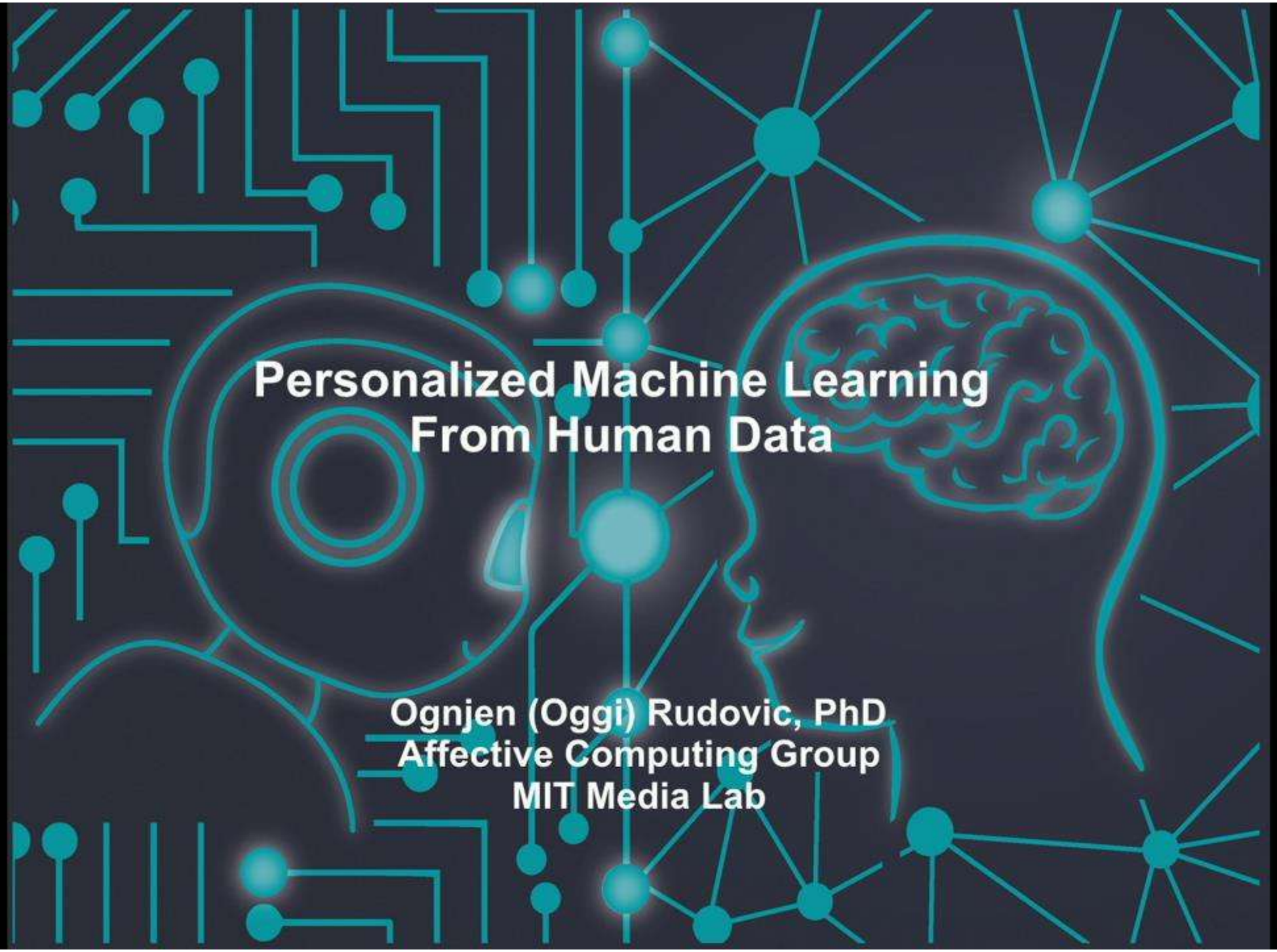

Personalized Machine Learning
From Human Data
Ognjen (Oggi) Rudovic, PhD
Affective Computing Group
MIT Media Lab

Research Path

Faculty of Electrical Eng., Uni. of Belgrade (BA, 2007)

Research Path

Faculty of Electrical Eng., Uni. of Belgrade (BA, 2007)

Computer Vision Center
UAB, Barcelona (MSc, 2009)

Computing Dept.,
Imperial College London (PhD, 2013)

Research Path

Faculty of Electrical Eng., Uni. of Belgrade (BA, 2007)

Media Lab
MIT (MC Fellow, 2016)

Computer Vision Center
UAB, Barcelona (MSc, 2009)

Computing Dept.,
Imperial College London (PhD, 2013)

Research Path

Faculty of Electrical Eng., Uni. of Belgrade (BA, 2007)

Media Lab
MIT (MC Fellow, 2016)

Computer Vision Center
UAB, Barcelona (MSc, 2009)

Computing Dept.,
Imperial College London (PhD, 2013)

Talk Outline

- Previous work on Facial Behavior Modeling
- Robot-assisted Autism Therapy
- Personalized Deep Networks
- Human Data and Future Directions

Modeling of Human Facial Behavior

Application Domains

Modeling of Human Facial Behavior

Two main approaches to encoding of facial expressions (Ekman, 2005):

Message judgment (emotion perception)

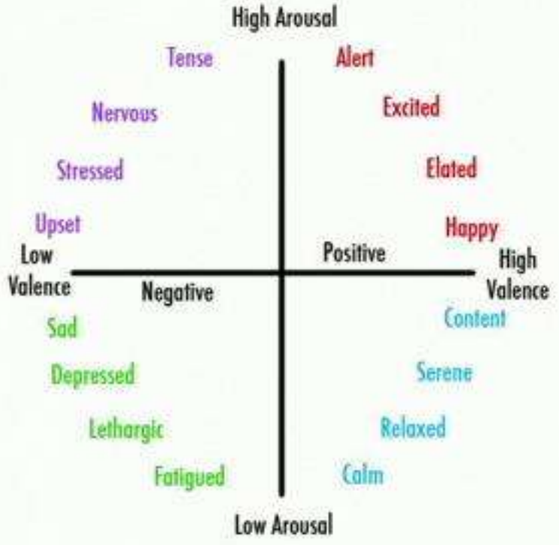
Modeling of Human Facial Behavior

Two main approaches to encoding of facial expressions (Ekman, 2005):

Message judgment (emotion perception)

Categorical

Dimensional



Ekman, P. (2005). Facial Expressions. John Wiley & Sons, Ltd.

Russell, J. A., and Pratt, G. (1980). A description of the affective quality attributed to environments. J. Pers. Soc. Psychol.

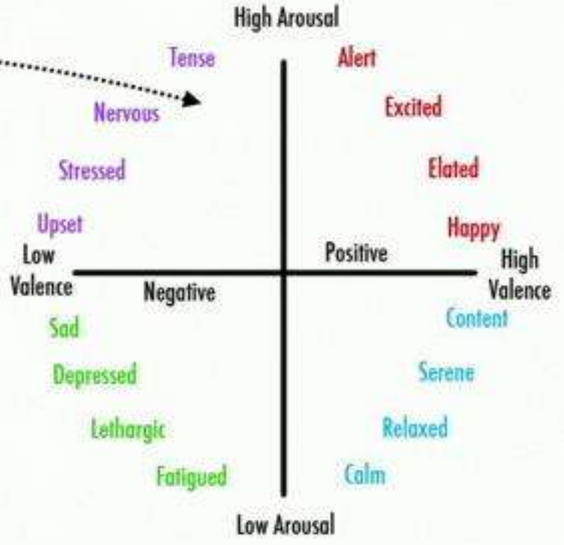
Modeling of Human Facial Behavior

Two main approaches to encoding of facial expressions (Ekman, 2005):

Message judgment (emotion perception)

Categorical

Dimensional



Ekman, P. (2005). Facial Expressions. John Wiley & Sons, Ltd.

Russell, J. A., and Pratt, G. (1980). A description of the affective quality attributed to environments. J. Pers. Soc. Psychol.

Modeling of Human Facial Behavior

Two main approaches to encoding of facial expressions (Ekman, 2005):

Sign judgment (face grammar)

Facial Action Unit Activations (0/1)

More than 32 Action Units

Ekman, P. (2005). Facial Expressions. John Wiley & Sons, Ltd.

Modeling of Human Facial Behavior

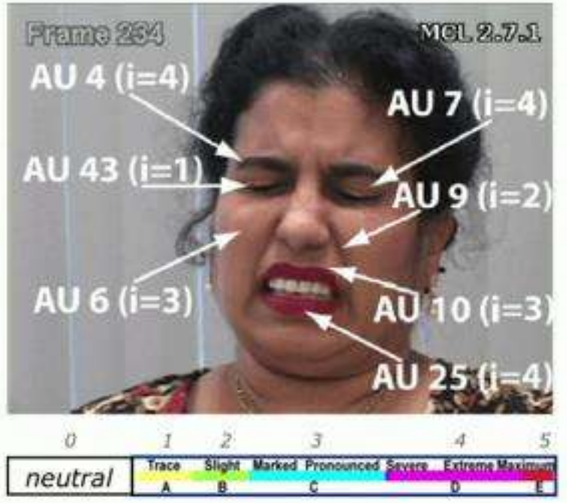
Two main approaches to encoding of facial expressions (Ekman, 2005):

Sign judgment (face grammar)

Facial Action Unit Activations (0/1)

More than 32 Action Units

Facial Action Unit Intensity (0-5)



Ekman, P. (2005). Facial Expressions. John Wiley & Sons, Ltd.

Modeling of Human Facial Behavior

Two main approaches to encoding of facial expressions (Ekman, 2005):

Sign judgment (face grammar)

Facial Action Unit Activations (0/1)

More than 32 Action Units

Direct mapping to discrete emotion categories (e.g., AU6+AU12 = happy)

Ekman, P. (2005). Facial Expressions. John Wiley & Sons, Ltd.

Modeling of Human Facial Behavior

Input

static images

multi-view images

image sequences (videos)

Machine Learning Computer Vision

multi-class classification

multi-view learning

structured-output regression

sequence classification

multi-task learning
(multi-class multi-output)

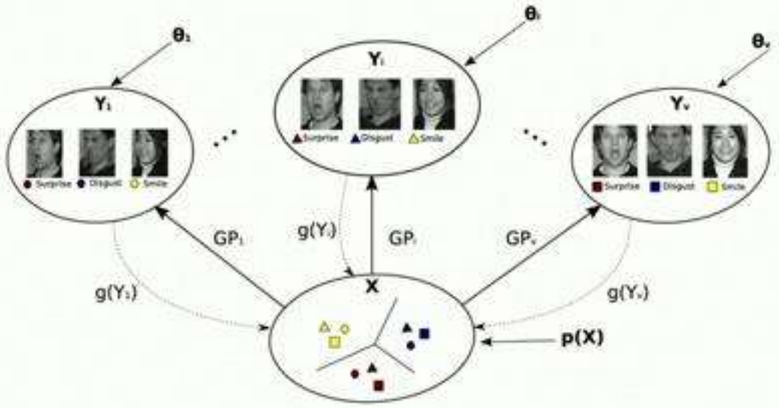
Output

Emotion category

FAU activations (0/1)

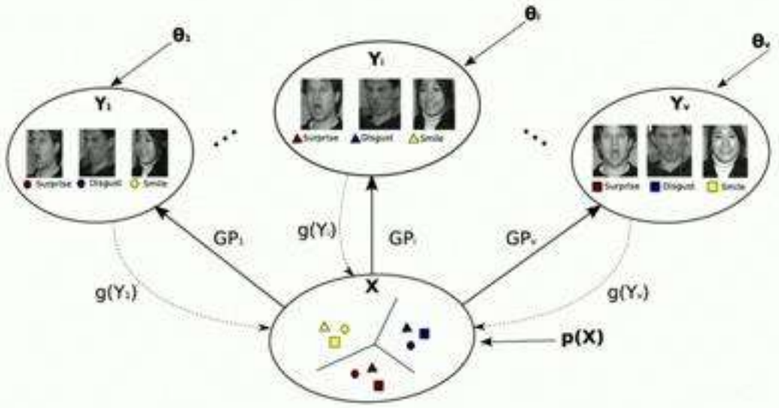
FAU intensities (0-5)

Modeling of Human Facial Behavior

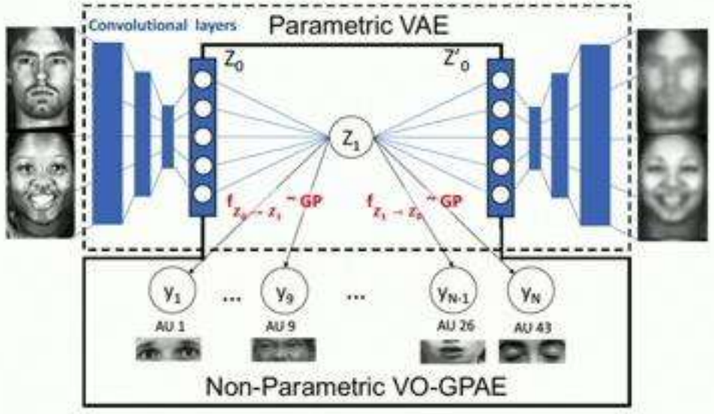


Discriminative Shared Gaussian Processes [TIP'15]

Modeling of Human Facial Behavior

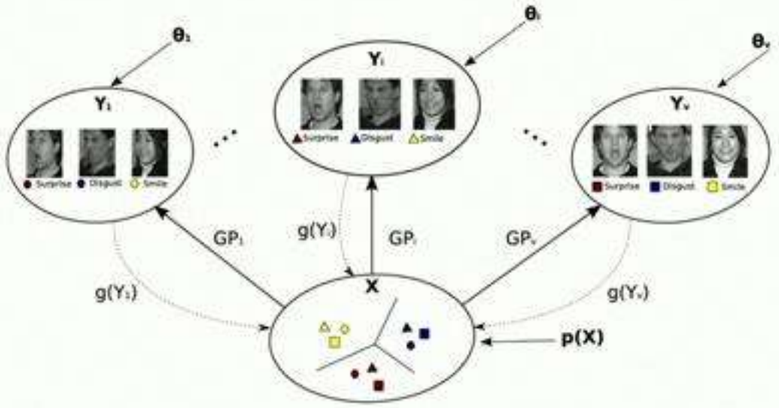


Discriminative Shared Gaussian Processes [TIP'15]

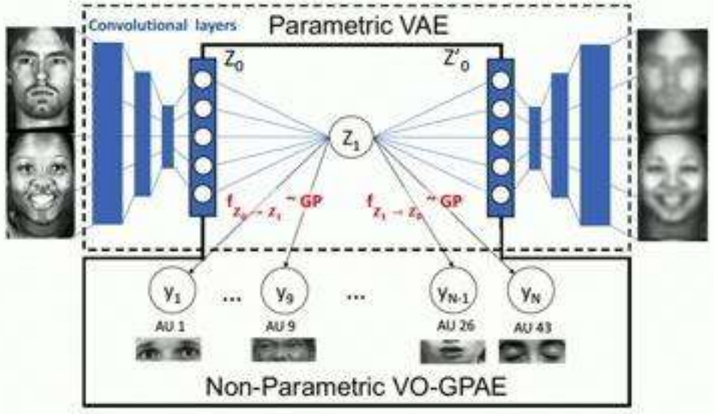


Semi-parametric Deep Learning [ICCV'17]

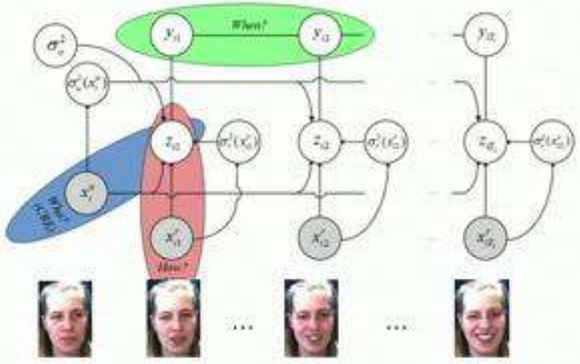
Modeling of Human Facial Behavior



Discriminative Shared Gaussian Processes [TIP'15]

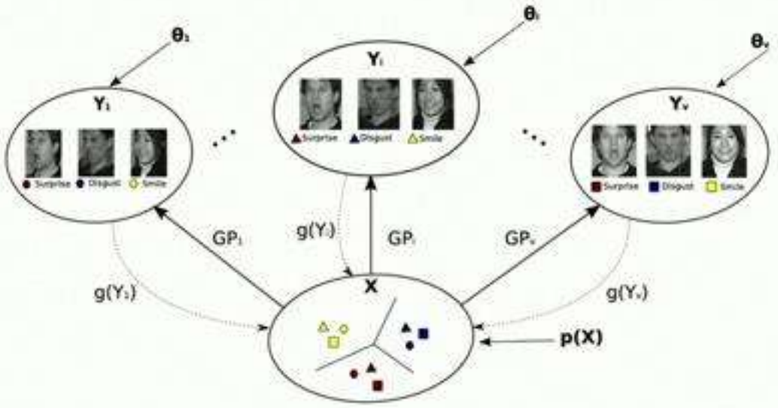


Semi-parametric Deep Learning [ICCV'17]

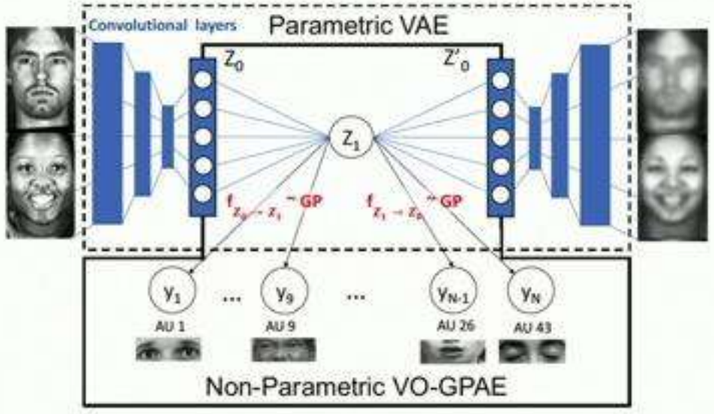


Context-sensitive Ordinal CRFs [TPAMI'14]

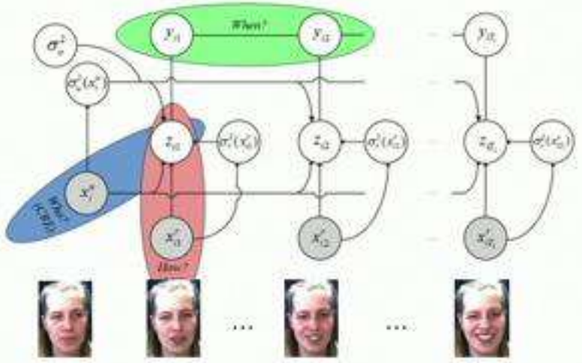
Modeling of Human Facial Behavior



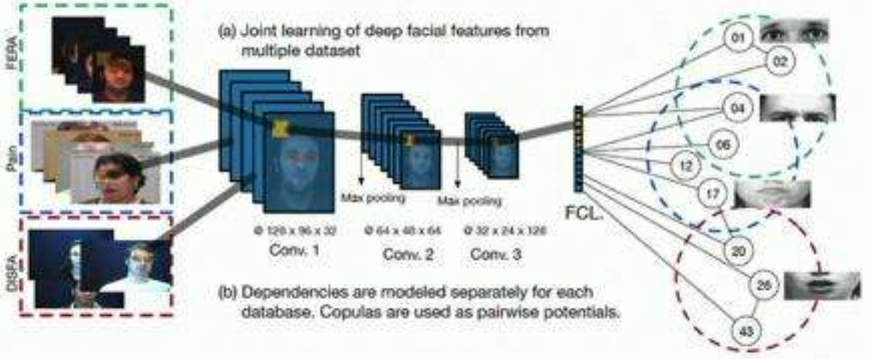
Discriminative Shared Gaussian Processes [TIP'15]



Semi-parametric Deep Learning [ICCV'17]

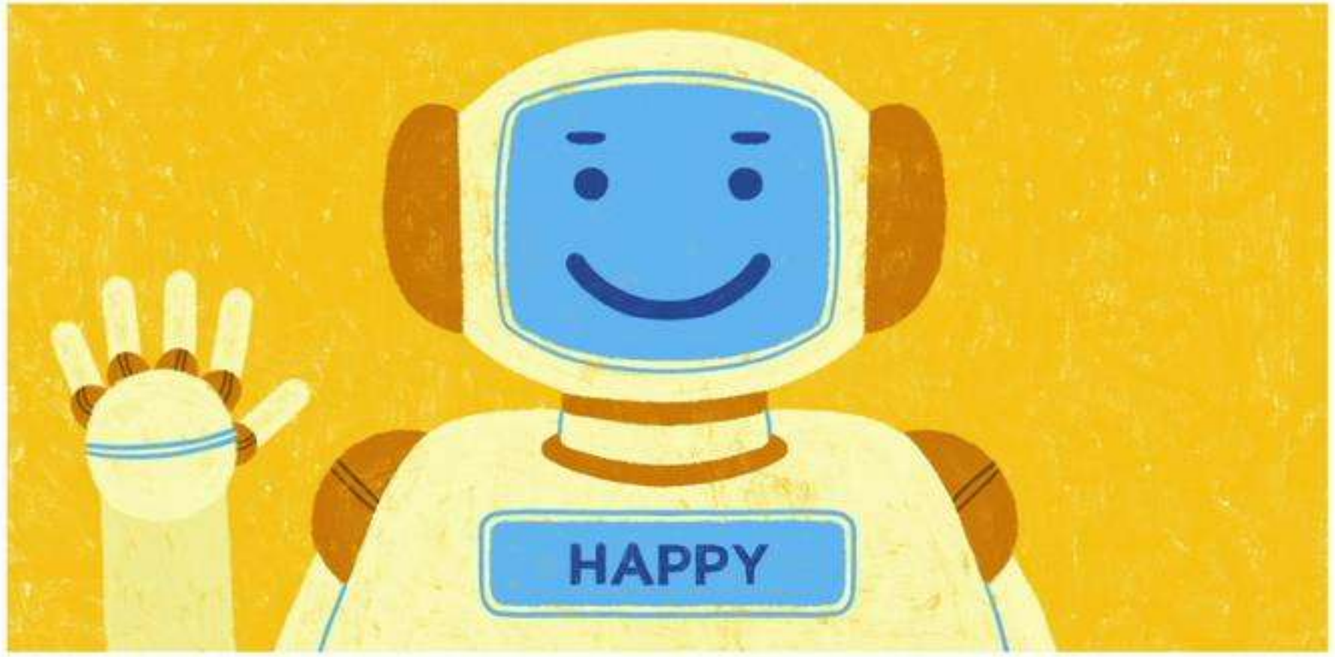


Context-sensitive Ordinal CRFs [TPAMI'14]



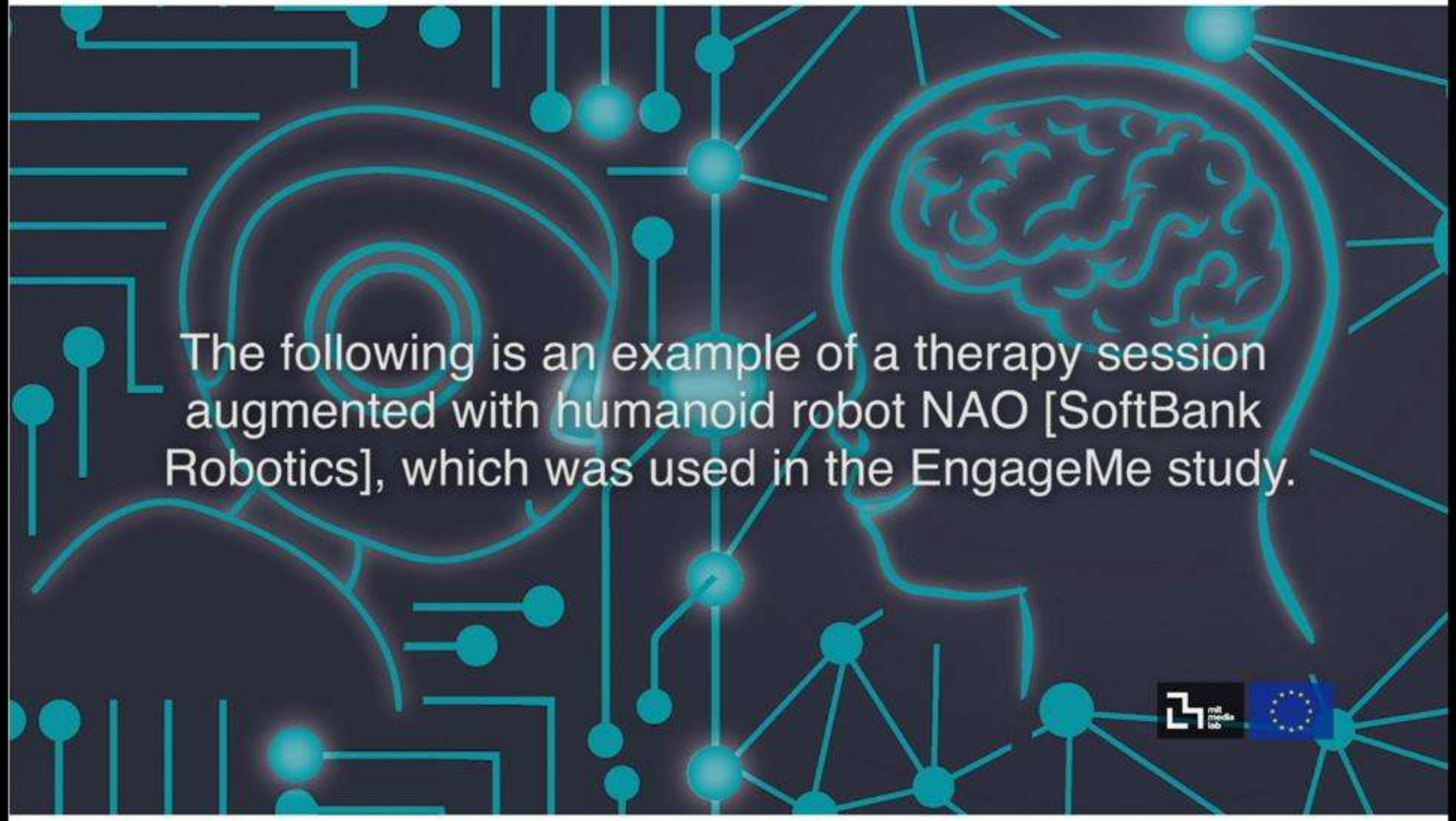
Deer Structured Learning [CVPR'17]

Affect and Engagement Estimation in Robot-Assisted Autism Therapy



O.Rudovic, J.Lee, M.Dai, B.Schuller, R.Picard (2018) "Personalized Machine Learning for Robot Perception of Affect and Engagement in Autism Therapy", Science Robotics

Example of a robot-assisted autism therapy



O.Rudovic, J.Lee, M.Dai, B.Schuller, R.Picard (2018) "Personalized Machine Learning for Robot Perception of Affect and Engagement in Autism Therapy", Science Robotics

Motivation

Robots provide a repetitive and predictable environment: *"This 'safe' environment can gently push a child with autism towards human interaction."*

One of the main challenges: How to effectively **elicit and maintain engagement** during interaction with robots?

Increased and sustained **engagement**: leads to improved learning opportunities!

Motivation

Augmenting the therapist with tools for monitoring the therapy progress

↓

Real-time indicators of child's affective states and engagement - consistent and unobtrusive measurements

↓

Personalized therapy content

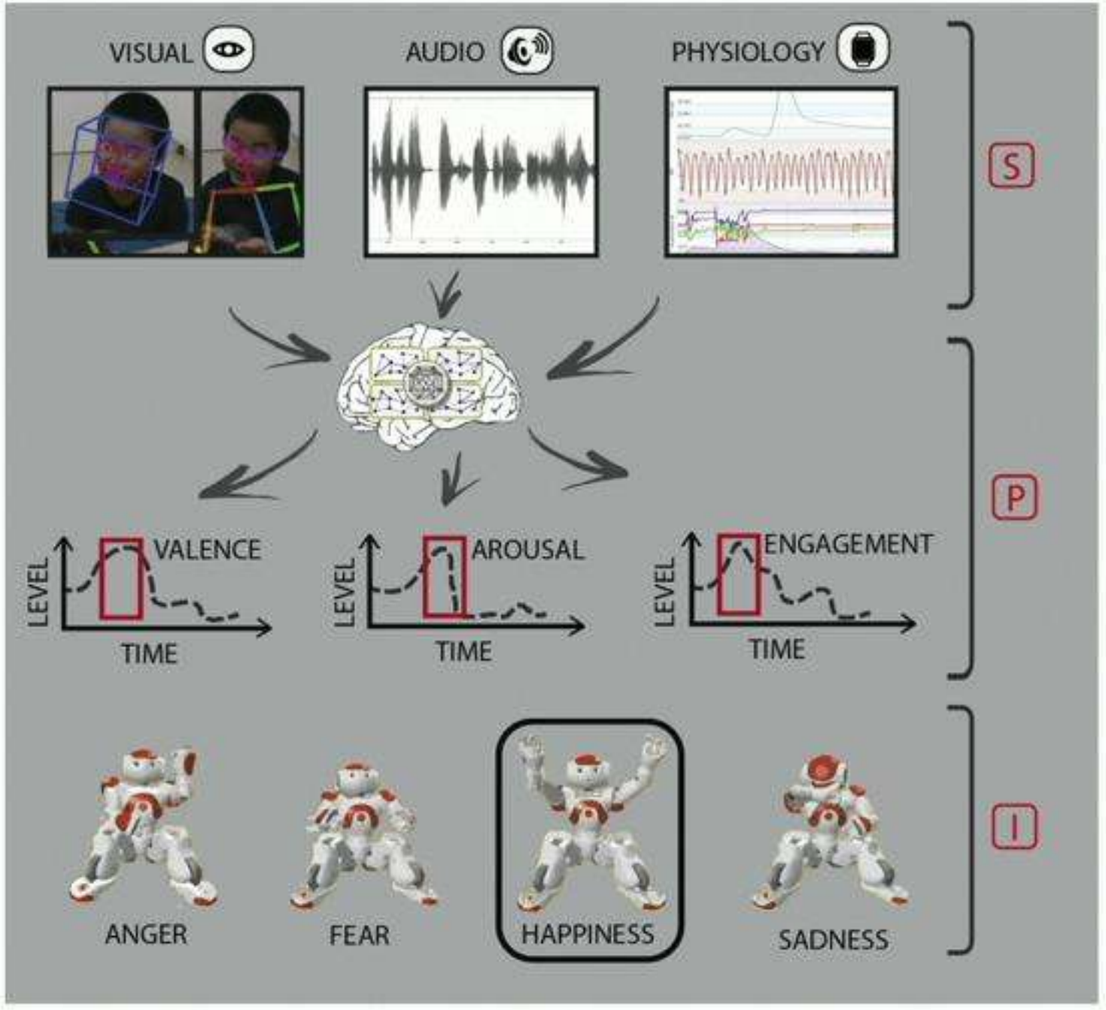
Problem Statement

Given multi-modal child-robot interaction videos of kids with autism, how do we make personalized estimation of their affect and engagement levels during the real-world autism therapy?

Dataset:

- 35 kids (17 Japan, 18 Serbia)
- age 3-13, ASC diagnosis
- 25 mins long therapy session
- continuous coding of valence, arousal and engagement in videos by five human experts

System Overview



Sensing

Perception

Interaction

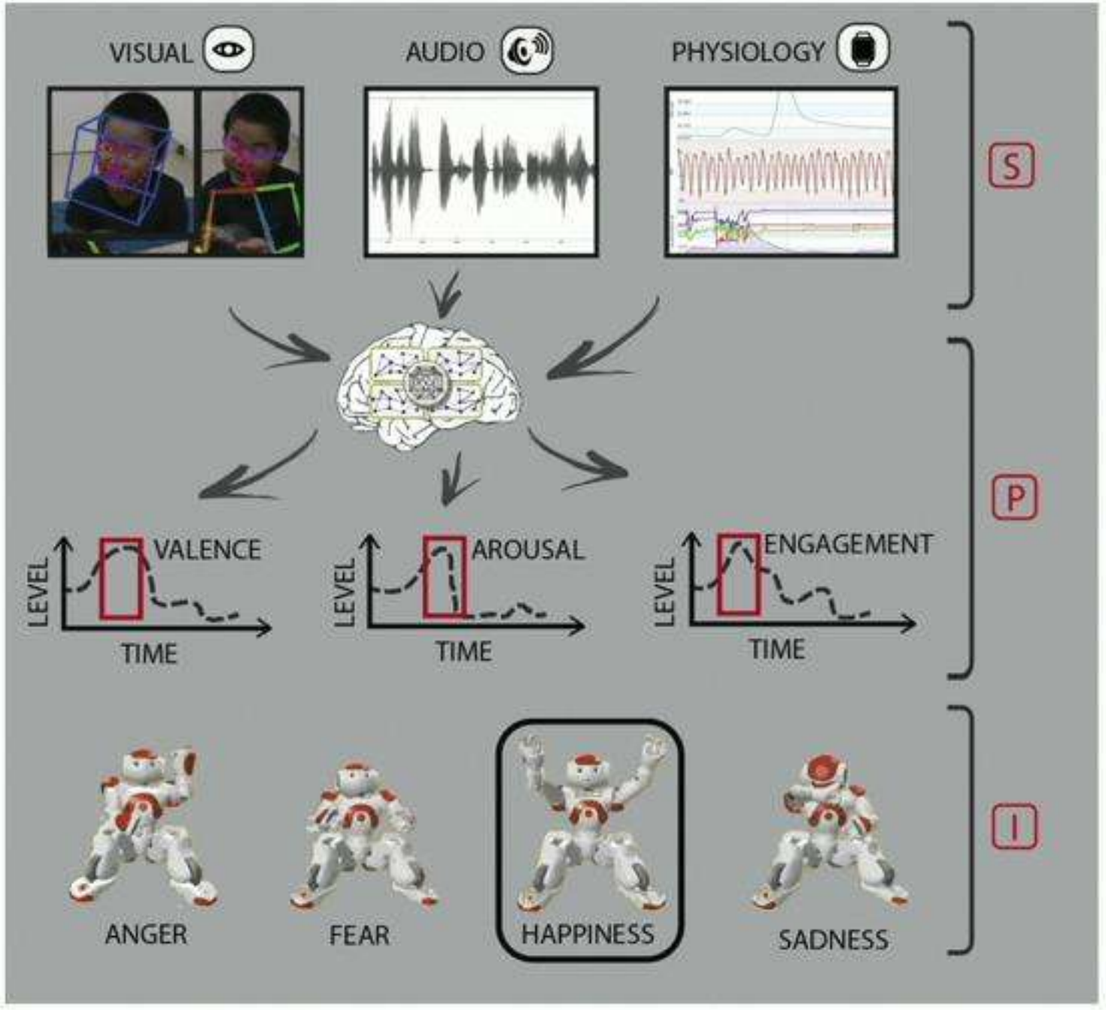
Problem Statement

Given multi-modal child-robot interaction videos of kids with autism, how do we make personalized estimation of their affect and engagement levels during the real-world autism therapy?

Dataset:

- 35 kids (17 Japan, 18 Serbia)
- age 3-13, ASC diagnosis
- 25 mins long therapy session
- continuous coding of valence, arousal and engagement in videos by five human experts

System Overview

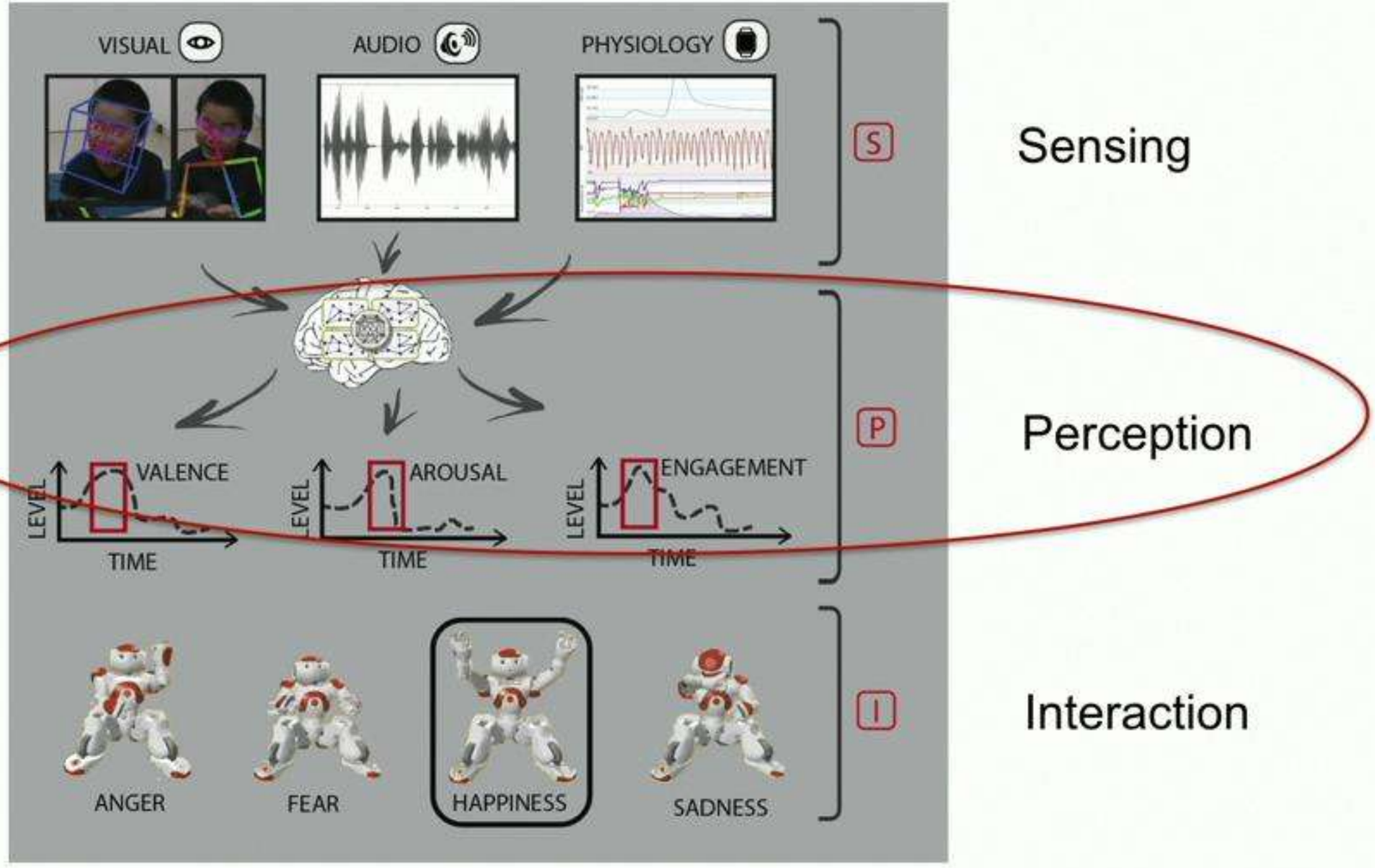


Sensing

Perception

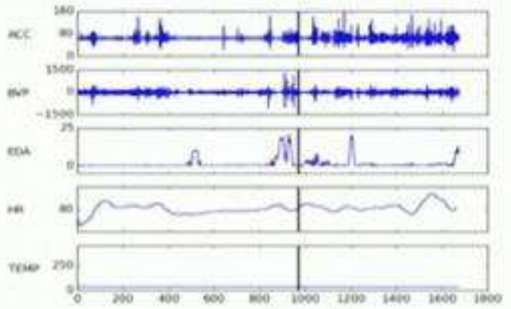
Interaction

System Overview

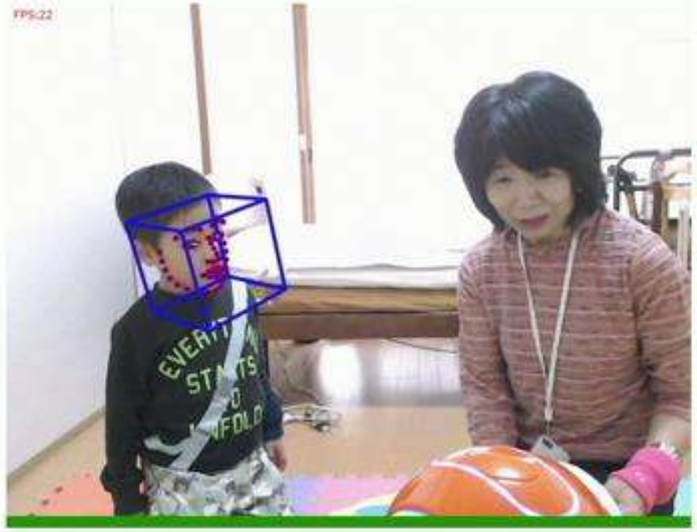


Multi-modal Data Sensing

E4



E4 Autonomic Physiology



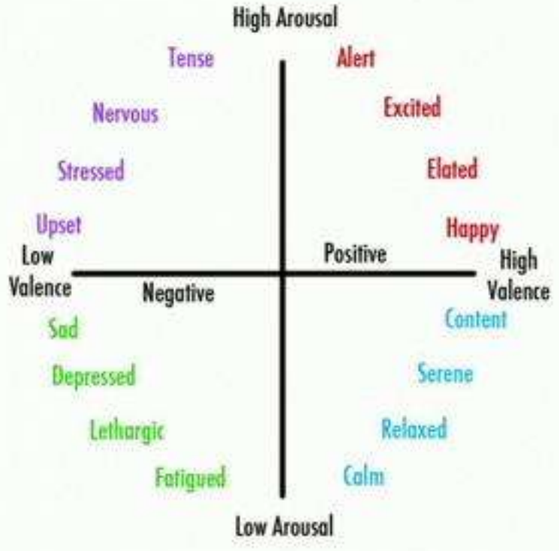
OpenFace (Baltrušaitis et al., 2016)

OpenPose (Cao et al., 2017)

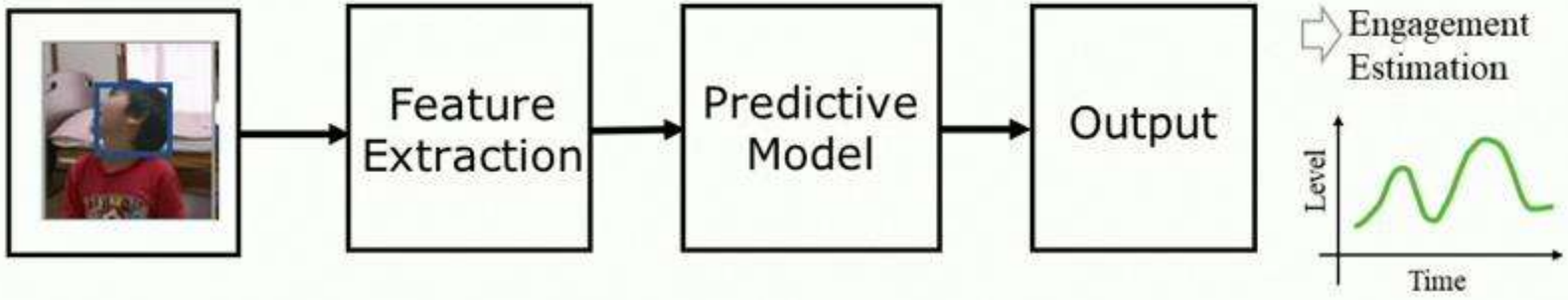
& LLDs (low-level audio descriptors) extracted using OpenSmile (Eyben et al., 2013)

Data Coding

- Coding of valence (pleasure-displeasure), arousal (alertness), and engagement in the task, on a continuous scale from -1 to +1
- 5 human experts rated the videos by inspecting the audio-visual recordings
- Combined annotations used as the ground truth for machine learning



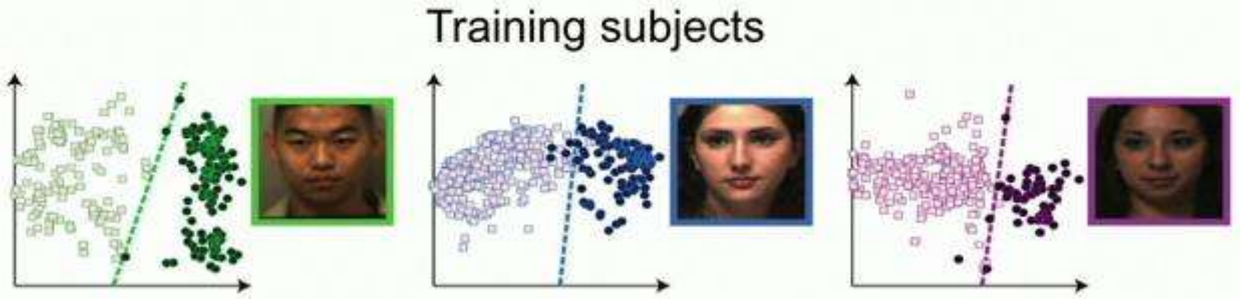
Traditional Machine Learning from Human Data ("one-size-fits-all")

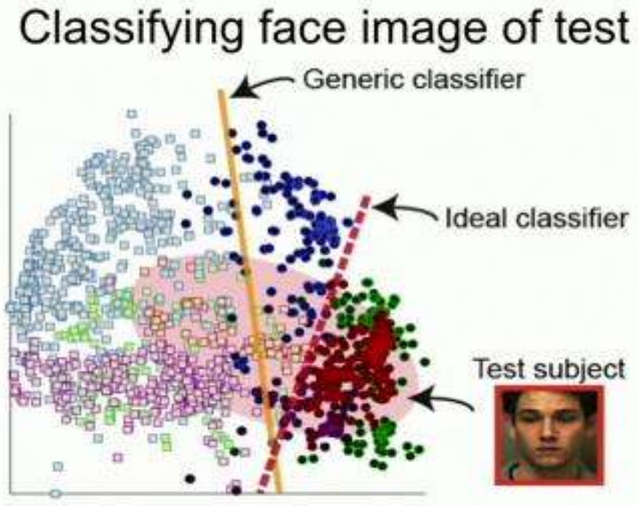


maximize average performance

Everybody is different

Example: Classification of Expressive vs. Neutral Faces

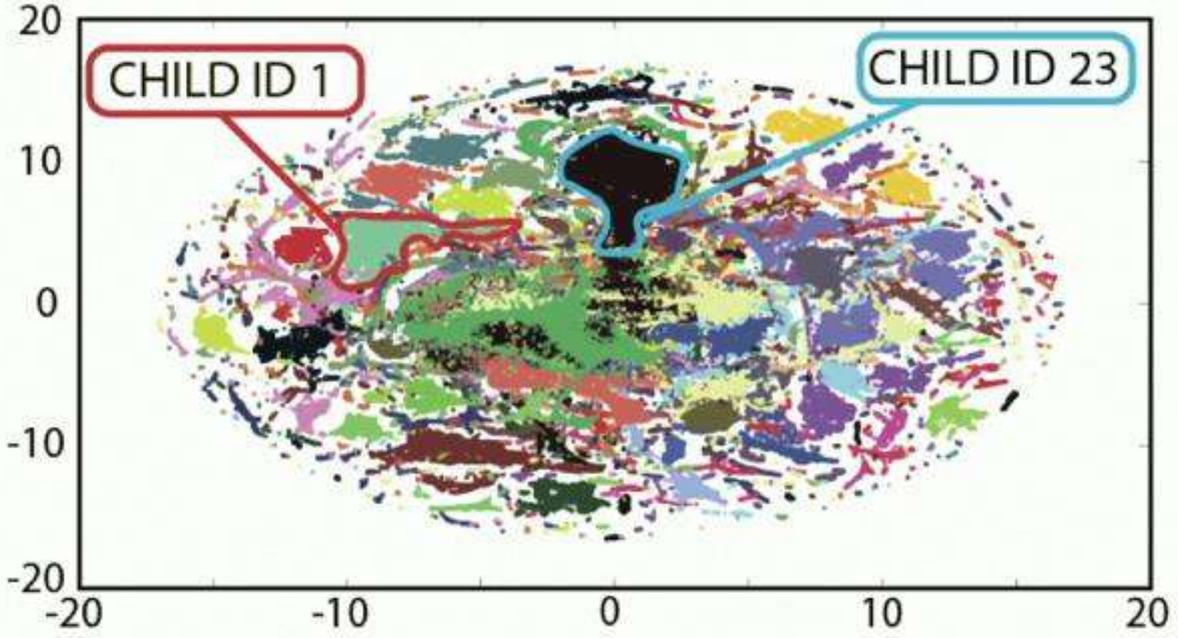




Chu, W. S., De la Torre, F., Cohn, J. F. (2016) "Selective transfer machine for personalized facial expression analysis" TPAMI

Everybody is different

Example: Clustering data representations of children with autism



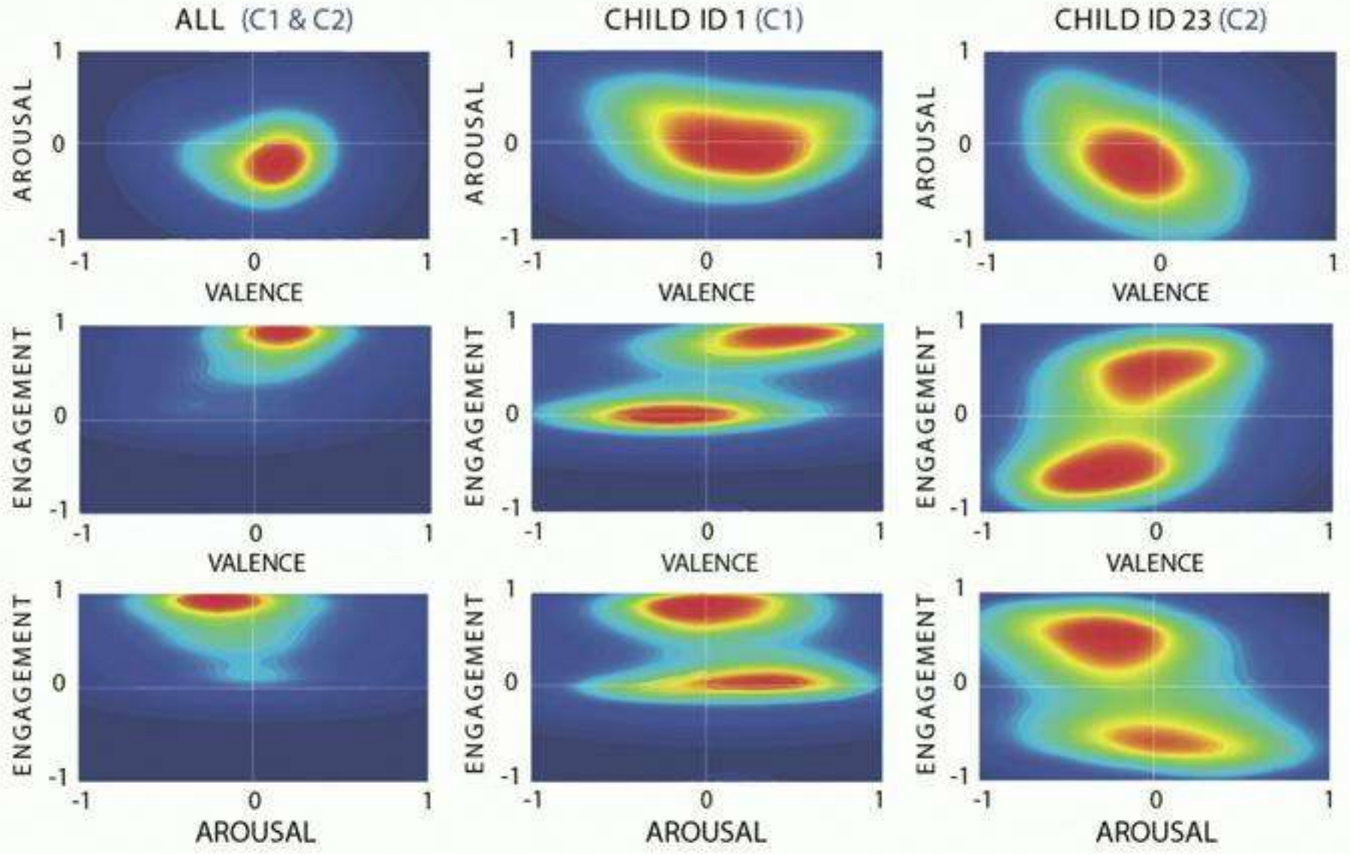
Clustering of multi-modal high-dimensional data of children with autism using the t-SNE, an unsupervised dimensionality reduction technique

L.J.P. van der Maaten and G.E. Hinton (2008) Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research

O.Rudovic, J.Lee, M.Dai, B.Schuller, R.Picard (2018) "Personalized Machine Learning for Robot Perception of Affect and Engagement in Autism Therapy", Science Robotics

Everybody is different

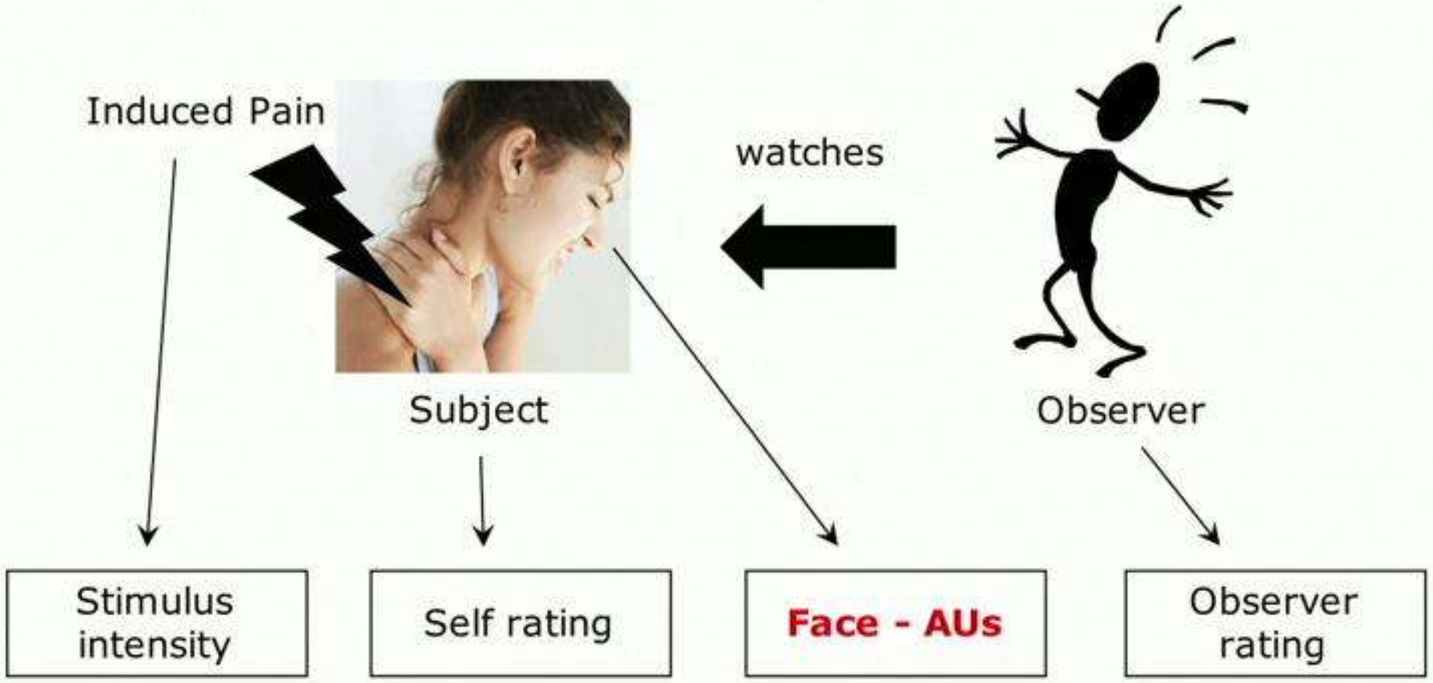
Example: Dependencies of affect and engagement ratings of data of children with autism



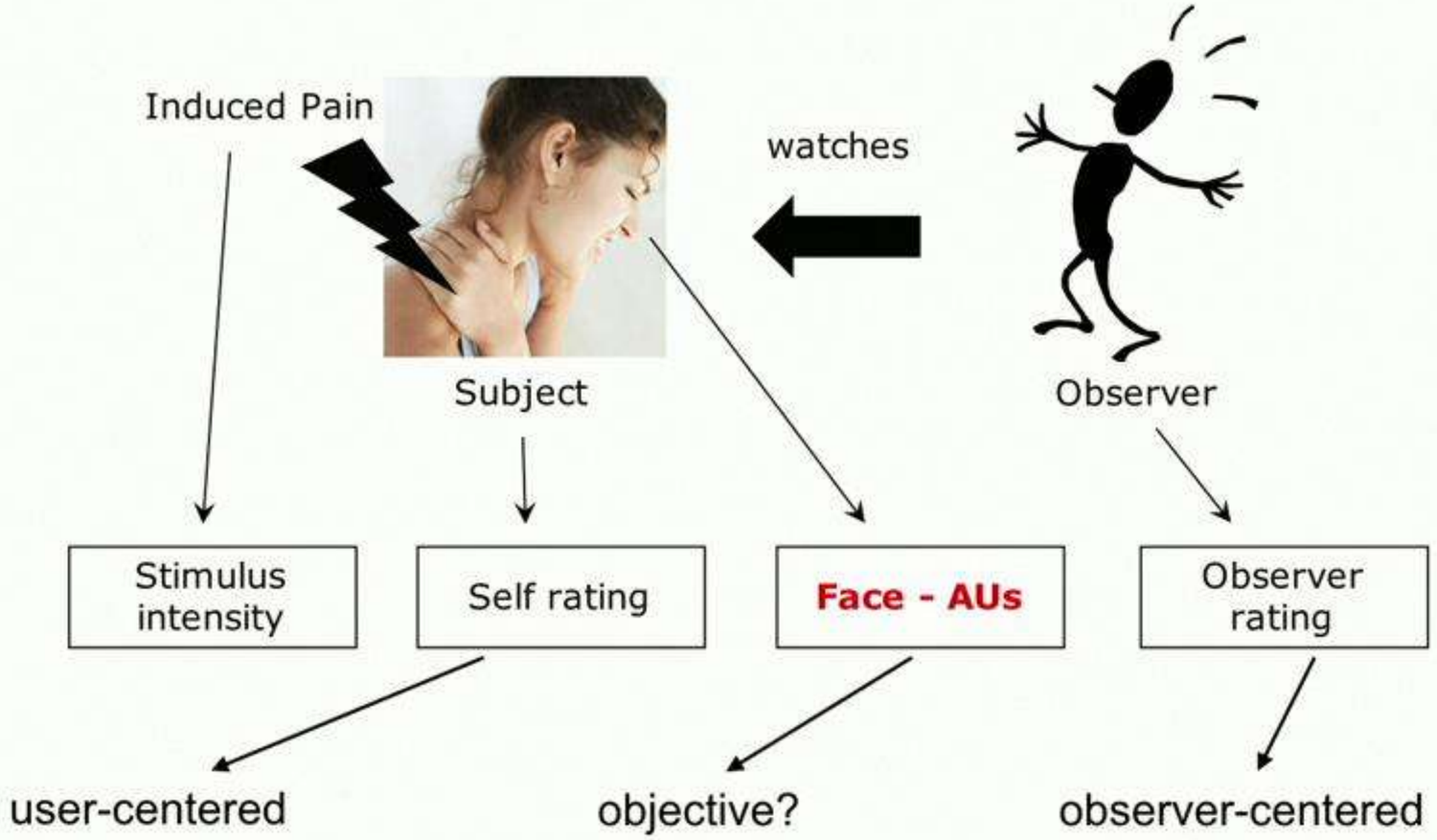
C1 - Japanese children
C2 - European children

O.Rudovic, J.Lee, M.Dai, B.Schuller, R.Picard (2018) "Personalized Machine Learning for Robot Perception of Affect and Engagement in Autism Therapy", Science Robotics

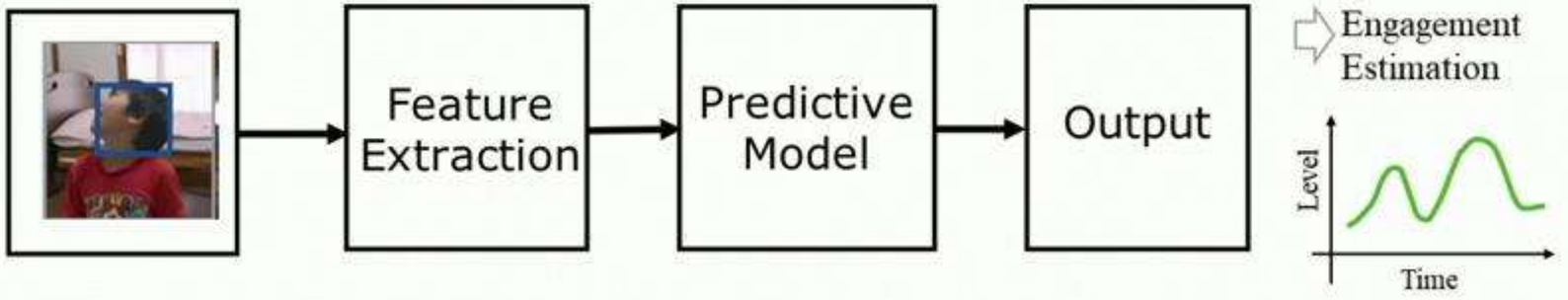
Everybody feels differently



Everybody feels differently

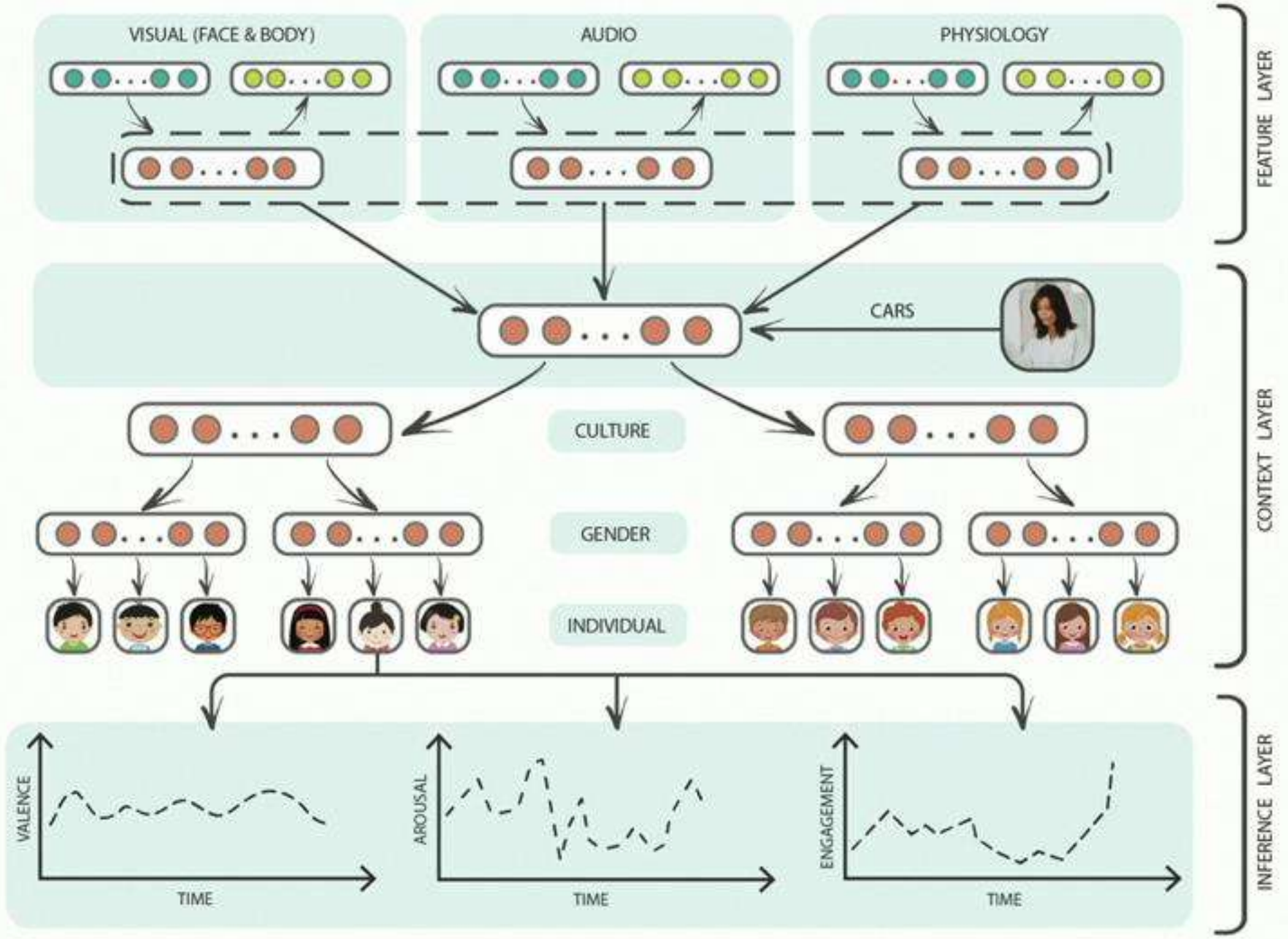


Personalized Machine Learning from Human Data



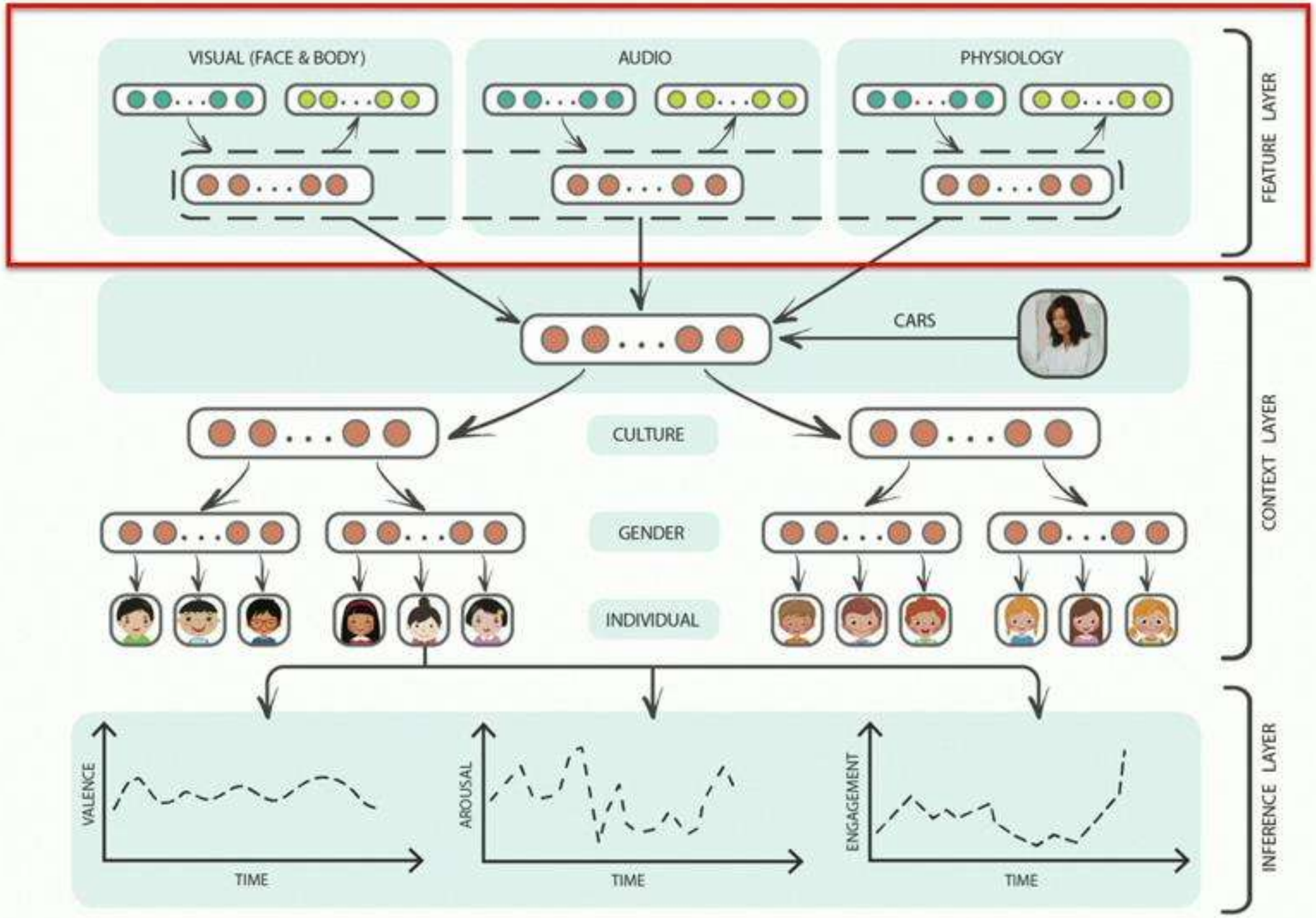
maximize individual performance

Personalized Perception of Affect Network (PPA-net)



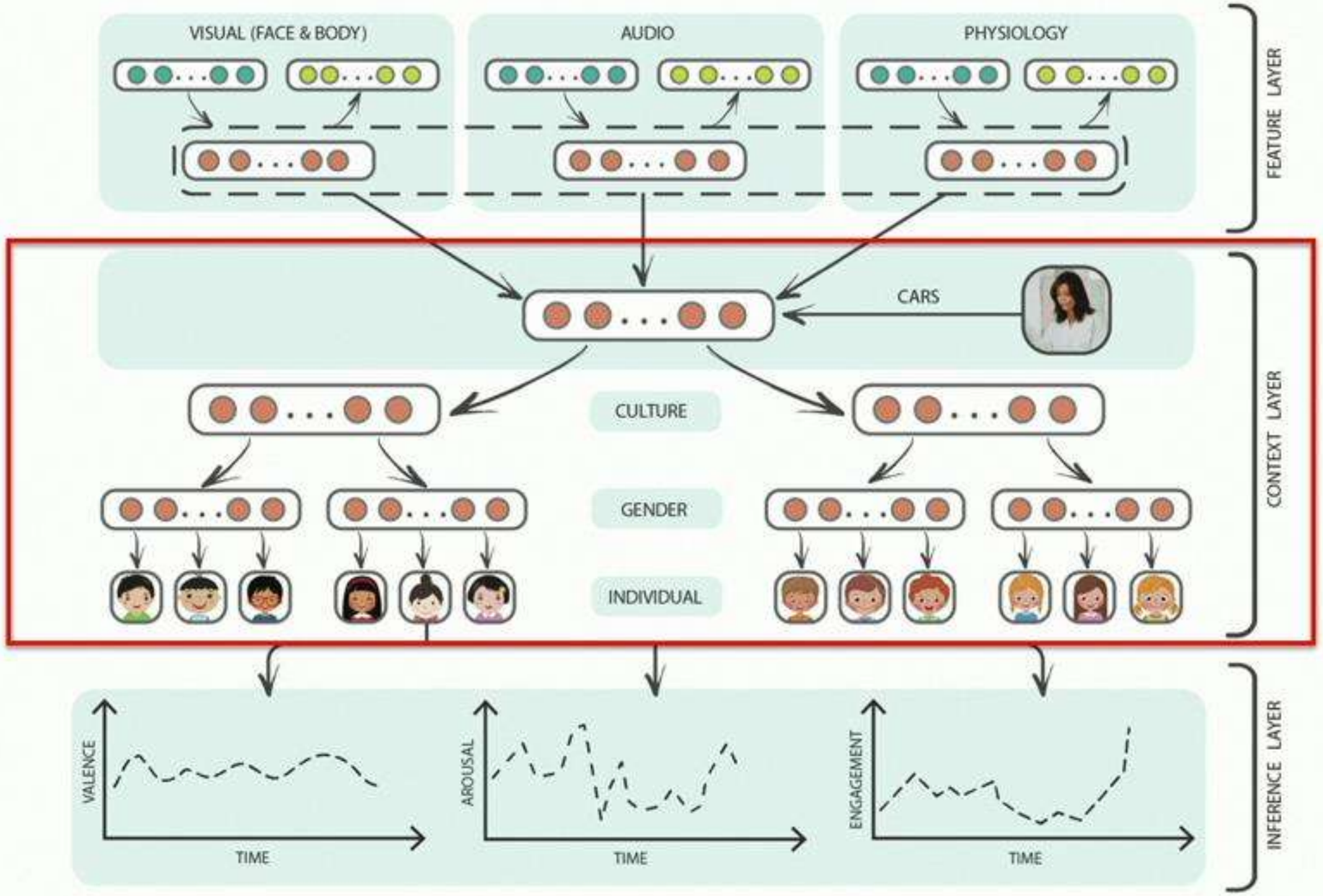
CARS (**C**hildhood **A**utism **R**ating **S**cale)

Personalized Perception of Affect Network (PPA-net)



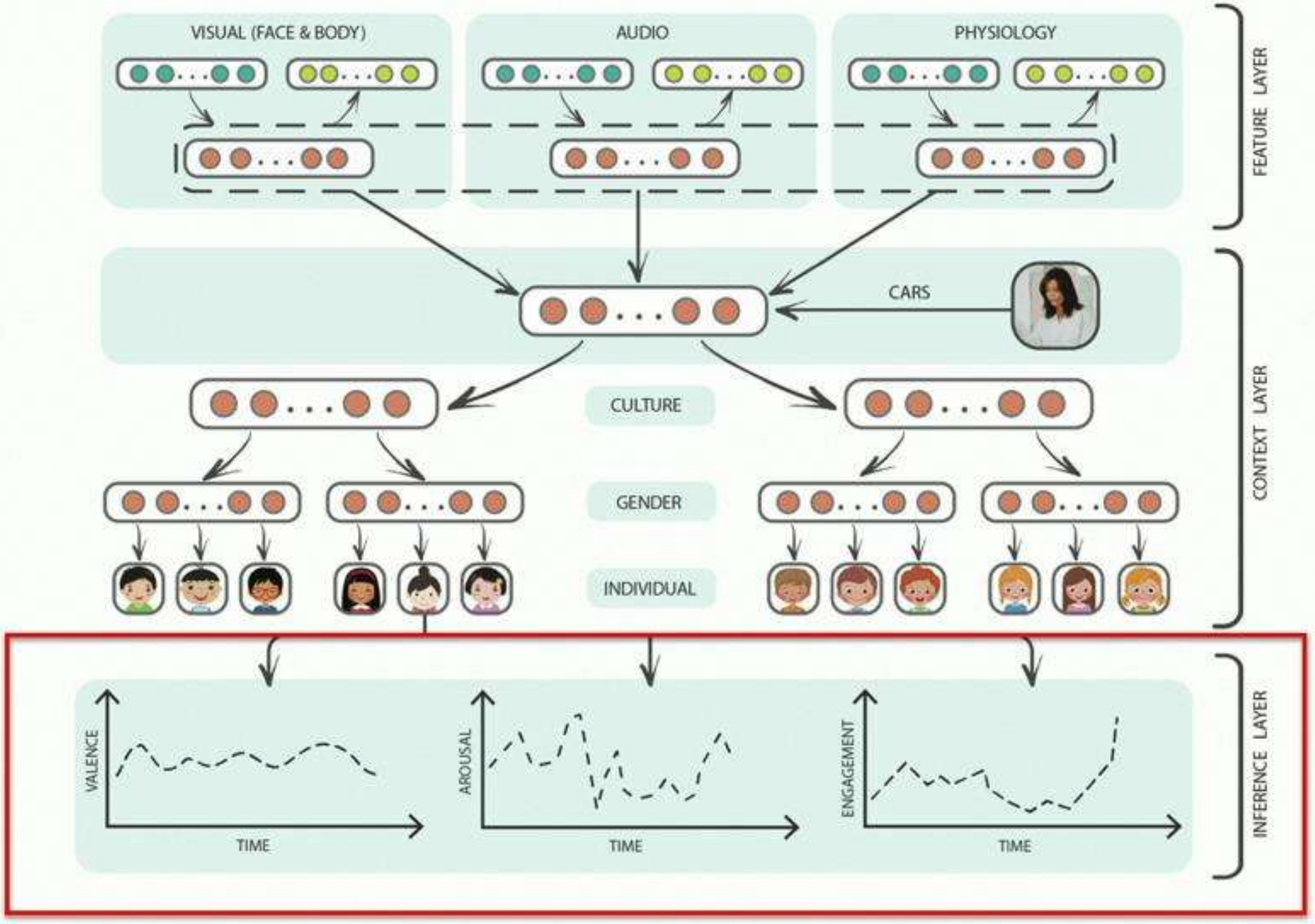
CARS (**C**hildhood **A**utism **R**ating **S**cale)

Personalized Perception of Affect Network (PPA-net)



CARS (**C**hildhood **A**utism **R**ating **S**cale)

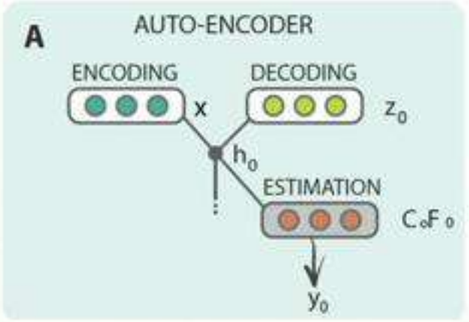
Personalized Perception of Affect Network (PPA-net)



CARS (**C**hildhood **A**utism **R**ating **S**cale)

PPA-net: Learning and Inference

Learning operators:



(A) Auto-encoding of multi-modal inputs:

$$h_0 = f_{\theta_0^{(e)}}(x) = W_0^{(e)}x + b_0^{(e)}, \quad \theta_0^{(e)} = \left\{W_0^{(e)}, b_0^{(e)}\right\}$$

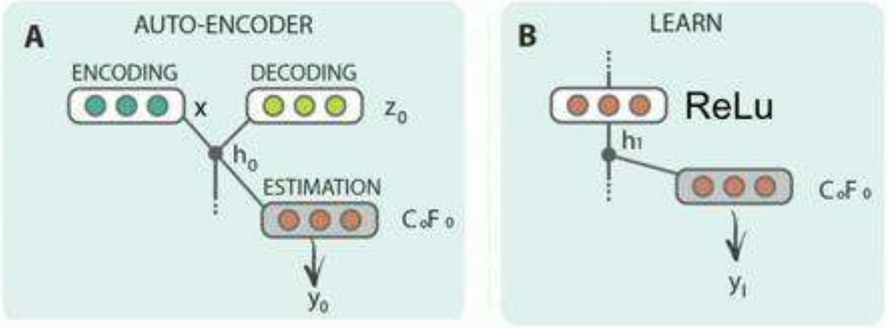
$$z_0 = f_{\theta_0^{(d)}}(h_0) = W_0^{(d)}h_0 + b_0^{(d)}, \quad \theta_0^{(d)} = \left\{W_0^{(d)}, b_0^{(d)}\right\}$$

LaF

MSE - Mean Squared Error, LaF - Linear Activation Function, ReLu - Rectified Linear Unit

PPA-net: Learning and Inference

Learning operators:



(A) Auto-encoding of multi-modal inputs:

$$h_0 = f_{\theta_0^{(e)}}(x) = W_0^{(e)} x + b_0^{(e)}, \quad \theta_0^{(e)} = \left\{ W_0^{(e)}, b_0^{(e)} \right\}$$

$$z_0 = f_{\theta_0^{(d)}}(h_0) = W_0^{(d)} h_0 + b_0^{(d)}, \quad \theta_0^{(d)} = \left\{ W_0^{(d)}, b_0^{(d)} \right\}$$

LaF

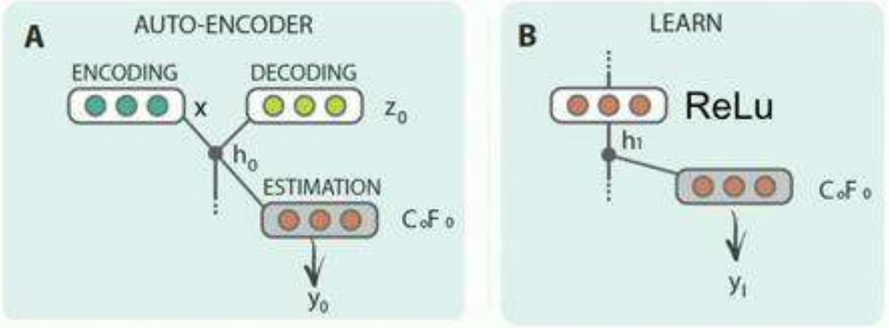
(B) CoFo -companion objective function:

$$y_0 = f_{\theta_0^{(c)}}(h_0) = W_0^{(c)} h_0 + b_0^{(c)}, \quad \theta_0^{(c)} = \left\{ W_0^{(c)}, b_0^{(c)} \right\}$$

MSE - Mean Squared Error, LaF - Linear Activation Function, ReLu - Rectified Linear Unit

PPA-net: Learning and Inference

Learning operators:



$$\omega_0^* = \underset{\omega_0}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left((1-\lambda)\alpha_d \underbrace{\left(x^{(i)}, z_0^{(i)} \right)}_{\text{MSE}} + \lambda \alpha_c \underbrace{\left(y_0^{(i)}, y^{(i)} \right)}_{\text{MSE}} \right)$$

(A) Auto-encoding of multi-modal inputs:

$$h_0 = f_{\theta_0^{(e)}}(x) = W_0^{(e)} x + b_0^{(e)}, \quad \theta_0^{(e)} = \left\{ W_0^{(e)}, b_0^{(e)} \right\}$$

$$z_0 = f_{\theta_0^{(d)}}(h_0) = W_0^{(d)} h_0 + b_0^{(d)}, \quad \theta_0^{(d)} = \left\{ W_0^{(d)}, b_0^{(d)} \right\}$$

LaF

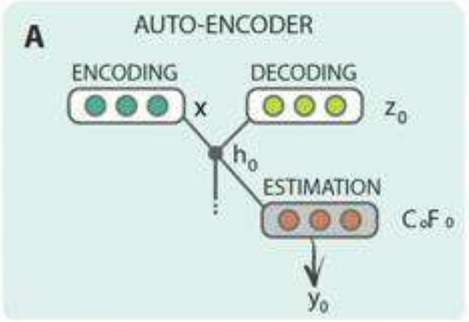
(B) CoFo -companion objective function:

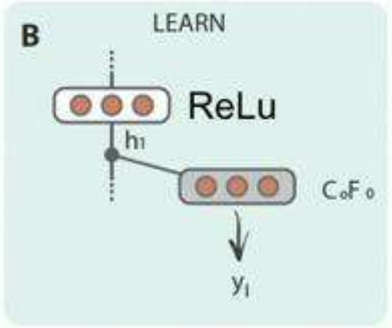
$$y_0 = f_{\theta_0^{(c)}}(h_0) = W_0^{(c)} h_0 + b_0^{(c)}, \quad \theta_0^{(c)} = \left\{ W_0^{(c)}, b_0^{(c)} \right\}$$

MSE - Mean Squared Error, LaF - Linear Activation Function, ReLu - Rectified Linear Unit

PPA-net: Learning and Inference

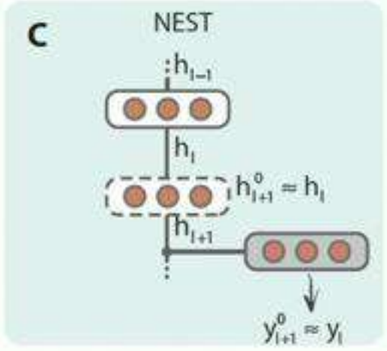
Learning operators:





$$\omega_0^* = \underset{\omega_0}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left((1-\lambda)\alpha_d \left(x^{(i)}, z_0^{(i)} \right) + \lambda \alpha_c \left(y_0^{(i)}, y^{(i)} \right) \right)$$

MSE MSE



(A) Auto-encoding of multi-modal inputs:

$$h_0 = f_{\theta_0^{(e)}}(x) = W_0^{(e)} x + b_0^{(e)}, \quad \theta_0^{(e)} = \left\{ W_0^{(e)}, b_0^{(e)} \right\}$$

$$z_0 = f_{\theta_0^{(d)}}(h_0) = W_0^{(d)} h_0 + b_0^{(d)}, \quad \theta_0^{(d)} = \left\{ W_0^{(d)}, b_0^{(d)} \right\}$$

LaF

(B) CoFo -companion objective function:

$$y_0 = f_{\theta_0^{(c)}}(h_0) = W_0^{(c)} h_0 + b_0^{(c)}, \quad \theta_0^{(c)} = \left\{ W_0^{(c)}, b_0^{(c)} \right\}$$

(C) Nesting a new layer:

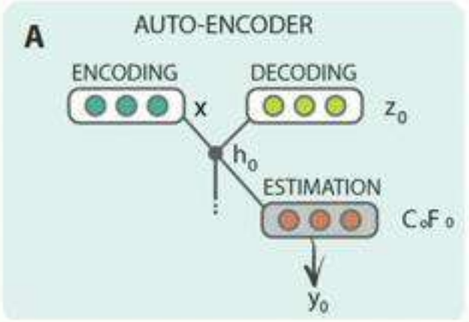
$$\theta_{l+1} = \left\{ W_{l+1} \leftarrow I + \epsilon, \quad b_{l+1} \leftarrow 0 \right\}$$

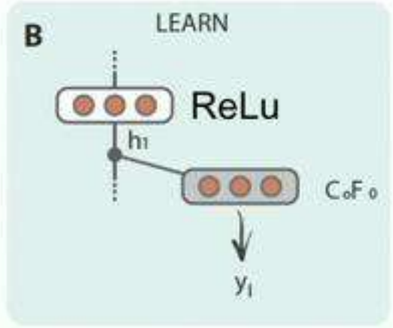
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

MSE - Mean Squared Error, LaF - Linear Activation Function, ReLu - Rectified Linear Unit

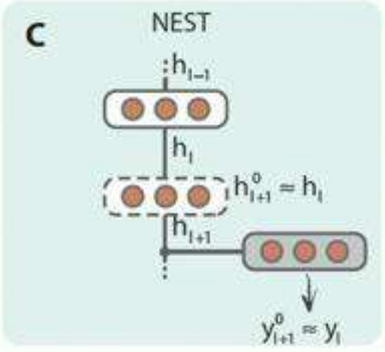
PPA-net: Learning and Inference

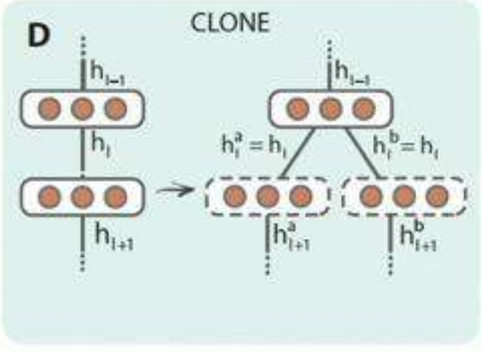
Learning operators:





$$\omega_0^* = \underset{\omega_0}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left((1-\lambda)\alpha_d \underbrace{\left(x^{(i)}, z_0^{(i)} \right)}_{\text{MSE}} + \lambda \alpha_c \underbrace{\left(y_0^{(i)}, y^{(i)} \right)}_{\text{MSE}} \right)$$





(A) Auto-encoding of multi-modal inputs:

$$h_0 = f_{\theta_0^{(e)}}(x) = W_0^{(e)} x + b_0^{(e)}, \quad \theta_0^{(e)} = \left\{ W_0^{(e)}, b_0^{(e)} \right\}$$

$$z_0 = f_{\theta_0^{(d)}}(h_0) = W_0^{(d)} h_0 + b_0^{(d)}, \quad \theta_0^{(d)} = \left\{ W_0^{(d)}, b_0^{(d)} \right\}$$

LaF

(B) CoFo -companion objective function:

$$y_0 = f_{\theta_0^{(c)}}(h_0) = W_0^{(c)} h_0 + b_0^{(c)}, \quad \theta_0^{(c)} = \left\{ W_0^{(c)}, b_0^{(c)} \right\}$$

(C) Nesting a new layer:

$$\theta_{l+1} = \left\{ W_{l+1} \leftarrow I + \epsilon, \quad b_{l+1} \leftarrow 0 \right\}$$

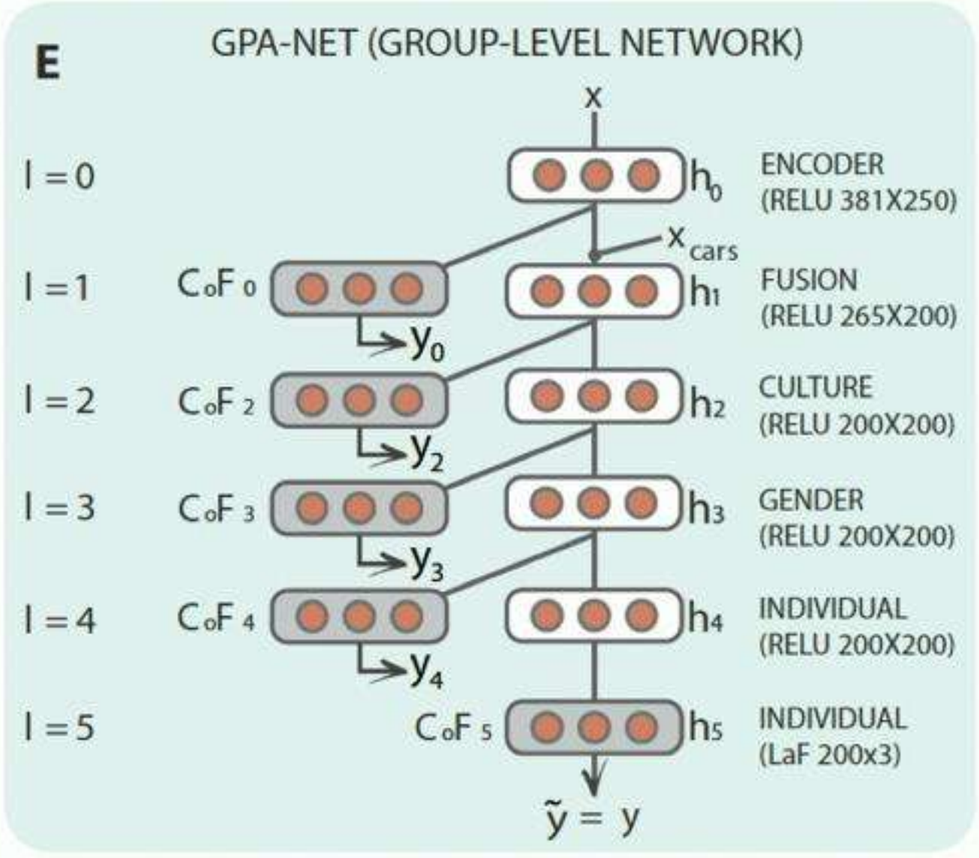
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

(D) Clone the nested layer :

$$\theta_{l+1}^{(c)} = \left\{ W_{l+1}^{(c)} \leftarrow W_l^{(c)}, \quad b_{l+1}^{(c)} \leftarrow b_l^{(c)} \right\}$$

MSE - Mean Squared Error, LaF - Linear Activation Function, ReLu - Rectified Linear Unit

PPA-net: Learning and Inference



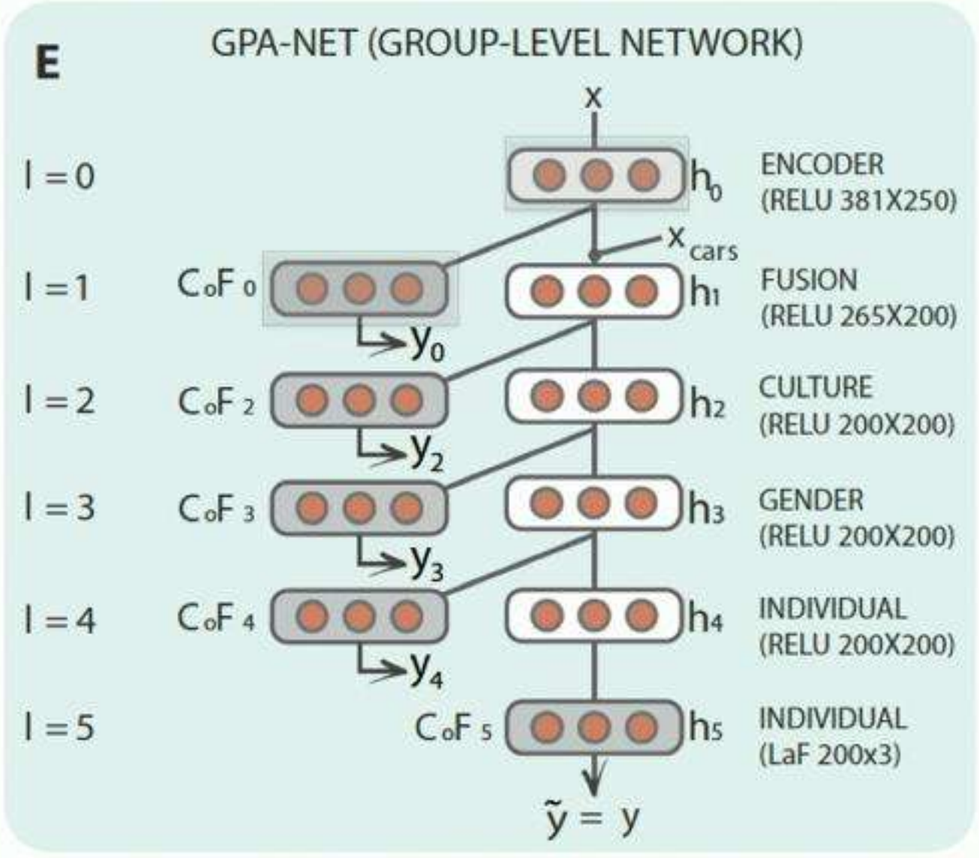
$$\omega_l^* = \underset{\omega_l}{\operatorname{argmin}}\, \alpha_c(h_l, y) = \underset{\omega_l}{\operatorname{argmin}}\, \frac{1}{N}\sum_{i=1}^{N} \underbrace{\alpha_c\left(y_{l+1}^{(i)}, y^{(i)}\right)}_{\text{MSE}}$$

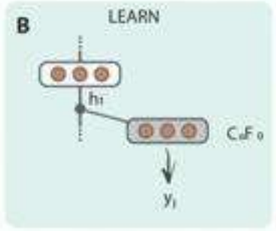
predicted true

Step 1: Train layer-wise
Step 2: Fine-tune the whole net
(and only the last CoFo)

Adadelta optimizer

PPA-net: Learning and Inference





$$\omega_l^* = \underset{\omega_l}{\operatorname{argmin}}\, \alpha_c(h_l, y) = \underset{\omega_l}{\operatorname{argmin}}\, \frac{1}{N}\sum_{i=1}^{N} \alpha_c\left(y_{l+1}^{(i)}, y^{(i)}\right)$$

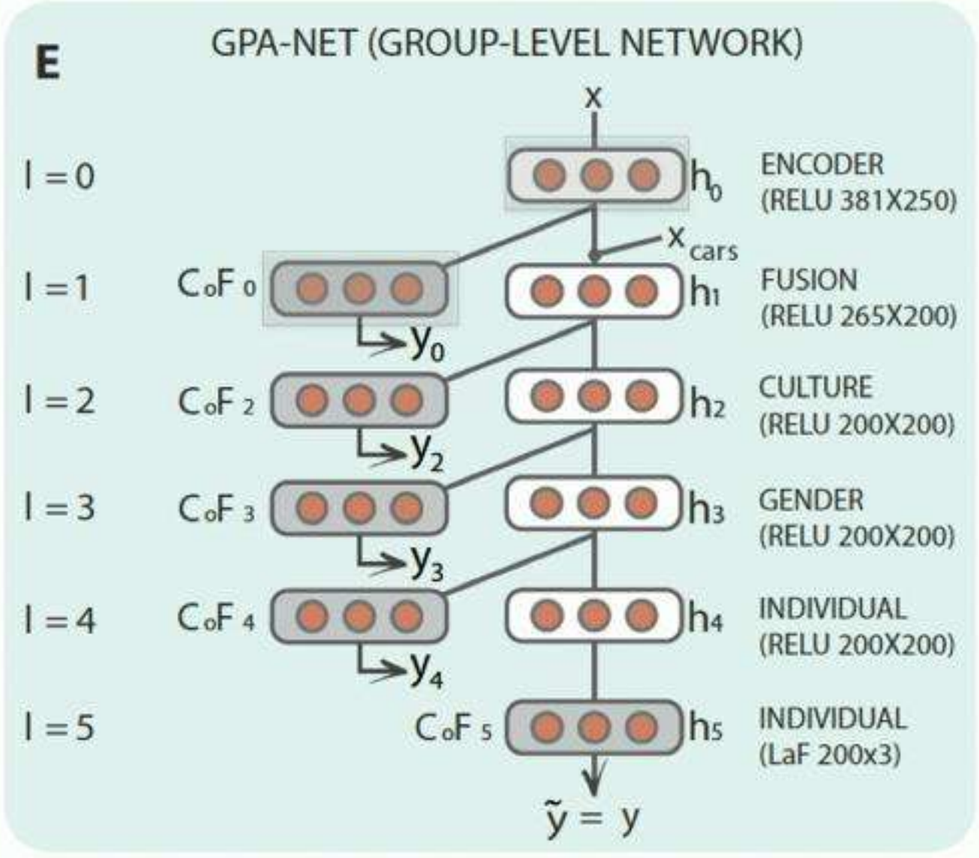
predicted true

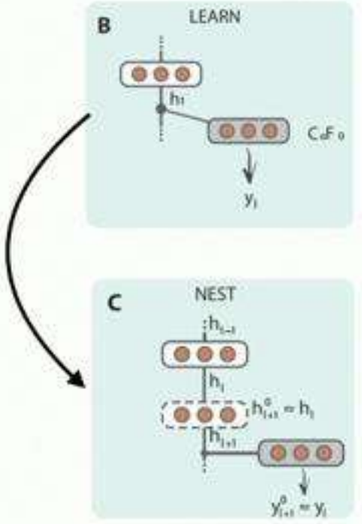
MSE

Step 1: Train layer-wise
Step 2: Fine-tune the whole net
(and only the last CoFo)

Adadelta optimizer

PPA-net: Learning and Inference





$$\omega_l^* = \underset{\omega_l}{\operatorname{argmin}}\, \alpha_c(h_l, y) = \underset{\omega_l}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \alpha_c \left(y_{l+1}^{(i)}, y^{(i)} \right)$$

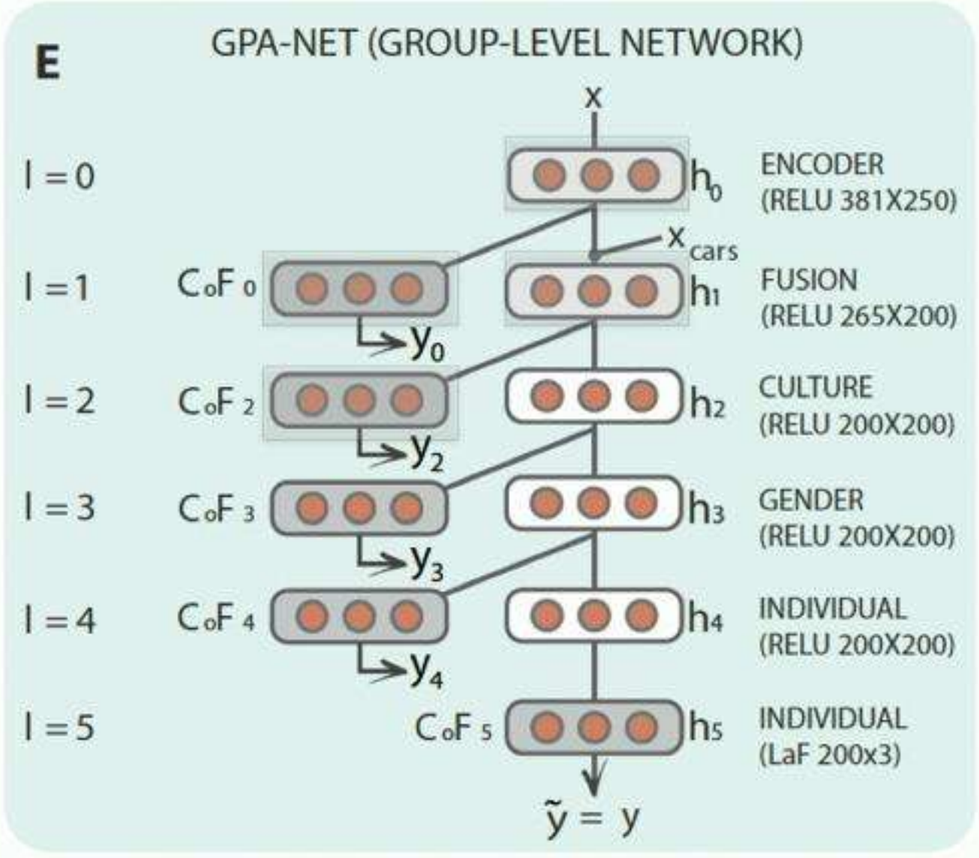
predicted true

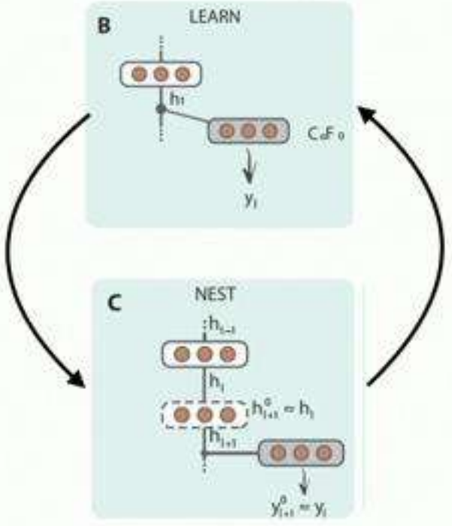
MSE

Step 1: Train layer-wise
Step 2: Fine-tune the whole net
(and only the last CoFo)

Adadelta optimizer

PPA-net: Learning and Inference





$$\omega_l^* = \underset{\omega_l}{\operatorname{argmin}}\, \alpha_c(h_l, y) = \underset{\omega_l}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \alpha_c \left(y_{l+1}^{(i)}, y^{(i)} \right)$$

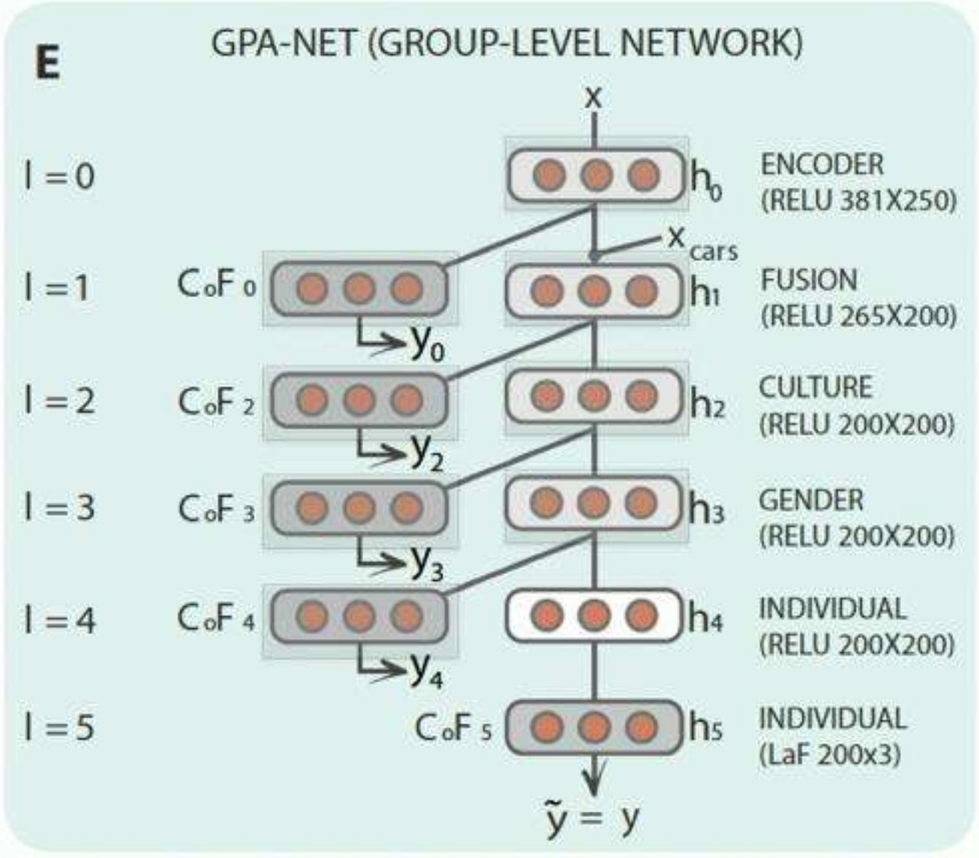
predicted true

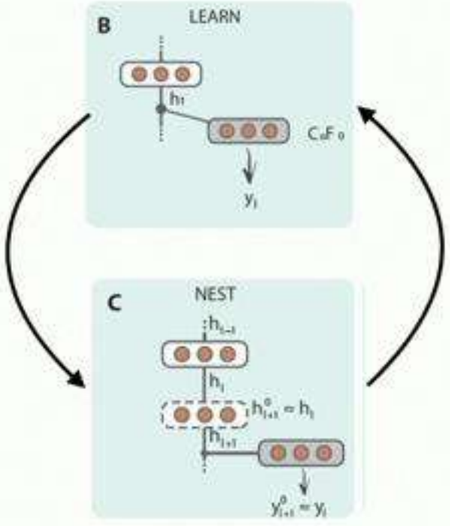
MSE

Step 1: Train layer-wise
Step 2: Fine-tune the whole net
(and only the last CoFo)

Adadelta optimizer

PPA-net: Learning and Inference





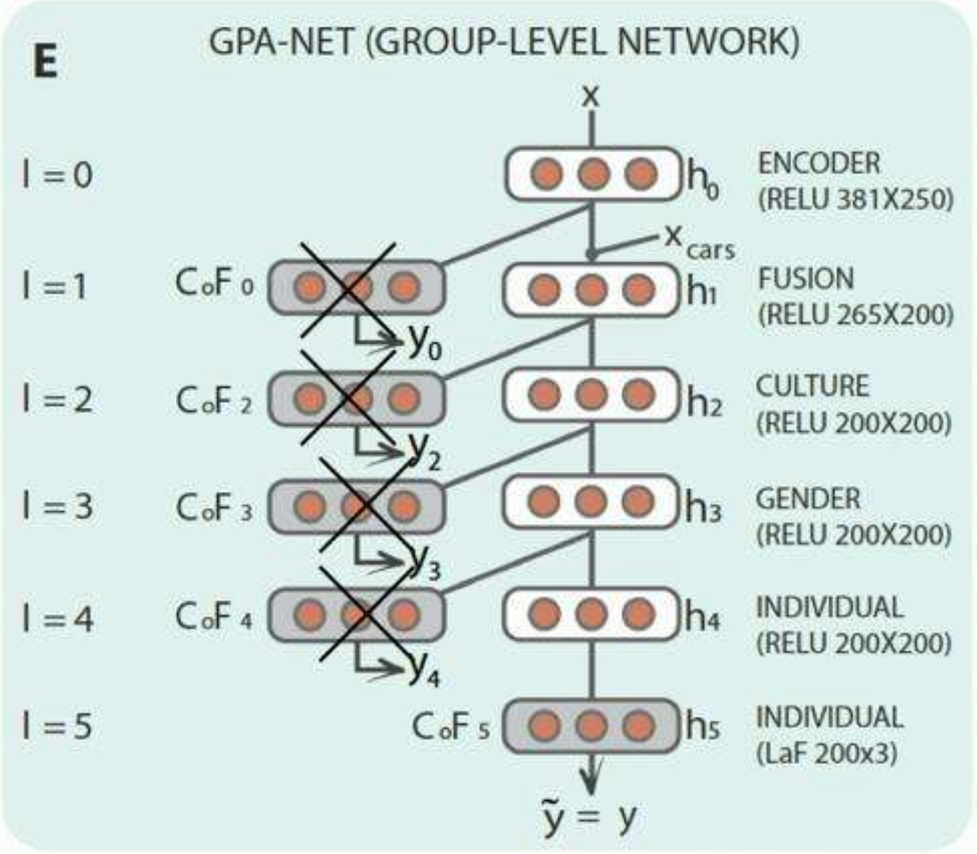
$$\omega_l^* = \underset{\omega_l}{\operatorname{argmin}}\, \alpha_c(h_l, y) = \underset{\omega_l}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \alpha_c \underbrace{\left(y_{l+1}^{(i)}, y^{(i)} \right)}_{\text{MSE}}$$

predicted true

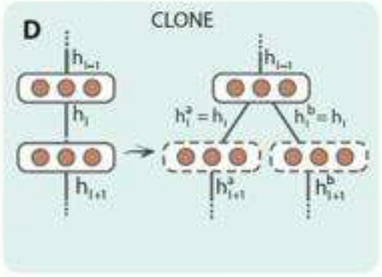
Step 1: Train layer-wise
Step 2: Fine-tune the whole net
(and only the last CoFo)

Adadelta optimizer

PPA-net: Learning and Inference

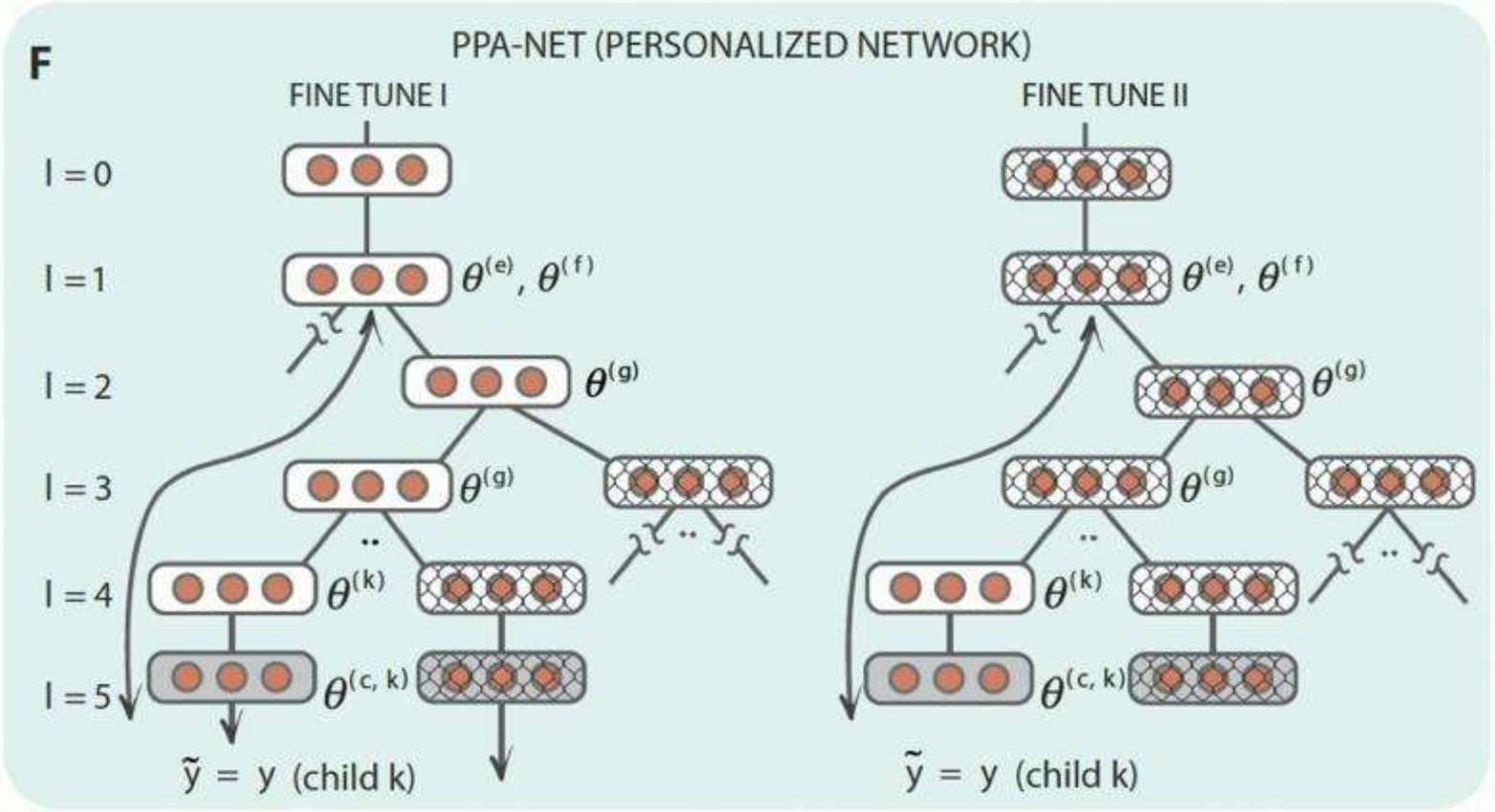


Form Personalized Network
by cloning the network layers



$$l = 2 : \theta_l^{(q)} \leftarrow \theta_l^0, \quad q = \{C_1, C_2\} \quad \textit{Culture ID}$$
$$l = 3 : \theta_l^{(g)} \leftarrow \theta_l^0, \quad g = \{\text{male}, \text{female}\}$$
$$l = 4 : \theta_l^{(k)} \leftarrow \theta_l^0, \quad k = \{1,..,K\} \quad \textit{Child ID}$$
$$l = 5 : \theta_{l-1}^{(c,k)} \leftarrow \theta_{l-1}^{(c,0)}, \quad k = \{1,..,K\}$$

PPA-net: Learning and Inference

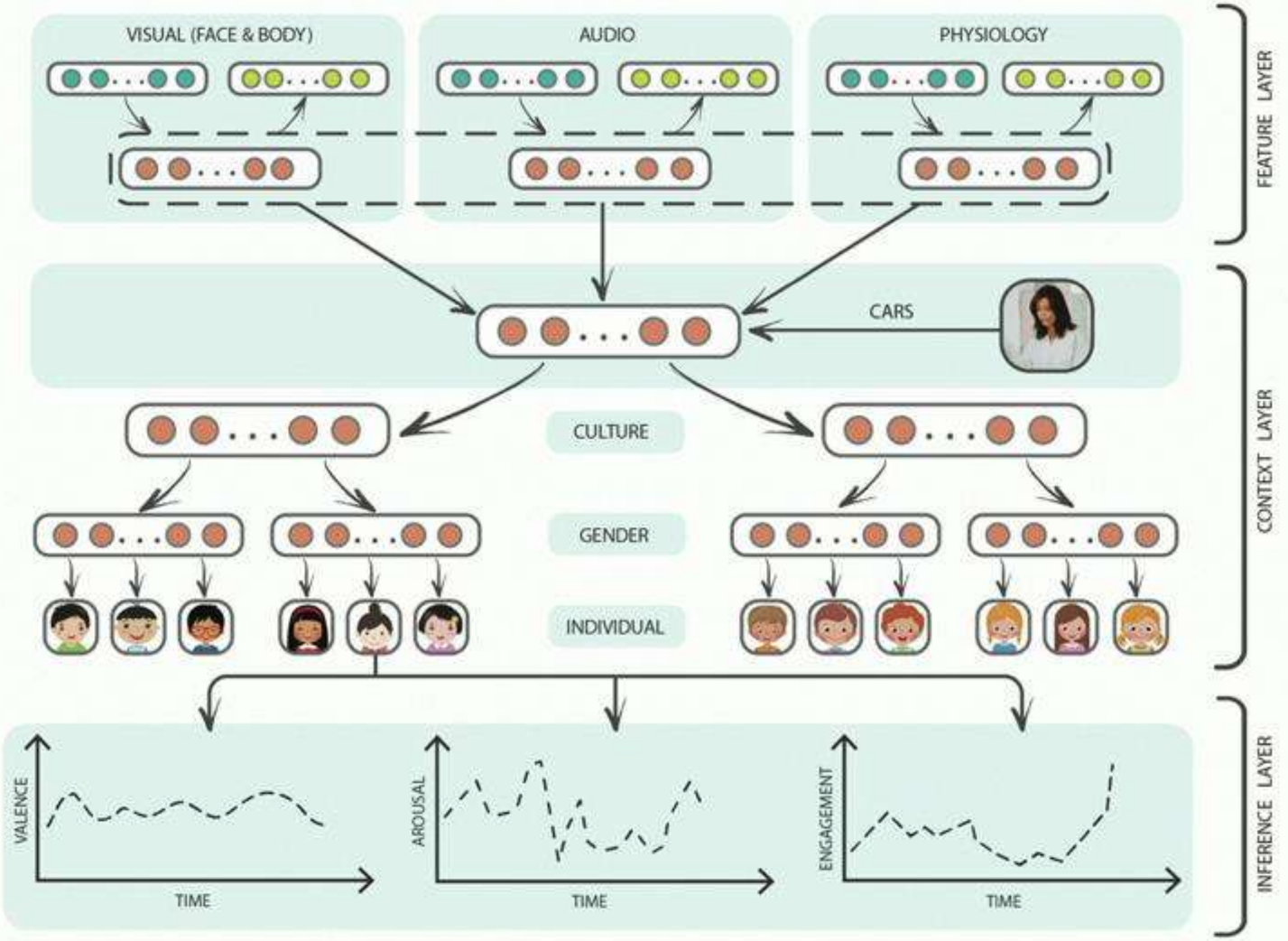


FINE TUNE I : Random sampling of the kid data (sub-chains)
FINE TUNE II: Final tuning using all data of the kid

SGD optimizer

Frozen layers

Personalized Perception of Affect Network (PPA-net)



CARS (**C**hildhood **A**utism **R**ating **S**cale)

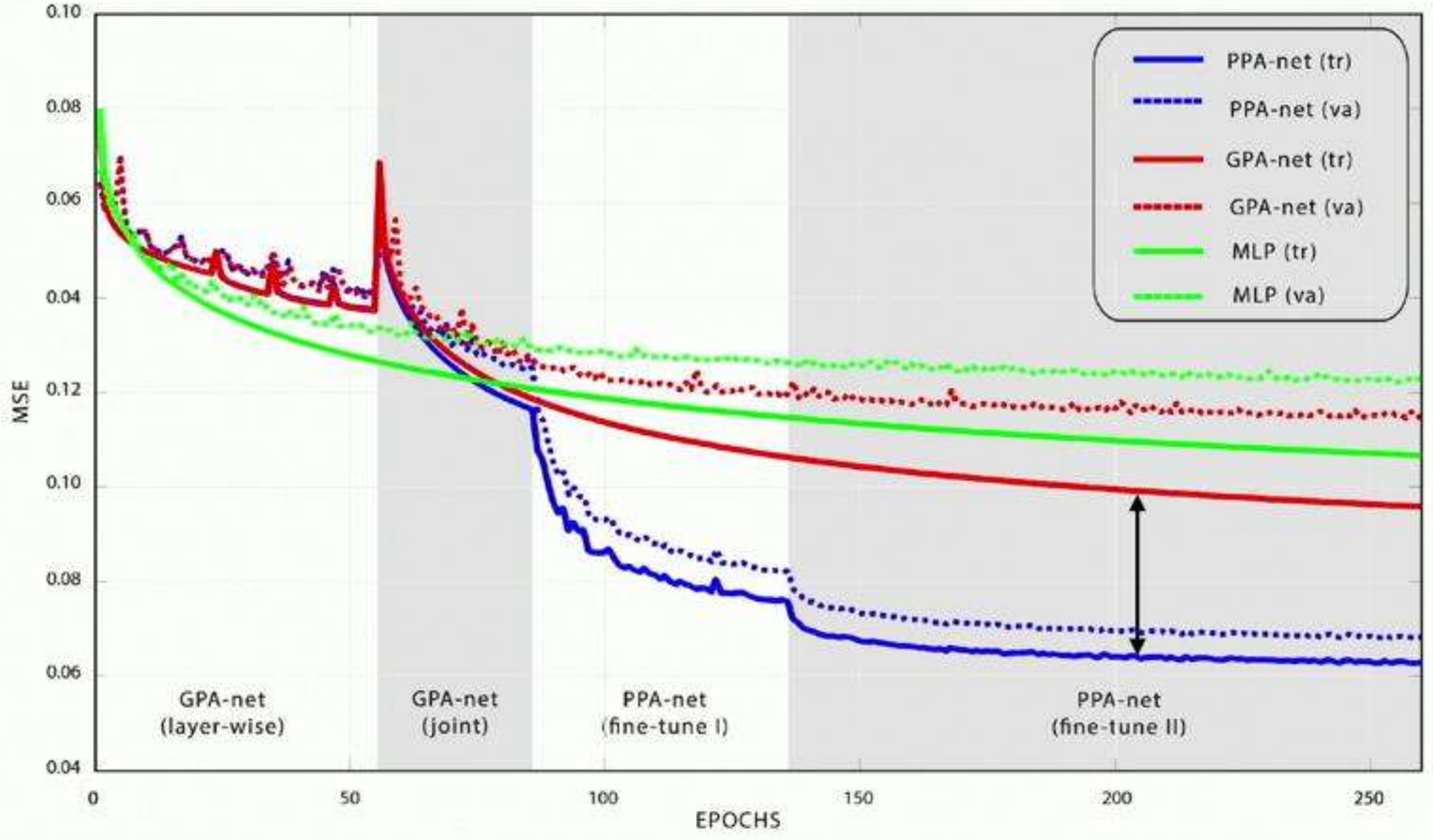
PPA-net: Learning and Inference

Learning operators:

MSE - Mean Squared Error, LaF - Linear Activation Function, ReLu - Rectified Linear Unit

PPA-net: Training

Convergence of the networks' cost during parameter optimization



Child data split:
40% train
20% validation
40% test
10-fold CV

MLP - multi-layer perceptron (group-level, the same depth as GPA-net)

Results: Average Test Performance

Tasks = 105 (35 kids x 3 outputs)

Models	Valence	Arousal	Engagement	Average	*TaskRank (in %)*
PPA-net	**52±21**	**60±16**	**65±24**	**59±21**	**46.5 (1)**
GPA-net	47±28	57±15	60±25	54±24	10.2 (4)
P-MLP	47±18	54±15	59±22	53±20	3.29 (5)
CD-MLP	43±22	52±15	57±23	51±20	2.81 (6)
LR	28±22	35±19	37±21	34±21	2.52 (7)
SVR	45±26	56±14	49±22	50±21	21.2 (2)
GBRT	47±21	51±15	49±22	49±20	12.0 (3)

Between human-raters
ICC: 50-55%

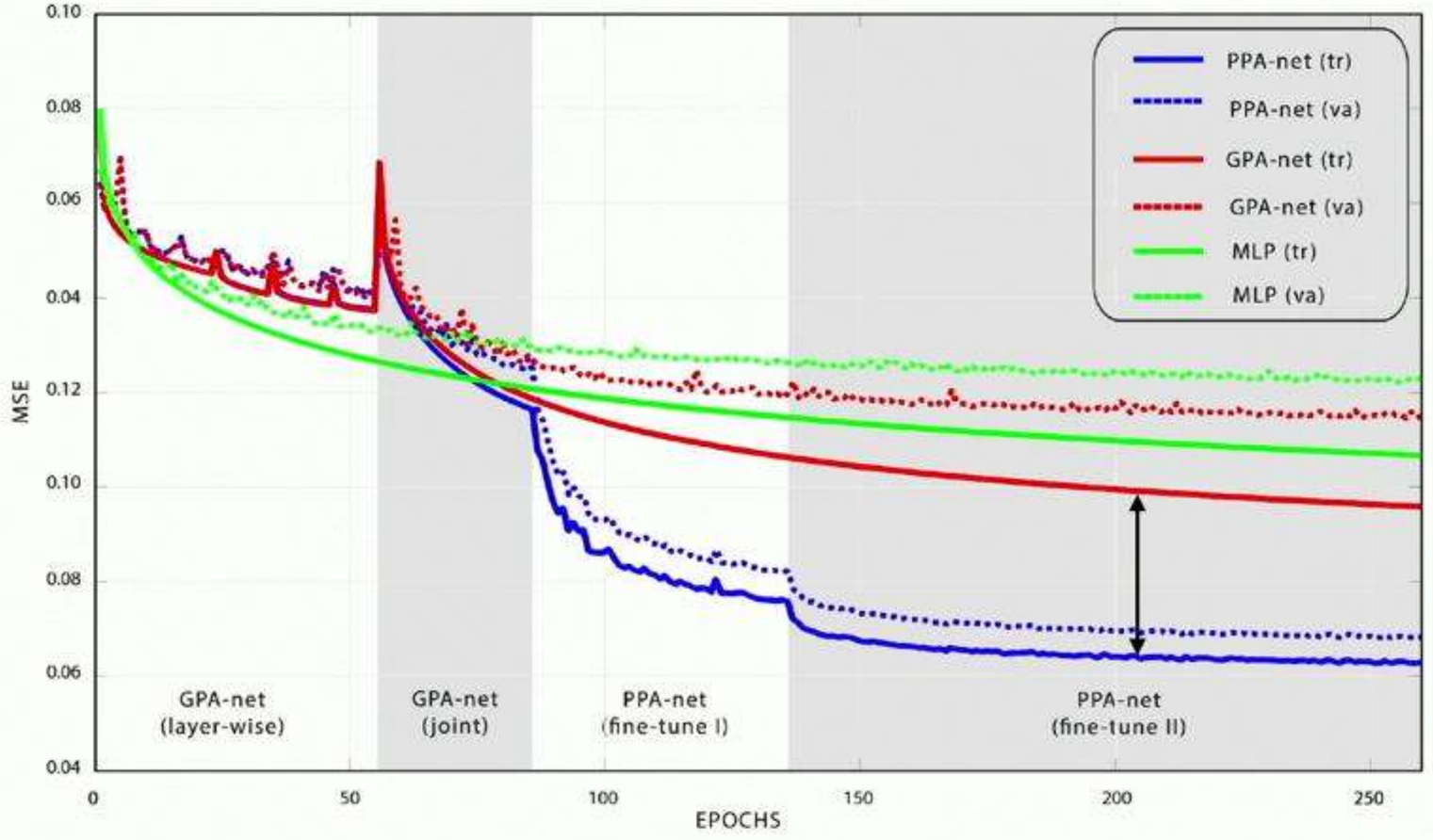
Performance measure: Intra-class Correlation (ICC) - type (3,1) [$\mu \pm 2\,std$ in %]

Consistency between model estimates and human raters
High when there is little variation between the scores given to data sample by the raters

MLP - multi-layer perceptron (P- personalized, CD -child dependent, CI -child-independent)
LR - lasso regression
SVR - support vector regression (RBF kernel)
GBRT - gradient boosted regression trees

PPA-net: Training

Convergence of the networks' cost during parameter optimization



Child data split:
40% train
20% validation
40% test
10-fold CV

MLP - multi-layer perceptron (group-level, the same depth as GPA-net)

Results: Average Test Performance

Tasks = 105 (35 kids x 3 outputs)

Models	Valence	Arousal	Engagement	Average	*TaskRank (in %)*
PPA-net	**52±21**	**60±16**	**65±24**	**59±21**	**46.5 (1)**
GPA-net	47±28	57±15	60±25	54±24	10.2 (4)
P-MLP	47±18	54±15	59±22	53±20	3.29 (5)
CD-MLP	43±22	52±15	57±23	51±20	2.81 (6)
LR	28±22	35±19	37±21	34±21	2.52 (7)
SVR	45±26	56±14	49±22	50±21	21.2 (2)
GBRT	47±21	51±15	49±22	49±20	12.0 (3)

Between human-raters
ICC: 50-55%

Performance measure: Intra-class Correlation (ICC) - type (3,1) [$\mu \pm 2\,std$ in %]

Consistency between model estimates and human raters
High when there is little variation between the scores given to data sample by the raters

MLP - multi-layer perceptron (P- personalized, CD -child dependent, CI -child-independent)
LR - lasso regression
SVR - support vector regression (RBF kernel)
GBRT - gradient boosted regression trees

Results: Average Test Performance

Tasks = 105 (35 kids x 3 outputs)

Models	Valence	Arousal	Engagement	Average	*TaskRank (in %)*
PPA-net	**52±21**	**60±16**	**65±24**	**59±21**	**46.5 (1)**
GPA-net	47±28	57±15	60±25	54±24	10.2 (4)
P-MLP	47±18	54±15	59±22	53±20	3.29 (5)
CD-MLP	43±22	52±15	57±23	51±20	2.81 (6)
LR	28±22	35±19	37±21	34±21	2.52 (7)
SVR	45±26	56±14	49±22	50±21	21.2 (2)
GBRT	47±21	51±15	49±22	49±20	12.0 (3)
CI-MLP	16±17	23±16	21±23	20±19	1.45 (8)

Between human-raters
ICC: 50-55%

Performance measure: Intra-class Correlation (ICC) - type (3,1) [$\mu \pm 2\,std$ in %]

Consistency between model estimates and human raters
High when there is little variation between the scores given to data sample by the raters

MLP - multi-layer perceptron (P- personalized, CD -child dependent, CI -child-independent)
LR - lasso regression
SVR - support vector regression (RBF kernel)
GBRT - gradient boosted regression trees

Results: Average Test Performance

Tasks = 105 (35 kids x 3 outputs)

Models	Valence	Arousal	Engagement	Average	*TaskRank (in %)*
PPA-net	**52±21**	**60±16**	**65±24**	**59±21**	**46.5 (1)**
GPA-net	47±28	57±15	60±25	54±24	10.2 (4)
P-MLP	47±18	54±15	59±22	53±20	3.29 (5)
CD-MLP	43±22	52±15	57±23	51±20	2.81 (6)
LR	28±22	35±19	37±21	34±21	2.52 (7)
SVR	45±26	56±14	49±22	50±21	21.2 (2)
GBRT	47±21	51±15	49±22	49±20	12.0 (3)

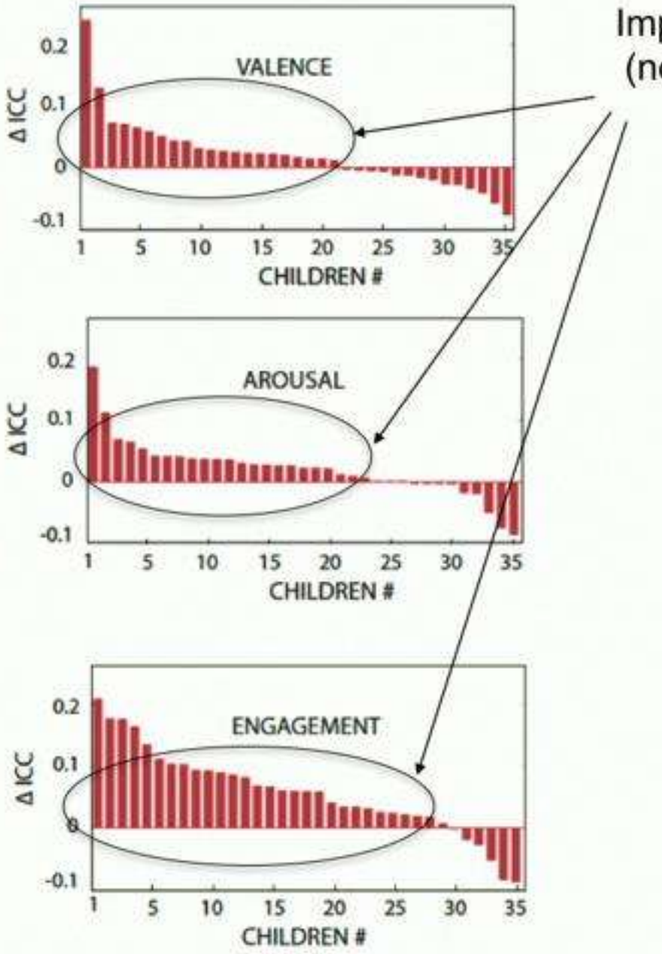
Between human-raters
ICC: 50-55%

Performance measure: Intra-class Correlation (ICC) - type (3,1) [$\mu \pm 2\,std$ in %]

Consistency between model estimates and human raters
High when there is little variation between the scores given to data sample by the raters

MLP - multi-layer perceptron (P- personalized, CD -child dependent, CI -child-independent)
LR - lasso regression
SVR - support vector regression (RBF kernel)
GBRT - gradient boosted regression trees

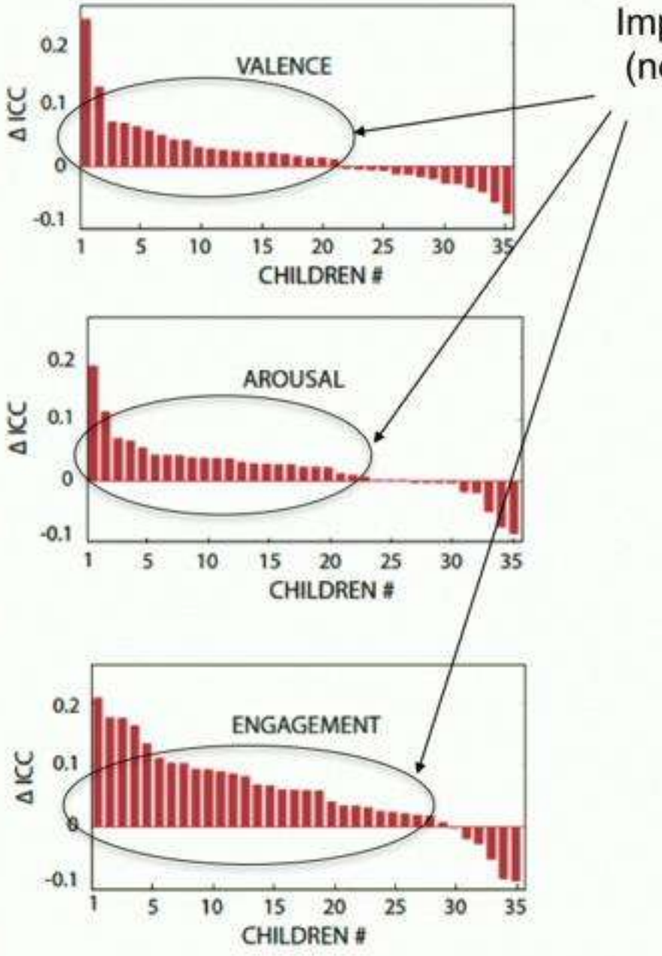
Results: Individual Performance



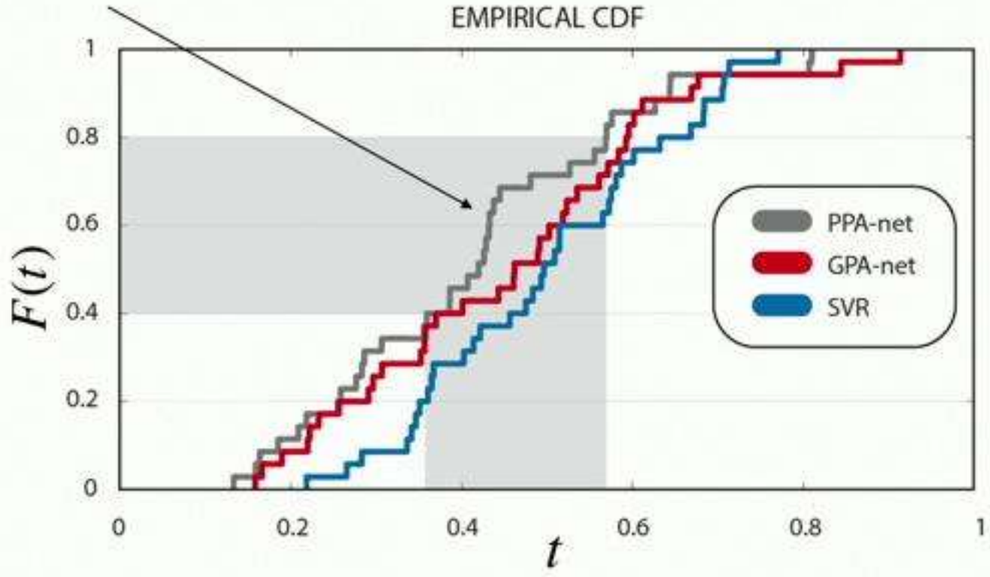
Improvements over the generic (non-personalized) perception

Improvements in the ICC performance (Δ ICC) of the PPA-net over GPA-net (per child)

Results: Individual Performance



Improvements over the generic (non-personalized) perception



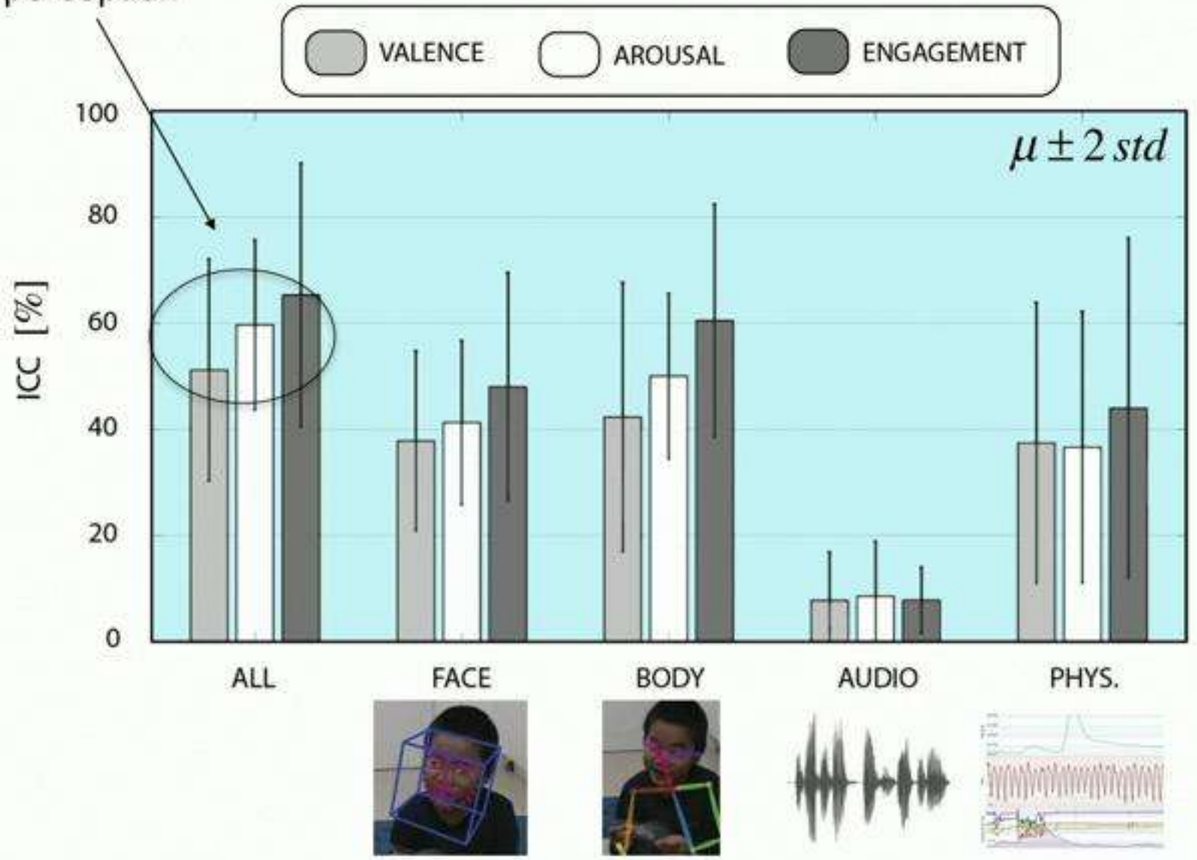
$$F(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i<t}$$

$$X_i = 1 - ICC_i \ , i = 1,..,n_{kids}$$

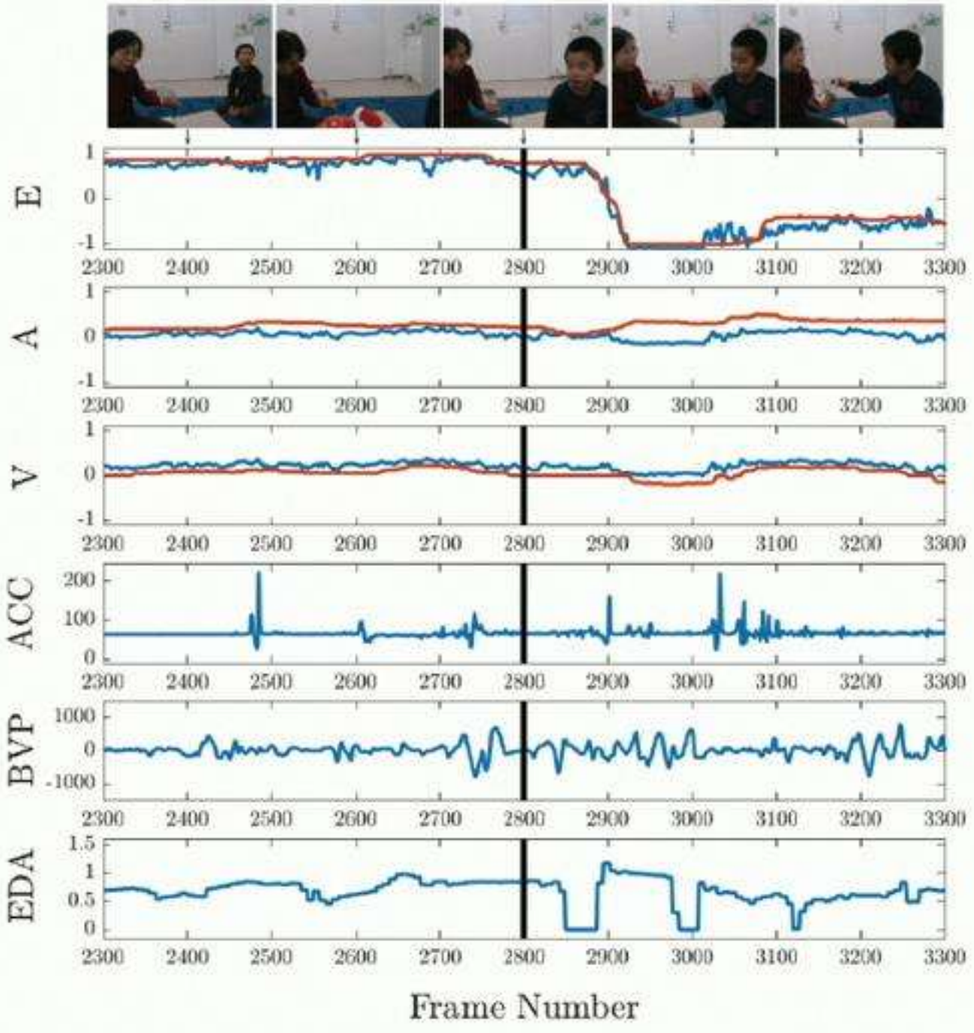
Improvements in the ICC performance (Δ ICC) of the PPA-net over GPA-net (per child)

Results: The Effects of Different Modalities

The fusion of different data modalities improves personalized perception



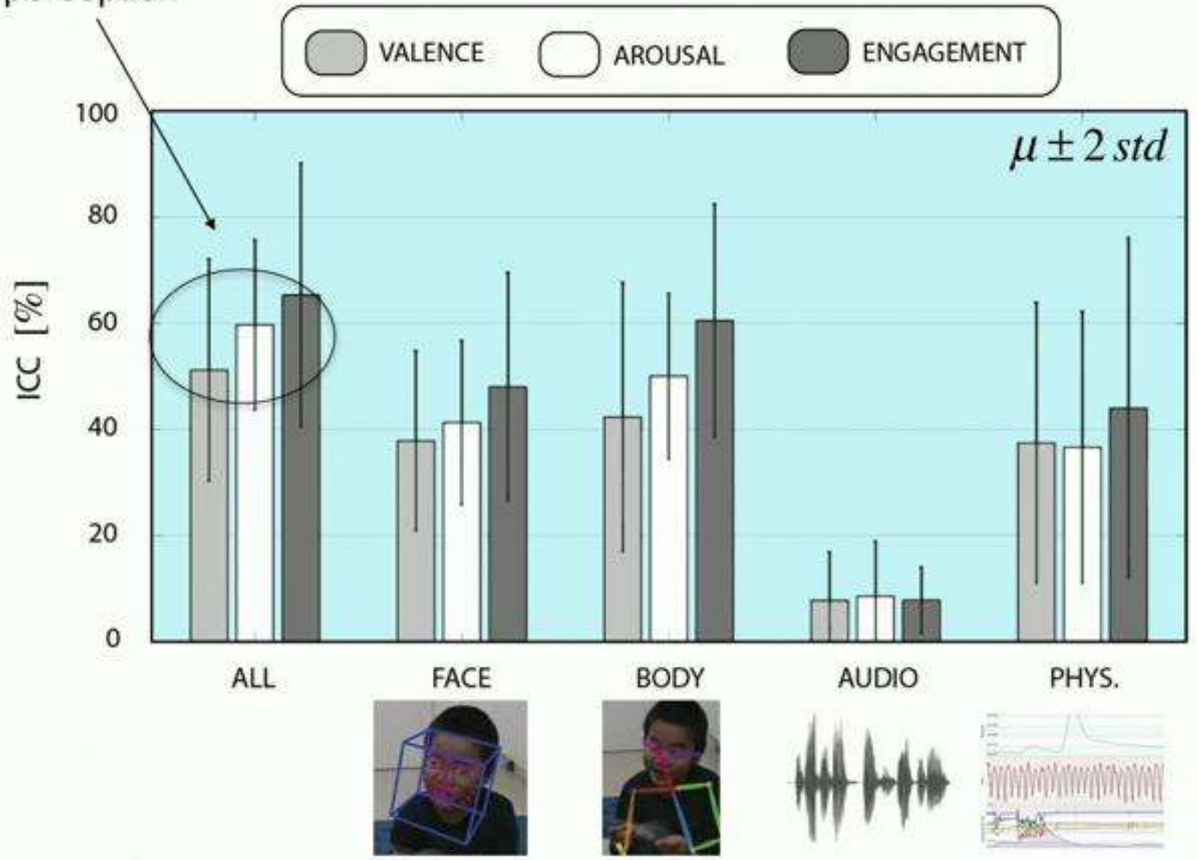
Utility: Real-time monitoring of child's responses



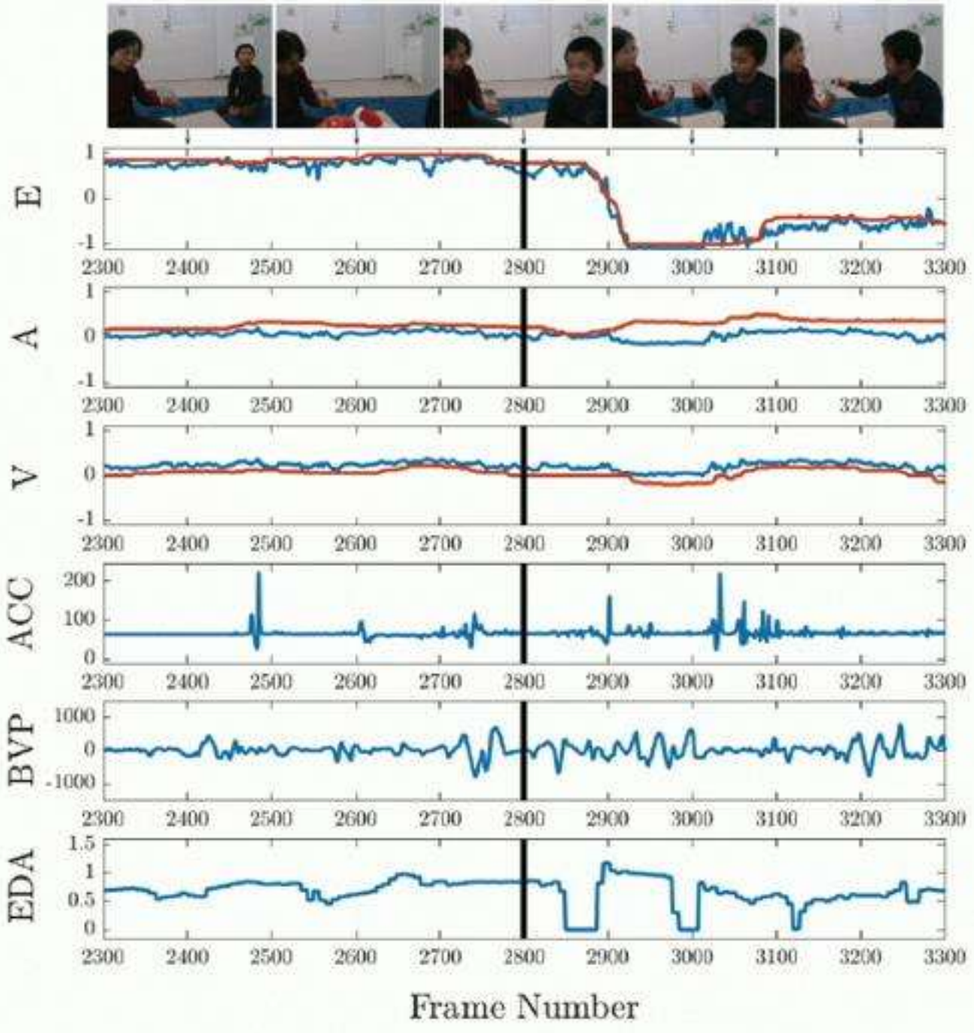
V - valence, **A** - arousal, **E**- engagement, **ACC** - accelerometer data (child's movements)
BVP - Blood Volume Pulse (heart rate), **EDA** - Electrodermal Activity

Results: The Effects of Different Modalities

The fusion of different data modalities improves personalized perception

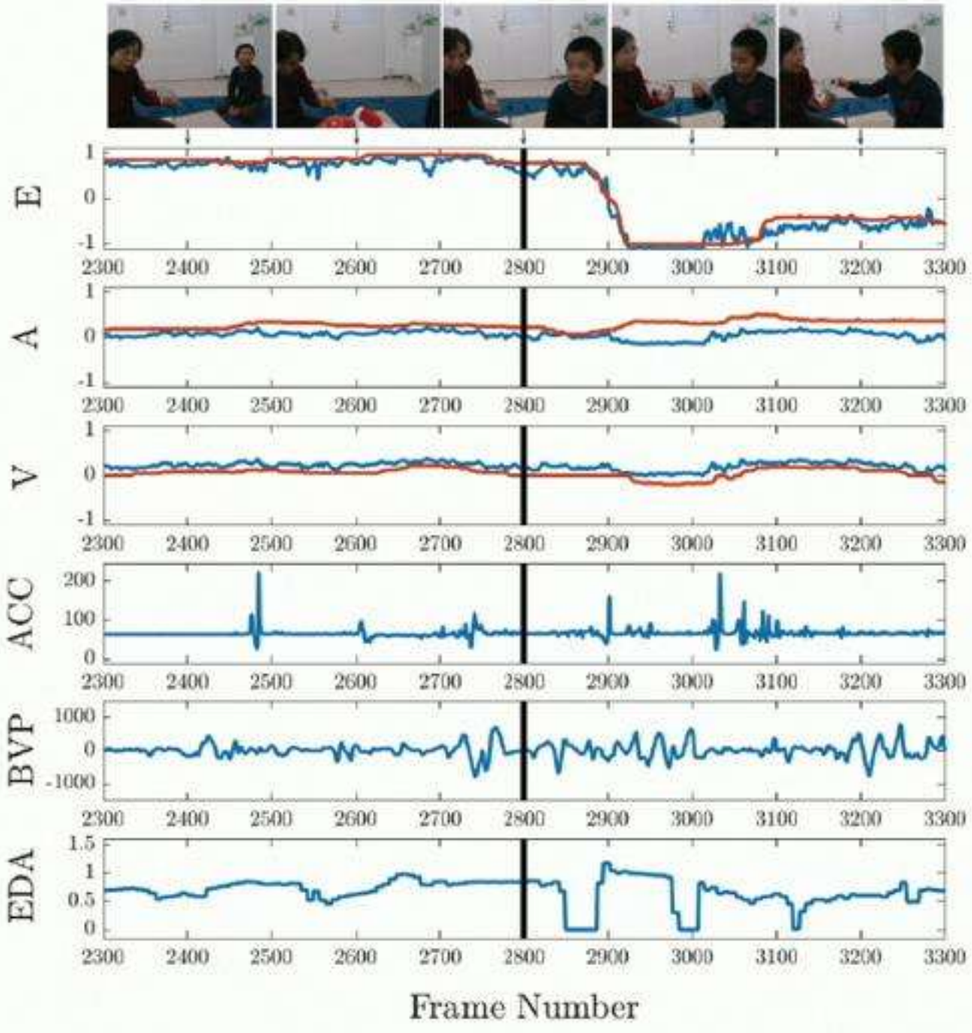


Utility: Real-time monitoring of child's responses



V - valence, **A** - arousal, **E**- engagement, **ACC** - accelerometer data (child's movements)
BVP - Blood Volume Pulse (heart rate), **EDA** - Electrodermal Activity

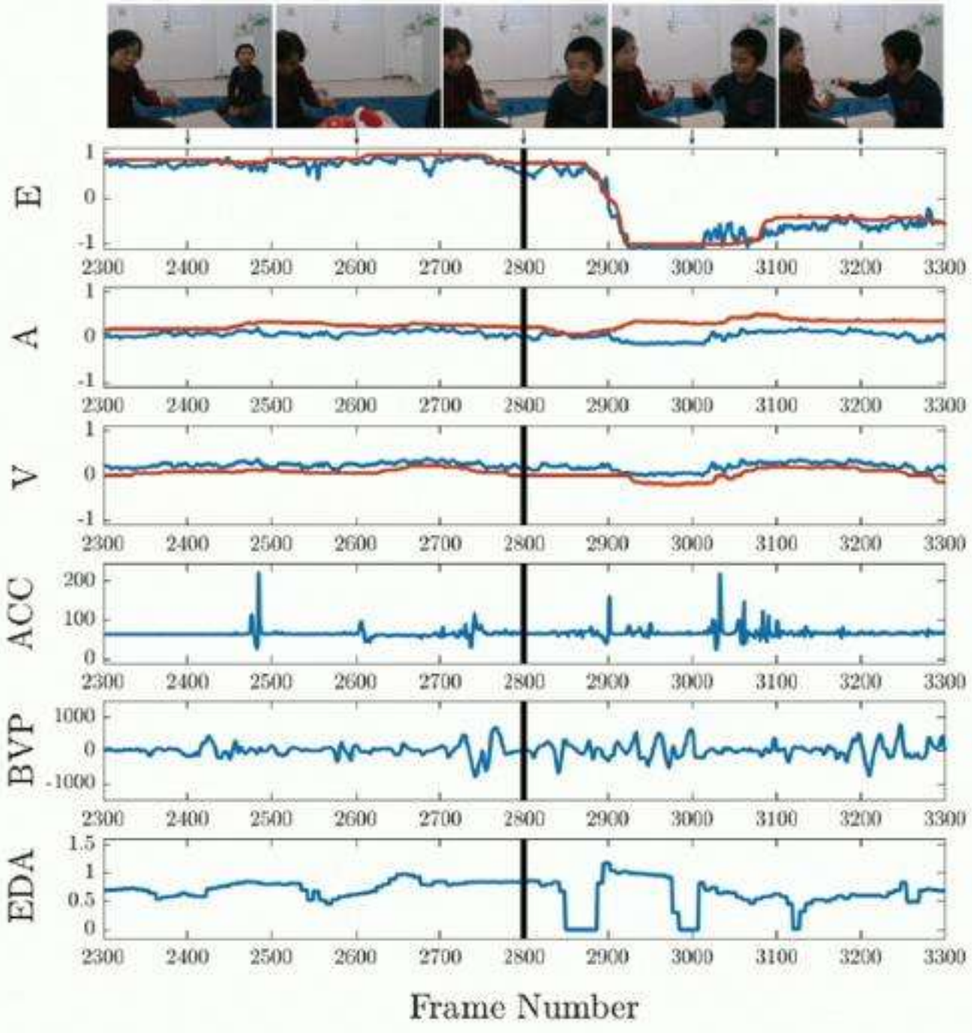
Utility: Real-time monitoring of child's responses

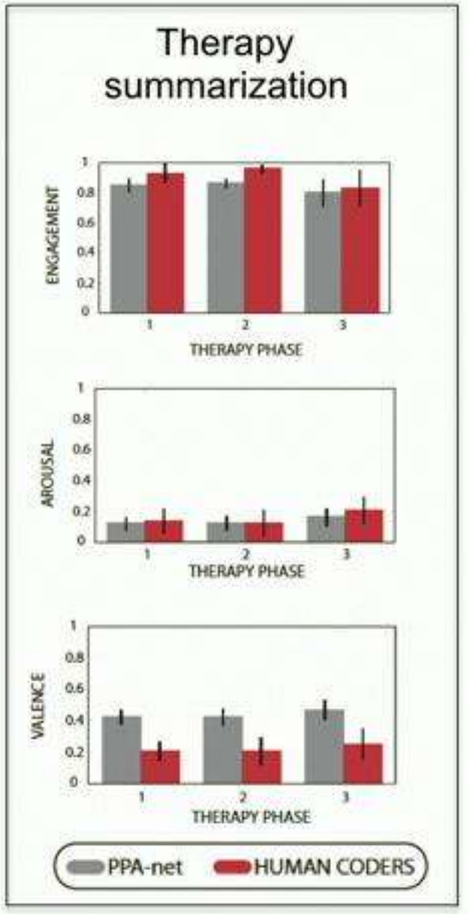


Intelligent interaction

V - valence, **A** - arousal, **E**- engagement, **ACC** - accelerometer data (child's movements)
BVP - Blood Volume Pulse (heart rate), **EDA** - Electrodermal Activity

Utility: Real-time monitoring of child's responses





V - valence, **A** - arousal, **E**- engagement, **ACC** - accelerometer data (child's movements)
BVP - Blood Volume Pulse (heart rate), **EDA** - Electrodermal Activity

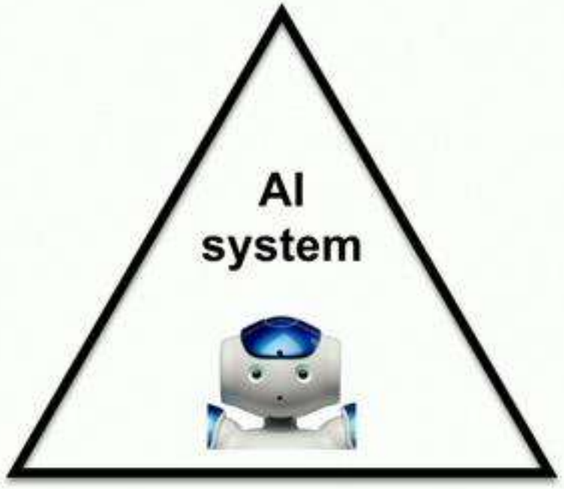
Summary

- The personalized network achieves estimation of affect and engagement that is consistent with human-experts
- Personalization brings large improvements for most of the kids, compared to the group-level network
- Enables the real-time monitoring and therapy summarization

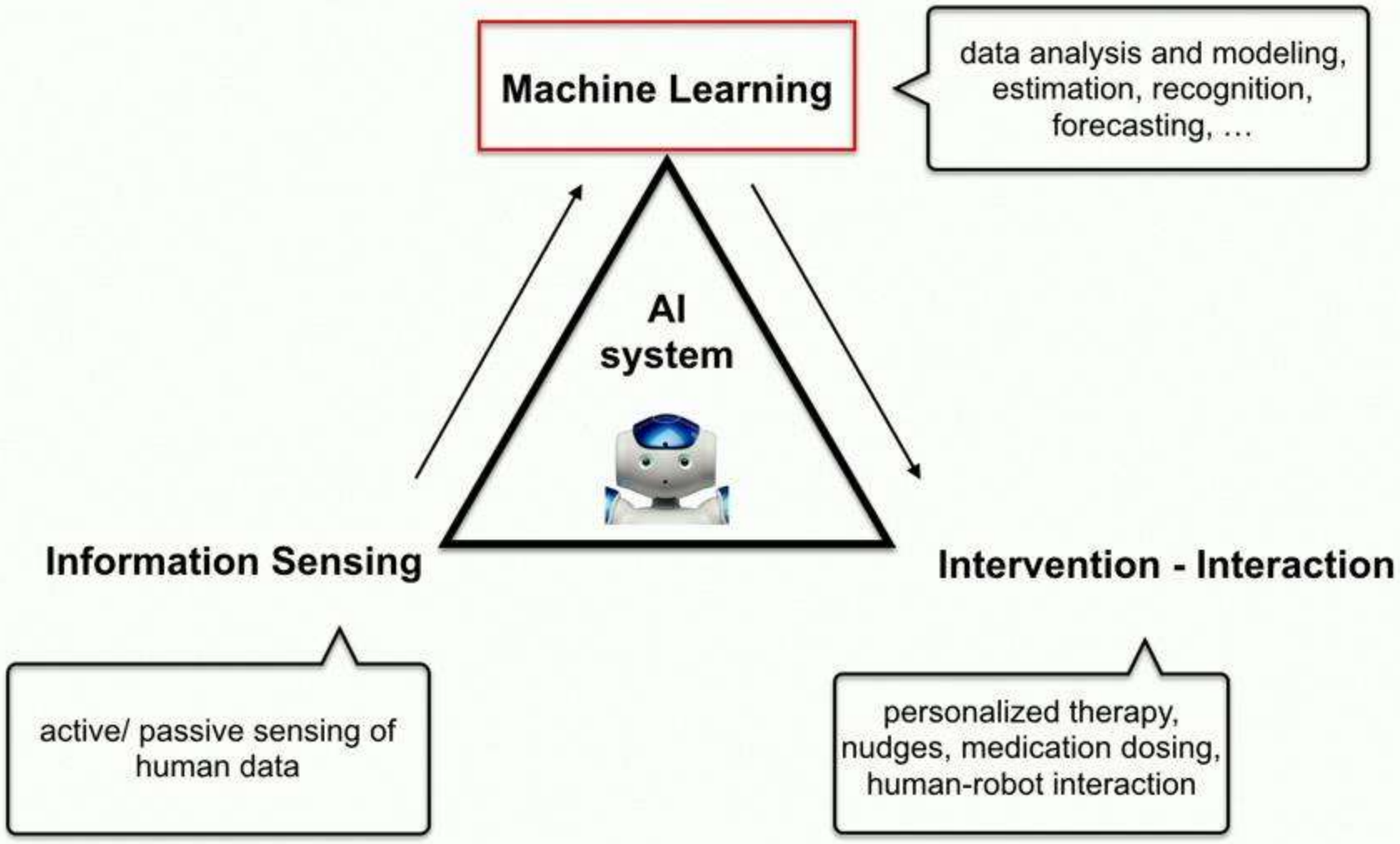
Limitations:

- *User-dependent* (data of target child needed during model training)
- *Static* (no information about dynamics of child's behavior)
- *Negative transfer* (the performance reduction for some kids)

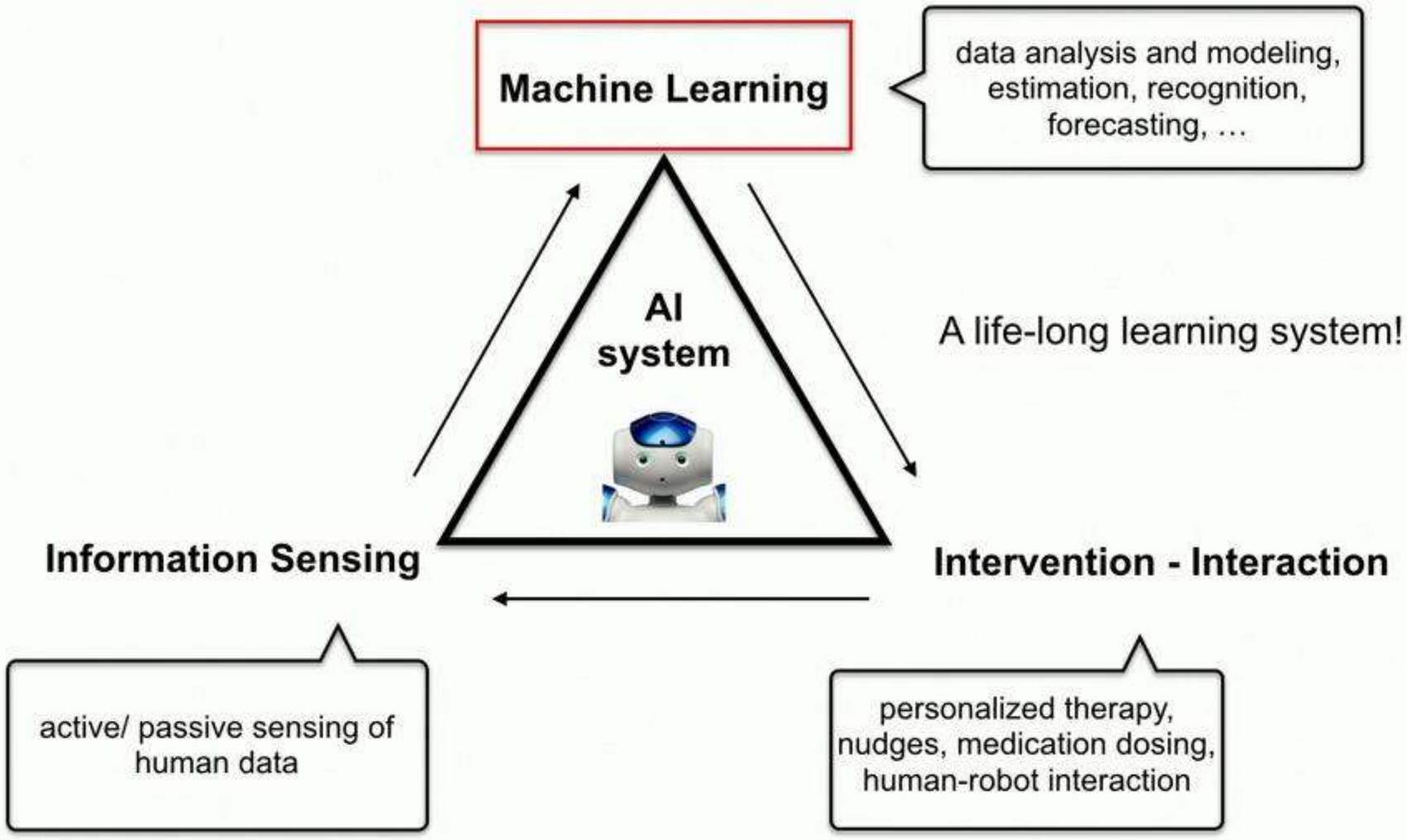
AI for Human Data: Ideal System



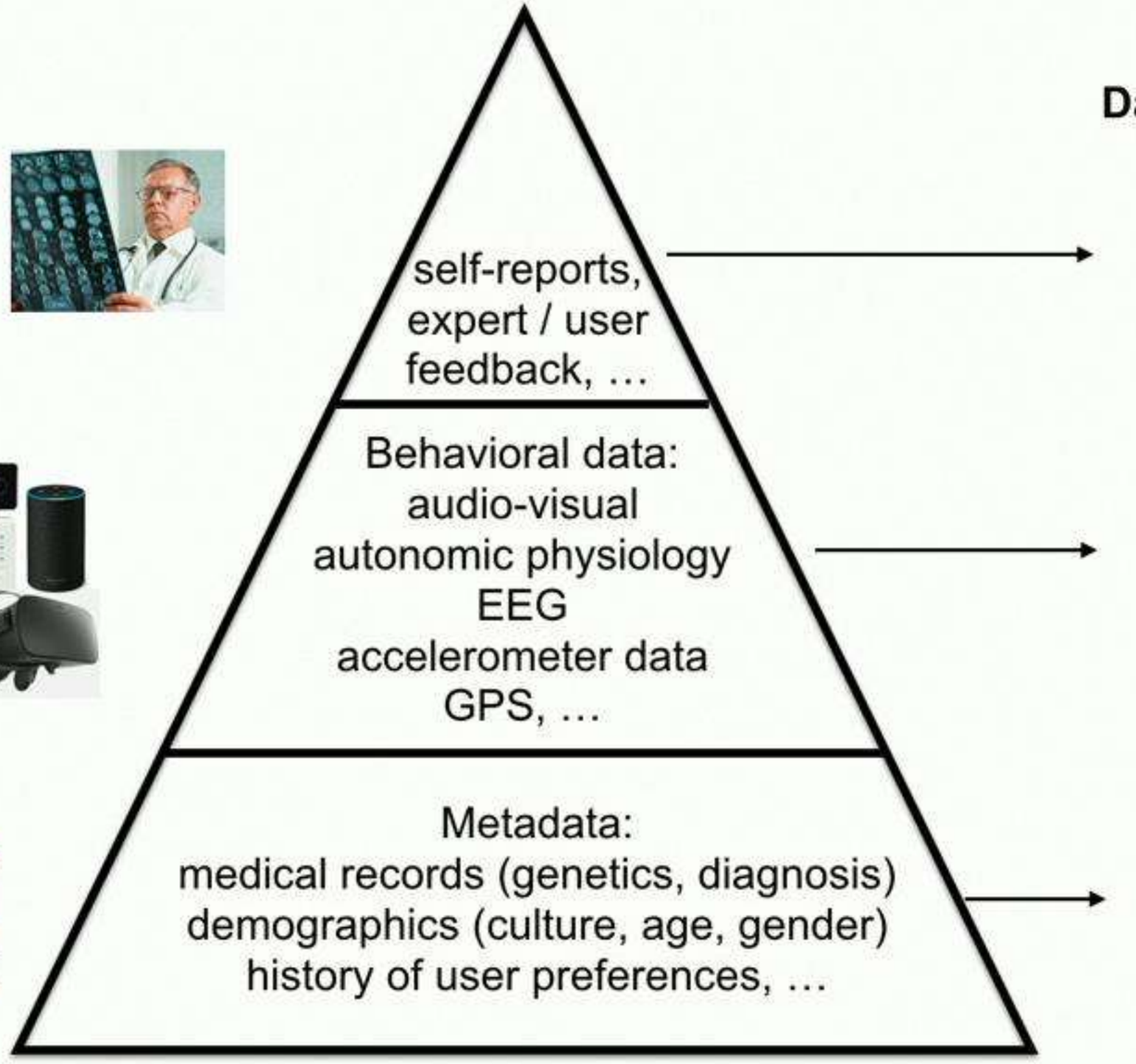
AI for Human Data: Ideal System



AI for Human Data: Ideal System



Human Data



Data sensing

active

passive

active / passive

Learning From Human Data

- To empower the user
- To protect the user
- To enable and sustain naturalistic and engaging interactions






Future Directions

- Active Learning for Continuous Personalization
- Scaling up and Data Privacy
- Efficient Deployment
- Context-sensitivity

Active Learning for Continuous Personalization

Rudovic, O., Park, H-W., Busche, J., Schuller, B. , Breazeal, C., Picard, R. W., (2019) " Personalized Estimation of Engagement from Videos Using Active Learning with Deep Reinforcement Learning," IEEE CVPR -AMFG W

Problem Statement

Given videos of child-robot learning activities, how do we make personalized estimation of their engagement levels using as few as possible labeled clips from the target child?

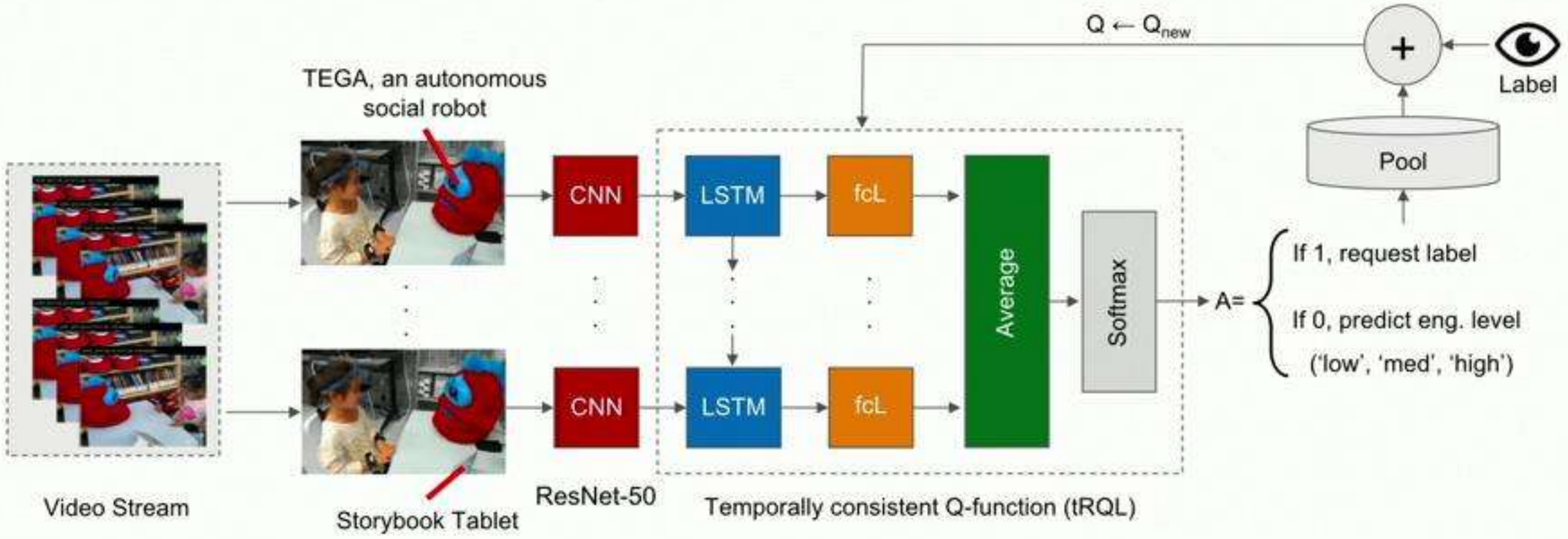
Dataset:

- 43 kids, age 4-6
- 8 sessions over 3 months
- discrete coding of 7.2k 5 sec clips in terms of engagement: low, medium, high, by three human experts
- fixed bird's view camera

Rudovic, O., Park, H-W., Busche, J., Schuller, B. , Breazeal, C., Picard, R. W., (2019) " Personalized Estimation of Engagement from Videos Using Active Learning with Deep Reinforcement Learning," IEEE CVPR -AMFG W

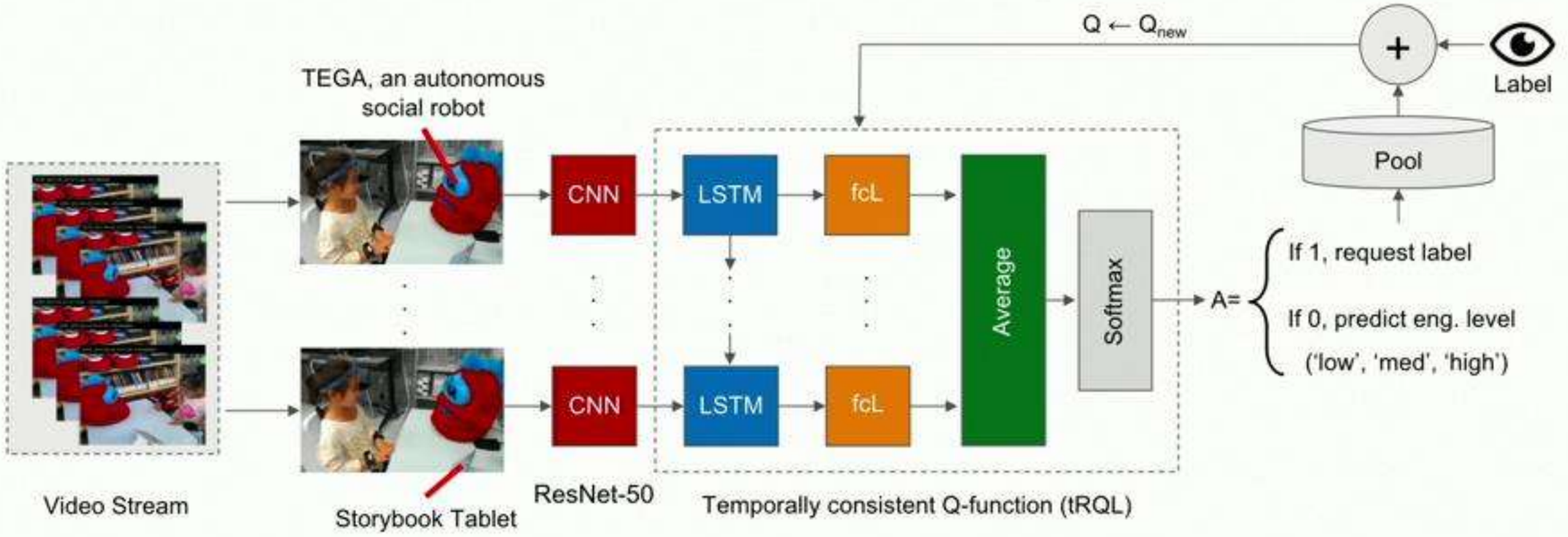
Personalized Deep Reinforcement Learning

Goal: Learn a personalized policy that "decides" whether to request a video label (for further learning) or estimate engagement!



Personalized Deep Reinforcement Learning

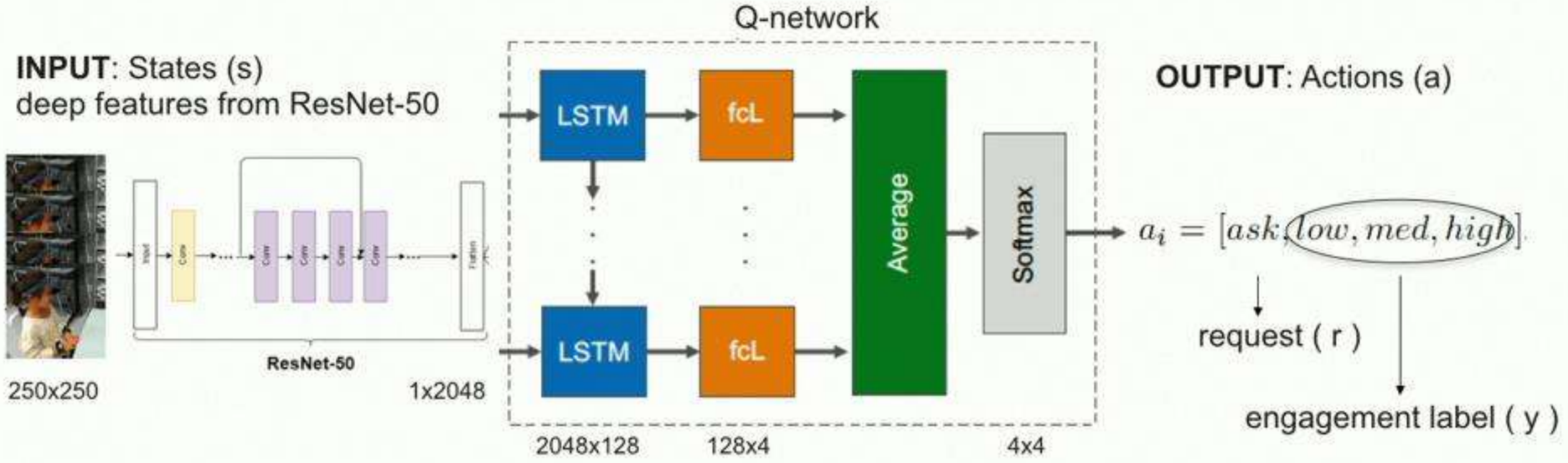
Goal: Learn a personalized policy that "decides" whether to request a video label (for further learning) or estimate engagement!



Rudovic, O., Park, H-W., Busche, J., Schuller, B. , Breazeal, C., Picard, R. W., (2019) " Personalized Estimation of Engagement from Videos Using Active Learning with Deep Reinforcement Learning," IEEE CVPR -AMFG W

Method: Personalized Deep Reinforcement Learning

Step 1: Learn a population-level policy from data of training children



Rewards:

$$R_i = \begin{cases} R_{req} = -0.05, \text{if } r_i = 1 \\ R_{cor} = +1, \quad \text{if } r_i = 0 \wedge y^* = y \\ R_{inc} = -1, \quad \text{if } r_i = 0 \wedge y^* \neq y \end{cases}$$

Minimize Bellman loss:

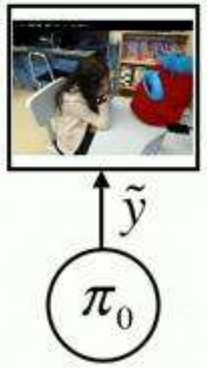
$$L_B^{(i)}(\Theta) = [Q_\Theta(s_i, a_i) - (R_i + \gamma \max_{a_{i+1}} Q_\Theta(s_{i+1}, a_{i+1}))]^2$$

↓

optimal policy π_g

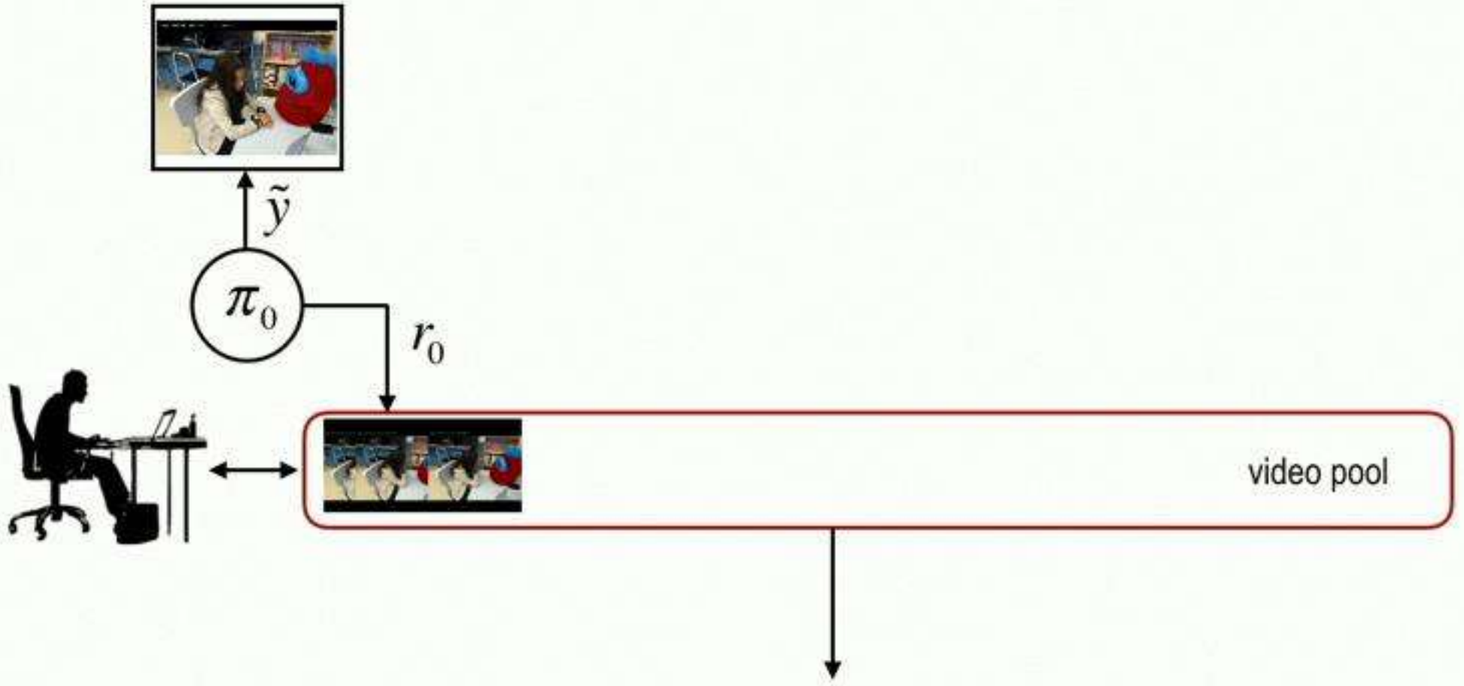
Personalized Deep Reinforcement Learning

Step 2: Personalize the population-level policy (π_0) to the target kid



Personalized Deep Reinforcement Learning

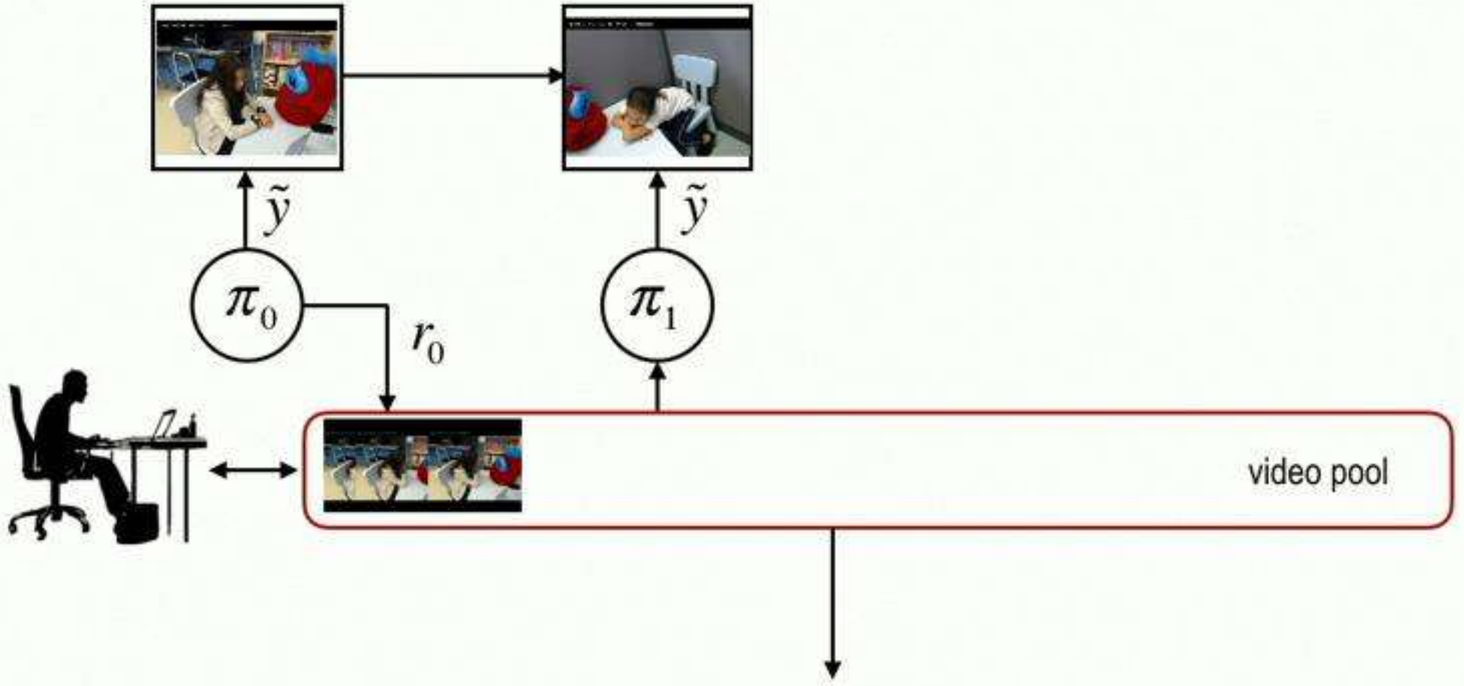
Step 2: Personalize the population-level policy (π_0) to the target kid



policy update π_j: $\hat{\Theta}_j \leftarrow \min_{\Theta_{j-1}} \sum_{v_r \in \{S_j\}} L_B^{(v_r)}(\Theta)$

Personalized Deep Reinforcement Learning

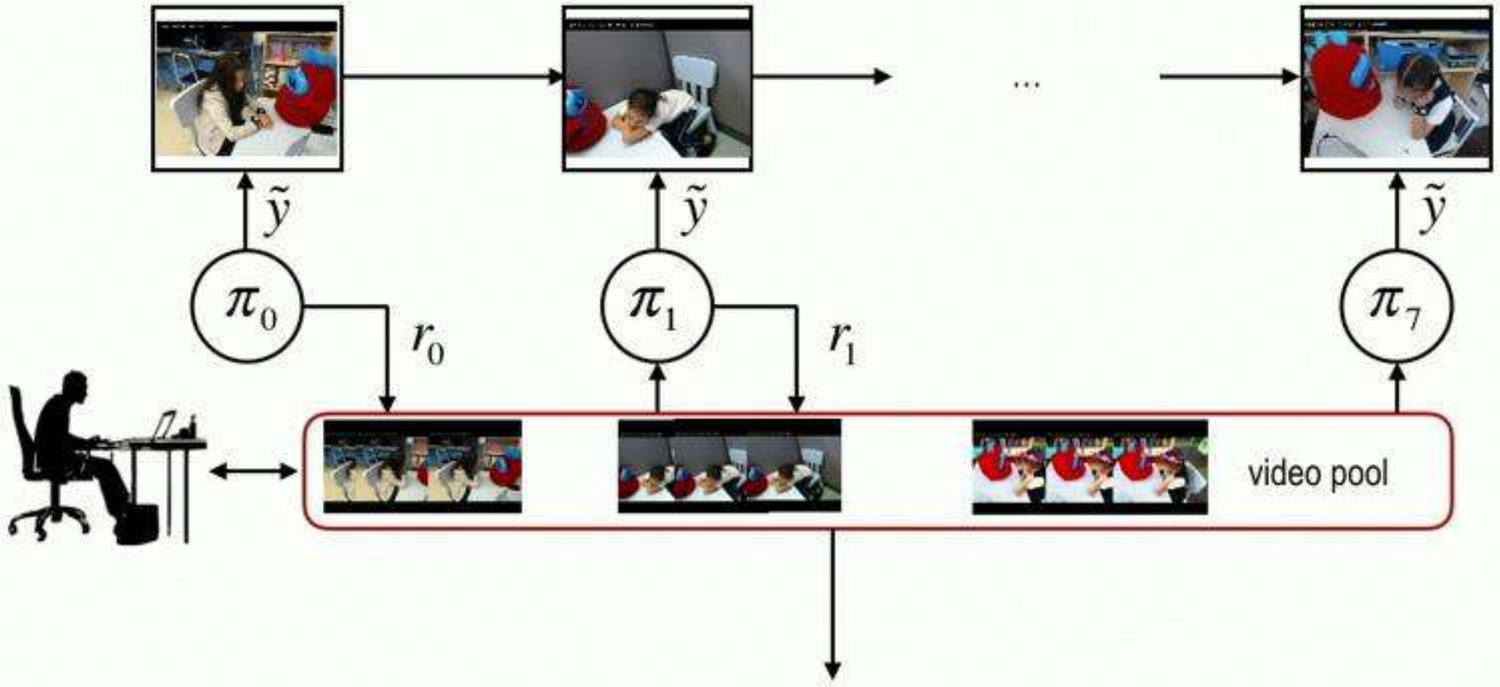
Step 2: Personalize the population-level policy (π_0) to the target kid



policy update π_j: $\hat{\Theta}_j \leftarrow \min_{\Theta_{j-1}} \sum_{v_r \in \{S_j\}} L_B^{(v_r)}(\Theta)$

Personalized Deep Reinforcement Learning

Step 2: Personalize the population-level policy (π_0) to the target kid



policy update π_j: $\hat{\Theta}_j \leftarrow \min_{\Theta_{j-1}} \sum_{v_r \in \{S_j\}} L_B^{(v_r)}(\Theta)$

Results: Average Performance

Model	Session	ACC [%]									F1 [%]								
		S1	S2	S3	S4	S5	S6	S7	S8	AVE.	S1	S2	S3	S4	S5	S6	S7	S8	AVE.
LSTM-S	INIT	65	59	61	60	62	60	48	67	60	31	40	36	37	38	40	32	35	35
	RND	65	**67**	69	65	67	64	55	68	65	31	**46**	40	35	40	41	**41**	33	38
	ENTR	65	61	65	**70**	73	65	**57**	**75**	66	31	42	**43**	37	38	43	40	35	39
	LCONF	65	66	62	69	72	69	52	**75**	66	31	40	39	**41**	37	42	**41**	35	38
	SMAR	65	66	**70**	66	**75**	**70**	55	73	**68**	31	45	41	36	**48**	**47**	37	**36**	**40**
TC-DQL	INIT	72	57	66	61	74	64	56	71	65	32	42	33	40	43	47	35	31	39
	RND	72	67	60	64	72	70	63	81	69	32	38	39	32	**46**	40	41	41	40
	RL	72	**70**	**74**	**70**	**82**	**75**	**65**	**85**	**74**	32	**45**	**42**	**46**	45	**50**	**48**	**46**	**44**
	REQ [%]	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6

average % of requests (~10 video clips per session)

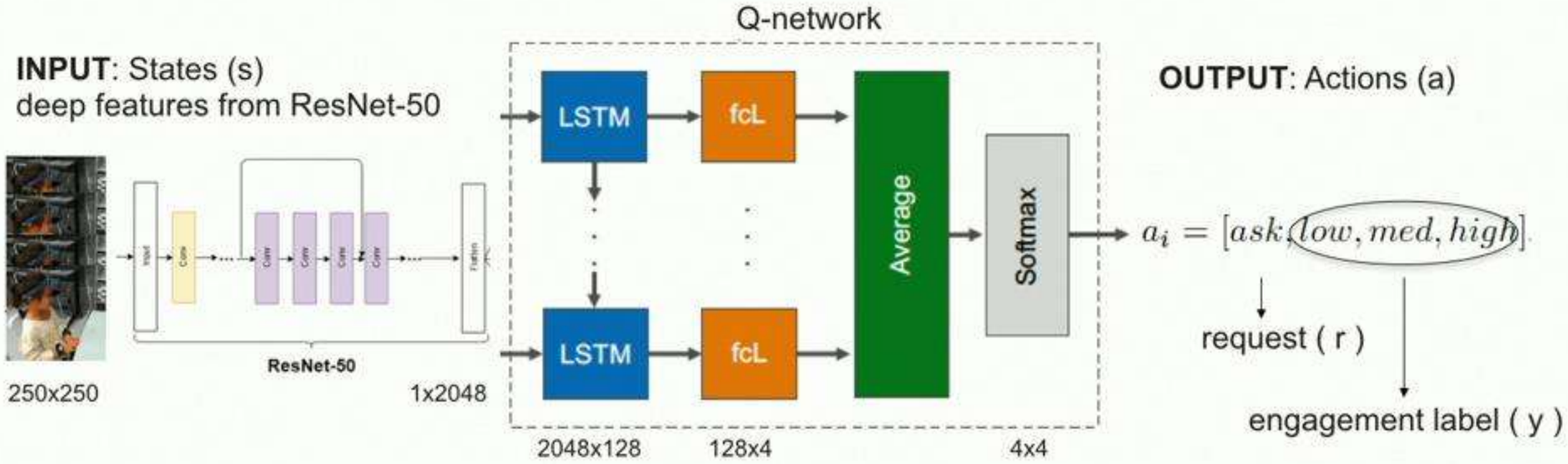
Setting: 22 kids for training/validation of the population policy (**INIT**), 21 kids for testing/policy adaptation

Supervised LSTM (LSTM-S) updated using data selected by heuristic Active Learning strategies:

RND - random requests
ENTR - max entropy
LCONF - least confidence
SMAR - smallest margin

Method: Personalized Deep Reinforcement Learning

Step 1: Learn a population-level policy from data of training children



Rewards:

$$R_i = \begin{cases} R_{req} = -0.05, \text{if } r_i = 1 \\ R_{cor} = +1, \quad \text{if } r_i = 0 \wedge y^* = y \\ R_{inc} = -1, \quad \text{if } r_i = 0 \wedge y^* \neq y \end{cases}$$

Results: Average Performance

Model	Session	ACC [%]									F1 [%]								
		S1	S2	S3	S4	S5	S6	S7	S8	AVE.	S1	S2	S3	S4	S5	S6	S7	S8	AVE.
LSTM-S	INIT	65	59	61	60	62	60	48	67	60	31	40	36	37	38	40	32	35	35
	RND	65	**67**	69	65	67	64	55	68	65	31	**46**	40	35	40	41	**41**	33	38
	ENTR	65	61	65	**70**	73	65	**57**	**75**	66	31	42	**43**	37	38	43	40	35	39
	LCONF	65	66	62	69	72	69	52	**75**	66	31	40	39	**41**	37	42	**41**	35	38
	SMAR	65	66	**70**	66	**75**	**70**	55	73	**68**	31	45	41	36	**48**	**47**	37	**36**	**40**
TC-DQL	INIT	72	57	66	61	74	64	56	71	65	32	42	33	40	43	47	35	31	39
	RND	72	67	60	64	72	70	63	81	69	32	38	39	32	**46**	40	41	41	40
	RL	72	**70**	**74**	**70**	**82**	**75**	**65**	**85**	**74**	32	**45**	**42**	**46**	45	**50**	**48**	**46**	**44**
	REQ [%]	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6

average % of requests (~10 video clips per session)

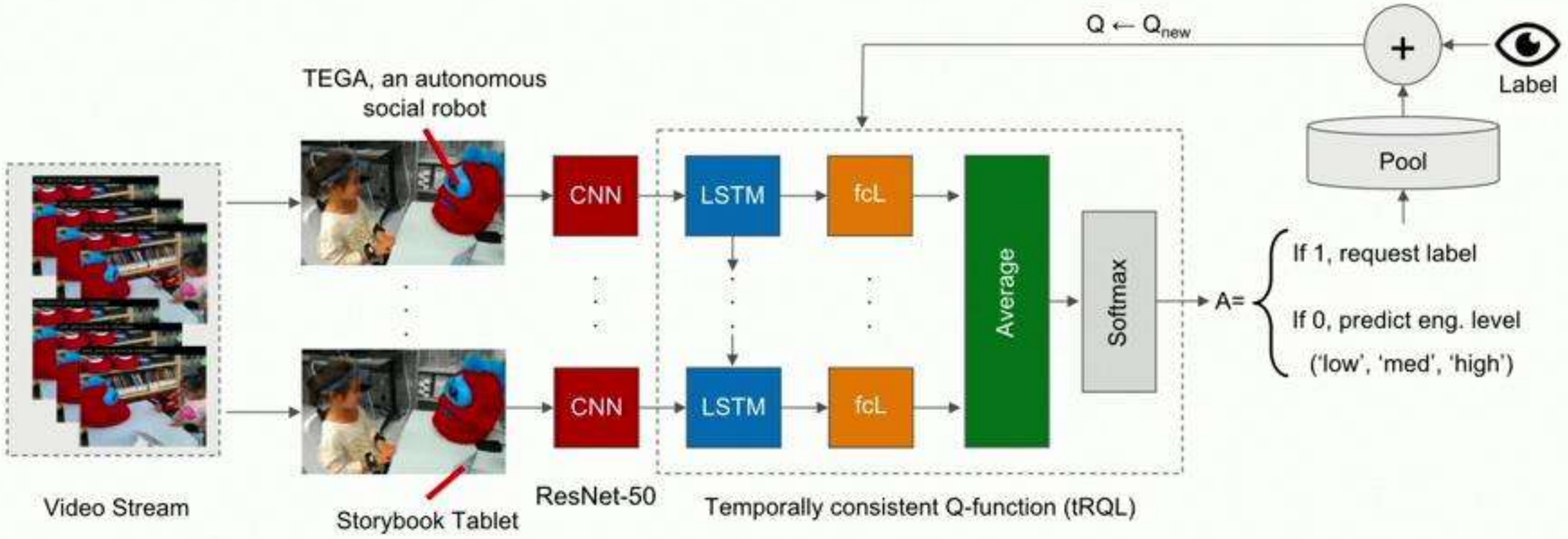
Setting: 22 kids for training/validation of the population policy (**INIT**), 21 kids for testing/policy adaptation

Supervised LSTM (LSTM-S) updated using data selected by heuristic Active Learning strategies:

RND - random requests
ENTR - max entropy
LCONF - least confidence
SMAR - smallest margin

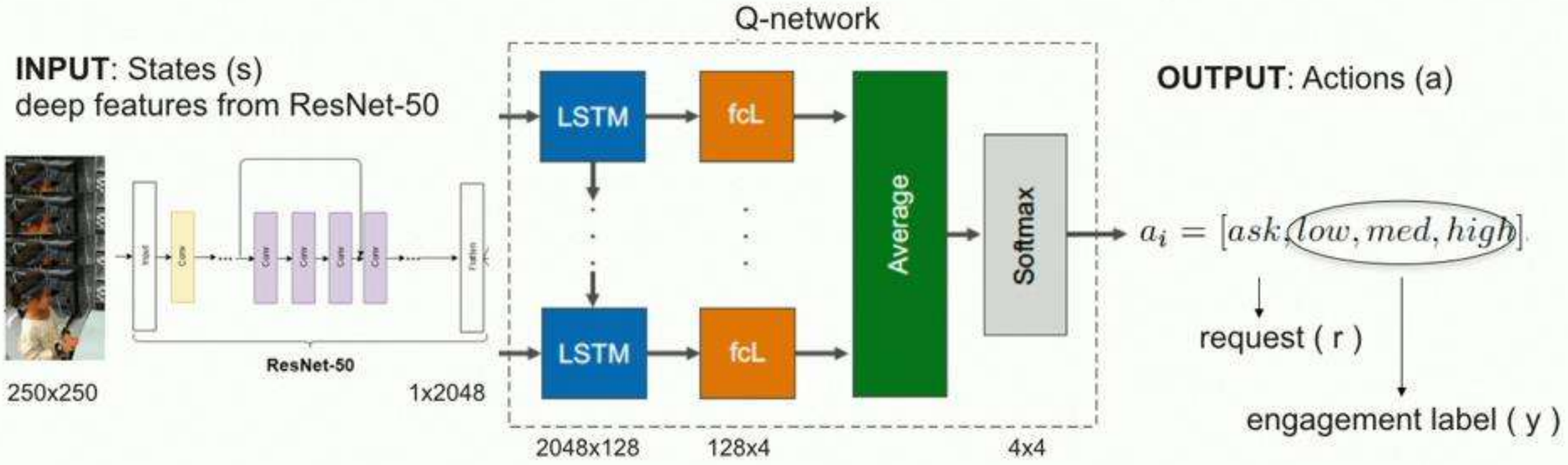
Personalized Deep Reinforcement Learning

Goal: Learn a personalized policy that "decides" whether to request a video label (for further learning) or estimate engagement!



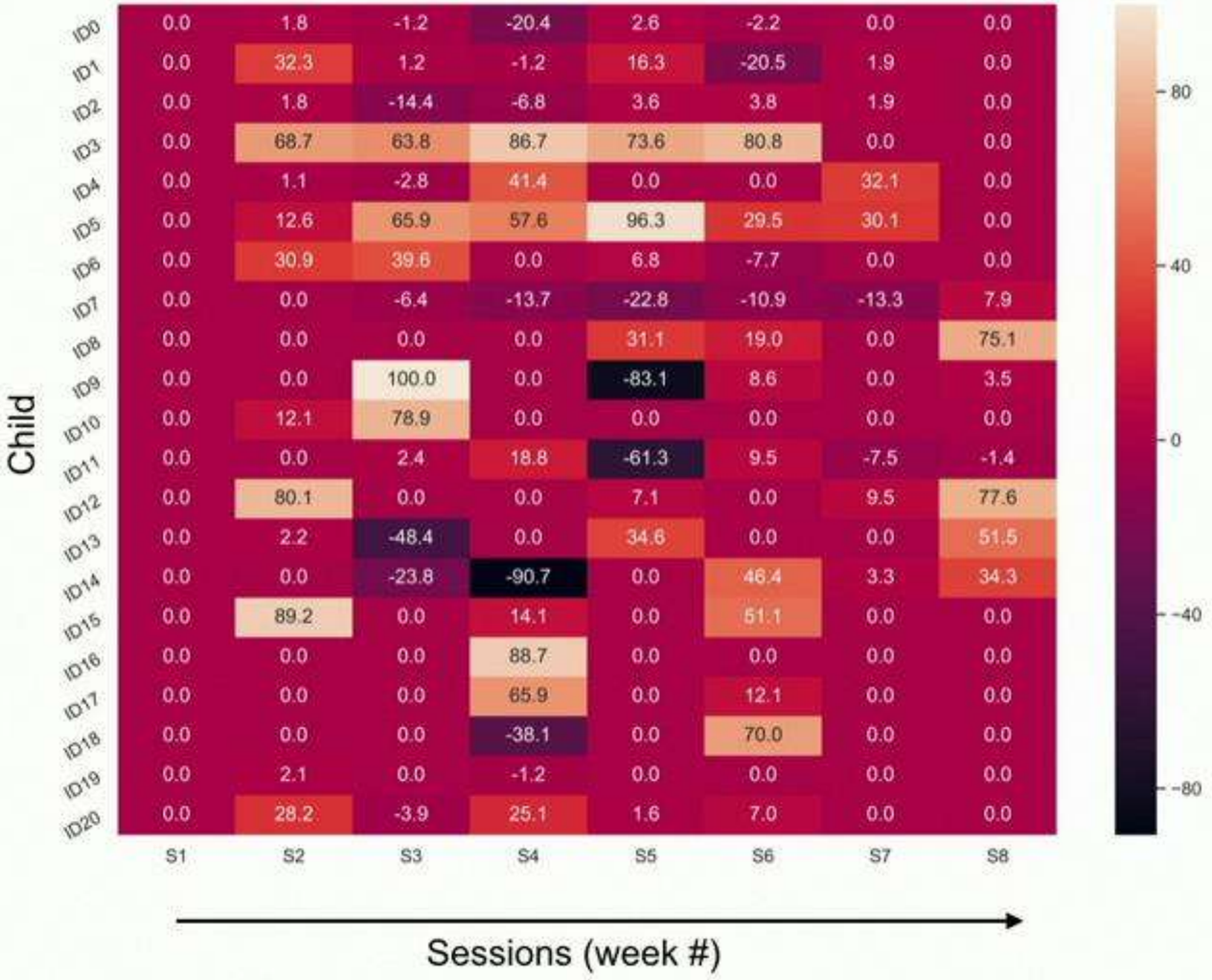
Method: Personalized Deep Reinforcement Learning

Step 1: Learn a population-level policy from data of training children



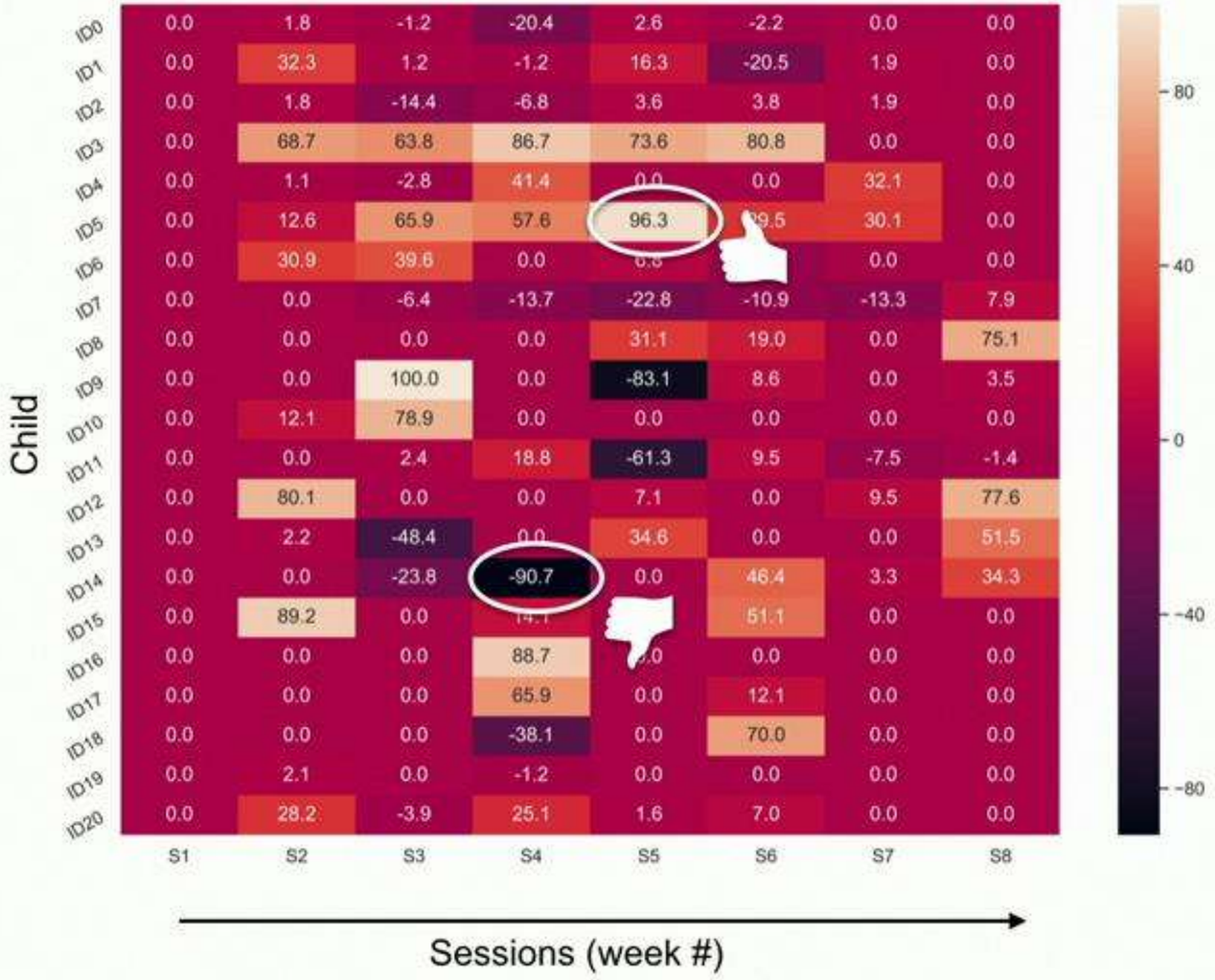
Results: Individual Performance

Relative improvement due to the policy personalization (Δ ACC [%])



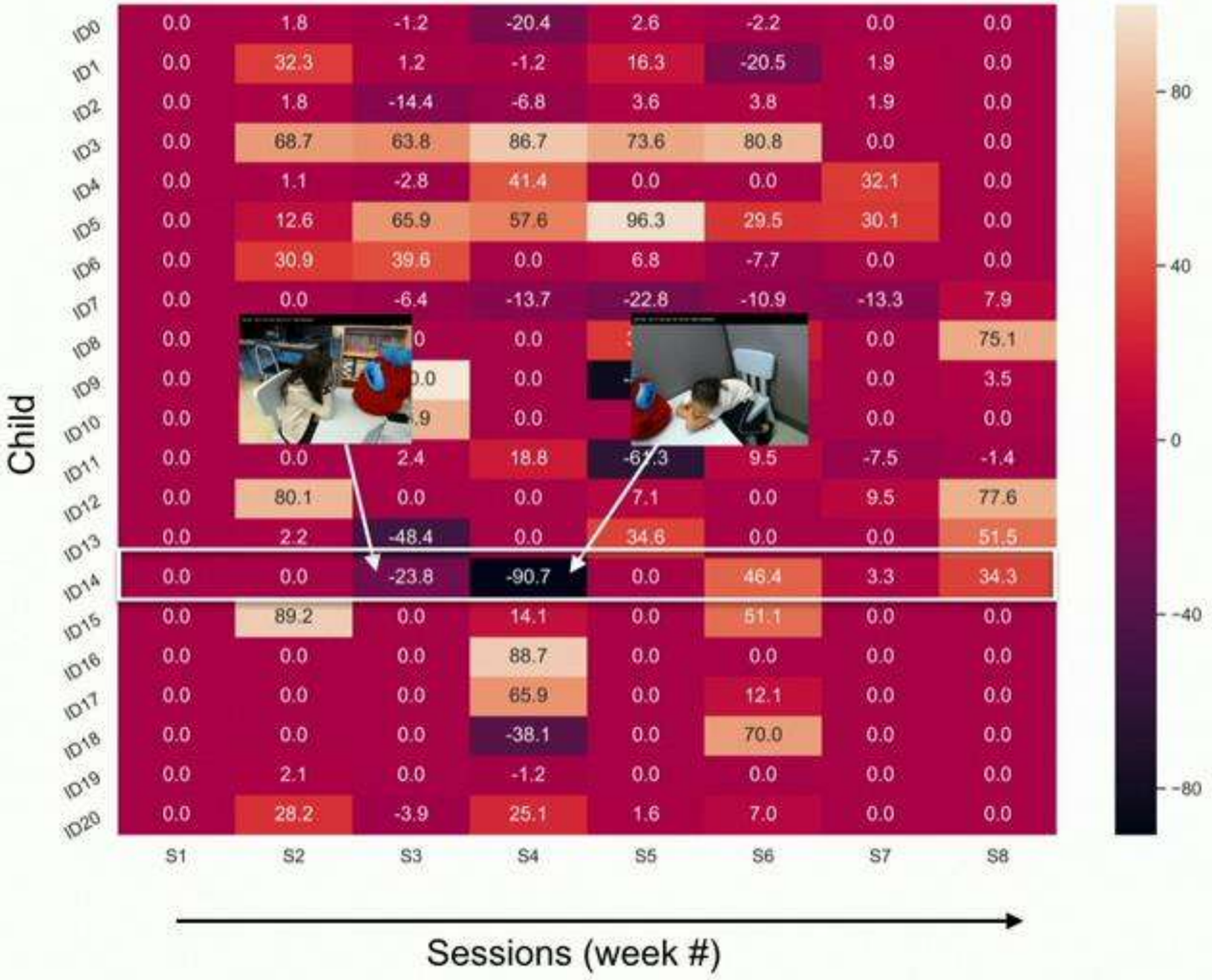
Results: Individual Performance

Relative improvement due to the policy personalization (Δ ACC [%])



Results: Individual Performance

Relative improvement due to the policy personalization (Δ ACC [%])



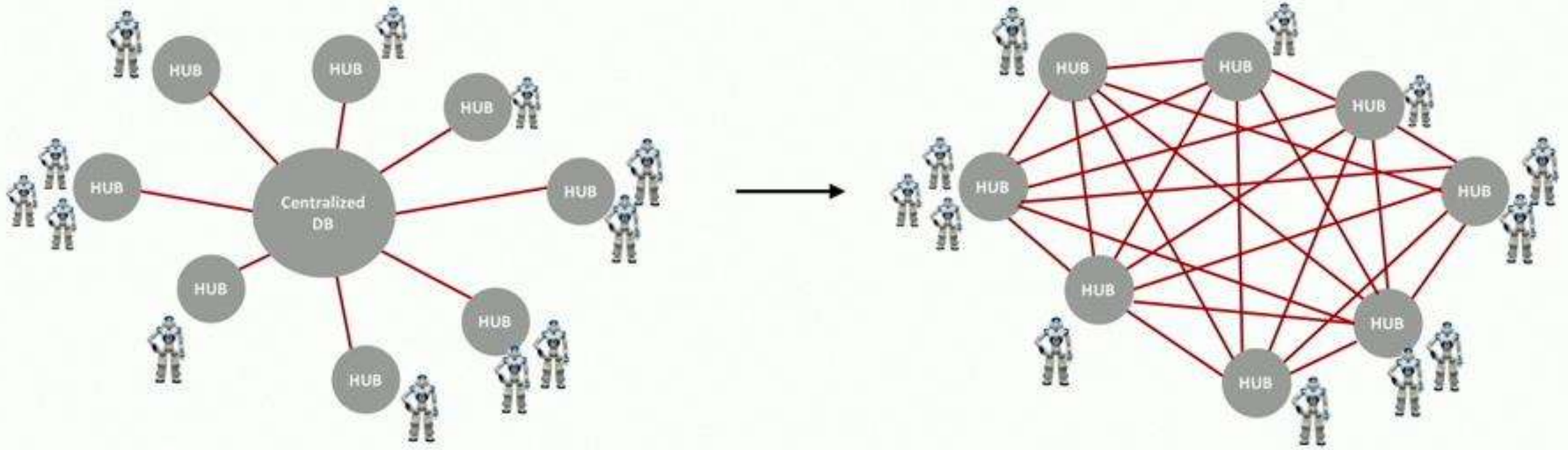
Scaling up and Data Privacy

Challenges:

- Information may be decentralized and stored in separate databases (hospitals)
- Information may be sensitive and required to be kept private (personal data)

Approach:

- Secure decentralized learning without sharing of raw personal data

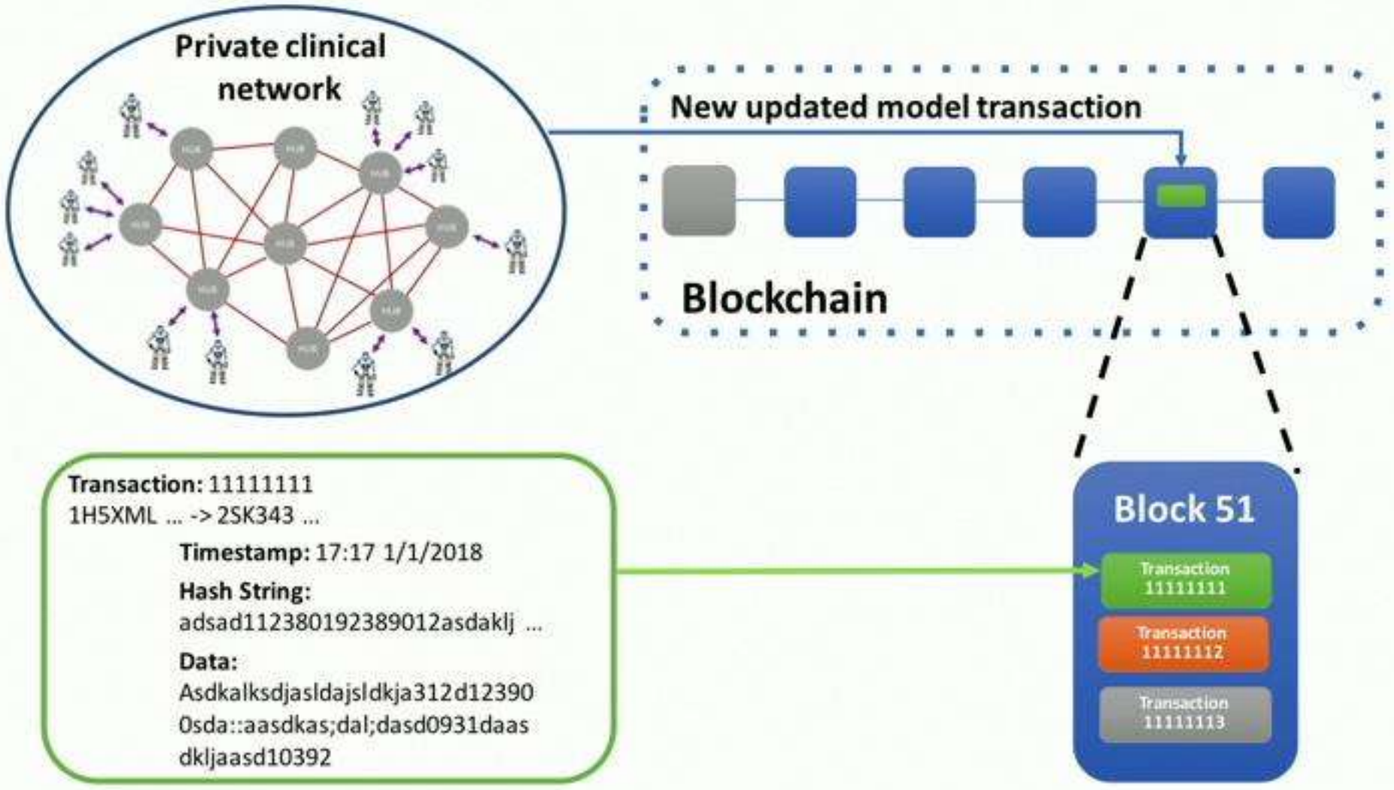


centralized learning

de-centralized learning

E. C., Ferrer, O., Rudovic, T., Hardjono, A., Pentland (2018) RoboChain: A Secure Data-Sharing Framework for Human-Robot Interaction. eTELEMED

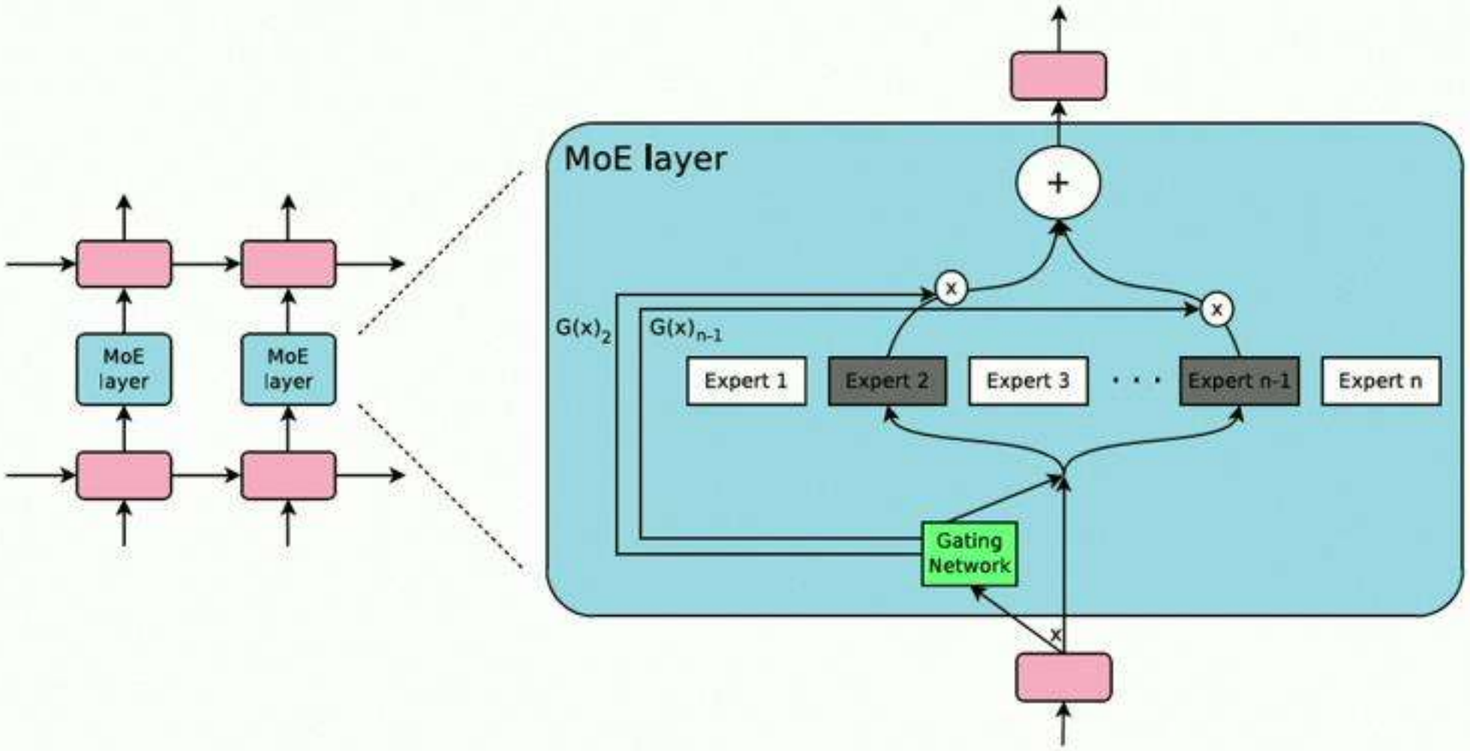
Scaling up and Data Privacy



- A robot sends a transaction to the blockchain
- The information is readable only by the participants of the private network

E. C., Ferrer, O., Rudovic, T., Hardjono, A., Pentland (2018) RoboChain: A Secure Data-Sharing Framework for Human-Robot Interaction. eTELEMED

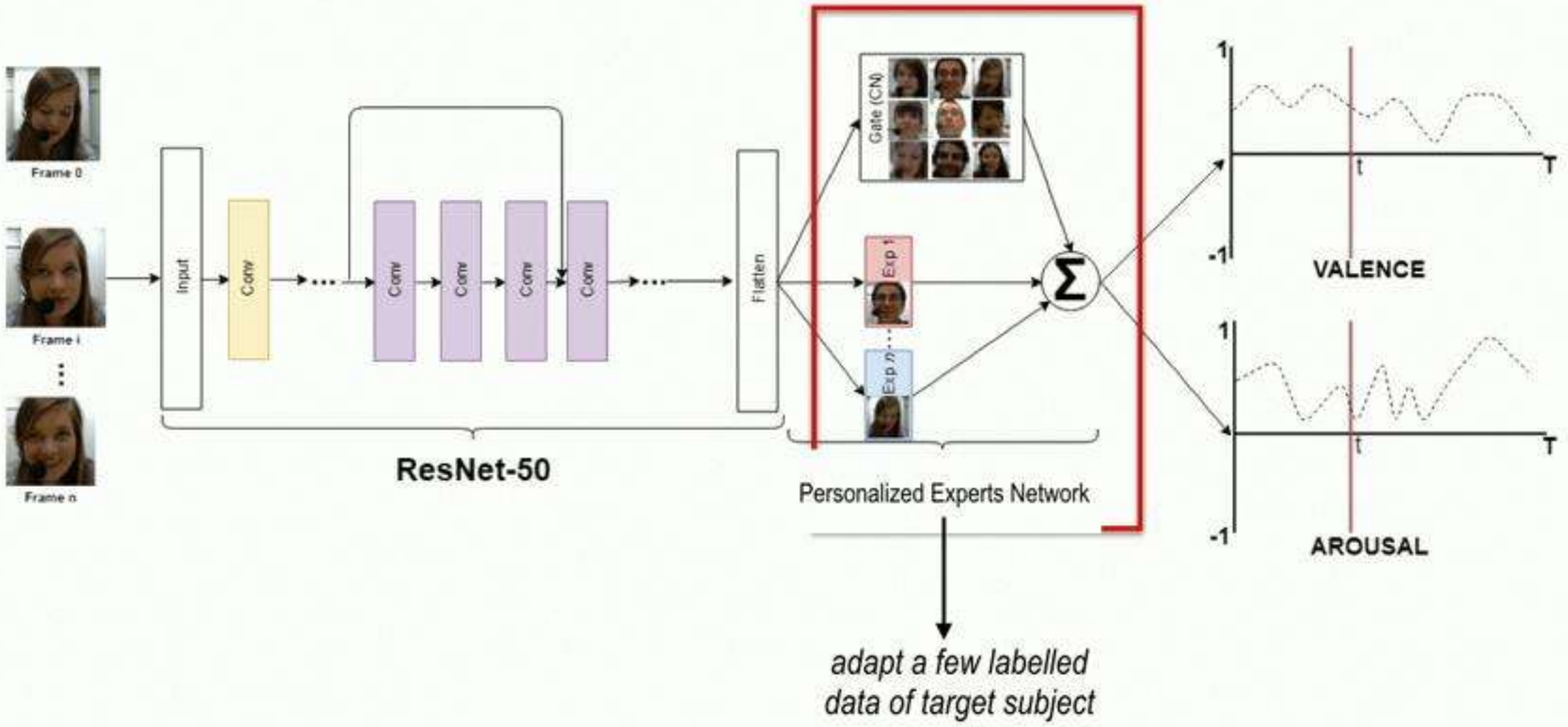
Efficient Deployment



Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." ICLR, 2017

Feffer, M., Rudovic, O., R.W. Picard. "A Mixture of Personalized Experts for Human Affect Estimation." ICMLDM. Springer, 2018.

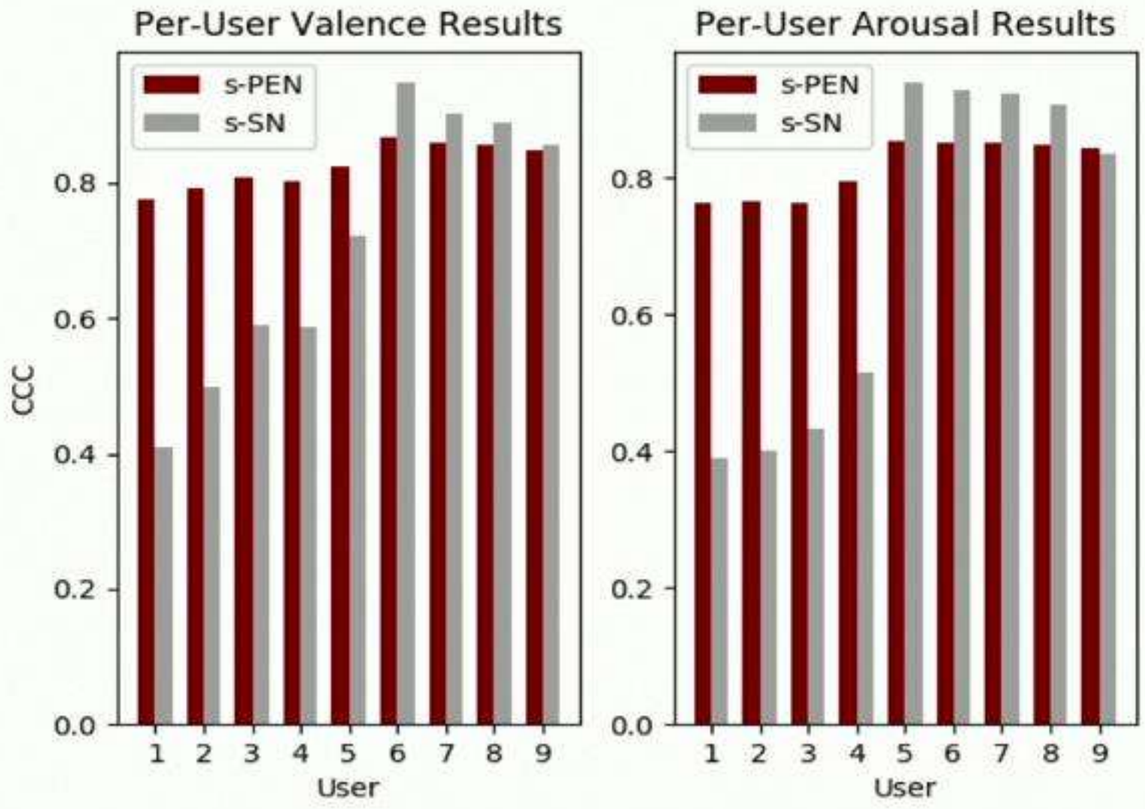
Efficient Deployment



Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." ICLR, 2017

Feffer, M., Rudovic, O., R.W. Picard. "A Mixture of Personalized Experts for Human Affect Estimation." ICMLDM. Springer, 2018.

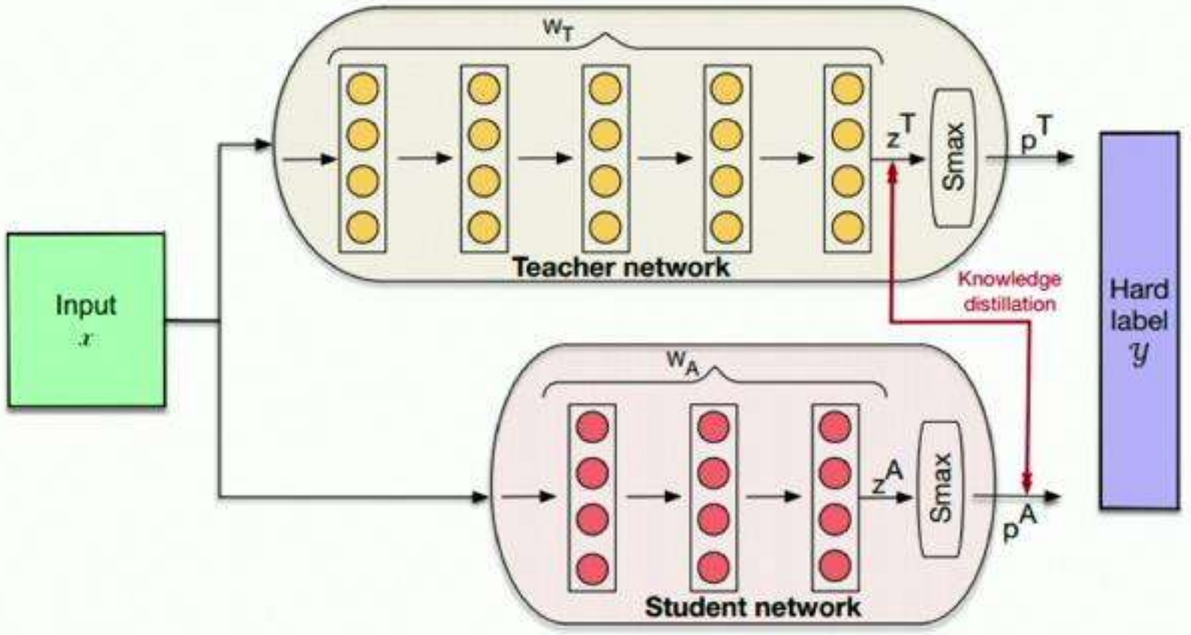
Efficient Deployment



Concordance Correlation Coefficient (CCC) - a measure of agreement between model predictions and human experts

Efficient Deployment

Setup: Apprentice

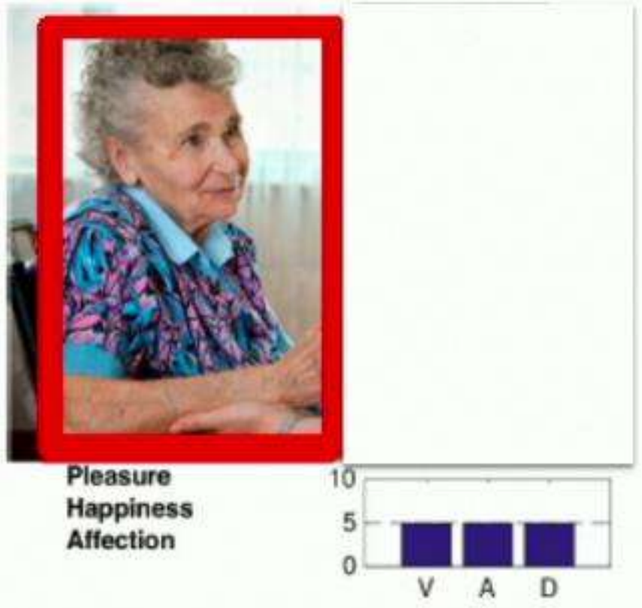


$$\mathcal{L}(x; W_T, W_A) = \alpha\mathcal{H}(y, p^T) + \beta\mathcal{H}(y, p^A) + \gamma\mathcal{H}(z^T, p^A)$$

Teacher loss Student loss Distillation

G., Hinton, O., Vinyals, J., Dean. (2015) "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531

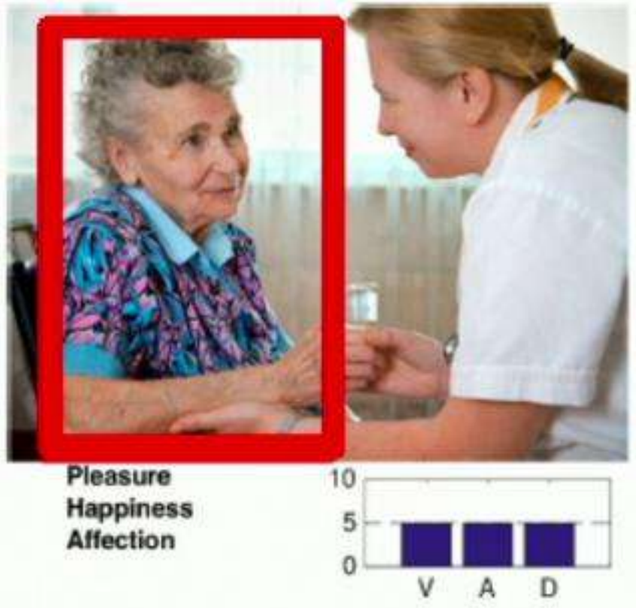
Context-sensitive Learning



Source: K. Ronak, J. M. Alvarez, A. Recasens, A. Lapedriza. "Emotion recognition in context." CVPR 2017

Context-sensitive Learning

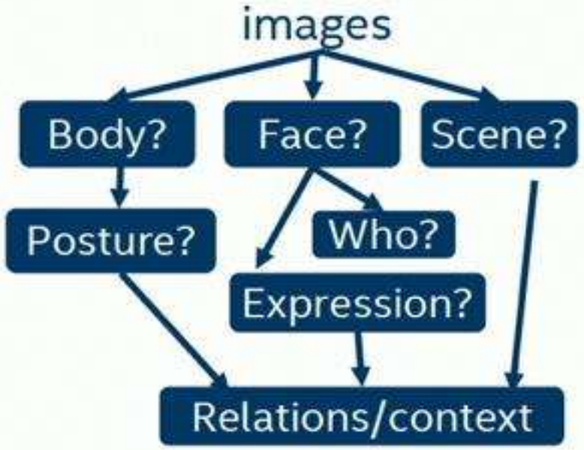
Context is critical for perception



Source: K. Ronak, J. M. Alvarez, A. Recasens, A. Lapedriza. "Emotion recognition in context." CVPR 2017

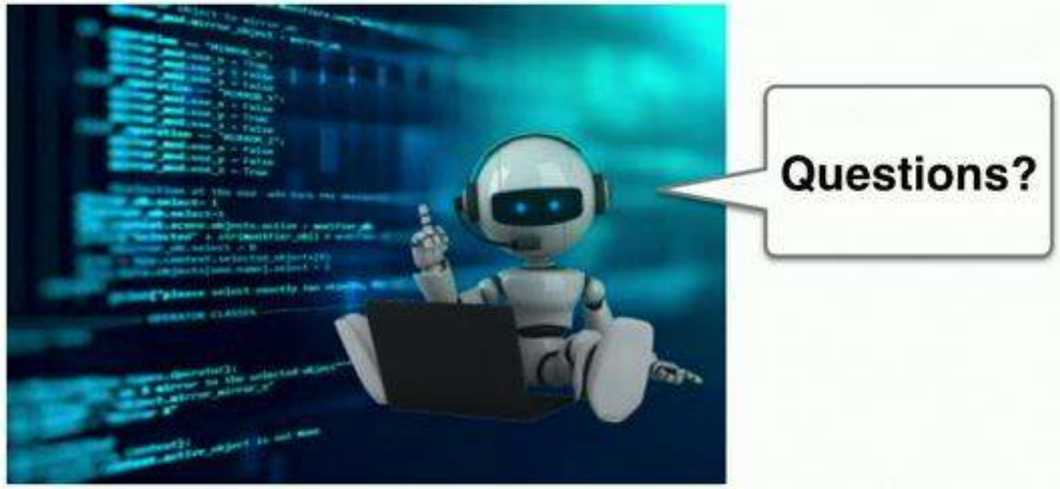
Context-sensitive Learning

Context is critical for perception



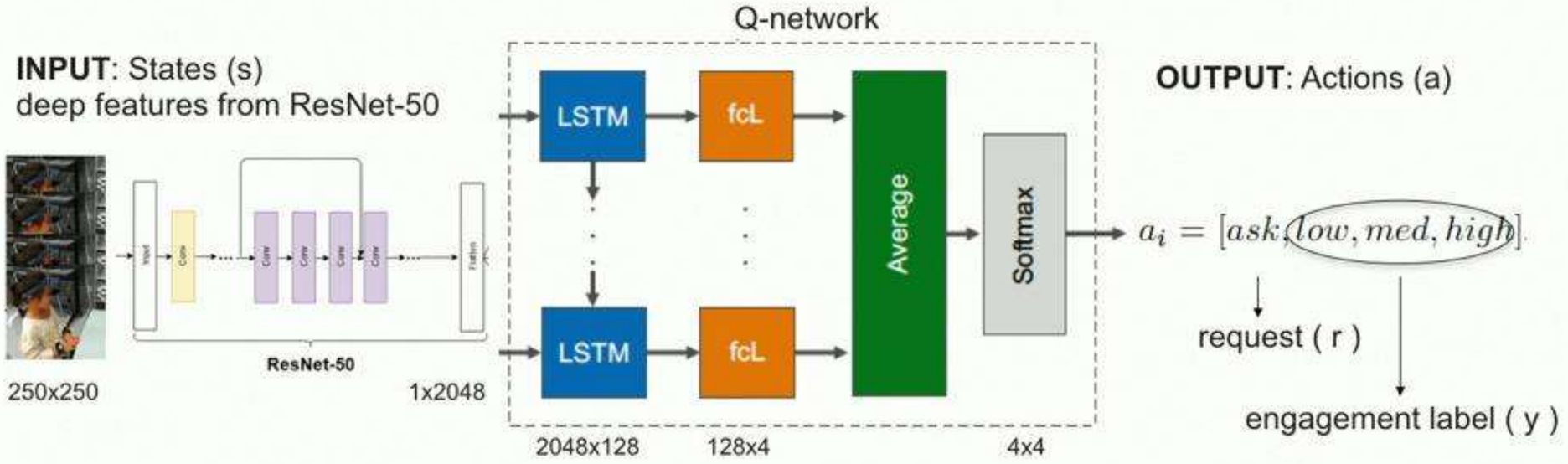
Source: K. Ronak, J. M. Alvarez, A. Recasens, A. Lapedriza. "Emotion recognition in context." CVPR 2017

Thank you!



Method: Personalized Deep Reinforcement Learning

Step 1: Learn a population-level policy from data of training children

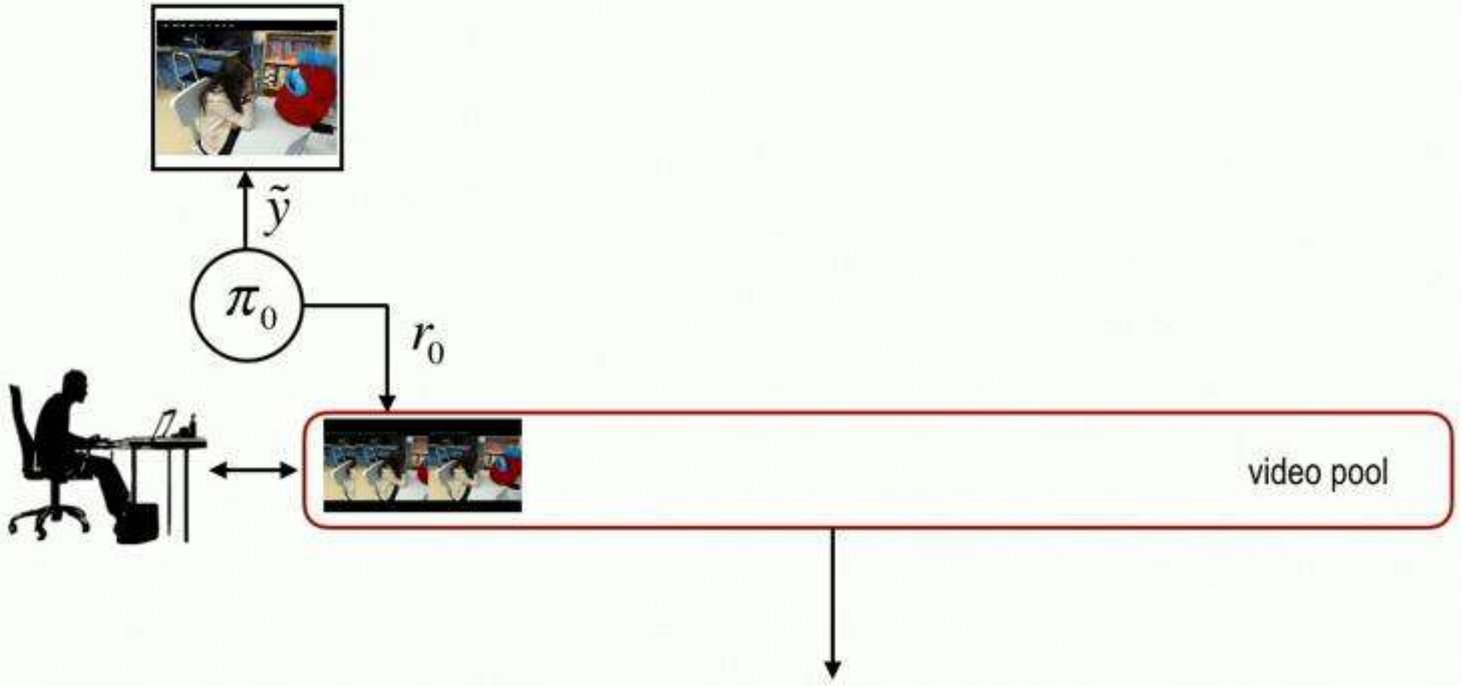


Rewards:

$$R_i = \begin{cases} R_{req} = -0.05, \text{if } r_i = 1 \\ R_{cor} = +1, \quad \text{if } r_i = 0 \wedge y^* = y \\ R_{inc} = -1, \quad \text{if } r_i = 0 \wedge y^* \neq y \end{cases}$$

Personalized Deep Reinforcement Learning

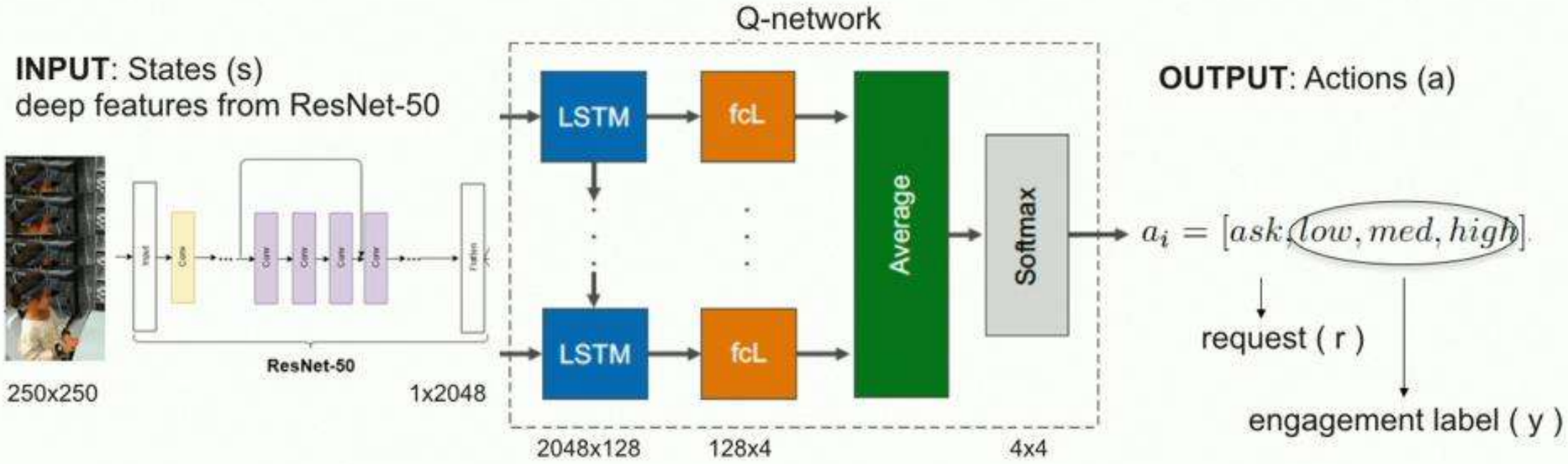
Step 2: Personalize the population-level policy (π_0) to the target kid



policy update π_j: $\hat{\Theta}_j \leftarrow \min_{\Theta_{j-1}} \sum_{v_r \in \{S_j\}} L_B^{(v_r)}(\Theta)$

Method: Personalized Deep Reinforcement Learning

Step 1: Learn a population-level policy from data of training children



Rewards:

$$R_i = \begin{cases} R_{req} = -0.05, \text{if } r_i = 1 \\ R_{cor} = +1, \quad \text{if } r_i = 0 \wedge y^* = y \\ R_{inc} = -1, \quad \text{if } r_i = 0 \wedge y^* \neq y \end{cases}$$

Results: Average Performance

Model	Session	ACC [%]									F1 [%]								
		S1	S2	S3	S4	S5	S6	S7	S8	AVE.	S1	S2	S3	S4	S5	S6	S7	S8	AVE.
LSTM-S	INIT	65	59	61	60	62	60	48	67	60	31	40	36	37	38	40	32	35	35
	RND	65	**67**	69	65	67	64	55	68	65	31	**46**	40	35	40	41	**41**	33	38
	ENTR	65	61	65	**70**	73	65	**57**	**75**	66	31	42	**43**	37	38	43	40	35	39
	LCONF	65	66	62	69	72	69	52	**75**	66	31	40	39	**41**	37	42	**41**	35	38
	SMAR	65	66	**70**	66	**75**	**70**	55	73	**68**	31	45	41	36	**48**	**47**	37	**36**	**40**
TC-DQL	INIT	72	57	66	61	74	64	56	71	65	32	42	33	40	43	47	35	31	39
	RND	72	67	60	64	72	70	63	81	69	32	38	39	32	**46**	40	41	41	40
	RL	72	**70**	**74**	**70**	**82**	**75**	**65**	**85**	**74**	32	**45**	**42**	**46**	45	**50**	**48**	**46**	**44**
	REQ [%]	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6

average % of requests (~10 video clips per session)

Setting: 22 kids for training/validation of the population policy (**INIT**), 21 kids for testing/policy adaptation

Supervised LSTM (LSTM-S) updated using data selected by heuristic Active Learning strategies:

RND - random requests
ENTR - max entropy
LCONF - least confidence
SMAR - smallest margin

Personalized Deep Reinforcement Learning

Goal: Learn a personalized policy that "decides" whether to request a video label (for further learning) or estimate engagement!