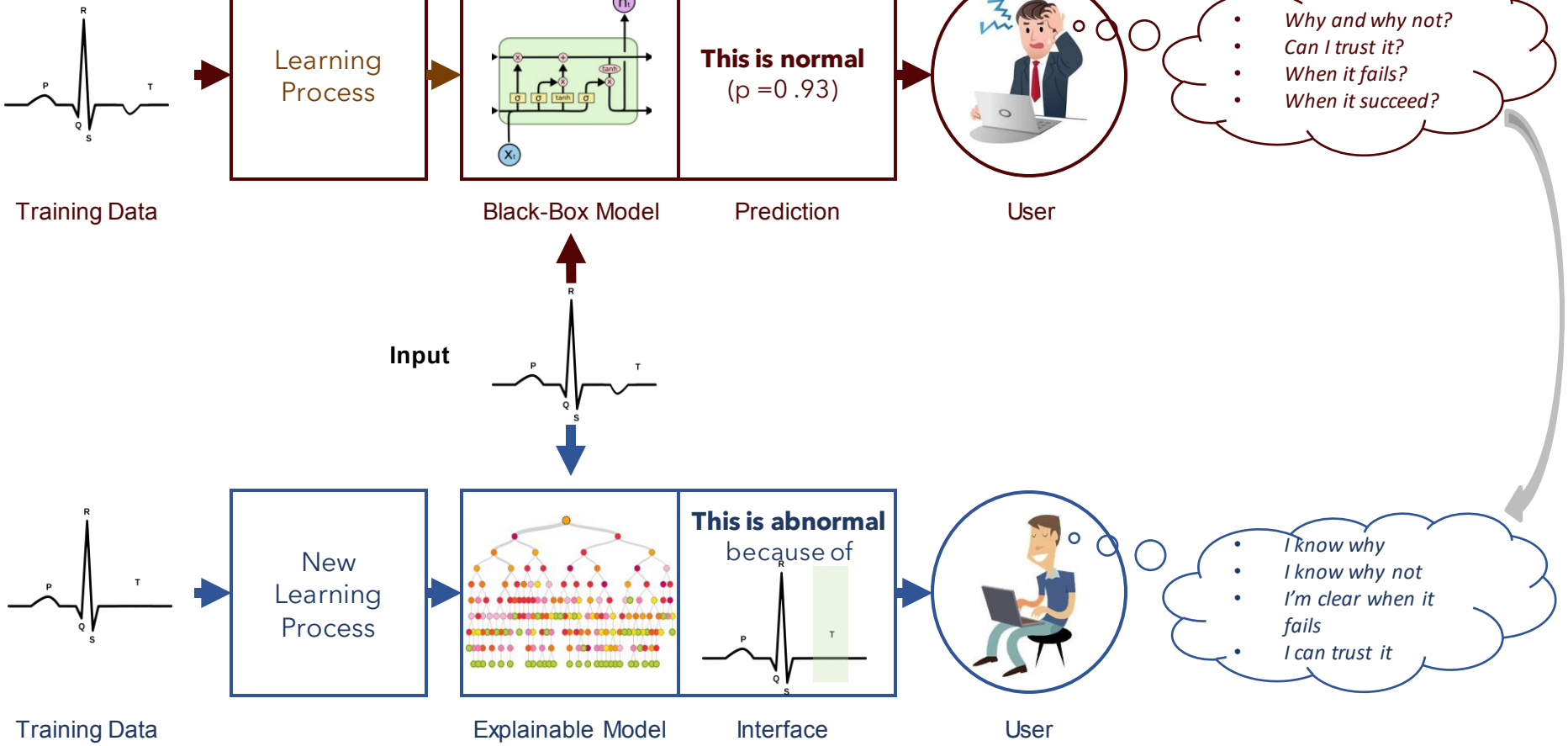
NEC Laboratories America
Relentless passion for innovation

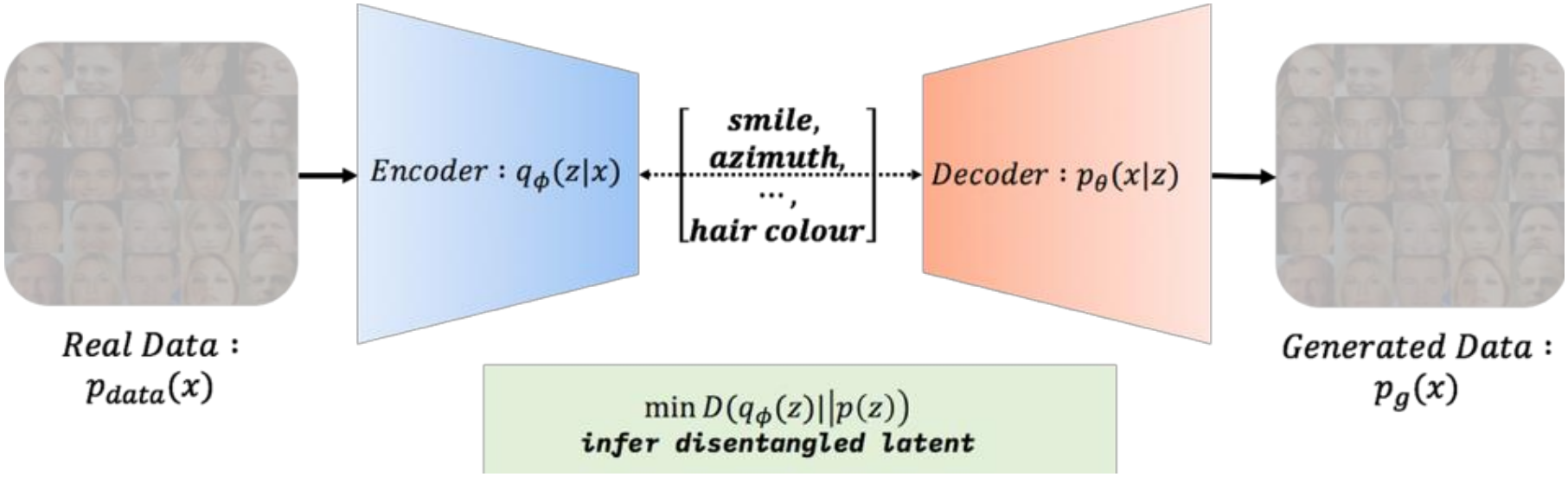
Towards Learning Disentangled Representations for Time Series

Yuening Li, Zhengzhang Chen, Daochen Zha, Mengnan Du,
Jingchao Ni, Denghui Zhang, Haifeng Chen, Xia Hu

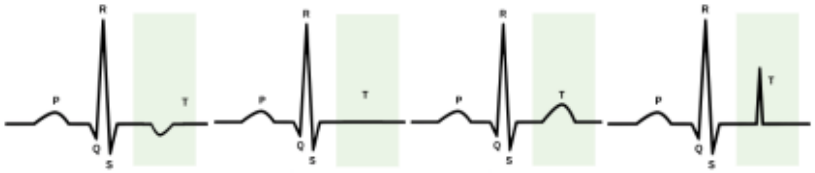
Interpretable Representation for Downstream Tasks



Decomposition: A Generalization of Disentanglement



A semantic factor controls the eye-glasses of a human facial image.



Can we decompose time series?

Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, **ICLR 2018**

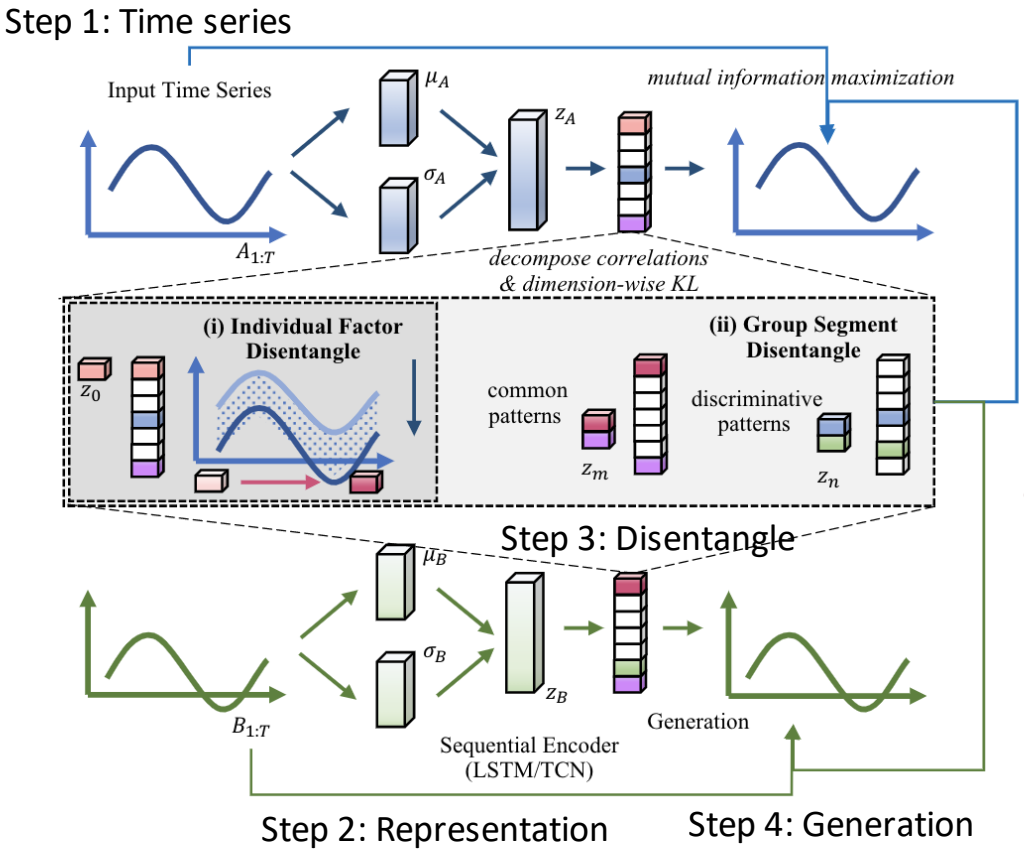
Challenges

Sequential data structure introduces complex temporal correlations; → Data Structure

How to represent time series with interpretability and expose semantic meanings?

Interpretable semantic concepts for time-series often rely on **multiple factors** instead of individuals. → Hierarchy

Overview of DTS



Individual Factor Disentangle

Latent semantic variables are independent if the change of the variable are relatively invariant to others.

Provide interpretations as fine-grained individuals

Group Segment Disentangle

All the pairs of latent segments are independent.

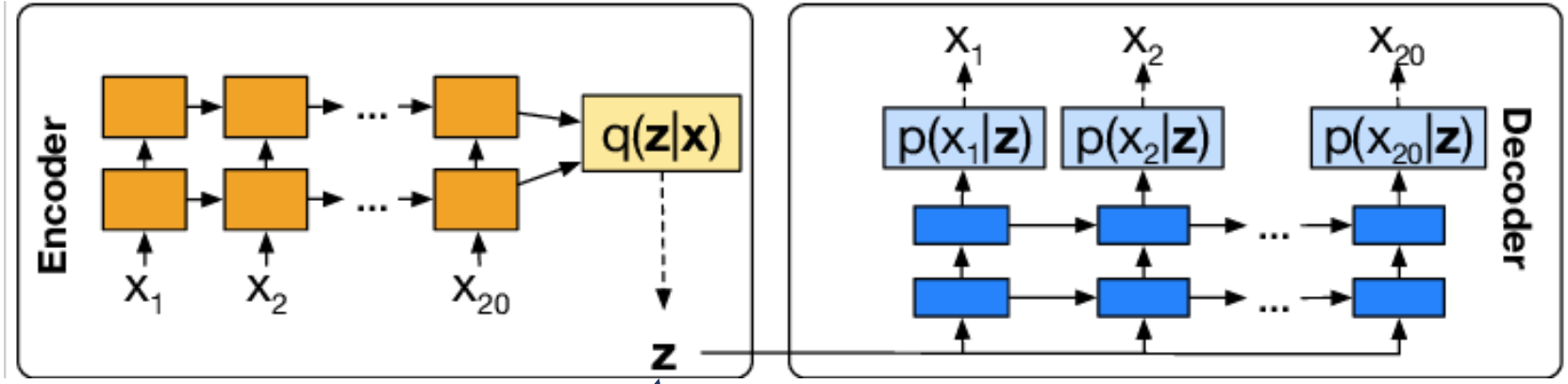
Provide interpretations as coarse-grained factors

Individual Factor Disentangle: KL-Vanishing Problem

1. Generative models (decoders) often have strong expressiveness;

2. The reconstruction term in the objective dominate the KL-divergence term;

$$\mathcal{L}_{\mathrm{ELBO}}(\mathrm{x}) = -D_{KL}\left(q_\phi(Z|\mathrm{x}_{1:T})||p(Z)\right) + \mathbb{E}_{q_\phi(Z|\mathrm{x}_{1:T})}[\log p_\theta(\mathrm{x}_{1:T}|Z)]$$



3. The model would generate time-series without making effective use of the latent codes;

Information Preference

4. Latent variables will become independent of the observations when the KL-divergence collapses to zero.

Individual Factor Disentangle: beta-VAE

- The lower bound to the log likelihood of vanilla VAE:

$$\mathcal{L}_{\text{ELBO}}(\mathrm{x}) = -D_{KL}\big(q_\phi(Z|\mathrm{x}_{1:T})||p(Z)\big) + \mathbb{E}_{q_\phi(Z|\mathrm{x}_{1:T})}[\log p_\theta(\mathrm{x}_{1:T}|Z)]$$

- Beta-VAE attempts to learn a disentangled representation by optimizing a heavily penalized objective with β on the KL term:

$$\mathcal{L}_{\beta-\text{ELBO}}(\mathrm{x}) = -\beta D_{KL}\big(q_\phi(Z|\mathrm{x}_{1:T})||p(Z)\big) + \mathbb{E}_{q_\phi(Z|\mathrm{x}_{1:T})}[\log p_\theta(\mathrm{x}_{1:T}|Z)]$$

heavier penalty

- However, pushing Gaussian clouds away from each other in the latent space becomes meaningless if latent distributions are unhooked with the observation space.

Our Individual Factor Disentanglement Strategy

ELBO Total Correlation-Decomposition:

$$D_{KL}(q(Z|\mathrm{x}_{1:T})||p(Z)) = \underbrace{D_{KL}(q(Z,\mathrm{x}_{1:T})||q(Z)p(\mathrm{x}_{1:T}))}_{\text{(i) index-code mutual information}} + \underbrace{D_{KL}\big(q(Z)||\prod_j q(z_j)\big)}_{\text{(ii) total correlation}} + \underbrace{\sum_j D_{KL}\big(q(z_j)||p(z_j)\big)}_{\text{(iii) dimension-wise KL}}$$

A new perspective of KL vanishing:

- Heavier penalty on the ELBO tends to neglect the mutual information between Z and x;
- Mutual information becomes vanishingly small;
- Increasing β may intensify the mutual information vanishing problem: better quality of disentanglement companion with heavier penalty on the mutual information;

Can we alleviate the mutual information vanishing?

Our Individual Factor Disentanglement Strategy

- To encourage the model to use the latent codes, we add a MI-maximization term as:

$$\mathcal{L}_{\text{ELBO}}(\mathrm{x}) = -D_{KL}\big(q_\phi(Z|\mathrm{x}_{1:T})||p(Z)\big) + \alpha I_{q_\varphi}(\mathrm{x}_{1:T};Z) + \mathbb{E}_{q_\phi(Z|\mathrm{x}_{1:T})}\left[\log p_\theta(\mathrm{x}_{1:T}|Z)\right]$$

- Compare with the TC decomposition, we found:

$$D_{KL}(q(Z|\mathrm{x}_{1:T})||p(Z)) = D_{KL}(q(Z,\mathrm{x}_{1:T})||q(Z)p(\mathrm{x}_{1:T})) + D_{KL}\big(q(Z)||\textstyle\prod_j q(z_j)\big) + \sum_j D_{KL}\big(q(z_j)||p(z_j)\big)$$

index-code mutual information

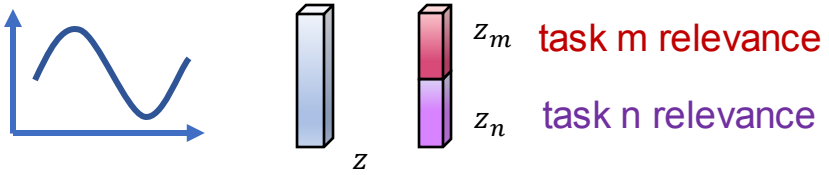
Play the same role, but the optimization directions are contrary!

To enforce the model to preserve the disentangle property while alleviating the KL vanishing, we have:

$$\mathcal{L}_{\text{ELBO}}(\mathrm{x}) = -\beta D_{KL}\big(q(Z)||\textstyle\prod_j q(z_j)\big) - \beta \sum_j D_{KL}\big(q(z_j)||p(z_j)\big) + (\alpha-\beta)D_{KL}\big(q_\phi(Z)||p(Z)\big) + \mathbb{E}_{q_\phi(Z|\mathrm{x}_{1:T})}\left[\log p_\theta(\mathrm{x}_{1:T}|Z)\right]$$

Latent Group Segment Disentanglement

- **Goal**: to learn decomposed semantic segments that contain batches of latent variables.
- **Solution**: Gradient Reversal Layer (GRL)



$$\mathbb{E}_m(\varphi_y, \theta_m, \theta_n) = \mathbb{E}(C_m(z_m; \theta_m), y_m) - \lambda \mathbb{E}(C_n(z_m; \theta_n), y_n) \quad (\widehat{\theta_f}, \widehat{\theta_y}) = \arg\min_{\theta_f, \theta_y} \mathbb{E}(\theta_f, \theta_y, \widehat{\theta_d})$$

$$\mathbb{E}_n(\varphi_y, \theta_m, \theta_n) = \mathbb{E}(C_n(z_n; \theta_n), y_n) - \lambda \mathbb{E}(C_m(z_n; \theta_m), y_m) \qquad \widehat{\theta_d} = \arg\max_{\theta_d} \mathbb{E}(\widehat{\theta_f}, \widehat{\theta_y}, \theta_d)$$

- Transferable Anomaly Detection from different domains as a concrete example, the empirical errors are:

error on source $\epsilon_S(h) = \mathbb{E}_{z_y \sim Z_S}[C(z_y) - h(z_y)] + \mathbb{E}_{z_d \sim Z_S}[C(z_d) - h(z_d)]$

error on target $\epsilon_T(h) = \mathbb{E}_{z_y \sim Z_T}[C(z_y) - h(z_y)] + \mathbb{E}_{z_d \sim Z_T}[C(z_d) - h(z_d)]$

anomaly detector — detector hypothesis — domain classifier — classifier hypothesis

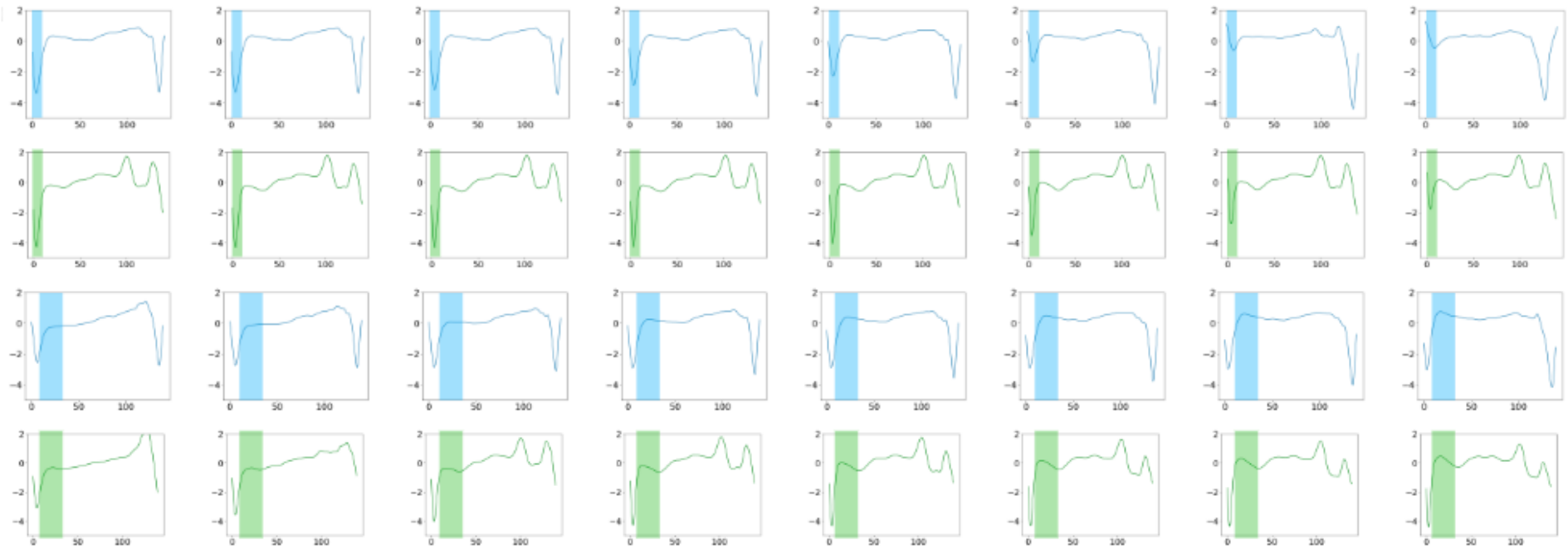
Quantitative Results for Transferable Adaptation Classifacation

Problem	No Adaptation	R-DANN	VRADA	CoDATS	DTS	Train on Target
HAR 2 → 11	83.3	80.7	64.1	74.5	**84.3**	100
HAR 7 → 13	89.9	75.3	78.3	96.5	**98.1**	100
HAR 12 →16	41.9	35.1	61.7	**77.5**	72.9	100
HAR 9 → 18	31.1	56.6	59.8	85.8	**89.8**	100
HAR 18 →23	89.3	78.2	72.9	86.2	**94.9**	100
HAR 6 → 23	52.9	79.1	78.2	94.7	**94.9**	100
HAR Average	69.2	70.2	70	88.4	**93.5**	100
HHAR 1 →3	77.8	85.1	81.3	93.2	**93.7**	99.2
HHAR 3 →5	68.8	85.4	82.3	95.6	**95.9**	99
HHAR 4 →5	60.4	70.4	71.6	94.2	**94.9**	99
HHAR 3 →8	77.8	82.8	82.2	93.4	**94.7**	99.3
HHAR 5 →8	95.3	82.5	87.5	97.1	**97.9**	99.3
HHAR Average	64.8	68.7	68.3	88.3	**89.9**	99
WISDM AR1 → 11	71.7	55.6	55	71.7	**91.7**	98.3
WISDM AR4 → 15	78.2	69.2	82.7	81.4	**82.9**	100
WISDM AR2 → 32	60.1	49	66.7	67.3	**70.7**	100
WISDM AR1 → 7	68.5	44.8	63	70.9	**72.7**	96.4
WISDM Average	56.8	48.3	61.2	70	**81.7**	98.5
uWave 3 → 5	82.7	63.7	32.4	93.8	**95.6**	100
uWave 2 → 7	85.1	53.9	12.2	91.4	**98.9**	100
uWave 3 → 7	95.5	64	30.4	92	**98.9**	100
uWave 4 → 5	83.3	35.4	12.8	**99.1**	96.7	100
uWave 7 → 8	95.2	49.7	12.5	93.8	**96.7**	100
uWave Average	91	48.4	19.7	94.3	**97.7**	100

Target detection accuracy for time-series domain adaptation (from source to target) between different participants.

- Each dataset (HAR, HHAR, WISDM AR and uWave) contains sequential accelerometer data from different participants.
- DTS boosts the performance by obtaining domain-invariant transferable components as common knowledge.

Traversal Results



Latent traversal plots from DTS on ECG. All figures of latent codes traversal each block corresponds to the traversal of a single latent variable while keeping others fixed. Blue and green denote two time-series with different sequential patterns.

Thank Everyone for Attending!

NEC Laboratories America

Relentless passion for innovation