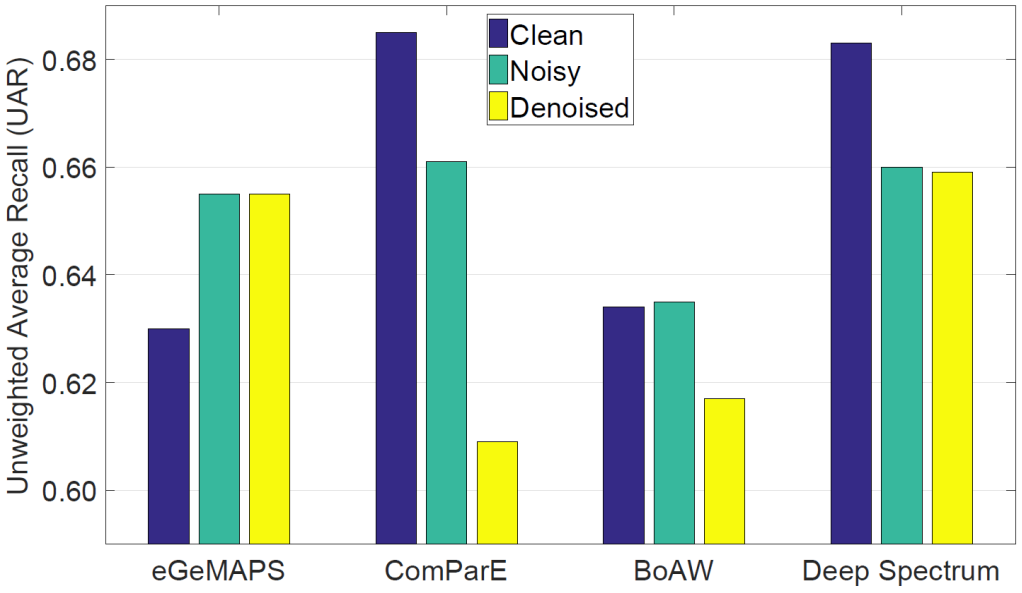
CCC Music	Arousal	Valence
M2M	.317	.090
S2M	.276	.133
S2M-TL	.269	.117



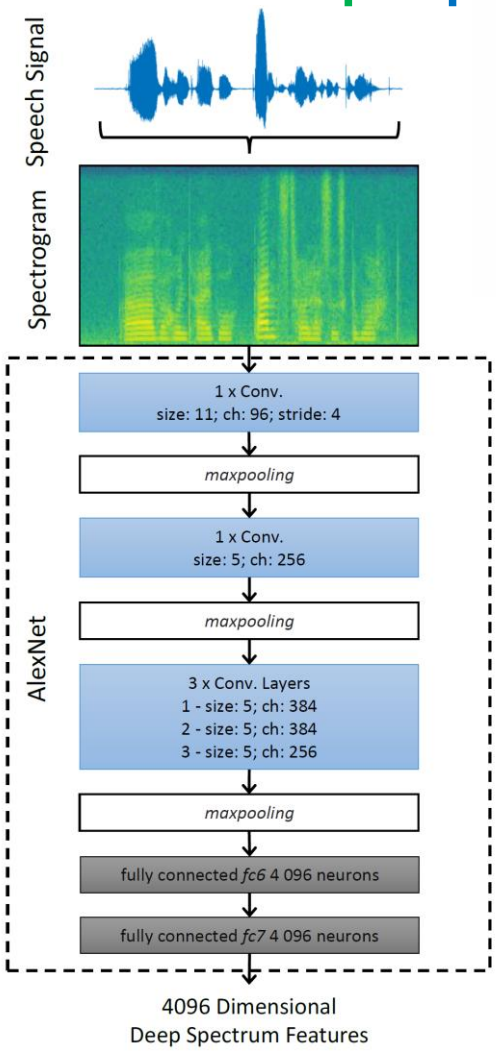
Speech = Images?

- **Emotion with Image Nets**

  IS Emotion Challenge task – 2 classes



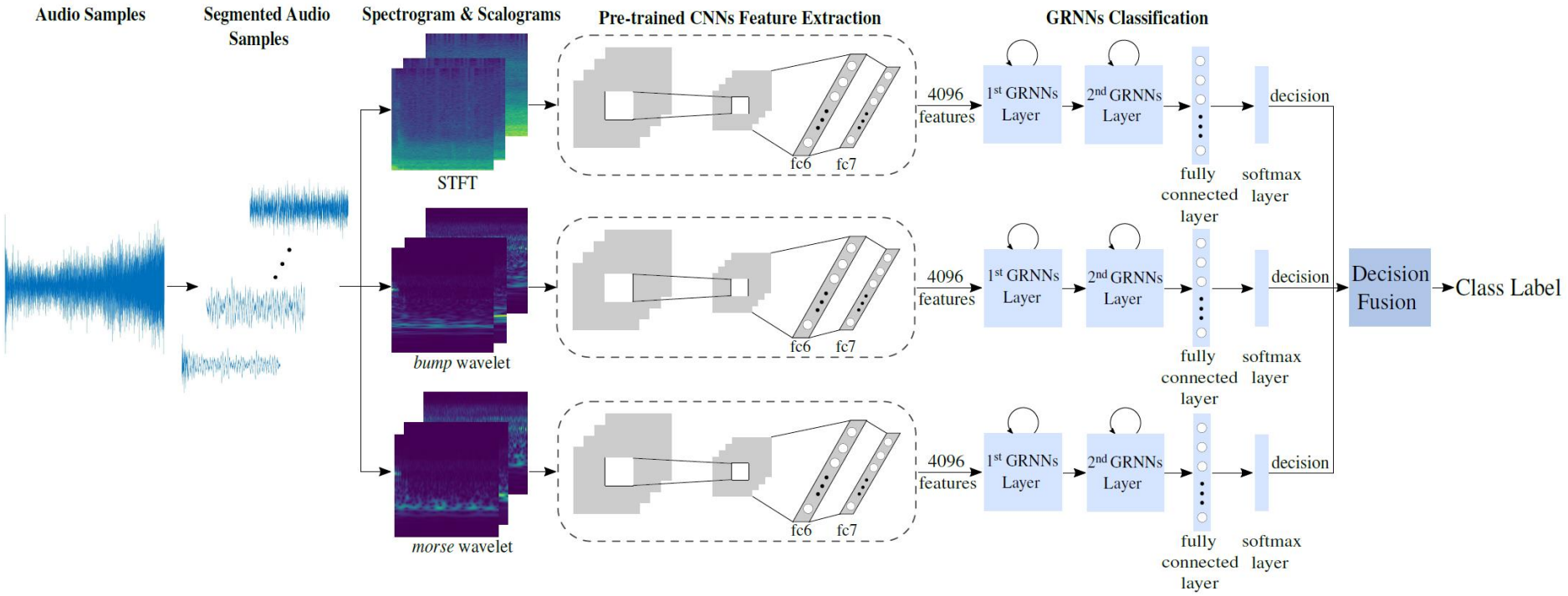
DeepSpectrum



Audio = Images?

- **Wavelets vs STFT via VGG16**

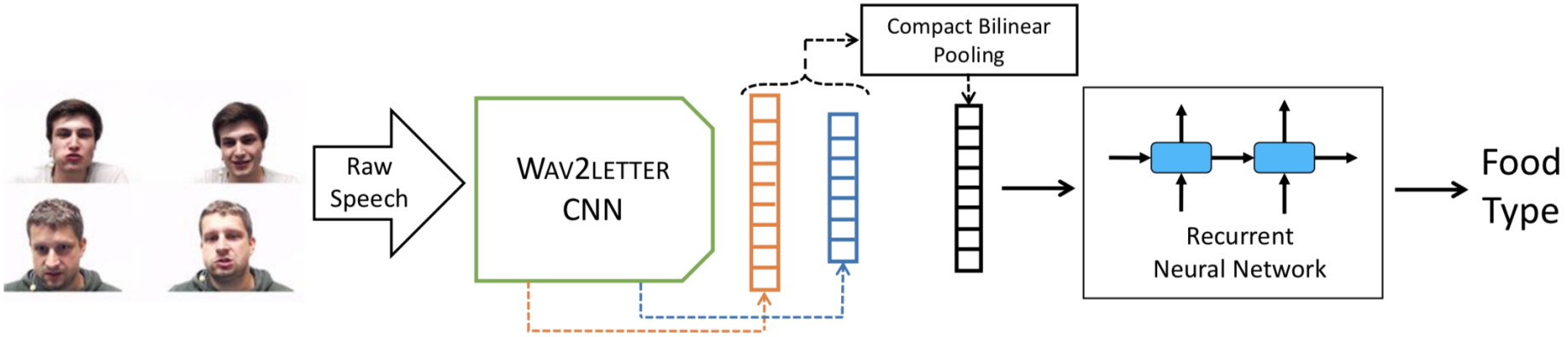
DCASE 2017	%WA
STFT	76.5
STFT+bump	79.8
STFT+morse	76.9
All	80.9



Speech = Speech …

- **Wav2Letter**
  (pre-trained on 1000h speech)

EAT Food-type – UAR [%]	LOSO-CV
Baseline End-to-End CNN-LSTM	-
Baseline BoAW & SVM	64.3
2-layer NN + ReLU + momentum [5]	68.6
Pre-trained CNN & LSTM [6]	67.2
Proposed method	76.4



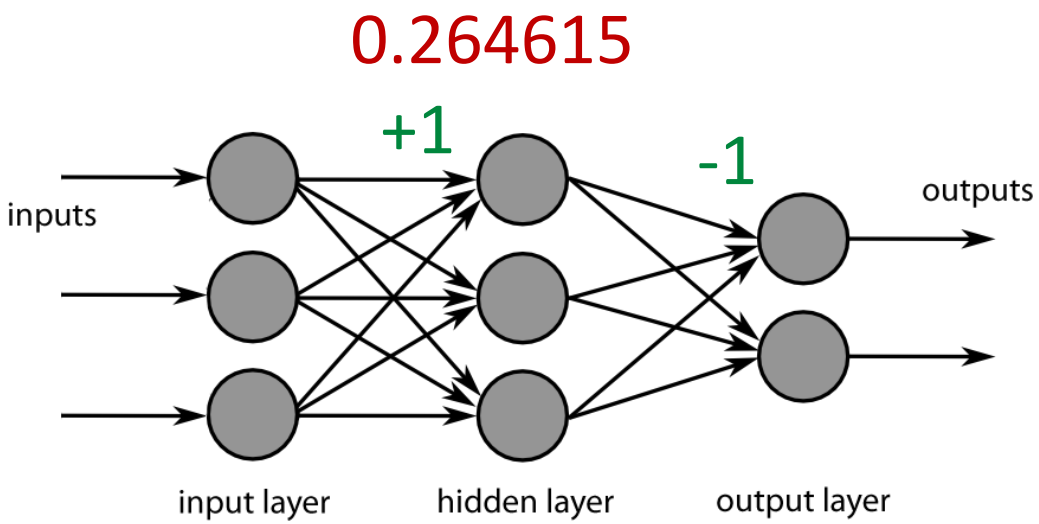
Hacks?

Squeezing.

- **Low precision networks**

  Less precision in weights and activations

  - Binary Neural Networks {-1,1}
  - Ternary neural networks {-1,0,1}
  - Or… 3 bits input, 4 bits weight, 4bits activation, …



Residuum Learning.

- **Random Initialisation of Weights:**
  Information loss at start of training in each layer

  Each Layer adds random noise

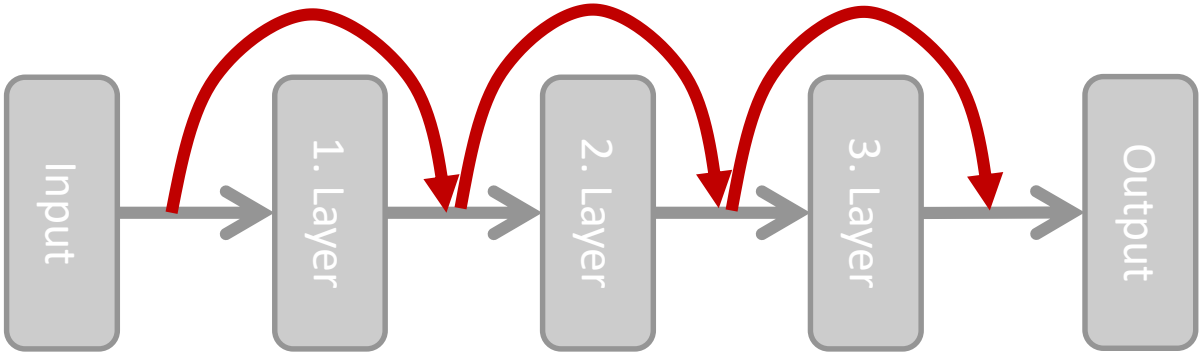
  Problem with very deep nets

  Later layers can only start to learn once
  earlier layers start produce reasonable output

Residuum Learning.

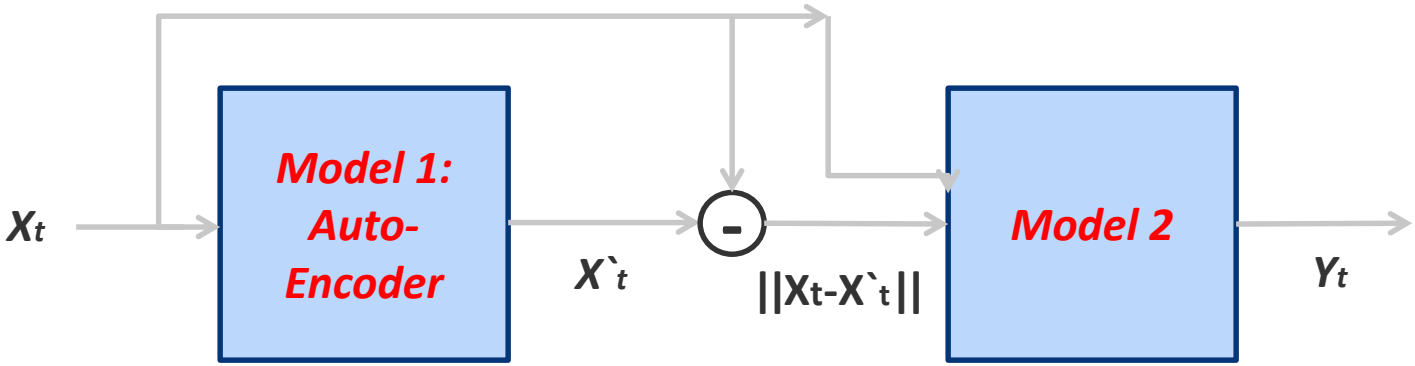
- **How can later layers receive useful information already at start of training?**

- **Solution: Short-cuts for data**



Learning from Errors.

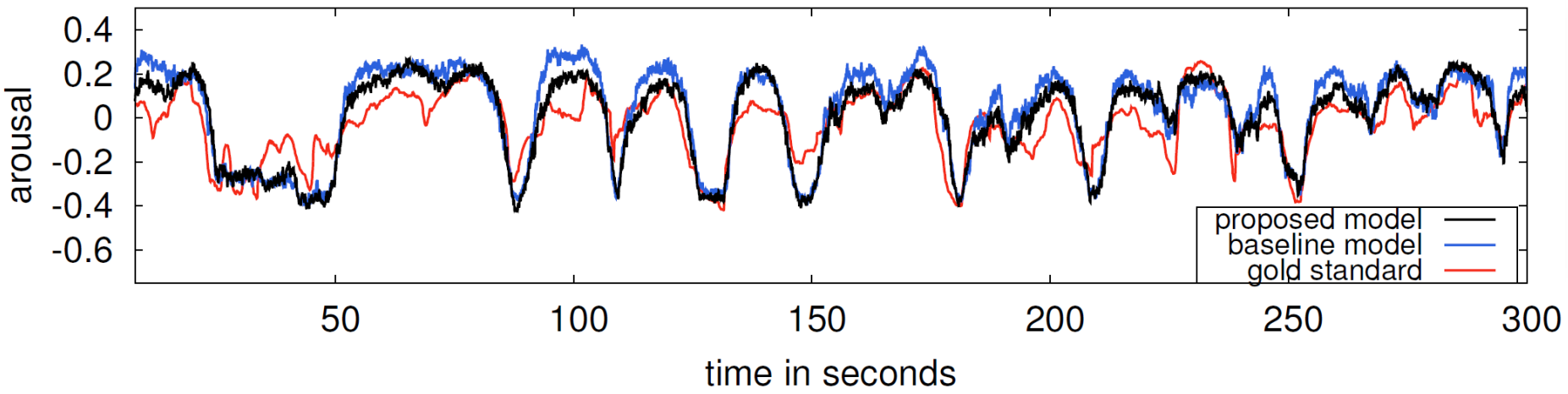
- **Reconstruction Error (RE) in 2 Levels**
  RE of Auto-Encoder as additional input feature

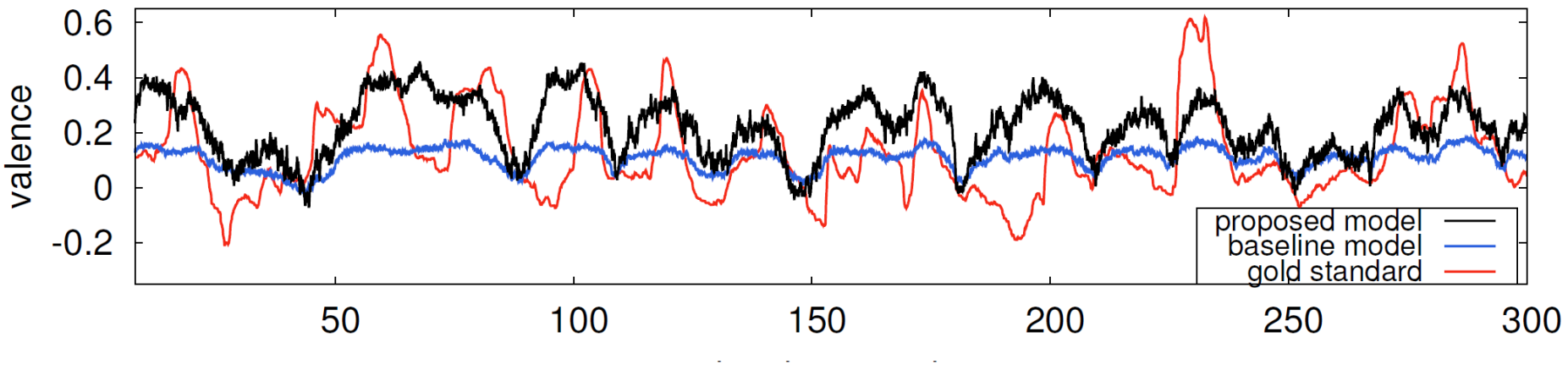


Either: Low Level Descriptors (LLD) or Statistical funtionals

Deep BLSTM RNN

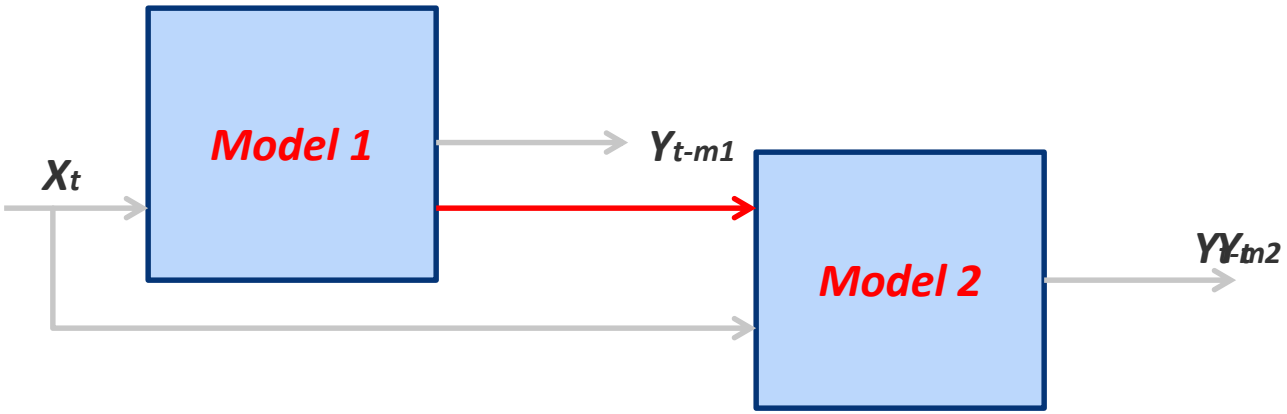
Learning from Errors.





Prediction-based Learning.

concatenate two models for combined strengths

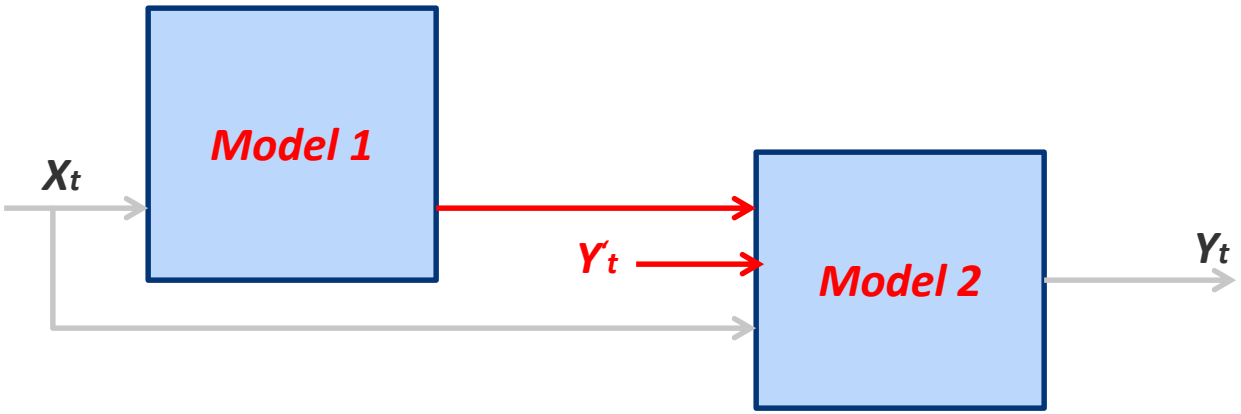


tandem structure:

outputs predicted by first model combined w/ features as input

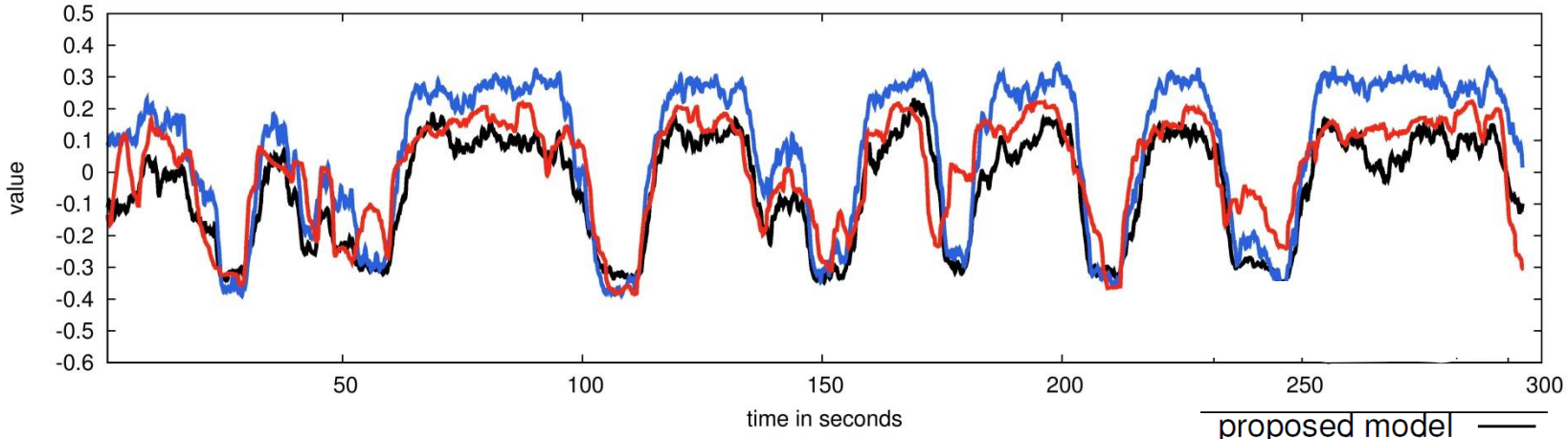
Prediction-based Learning.

(Pseudo prediction: simulated by applying noise to the true label)

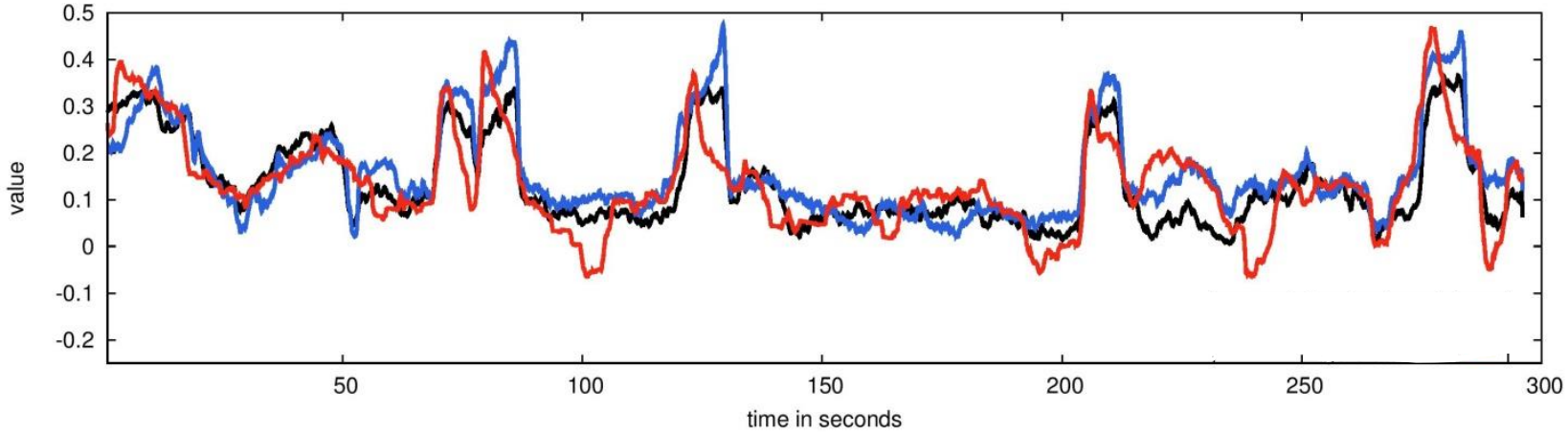


- **SVR**
  - \+ more likely to achieve global optimal solution
  - \- Not context-sensitive
- **BLSTM-RNN**
  - \+ context-sensitive
  - \- Easily trapped in local minimum and risk of overfitting

CCC (Test)	Arousal	Valence
		Baseline
SVR	.726	.300
RNN(2 layers)	.738	.278
RNN (4 layers)	.708	.305
		*Trained with **true** predictions*
RNN-SVR	.726	.387*
SVR-RNN	.730	**.393***
RNN-RNN	.726	.369*
		*Trained with **pseudo** predictions*
RNN-SVR	.729	.301
SVR-RNN	.743*	.373*
RNN-RNN	**.744***	.377*
		State-of-the-art
CCC-objected[1]	.350	.199
End-to-End[2]	.686	.261



arousal

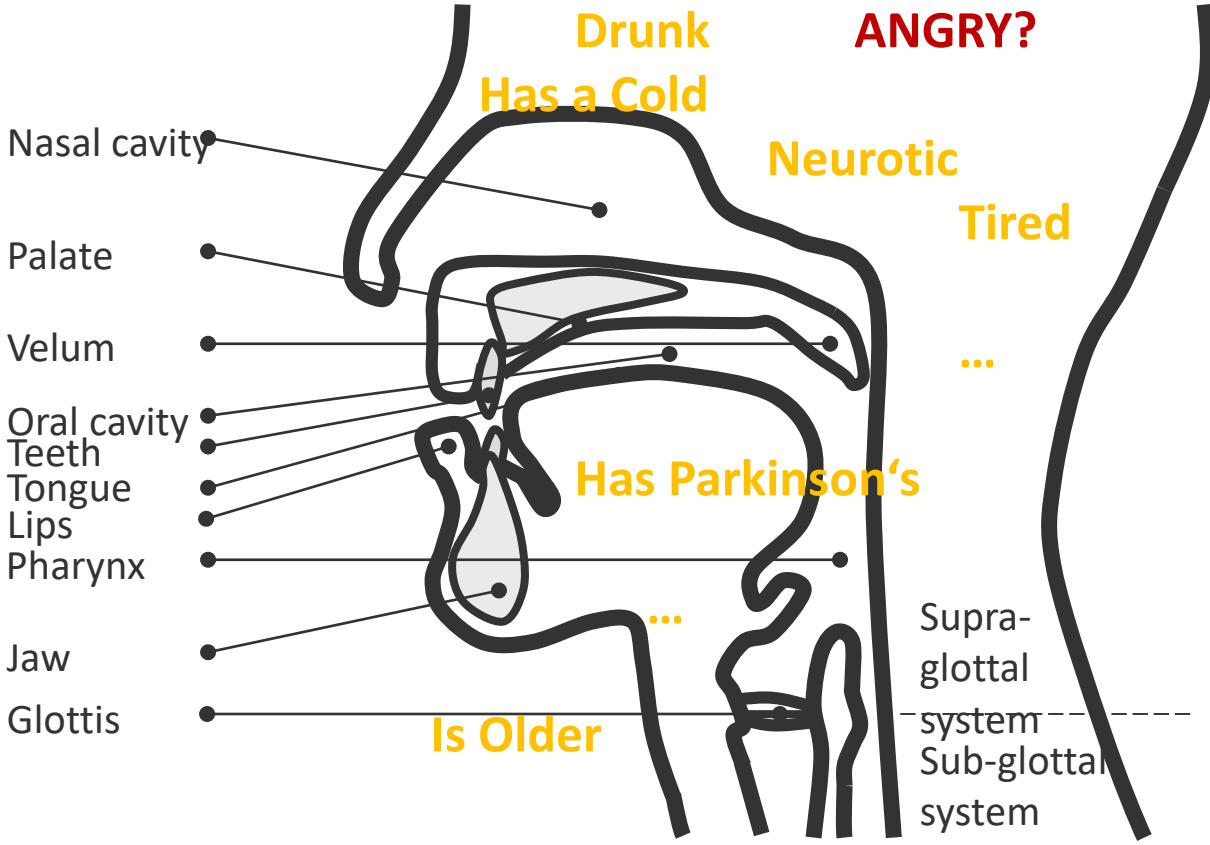


valence

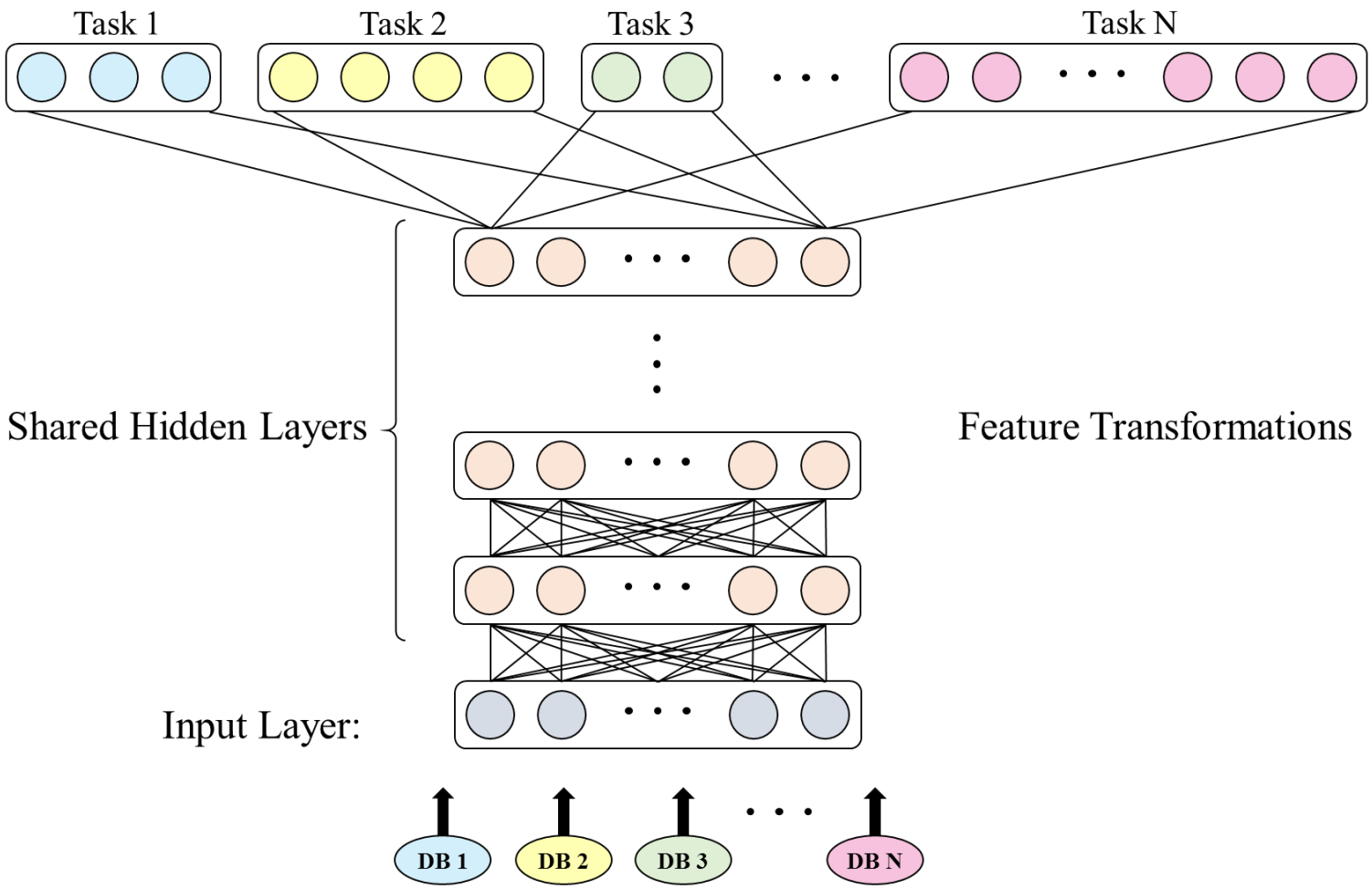
Holistic Modelling.

Modelling: Holism!

- **Multiple-Targets**
- **1 Voice. Face. …**

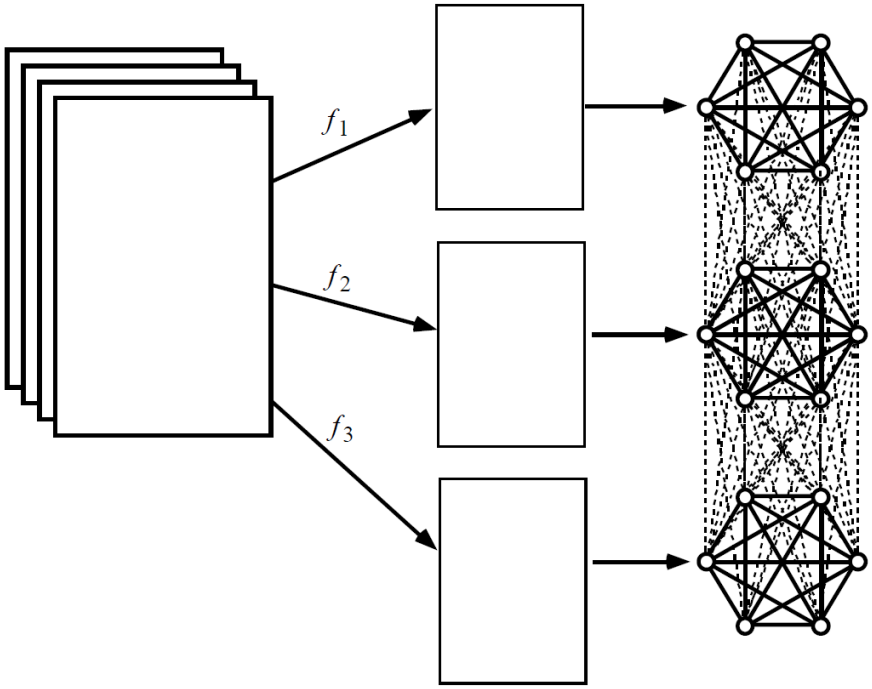


NNs: Multi-target!



Subdivided Networks

Model of sensory processing:



→ Weakly connected Sub Nets?

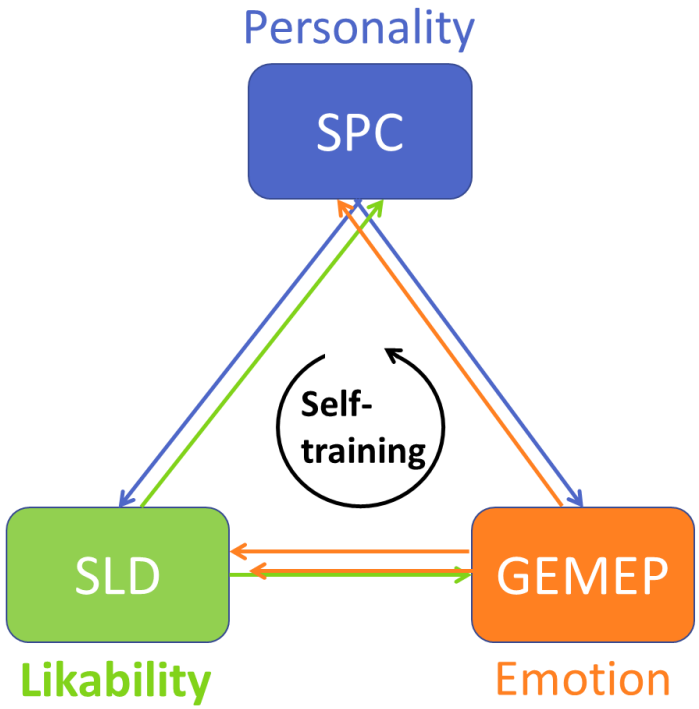
NNs: Multi-target!

- **Cross-Task Self-Labelling**

%UA	Base	CTL
Extraversion	71.7	+1.8
Agreeableness	58.6	+4.5
Neuroticism	63.3	+3.0
Likability	57.2	+2.9

Algorithm: *Cross-Task Labelling*
Repeat for each task:
Repeat until $\mathcal{U} \in \{\}$:

1. (Optional) Upsample training set $\mathcal{L}$ to even class distribution $\mathcal{L}_D$
2. Use $\mathcal{L}/\mathcal{L}_D$ to train classifier $\mathcal{H}$, then classify $\mathcal{U}$
3. Select a subset $\mathcal{N}_{st}$ that contains those instances predicted with the highest confidence values
4. Remove $\mathcal{N}_{st}$ from the unlabelled set $\mathcal{U}, \mathcal{U} = \mathcal{U} \setminus \mathcal{N}_{st}$
5. Add $\mathcal{N}_{st}$ to the labelled set $\mathcal{L}$, $\mathcal{L} = \mathcal{L} \cup \mathcal{N}_{st}$



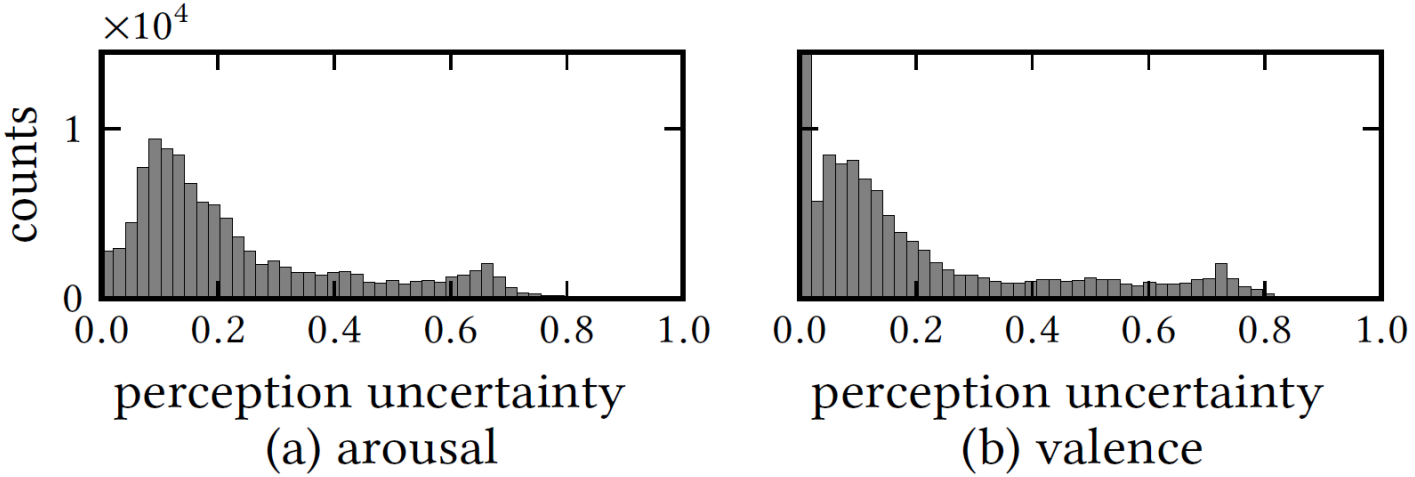
Co-Learning Uncertainty.

- **Multi-task Learning of Subjective / Uncertain Ground Truth**

Example: Arousal / Valence (SEWA data of AVEC 2017)

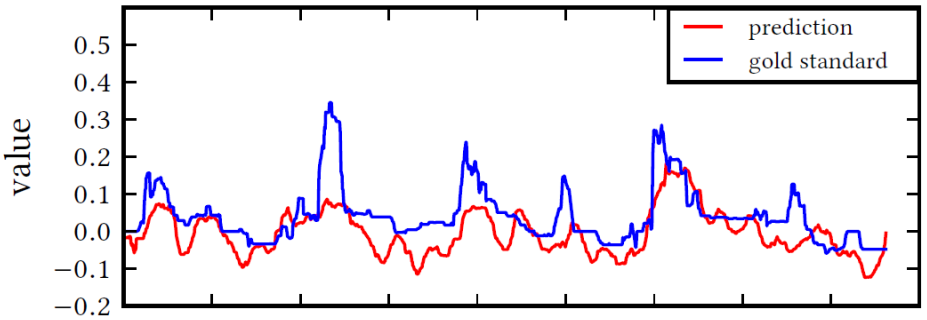
Perception uncertainty (K ratings):

$$\sigma_n^{(i)} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(e_{n,k}^{(i)} - e_n^{\mathrm{MLE},(i)})^2}$$

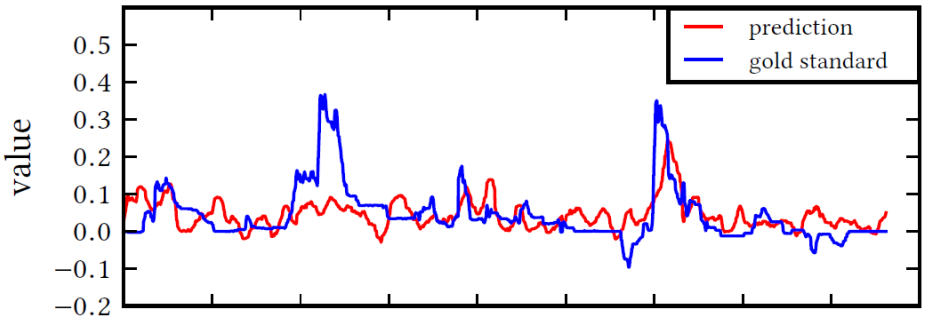


Co-Learning Uncertainty.

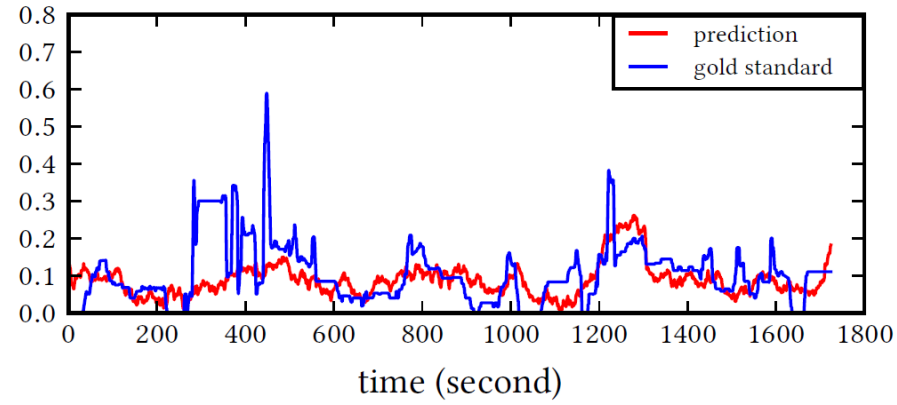
CCC		dev.		test		dev.		test	
modality	task	$E^{(A)}$	$\sigma^{(A)}$	$E^{(A)}$	$\sigma^{(A)}$	$E^{(V)}$	$\sigma^{(V)}$	$E^{(V)}$	$\sigma^{(V)}$



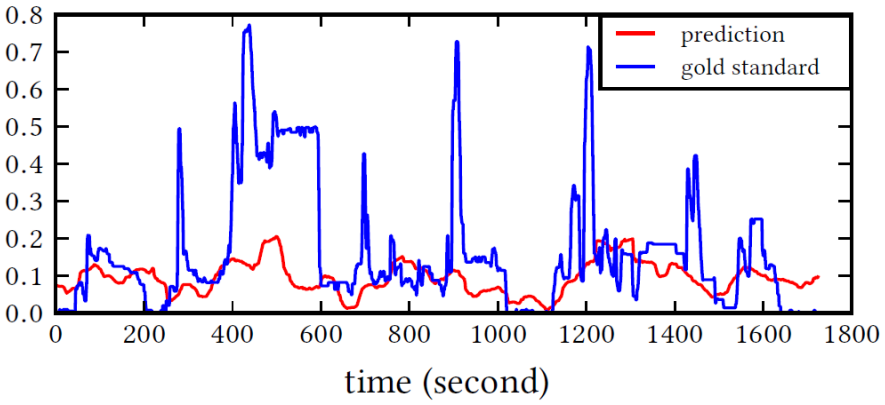
(a) arousal, emotional state $E^{(A)}$



(b) valence, emotional state $E^{(V)}$



(c) arousal, perception uncertainty $\sigma^{(A)}$



(d) valence, perception uncertainty $\sigma^{(V)}$

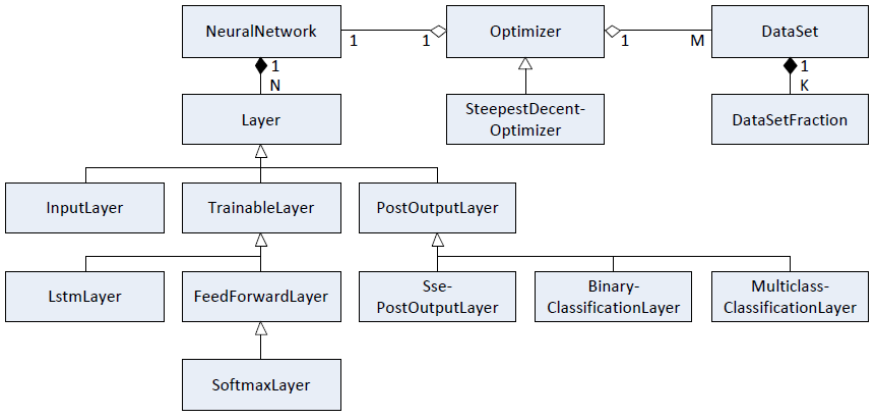
Labels?

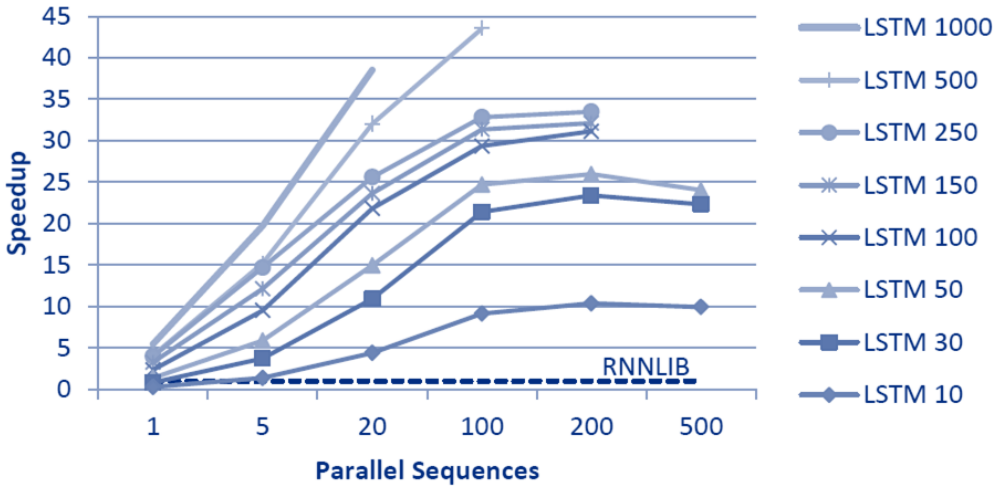
Big Data.

CURRENNT

- **GPU-Learning**
  - 10 – 1k LSTM cells.
  - 2k – 4Mio parameters
  - GPGPU

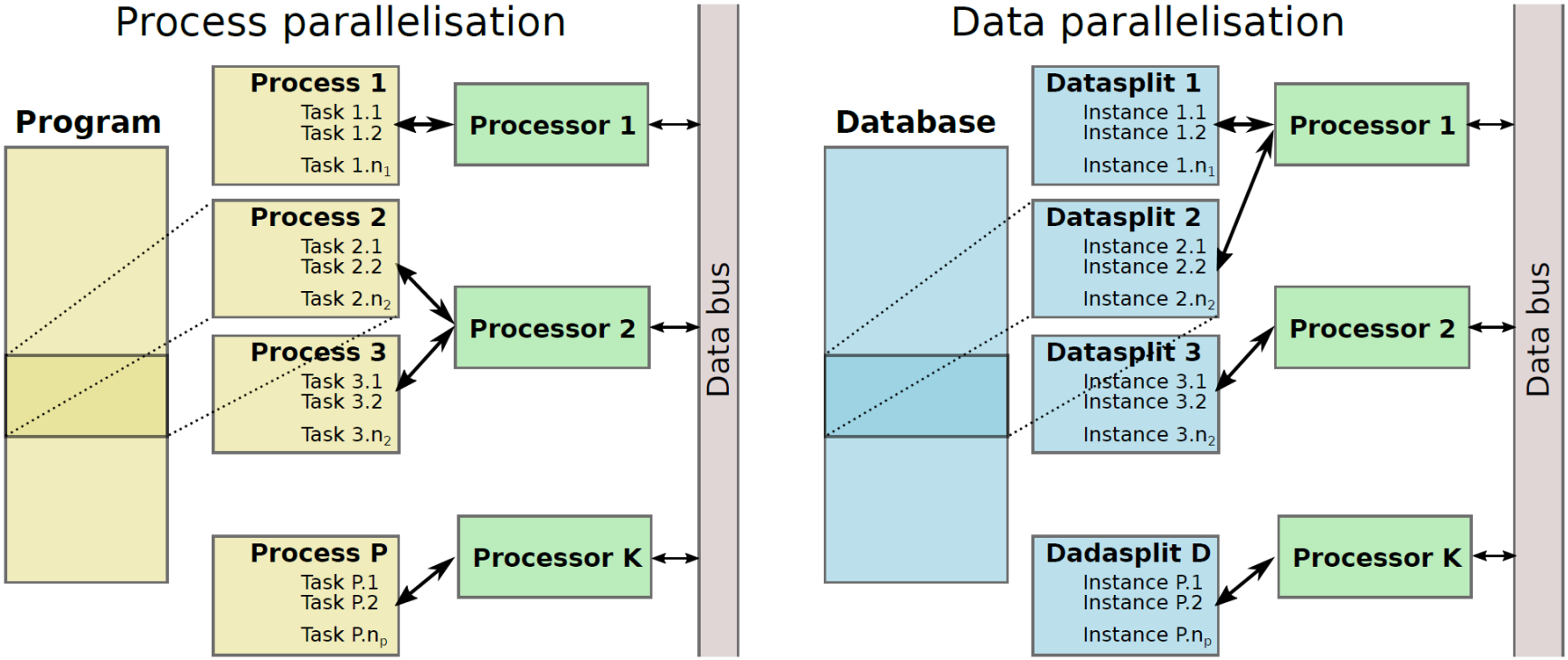
CHiME 2	RNNLIB	CURRENNT		
#seq.	1	1	10	200
speedup	(1)	2	13	22





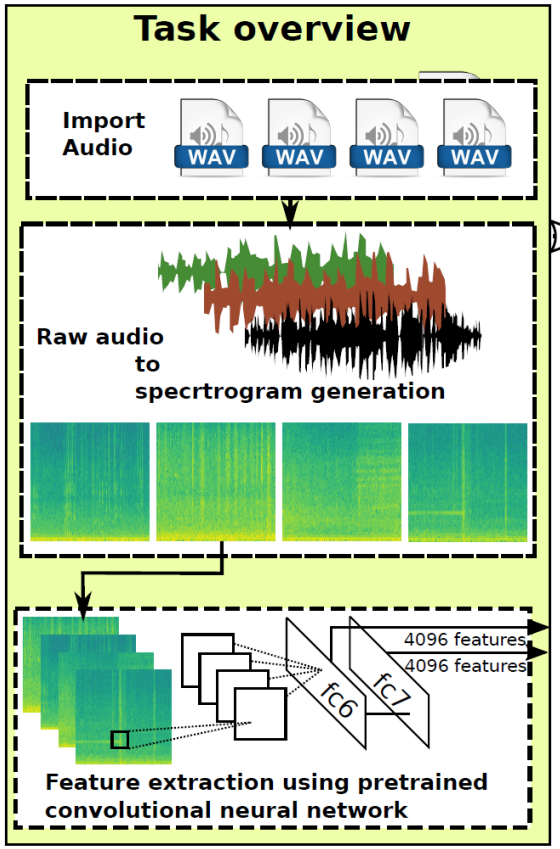
Big Data.

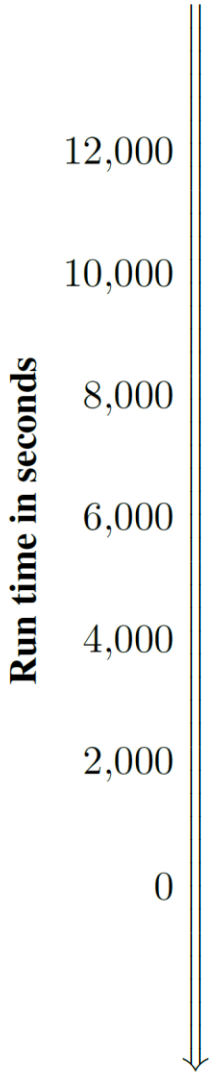
- **Parallelisation**



Big Data.

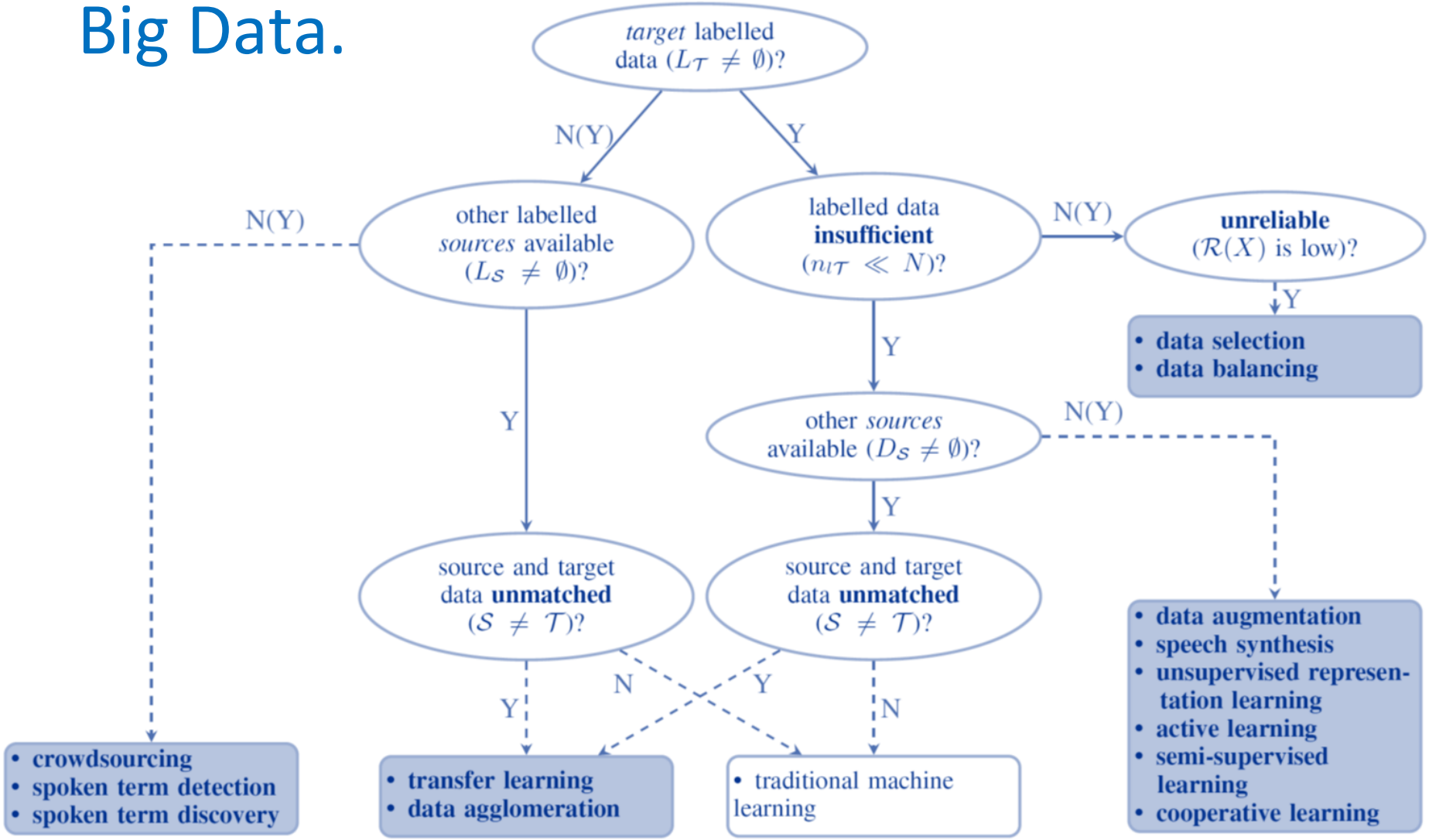
- **Parallel…**





AlexNet	VGG19
Input = RGB image	
size: 227 × 227 pixels	size: 224 × 224 pixels
1×Convolution size: 11; ch: 96; stride: 4	2×Convolution size: 3; ch: 64; stride: 1
Maxpooling	
1×Convolution size: 5; ch: 256	2×Convolution size: 3; ch: 128
Maxpooling	
1×Convolution size: 3; ch: 384	4×Convolution size: 3; ch: 256
	Maxpooling
1×Convolution size: 3; ch: 384	4×Convolution size: 3; ch: 512
	Maxpooling
1×Convolution size: 3; ch: 256	4×Convolution size: 3; ch: 512
Maxpooling	
Fully connected *fc6* layer, 4 096 neurons	
Fully connected *fc7* layer, 4 096 neurons	
Fully connected 1 000 neurons	
Output = Probabilities for 1 000 object classes through soft-max	

Big Data.


target labelled data ($L_T \neq \emptyset$)?
N(Y)
Y
other labelled sources available ($L_S \neq \emptyset$)?
labelled data insufficient ($n_{lT} \ll N$)?
N(Y)
unreliable ($\mathcal{R}(X)$ is low)?
Y
• data selection
• data balancing
N(Y)
Y
Y
other sources available ($D_S \neq \emptyset$)?
N(Y)
Y
source and target data unmatched ($\mathcal{S} \neq \mathcal{T}$)?
source and target data unmatched ($\mathcal{S} \neq \mathcal{T}$)?
• data augmentation
• speech synthesis
• unsupervised representation learning
• active learning
• semi-supervised learning
• cooperative learning
Y
N
Y
N
• crowdsourcing
• spoken term detection
• spoken term discovery
• transfer learning
• data agglomeration
• traditional machine learning

Big Data.

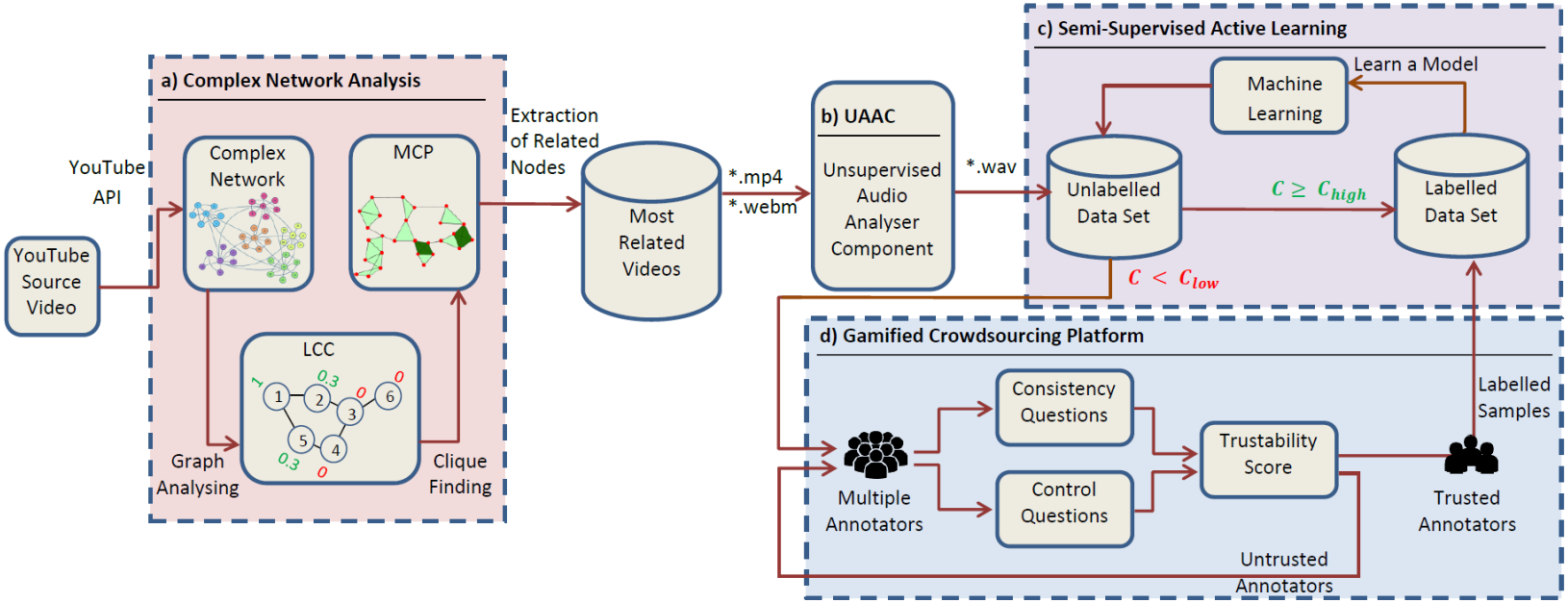
- **Targeted Data Acquisition**

  Small World Modelling: find highly related videos

  Local Clustering Coeff.+Maximum Clique Problem

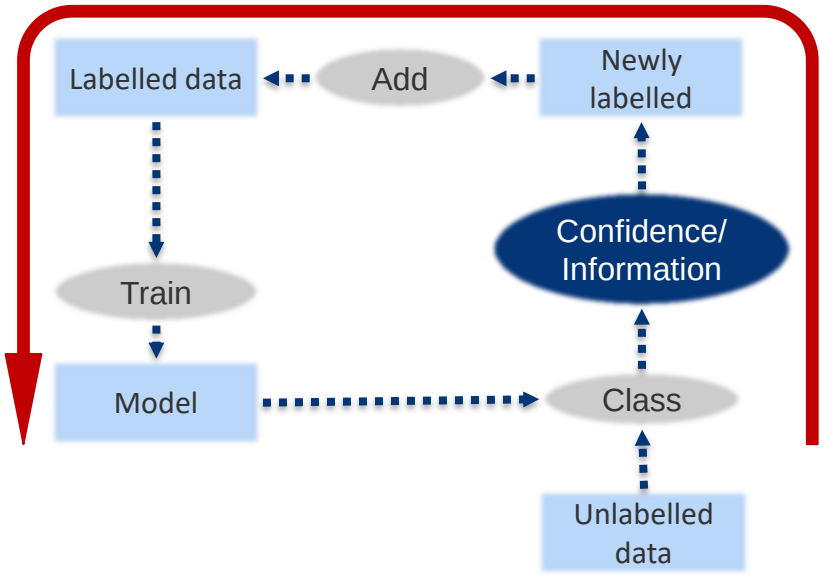
  Example: 3k videos for rapid training of new tasks

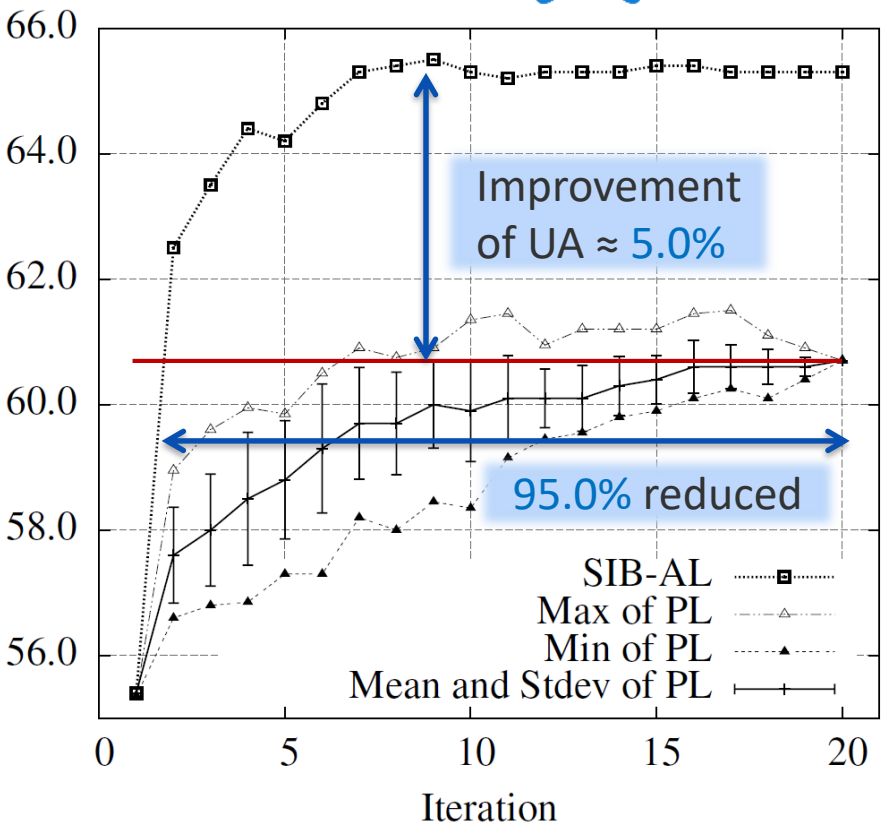
Task	%UA	BEST
Freezing	70.2	func.
Coughing	97.6	BoAW
Sneezing	85.2	NN
Intoxicat.	72.6	BoAW



Big Data.

- **Cooperative Learning**
  0) Transfer Learning
  1) Dynamic Active Learning
  2) Semi-Supervised Learning

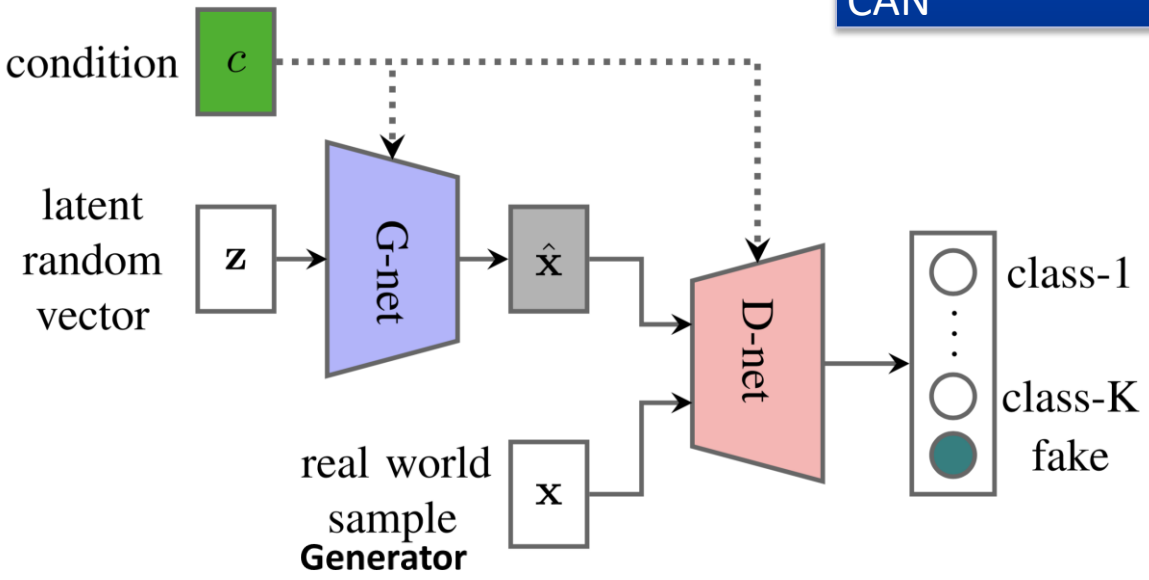


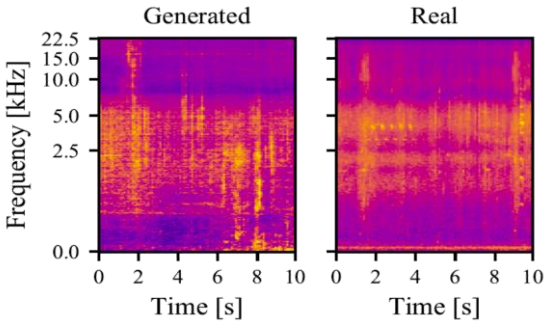


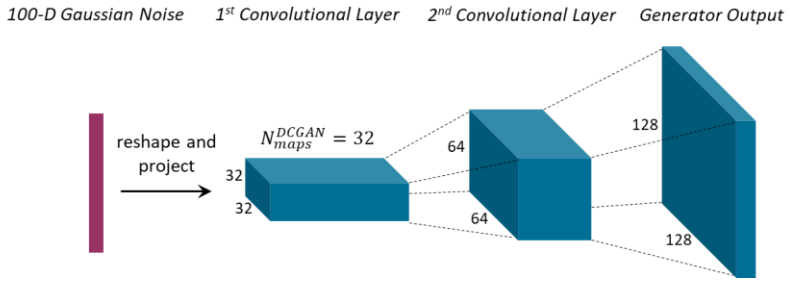
Adversarial Nets.

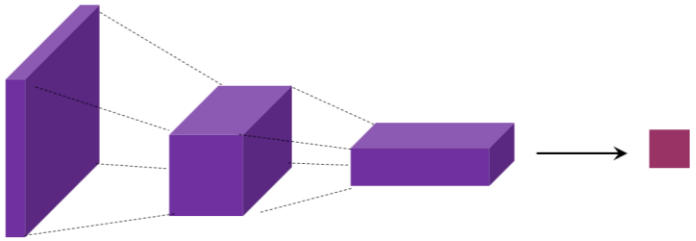
CCC Recola	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261
CAN	.737	.455

- **Conditional Adversarial Nets**



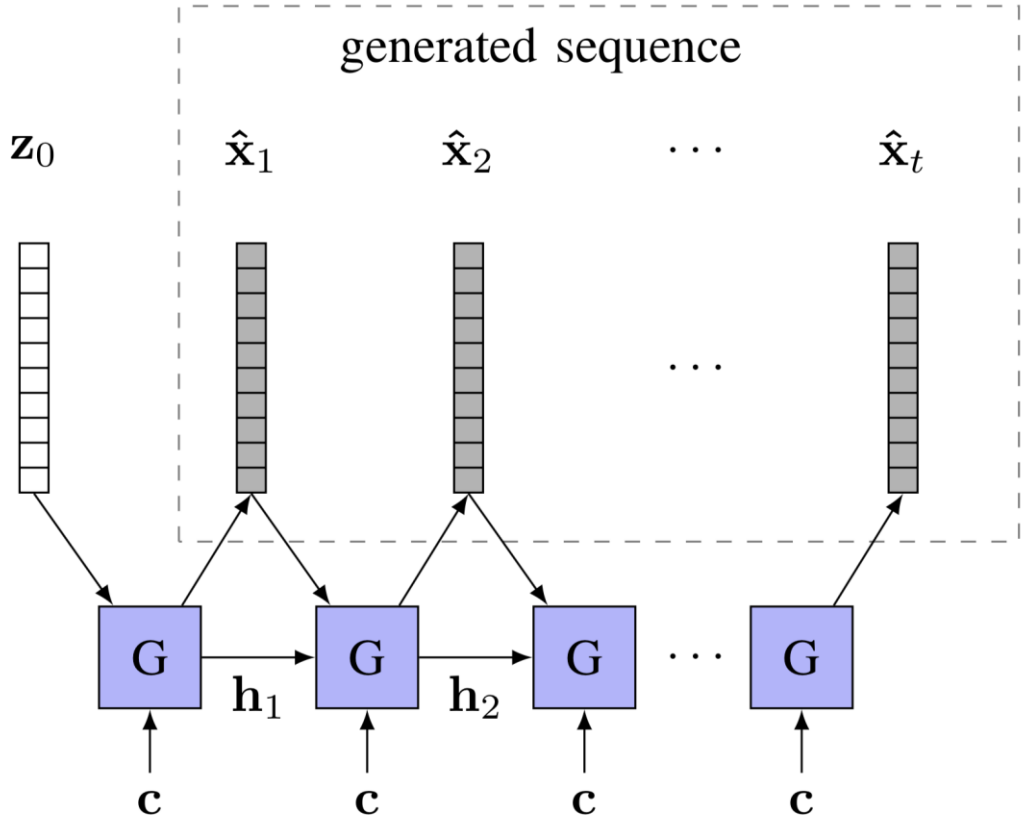






Adversarial Nets.

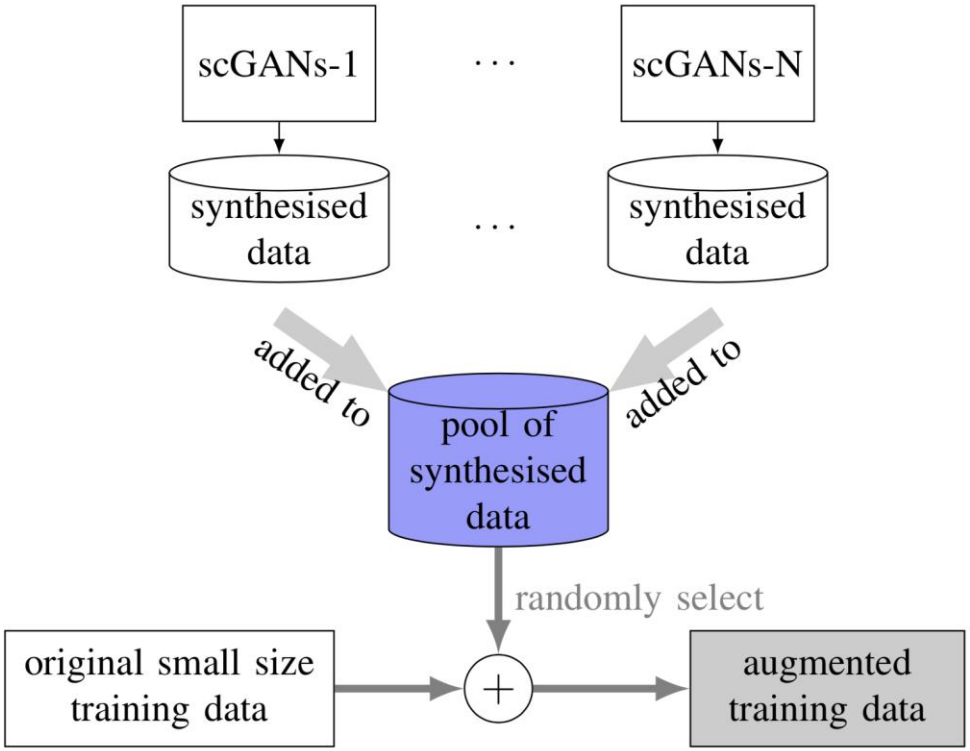
- **Sequence Generation by Recurrent Generator**



Adversarial Nets.

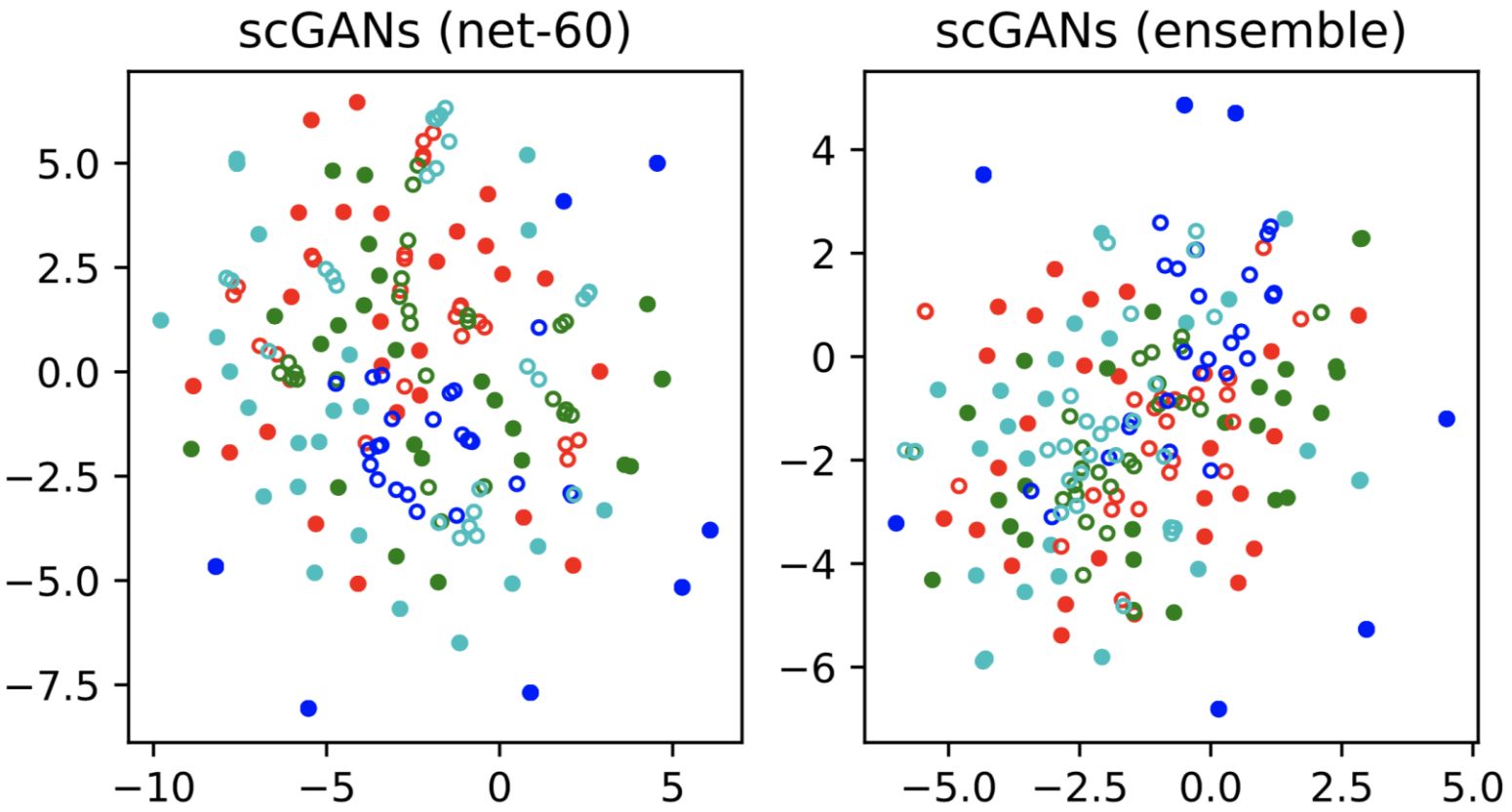
ComParE 17 Snoring	%UA
Baseline (BoAW)	48.2
scGAN	54.8
scGAN Ensemble	56.7

- **Ensemble semi-supervised CANs against Model Collapse**



Adversarial Nets.

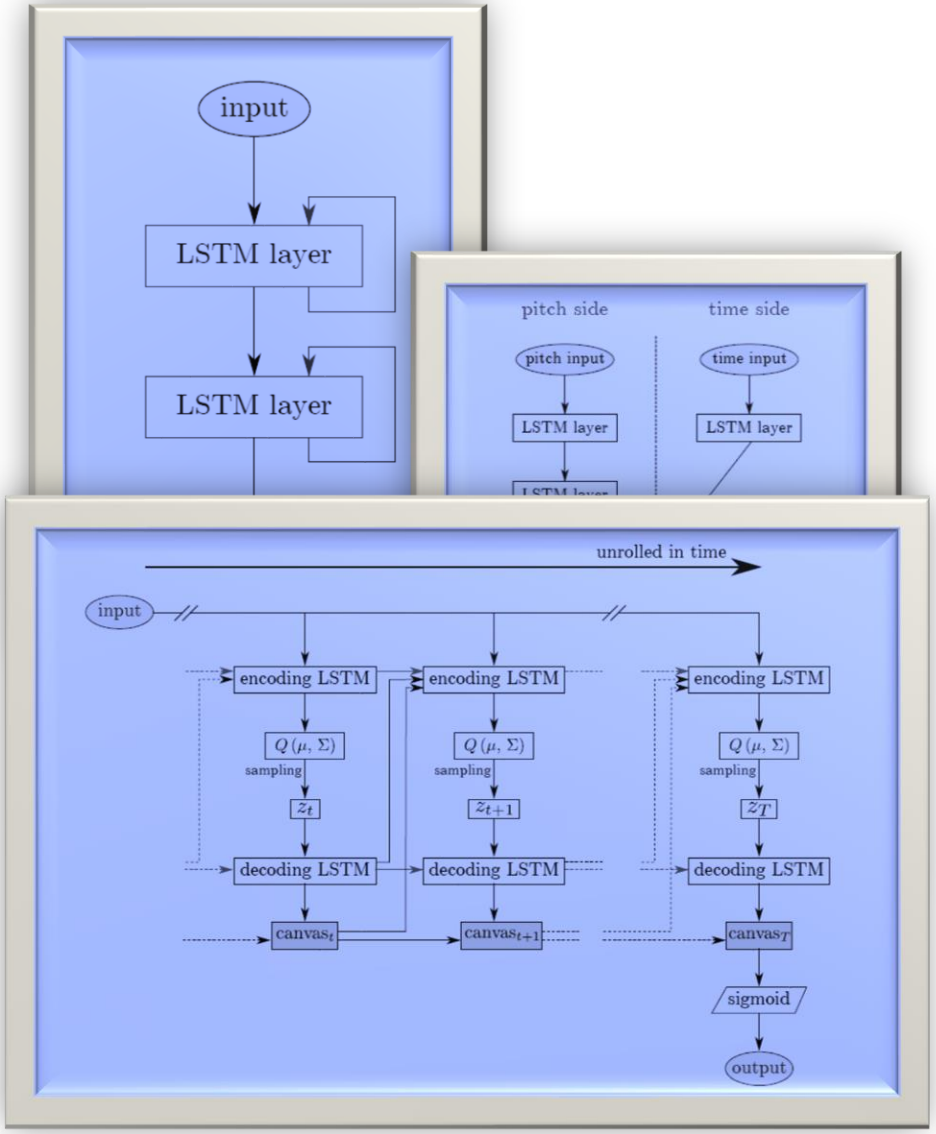
- **Ensemble GANs against Model Collapse: t-SNE Visualisation**



Generation?

Creativity.

- **Deep Creation**
- **Deep Matching**

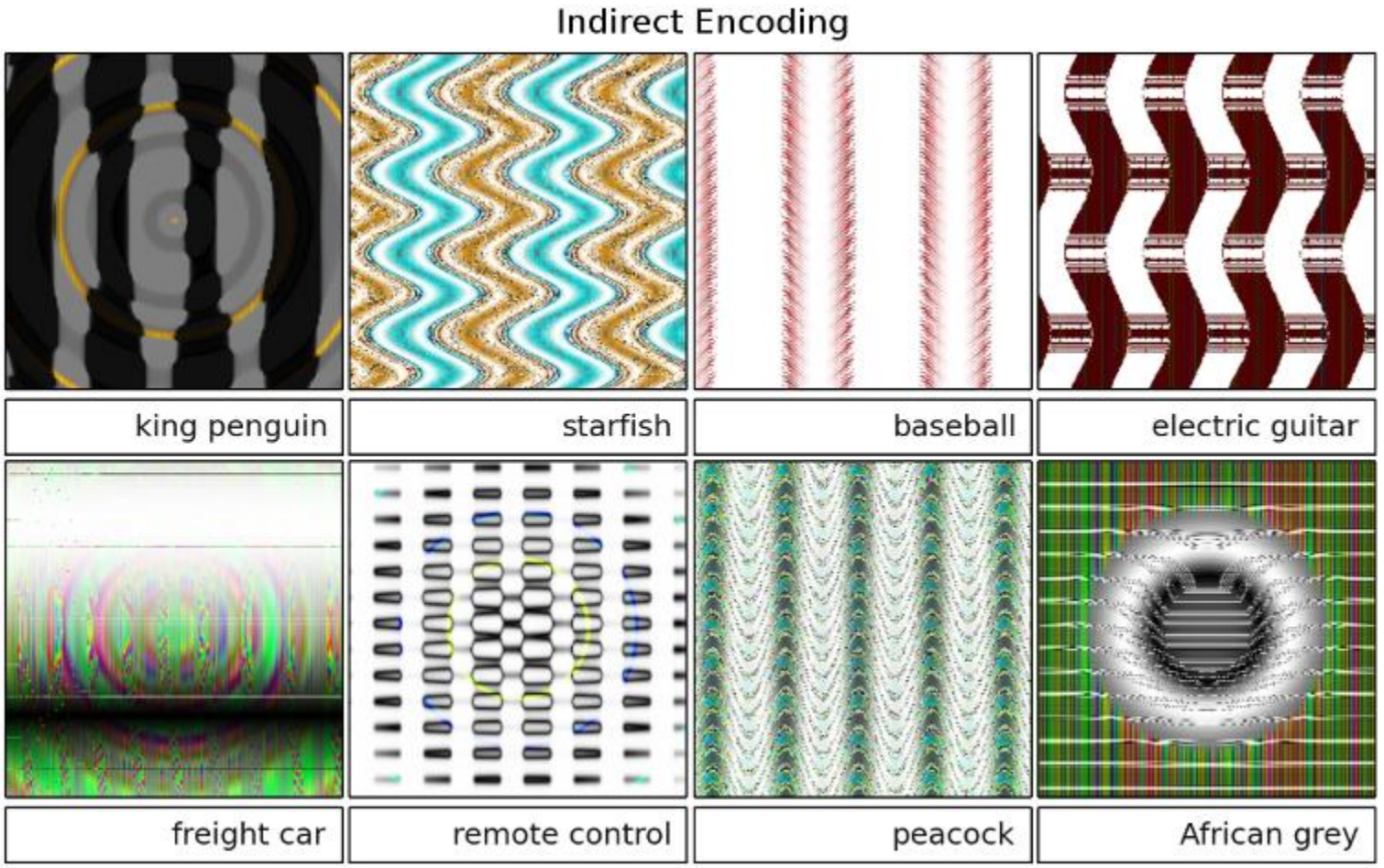


*"Deep Recurrent Music Writer: Memory-enhanced Variational Autoencoder-based Musical Score Composition and an Objective Measure". **IJCNN**. 2017.*

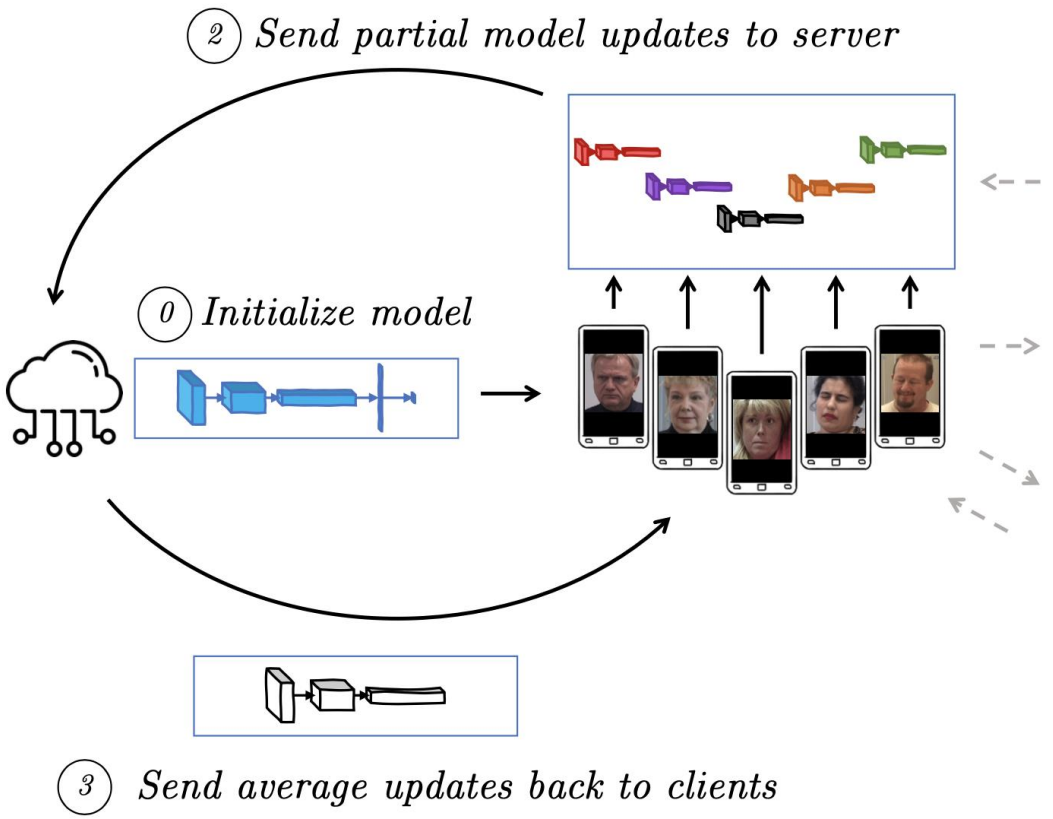
Problems?

Fooling Deep AI?

Nguyen.Yosinski. Clune (2015):
Deep



Federated Learning



Emotion	%UA
No Learn	44
Baseline	73
UAU	76

Ethical. Legal. Societal Implications?

- **Emergent Intelligence?**
  → Consciuosness?
- **Computer Night Mares?**
- **White Boxing?**
- **Responsible?**

Bottlenecks.

- **Sufficient (Labelled) Data**
  → Weakly supervised & transfer learning
- **Model complexity (fully model relationship in data) → Deep Learning**
- **Computing time**
  → NPUs, TPUs, Parallelisation, …

Deep Learning Limitations?

- **Just mapping input → output w/ continuous geometric transform**
- **w/ large human-annotated data**
- **far from human-alike A.I.: also reasoning & abstraction.**
- **→ Learn any program, in modular/reusable way**

Challenges.

- Learning from few examples
- Emotional/Human-alike Learning
- Socioemotional Intelligence
- Causal Relations?
- Logical Interference?
- Learning Grammar? / Production Rules?
  → “Reasonable Reasoning”?
- Explainable/Responsible/Accountable
- Reinforcement Learning w/ intrinsic motivation

- “Always Active” (Dreaming!)
- Emergent Intelligence? Consciousness?

Ways Out.

- New forms of learning beyond differentiable transforms → No BP but neuron activity?
- Models closer to general-purpose computer programs based on richer primitives (than differentiable layers) → Reasoning & Abstraction
- Less human involvement → Automatic ML
- Reuse of learnt features, models, architectures

Universität Augsburg
Embedded Intelligence for
Health Care and Wellbeing