A Neural Network Alternative to Non-negative Audio Models

Paris Smaragdis[#,*]

Shrikant Venkataramani[*]

[*]*University of Illinois at Urbana Champaign*
[#]*Adobe Research*

ICASSP 2017

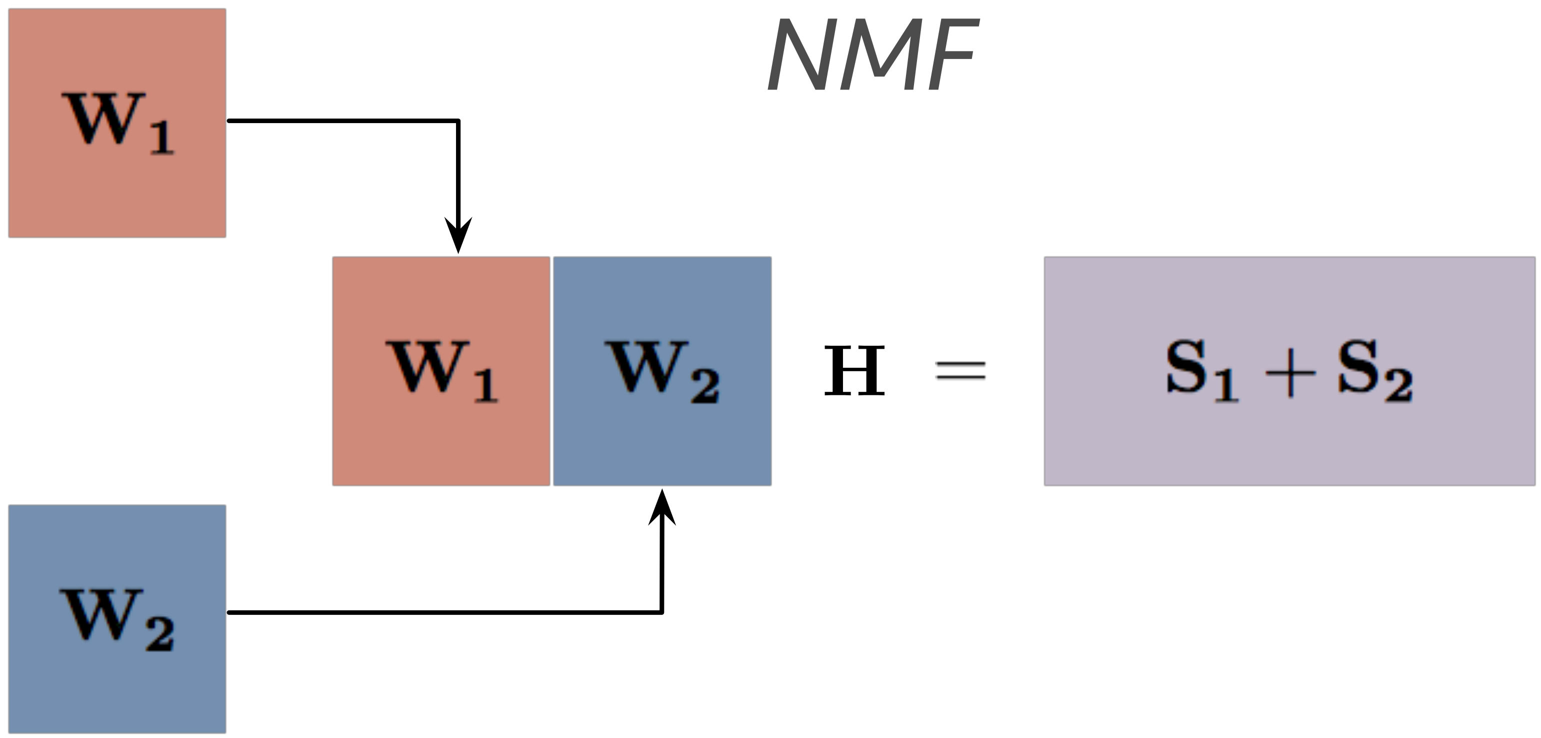
Motivation

- Supervised single channel source separation
  - Using models trained from clean sounds
- Two dominant approaches
  - Non-negative Matrix Factorization (NMF)
    - Reusable and interpretable models
  - Deep learning
    - State of the art results, Non-transferable models
- Neural network formulation of NMF models
  - Maintaining reusability, taking advantage of deep-learning structures

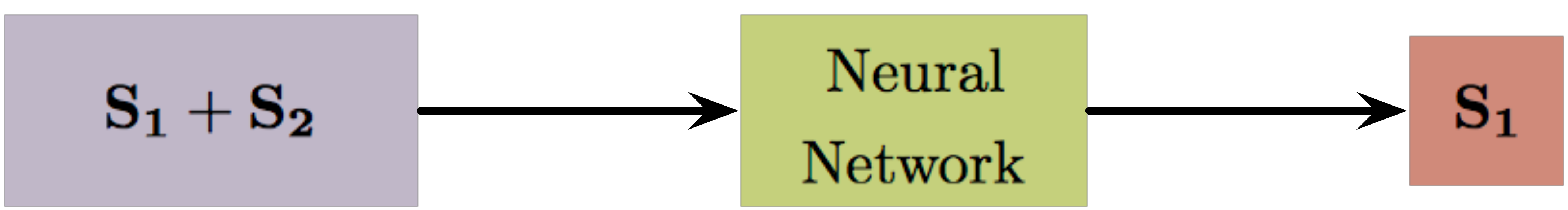
Transferable Models

- Being able to plug-in trained models

NMF



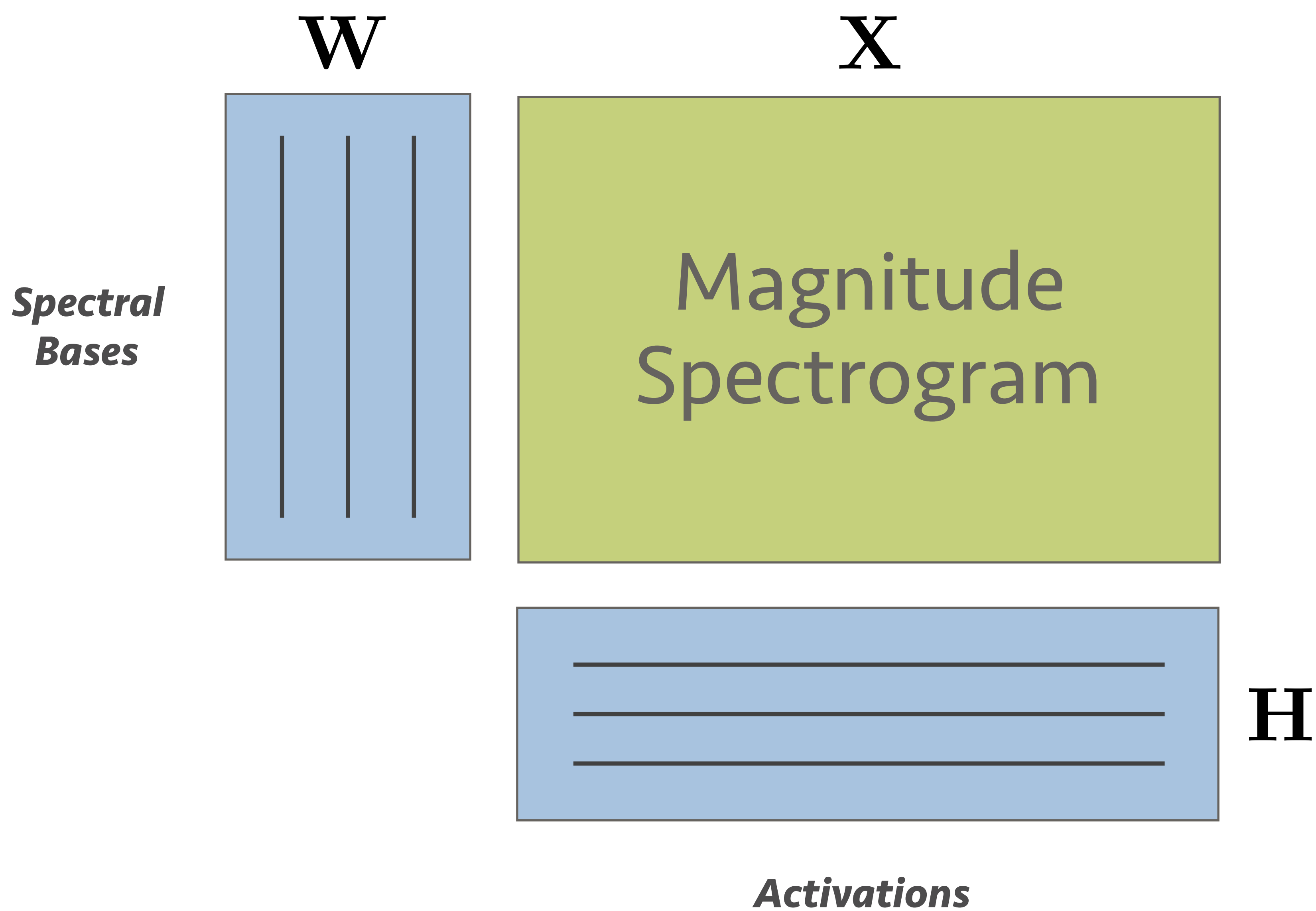
DNN



Learning an NMF model

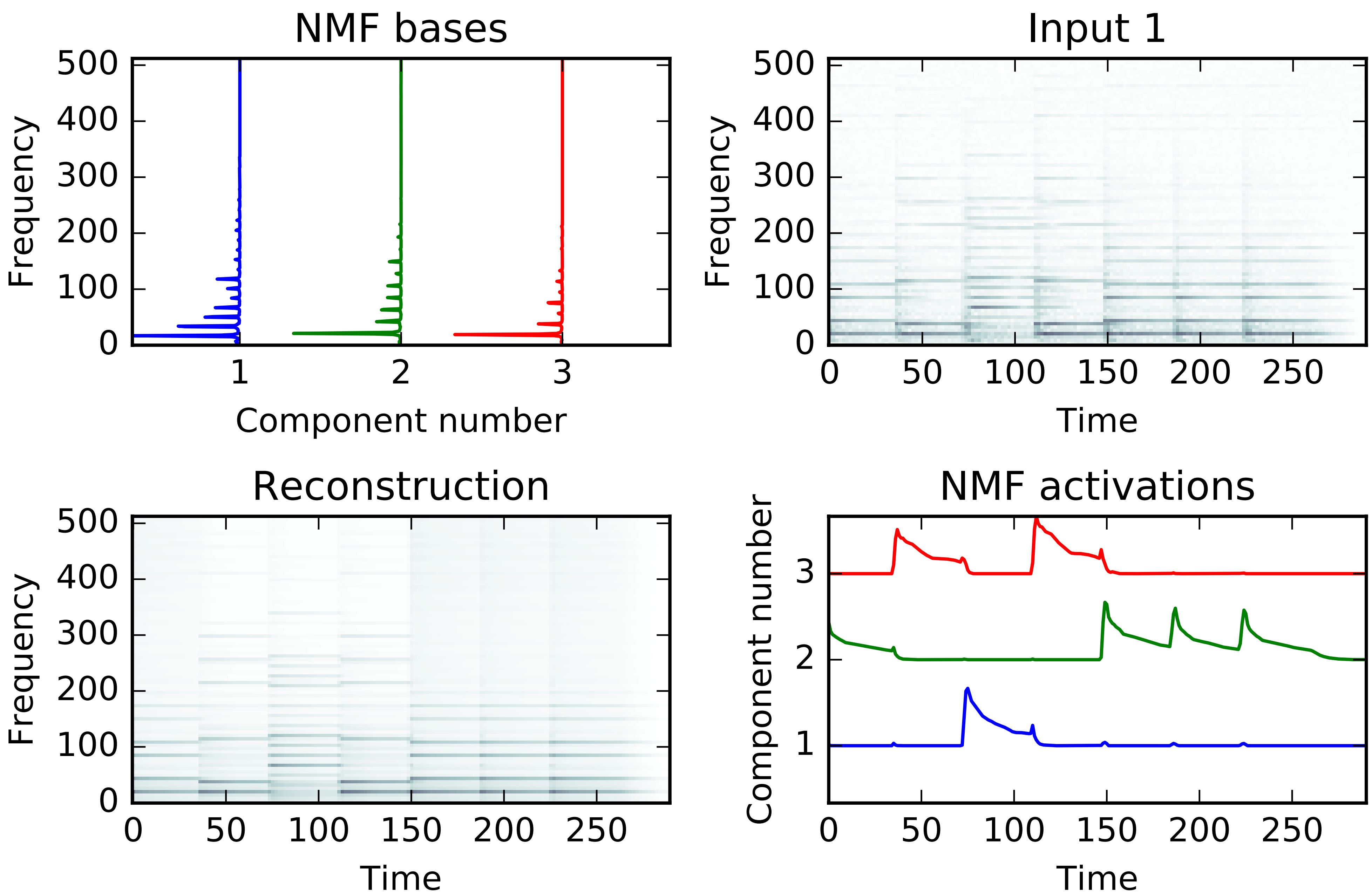
- Learning spectral bases from spectrograms.

$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H} \qquad \mathbf{X}, \mathbf{W}, \mathbf{H} \in \mathbb{R}^{+}$$



NMF in action

- Analyzing piano notes



NMF as an Auto-encoder

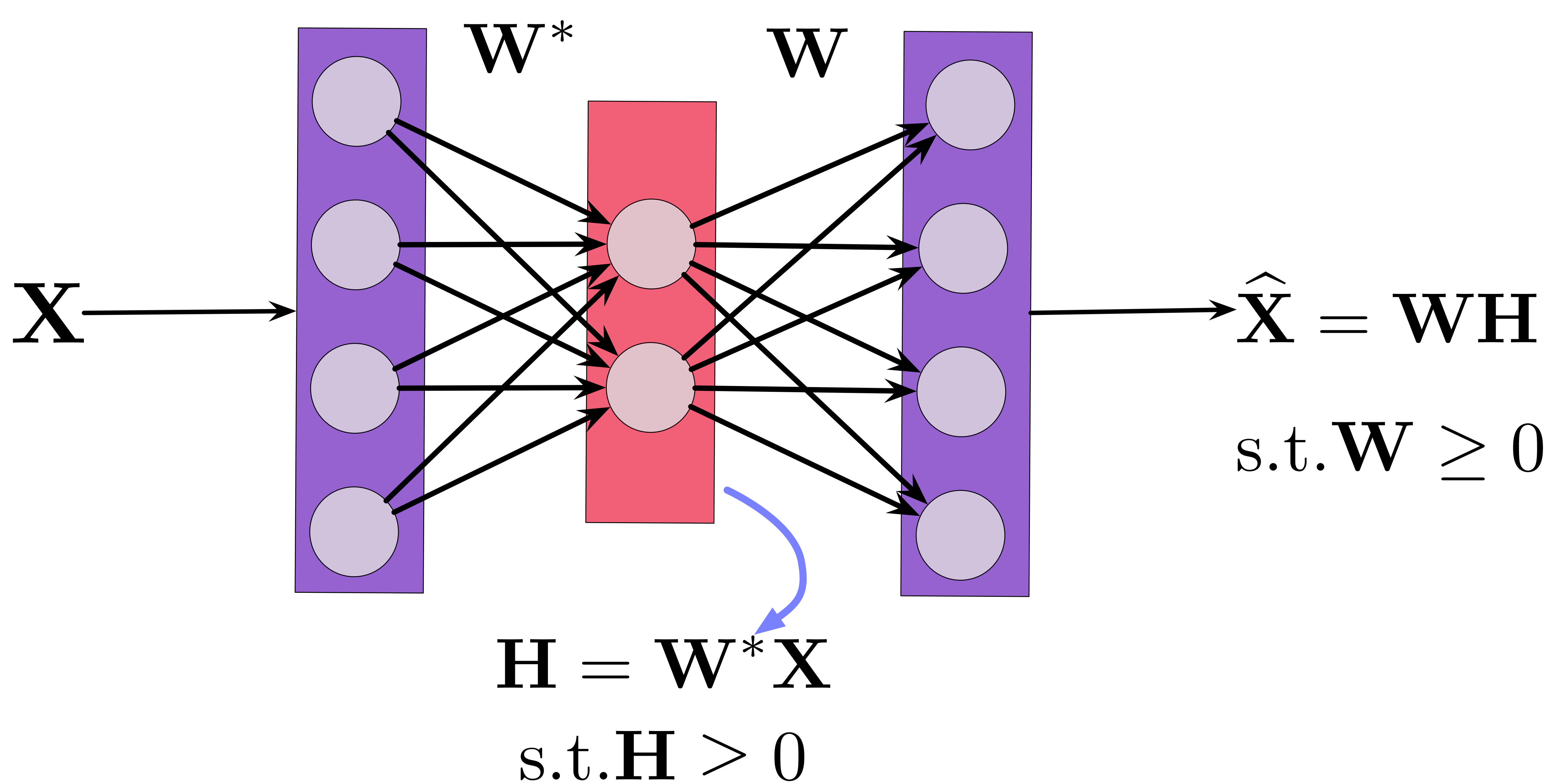
NMF

$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H}$$

Non-negative Auto-encoder (NAE)

$$\mathbf{H} = \mathbf{W}^* \cdot \mathbf{X} \text{ such that } \mathbf{H} \geq 0$$

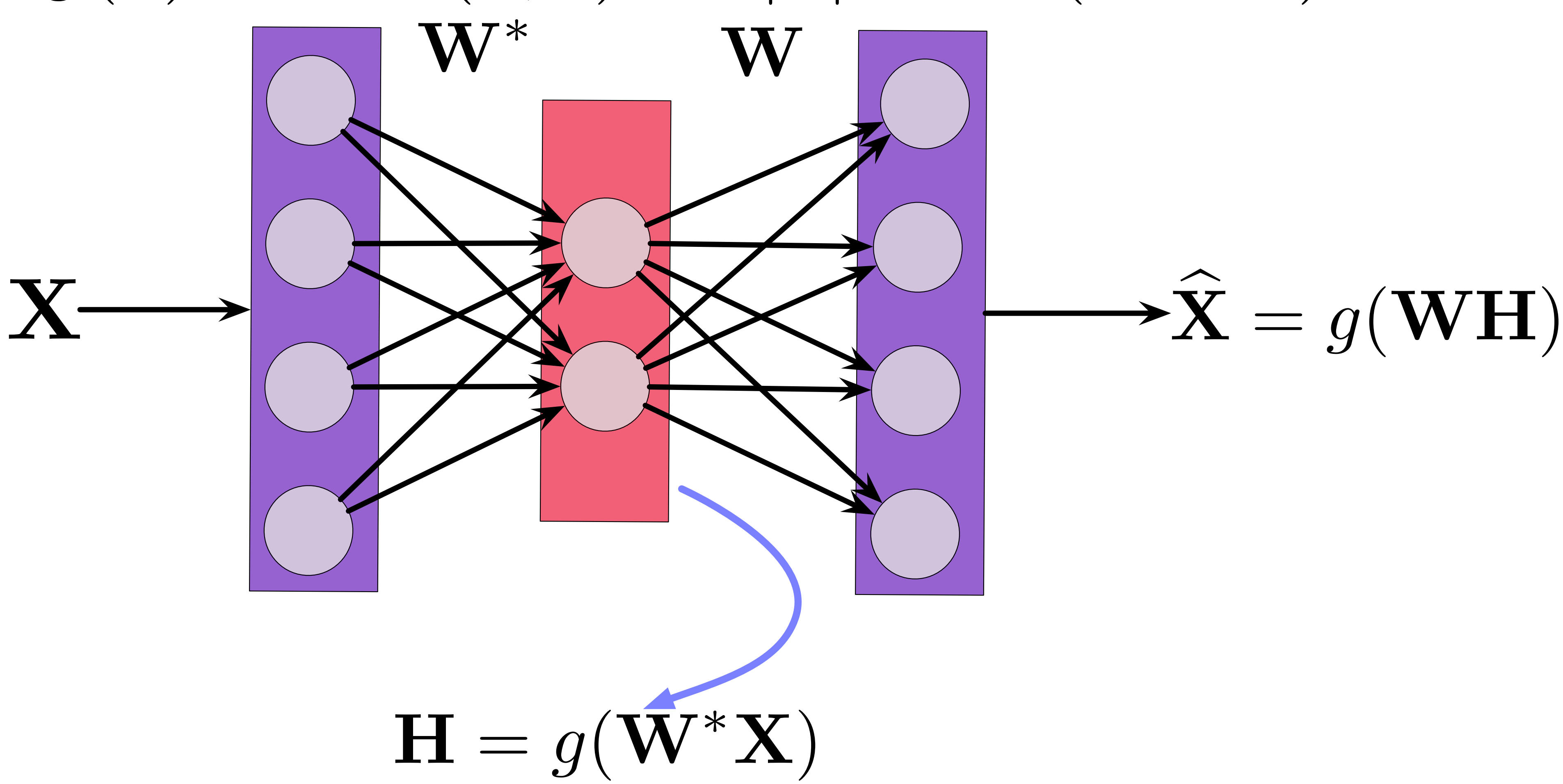
$$\widehat{\mathbf{X}} = \mathbf{W} \cdot \mathbf{H} \text{ such that } \mathbf{W} \geq 0$$



Removing Non-negativity constraints

- Enforcing non-negativity is cumbersome.
- Non-negative layer outputs
  - Results in a magnitude spectrogram at the output
  - Results in a non-negative activation at hidden layer
  - Can be enforced with an activation function

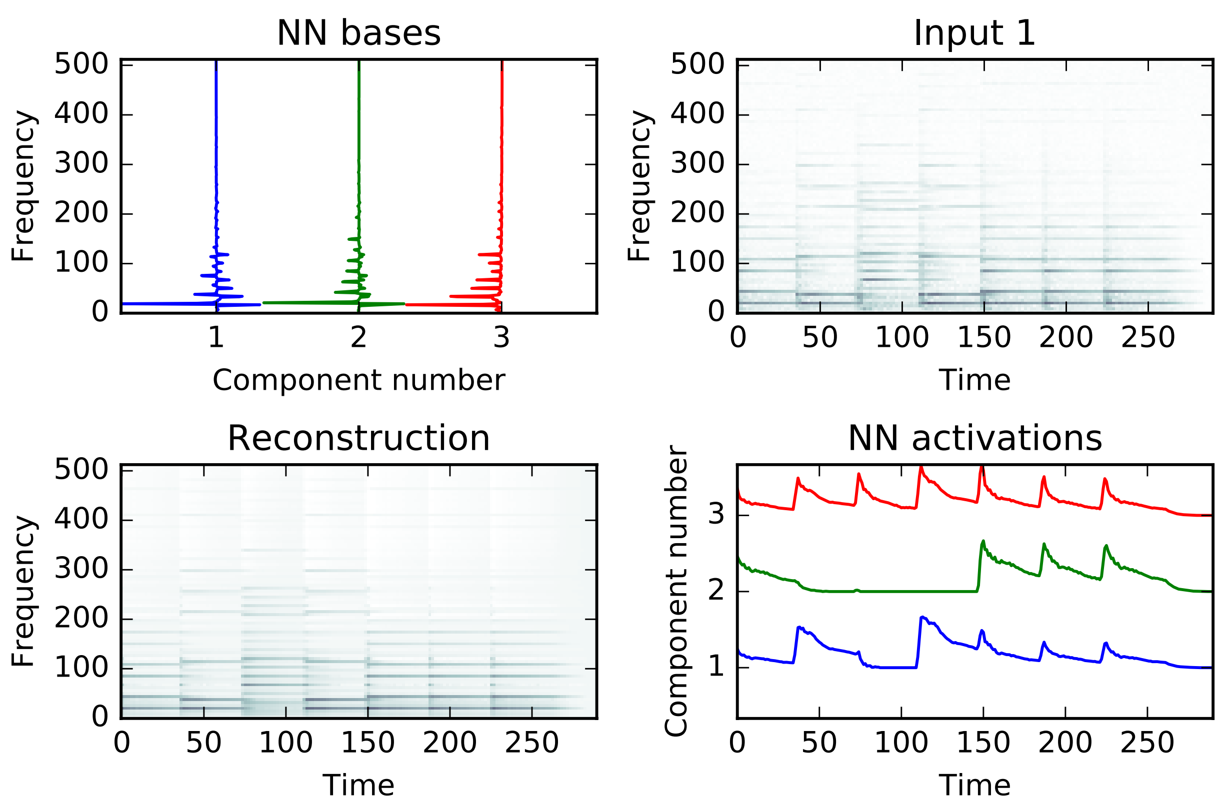
$$g(x) = \max(x, 0) \text{ or } |x| \text{ or } ln(1 + e^x)$$



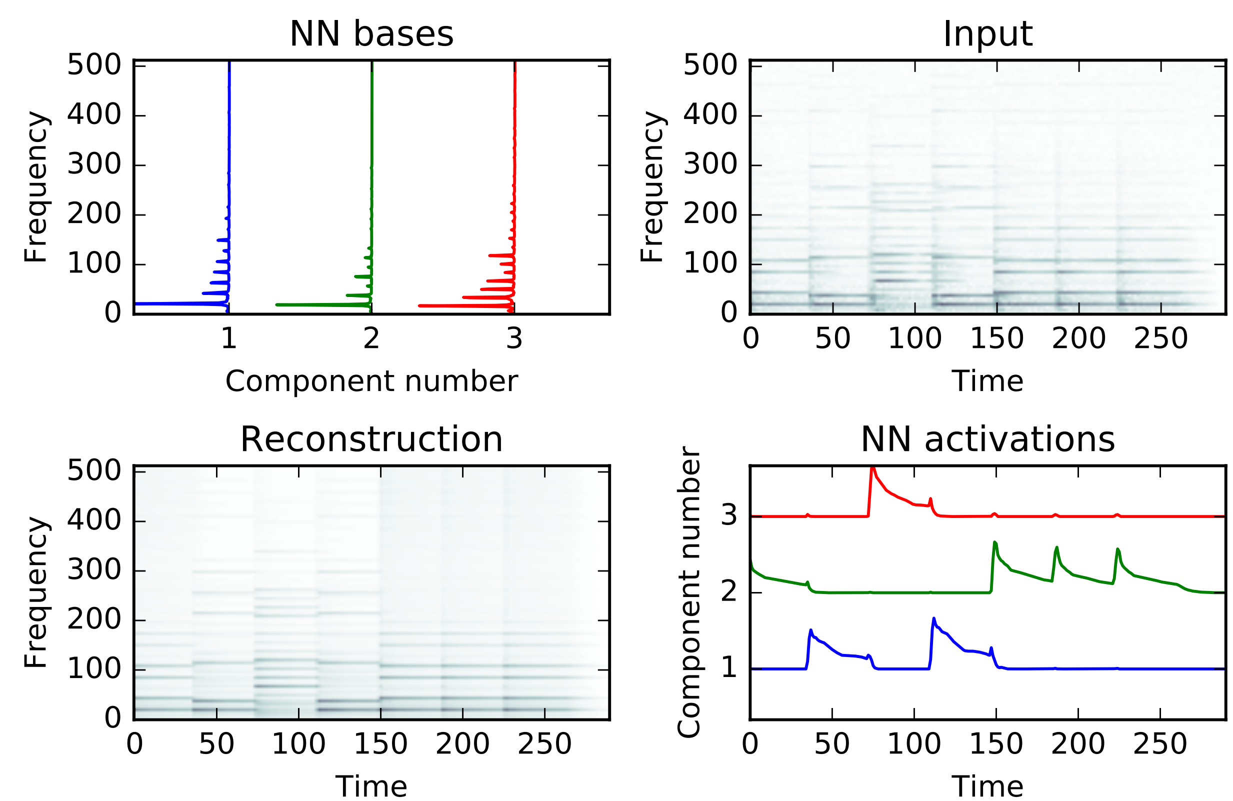
NAE in action

- Bases are not guaranteed to be non-negative
  - Results in dense activations
- Adding sparsity to hidden layer output
  - Results in sparse activations
    - Intuitive bases and activations

$$KL(\mathbf{X}||g(\mathbf{W}\cdot\mathbf{H}))$$



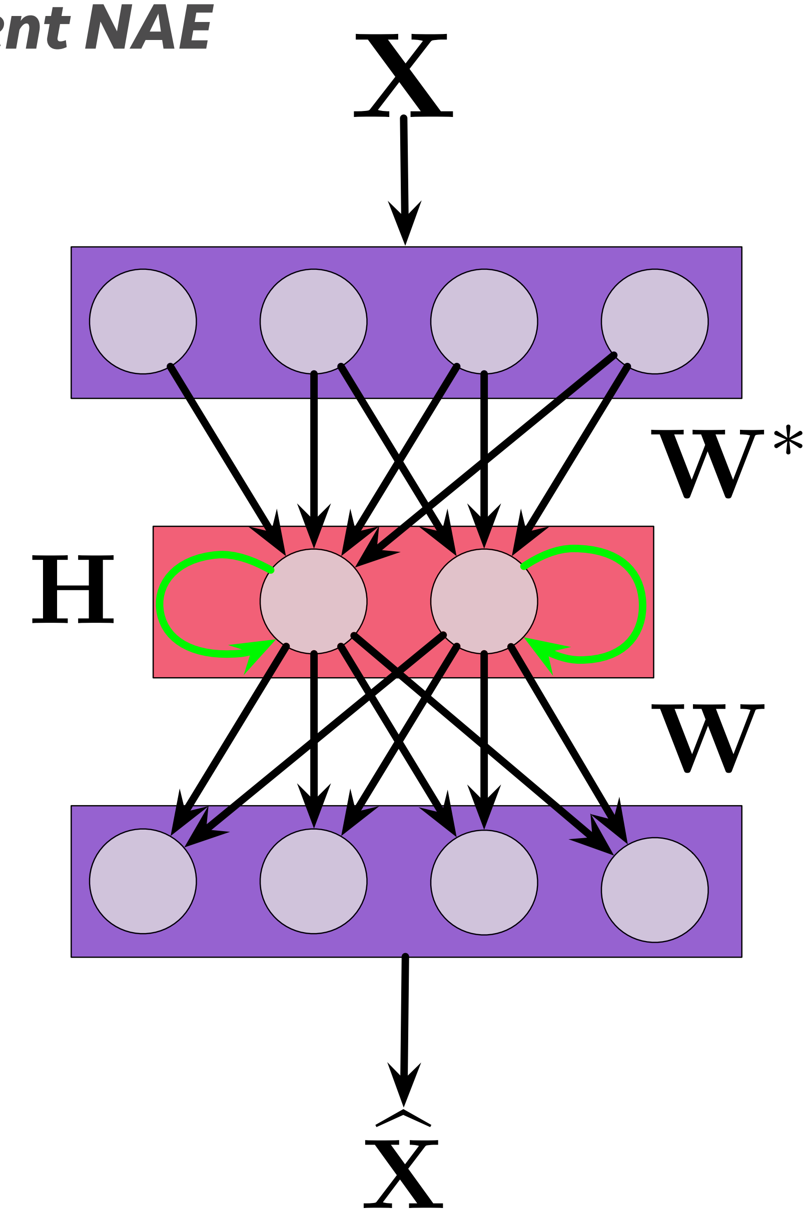
$$KL(\mathbf{X}||g(\mathbf{W}\cdot\mathbf{H})) + \lambda||\mathbf{H}||_1$$



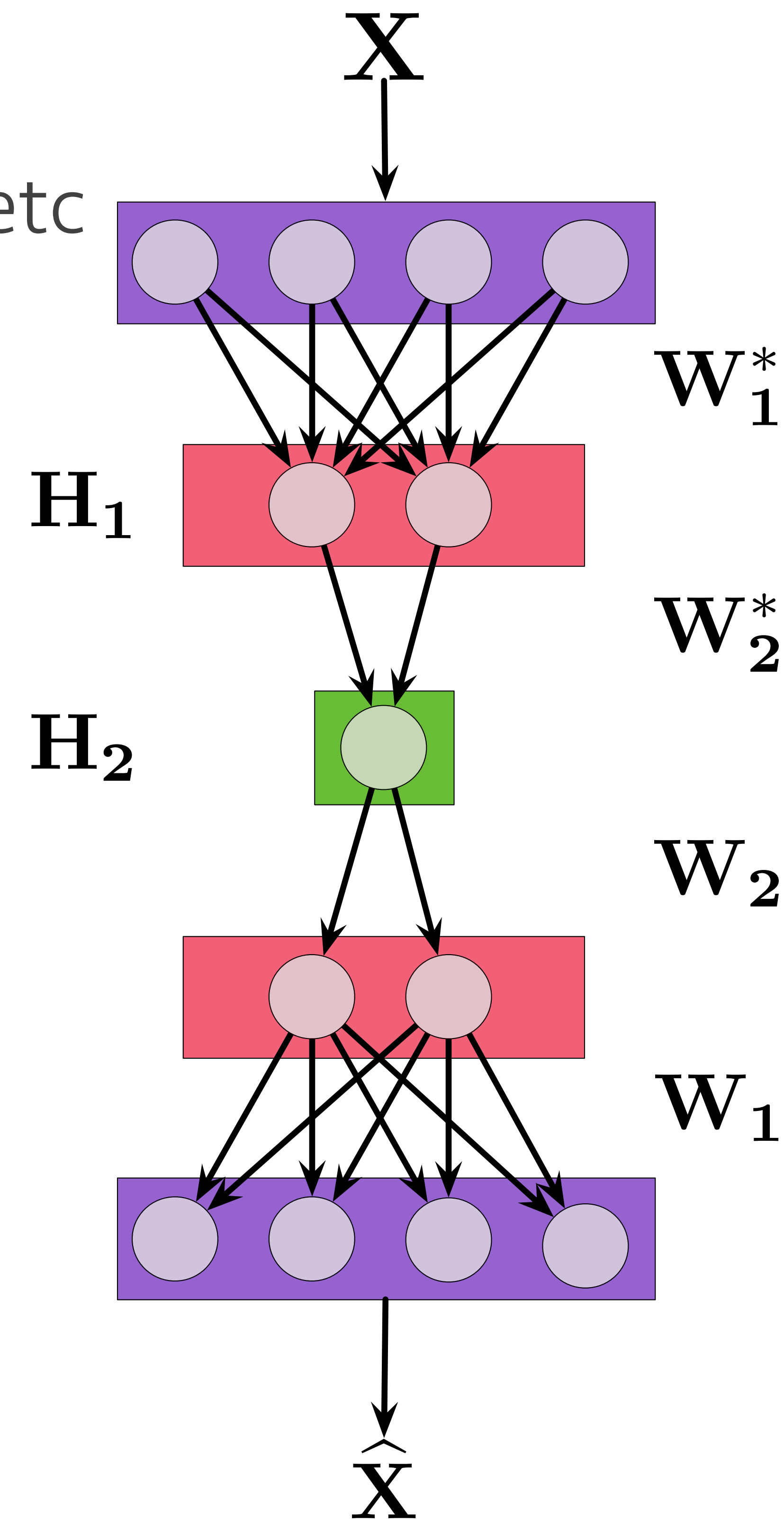
Advantages of NAE

- Difficult to extend NMF models
  - Easy to do with Neural nets.
    - LSTMs, CNNs, Multi-layer networks etc

Recurrent NAE

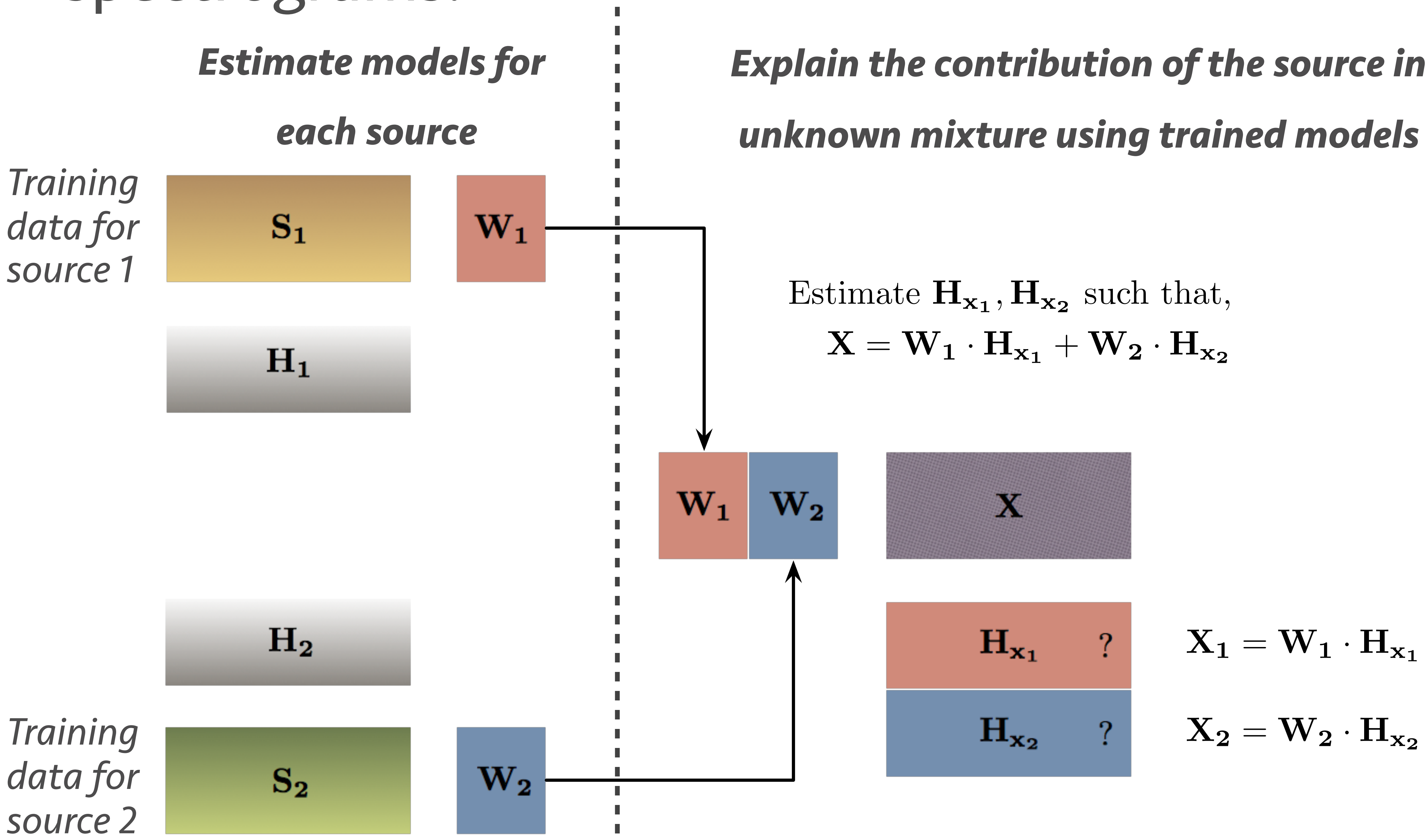


Multi-layer NAE



NMF source separation

- Spectrogram of the mixture is the sum of source spectrograms.



NAE source separation

Estimate models for each source

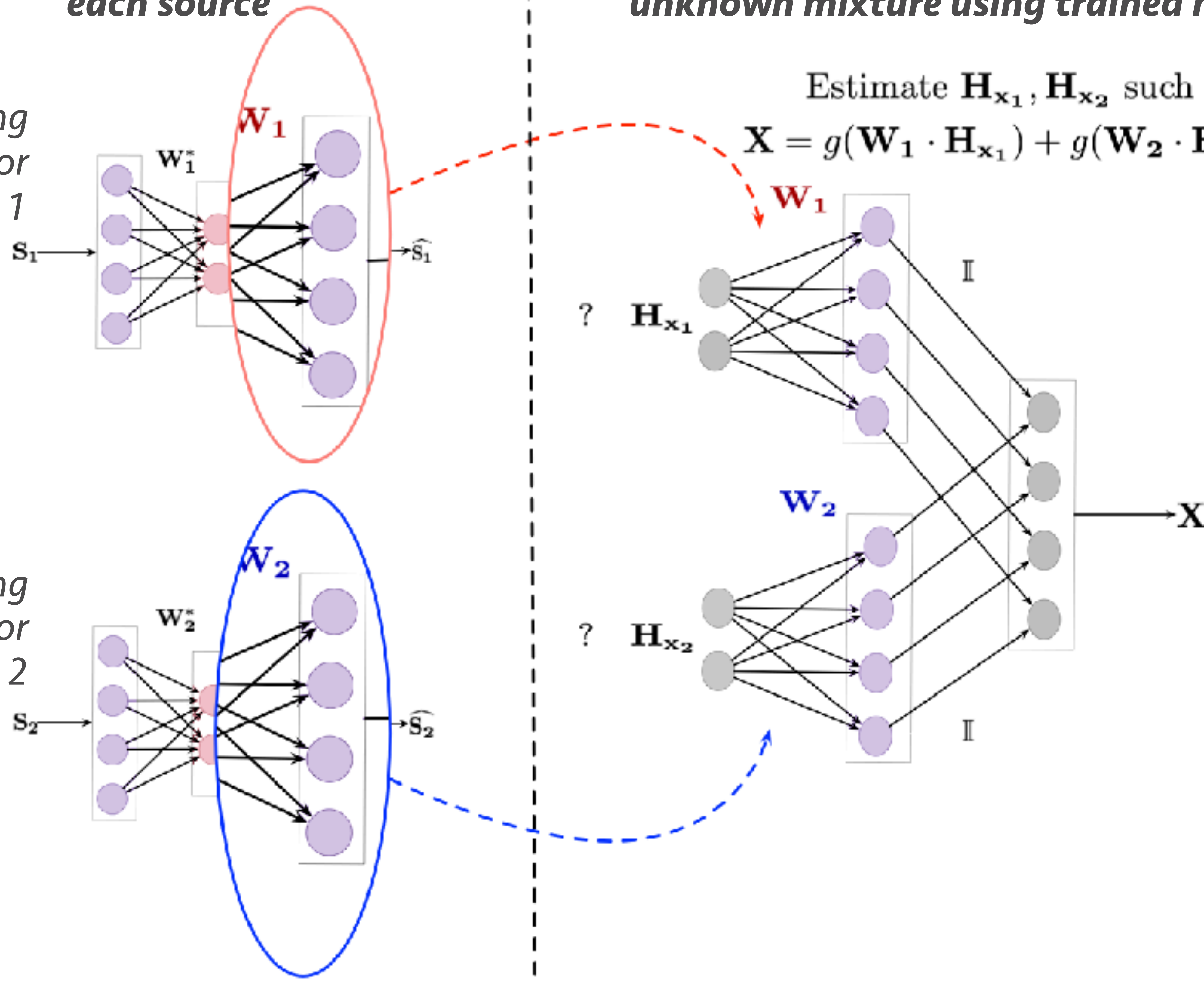
Explain the contribution of the source in unknown mixture using trained models

Estimate $\mathbf{H}_{\mathbf{x}_1}, \mathbf{H}_{\mathbf{x}_2}$ such that

$$\mathbf{X} = g(\mathbf{W_1} \cdot \mathbf{H}_{\mathbf{x}_1}) + g(\mathbf{W_2} \cdot \mathbf{H}_{\mathbf{x}_2})$$

Training data for source 1

Training data for source 2



NAE source separation

- Goal: Estimating network inputs instead of the weights

$$\mathbf{X} = g(\mathbf{W_1} \cdot \mathbf{H_{x_1}}) + g(\mathbf{W_2} \cdot \mathbf{H_{x_2}})$$

  - Gradient-descent/back-propagation to train the network

- Separated spectrograms

$$\mathbf{X_1} = g(\mathbf{W_1} \cdot \mathbf{H_{x_1}})$$

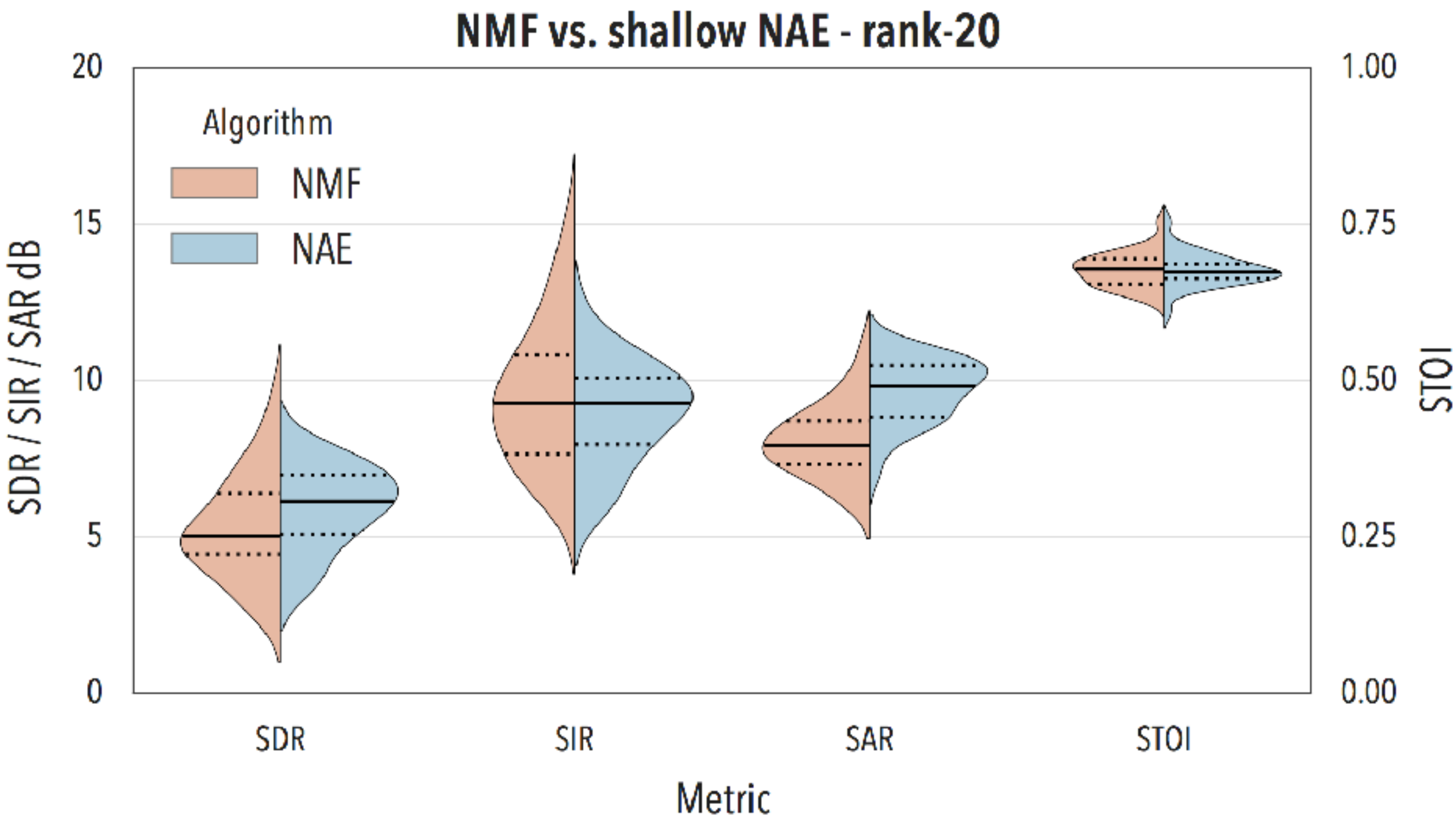
$$\mathbf{X_2} = g(\mathbf{W_2} \cdot \mathbf{H_{x_2}})$$

Evaluation

- Two-speaker mixtures
  - Training data ~ 20-25 seconds
  - Test data: Single sentence of known speakers
- Evaluation metrics
  - BSS_eval metrics (SDR, SIR, SAR)
  - STOI (intelligibility measure)
- Compared multilayer and shallow versions
  - With multiple ranks (number of hidden units)

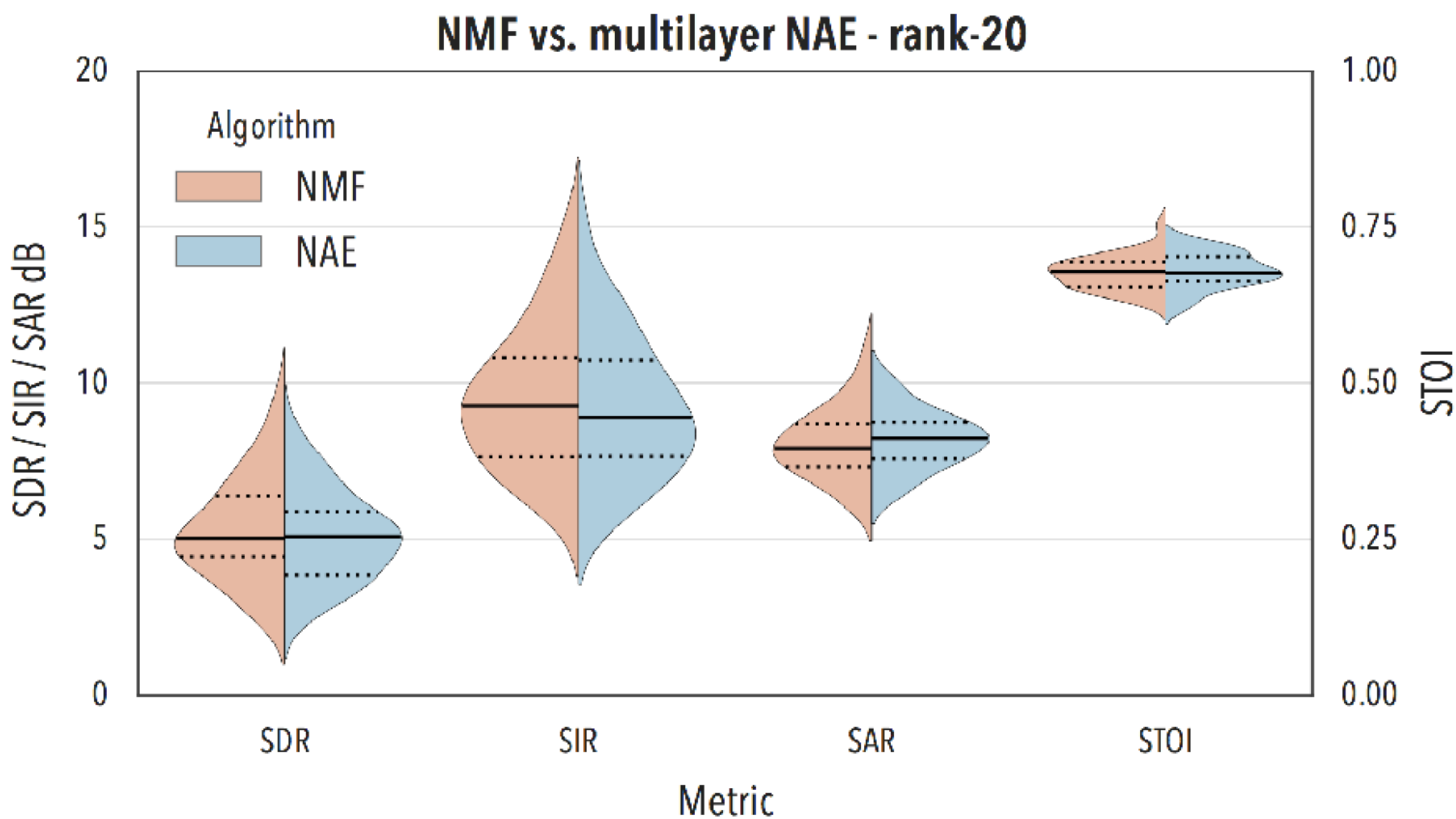
NMF vs NAE

- Number of NAE layers = 2
- Number of hidden units = 20



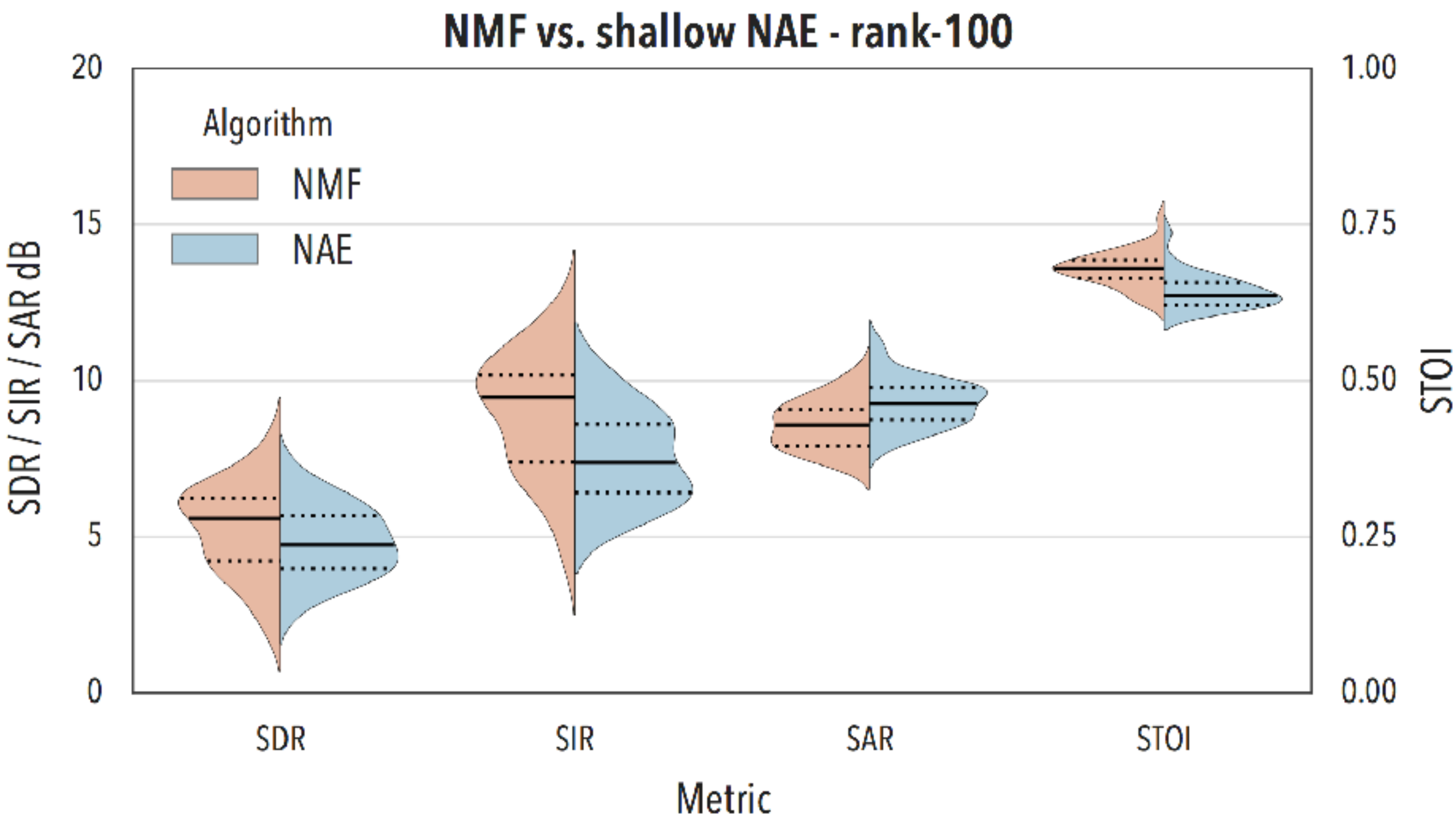
NMF vs NAE

- Number of NAE layers = 4
- Number of hidden units = 20



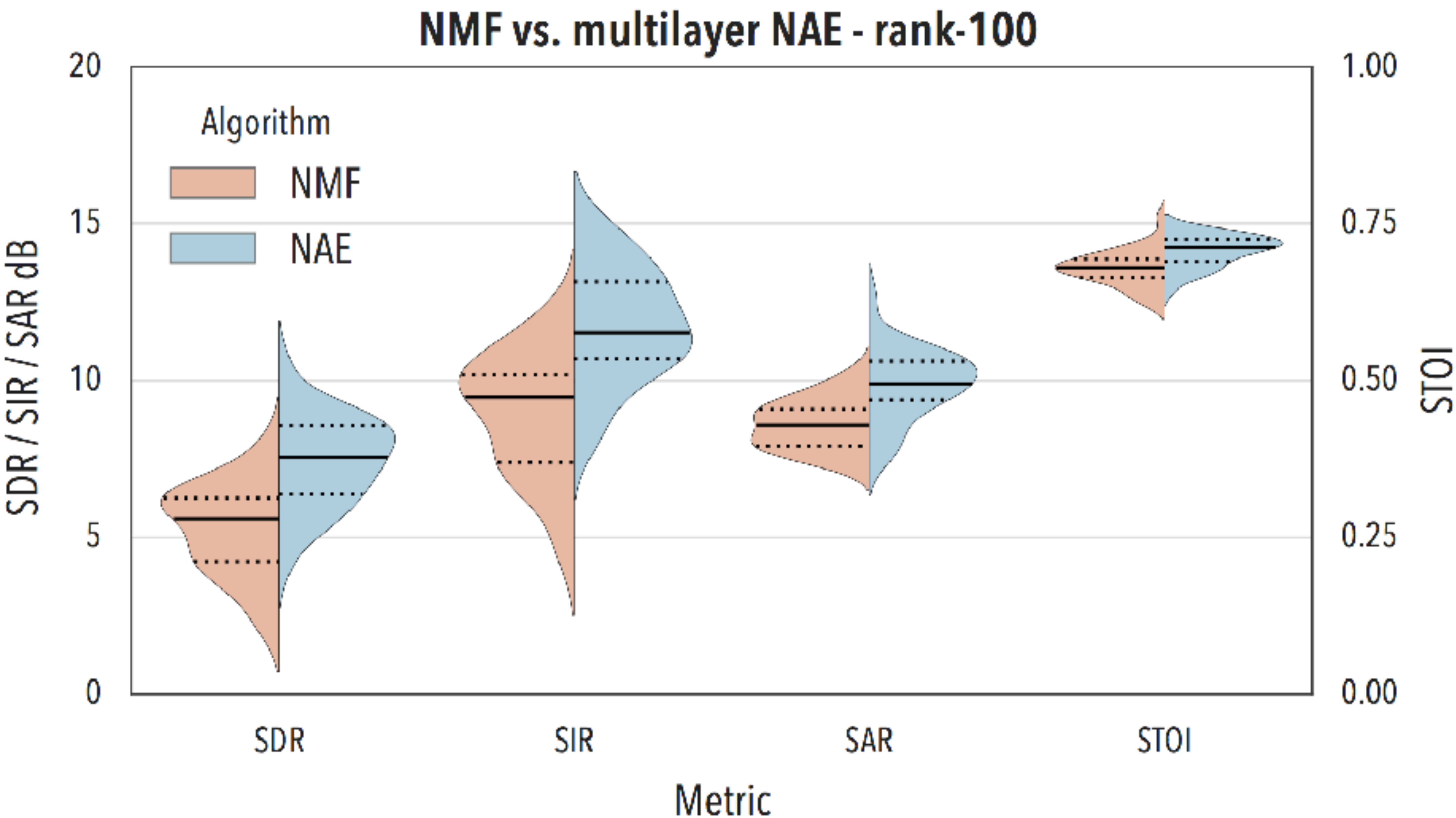
NMF vs NAE

- Number of NAE layers = 2
- Number of hidden units = 100



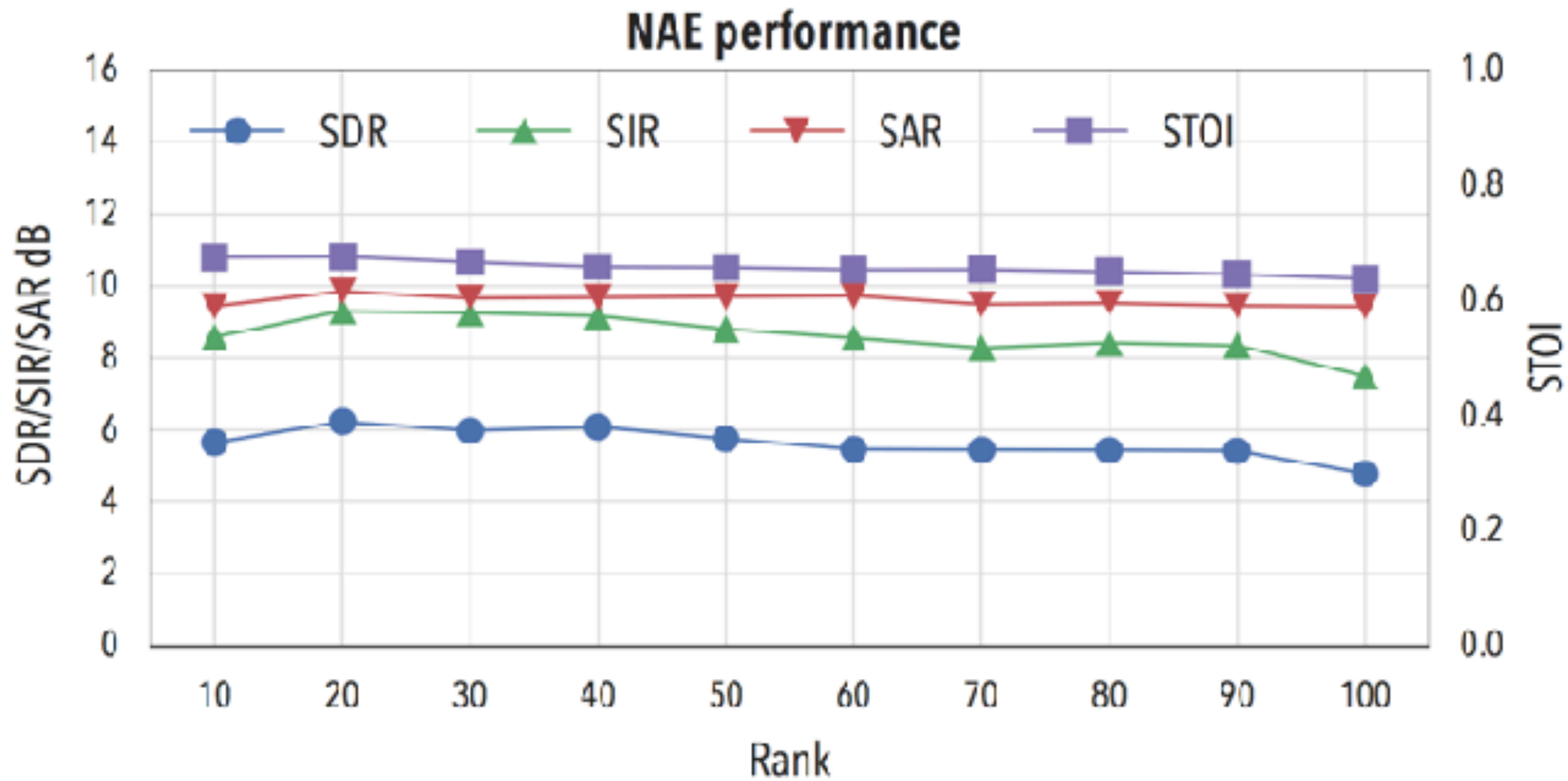
NMF vs NAE

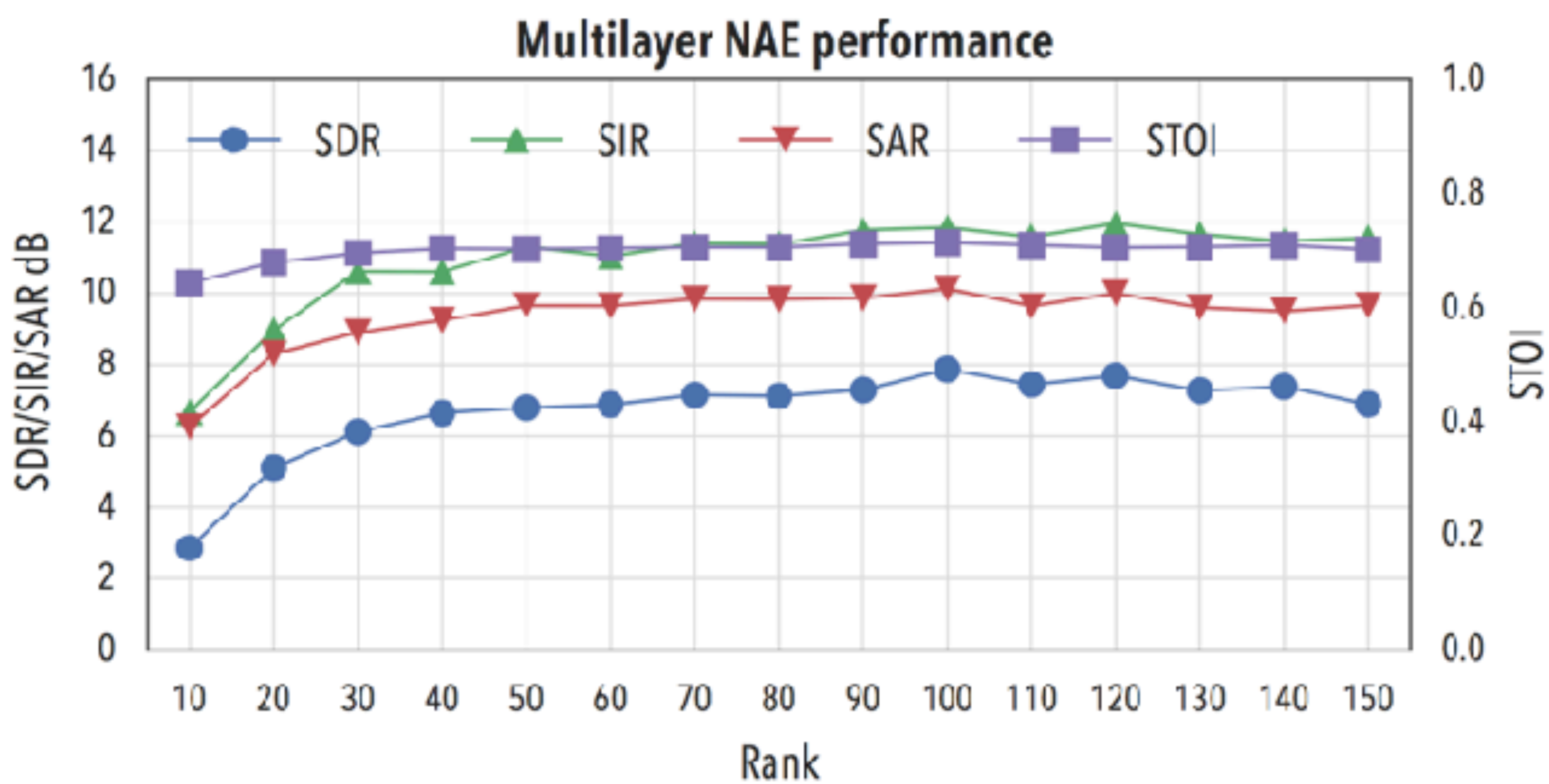
- Number of NAE layers = 4
- Number of hidden units = 100



Shallow vs Multi-layer NAE

- Shallow NAEs give comparable performance over all ranks
- Multi-layer NAEs require higher ranks





Conclusions

- NAE models can replace NMF
  - This allows us to generalize to complex structures
- NAE models superior to NMF models
  - Shallow NAEs comparable to NMF
  - Multi-layer NAEs outperform NMF significantly
- Future directions
  - Incorporating exotic neural models (LSTMs, CNNs etc)

THANK YOU