MITSUBISHI ELECTRIC
Changes for the Better

Stochastic Bottleneck:
Rateless Auto-Encoder for Flexible Dimensionality Reduction

Toshiaki Koike-Akino

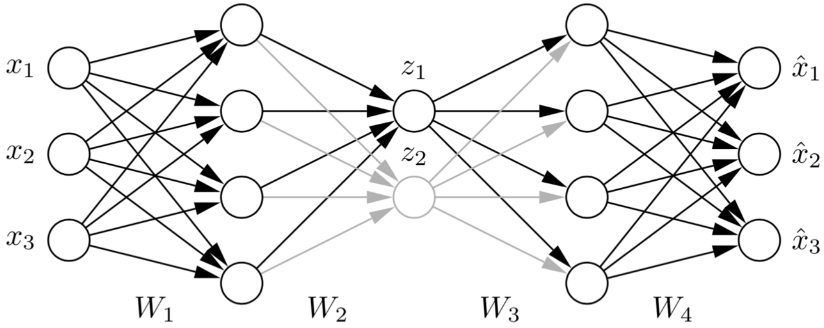
Ye Wang

June 2020

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Outline

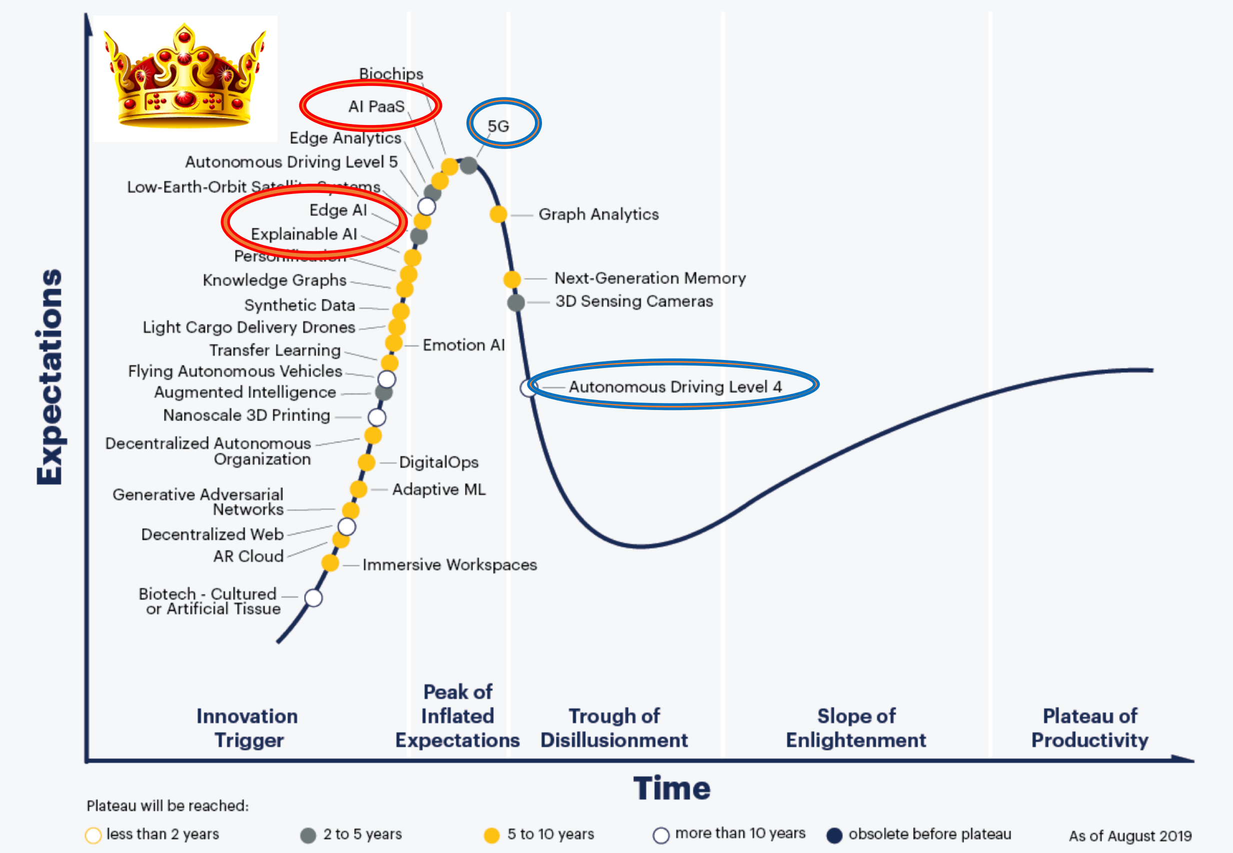
- Motivations
  - Machine learning for real-world data analysis
- Dimensionality reduction
  - Principal component analysis (PCA)
  - Auto-encoder (AE)
- **Rateless** property
  - Fountain codes
- Stochastic bottleneck
  - Stochastic Width vs. Stochastic Depth
  - **TailDrop** regularization
- Multi-objective learning
- Experiments
  - MSE
  - SSIM
  - Accuracy
- Summary



How many latent variables required?

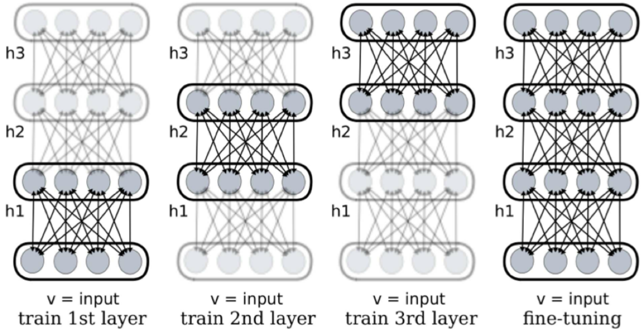
Emerging Technologies

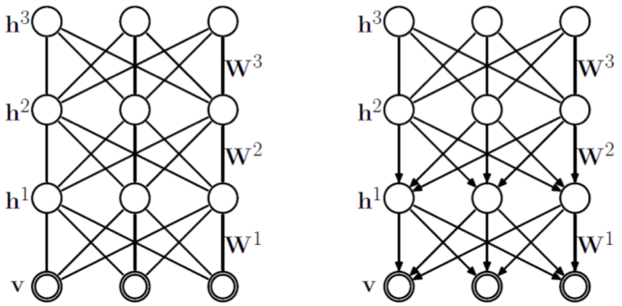
- Gartnar's Hype Cycle for Emerging Technologies, 2019 August



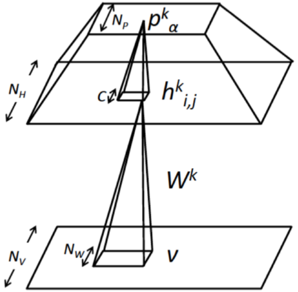
Deep Learning for AI

- Deep learning = fancy name of multi-layer perceptron neural networks.
  - 2006 Hinton: Many layers, layer-wise pre-training, massive data sets
- Massively parallel computation
  - Driver: graphic processor units, tensor processor units …
- Variants:
  - Deep belief networks
  - Deep convolutional networks
  - Deep recurrent networks
  - Deep Boltzmann machines
  - Deep autoencoder

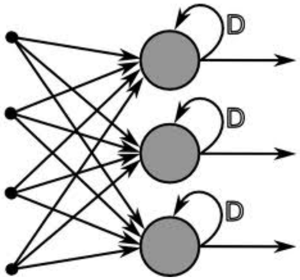




Deep Boltzmann Machine   Deep Belief Network



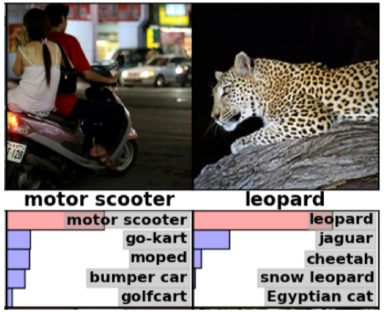
Convolutional Networks

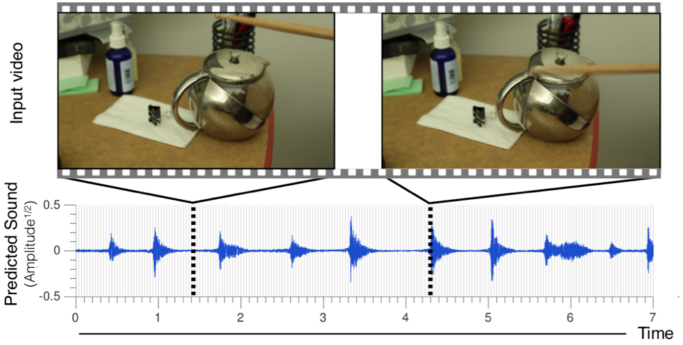


Recurrent Networks

Deep Learning for Media Signal Processing

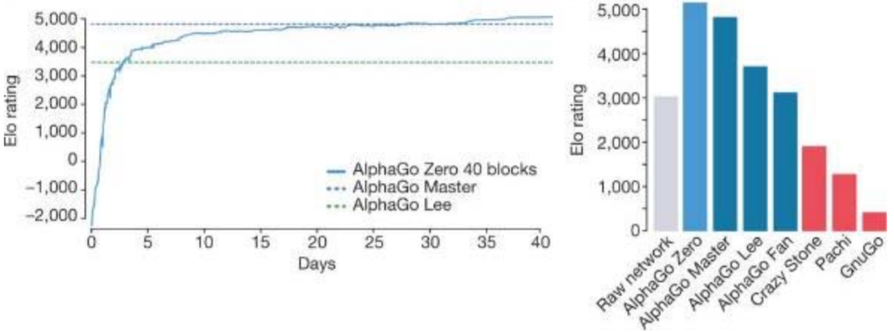
- Audio & Visual Applications





"man in black shirt is playing guitar."

AI Surpassing Human-Level Performance



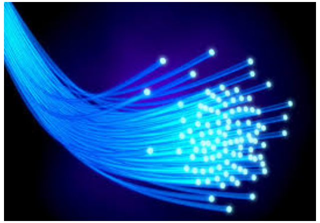
Applied Deep Learning

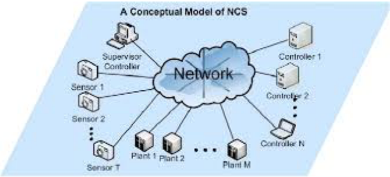
- AI has been applied to various fields

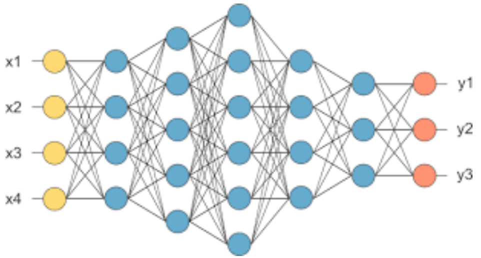
Wireless Communication

Networked Control

Optical Communication

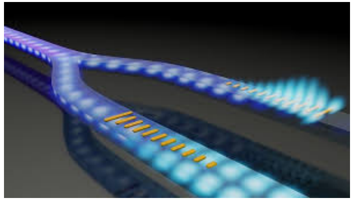




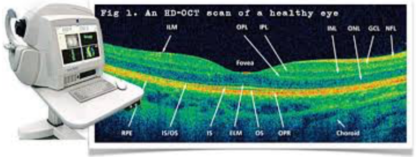


Localization Navigation

Device / Integrated Circuit



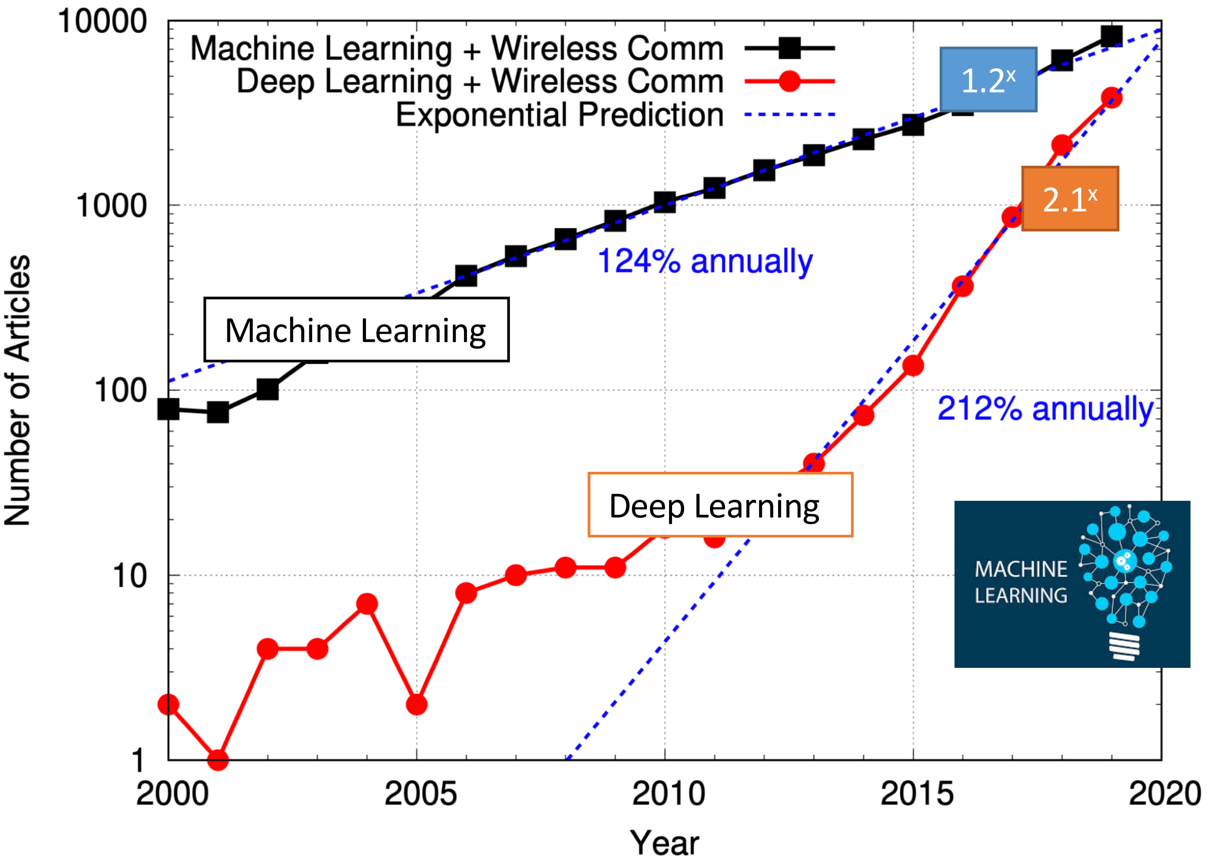
Tomography Imaging



Bio-Sensing Human Interface

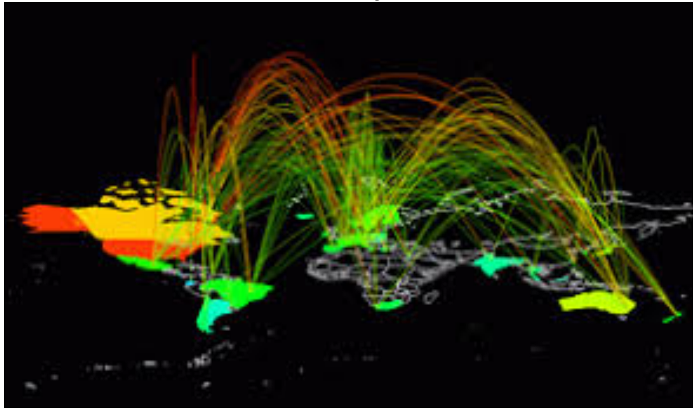
Moore’s Law: Exponential Grow in Applications

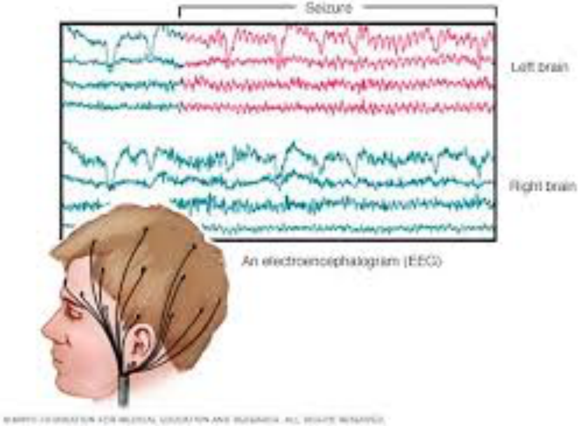
- The hit count of articles per year in GoogleScholar; Wireless Communication applications

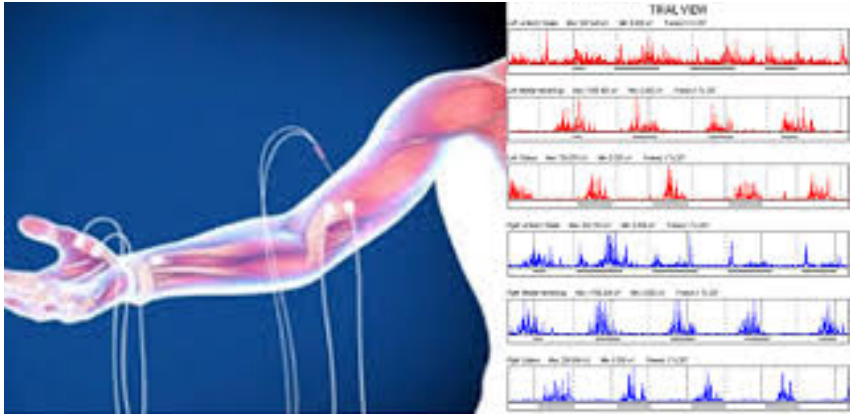


High-Dimensional Real-World Data

- Raw data dimensionality is often extremely large

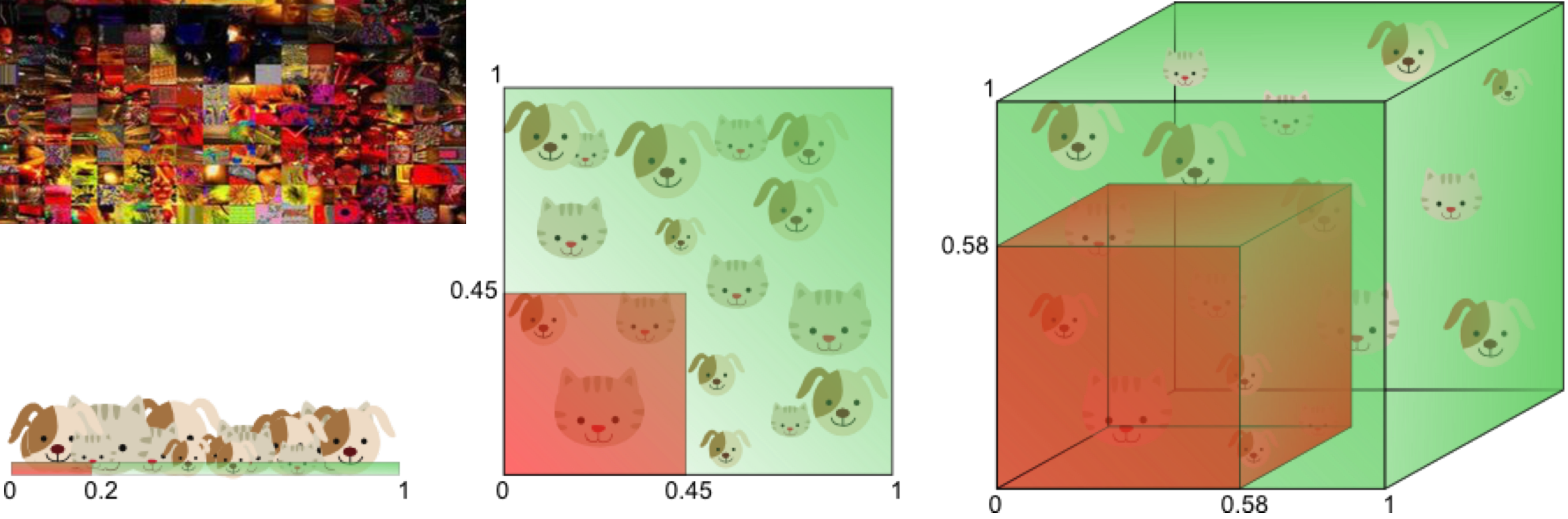






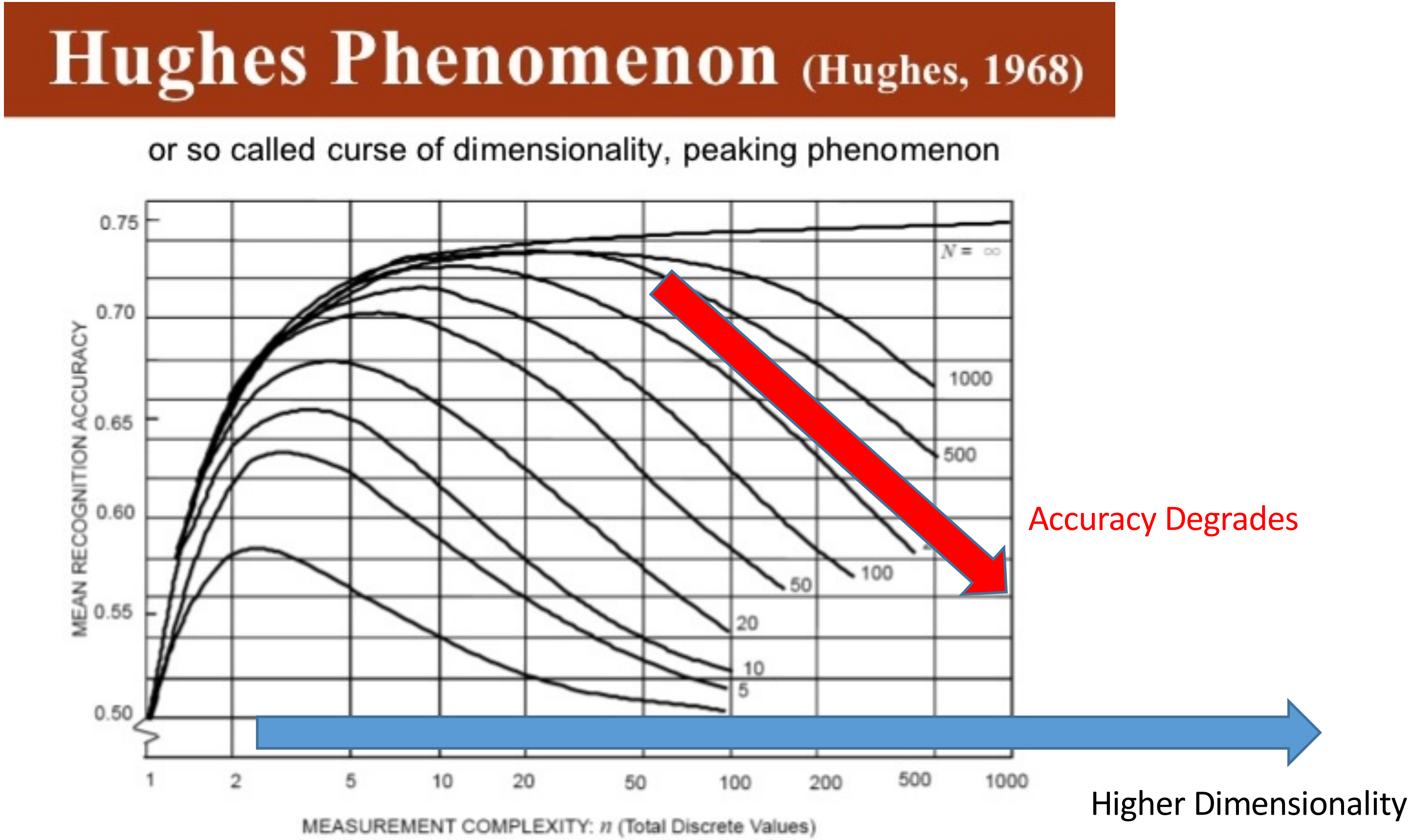
Curse of Dimensionality

- Data-space volume exponentially increases with dimensionality



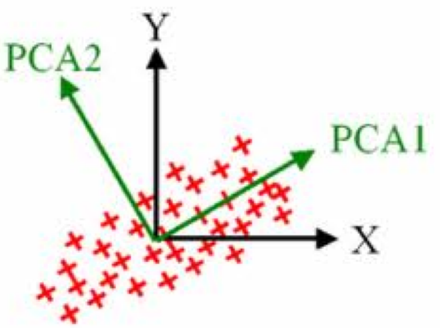
Hughes Phenomenon

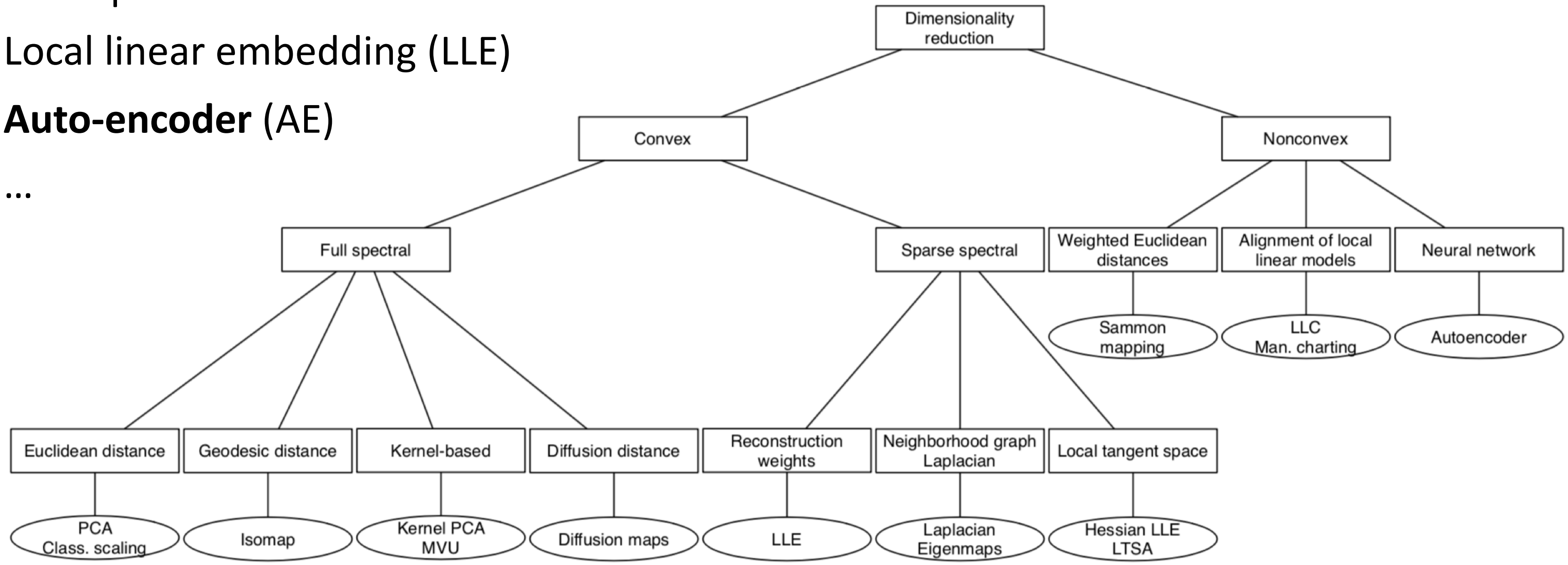
- Classifier performance drops for high-dimension data with finite training samples



Dimensionality Reduction

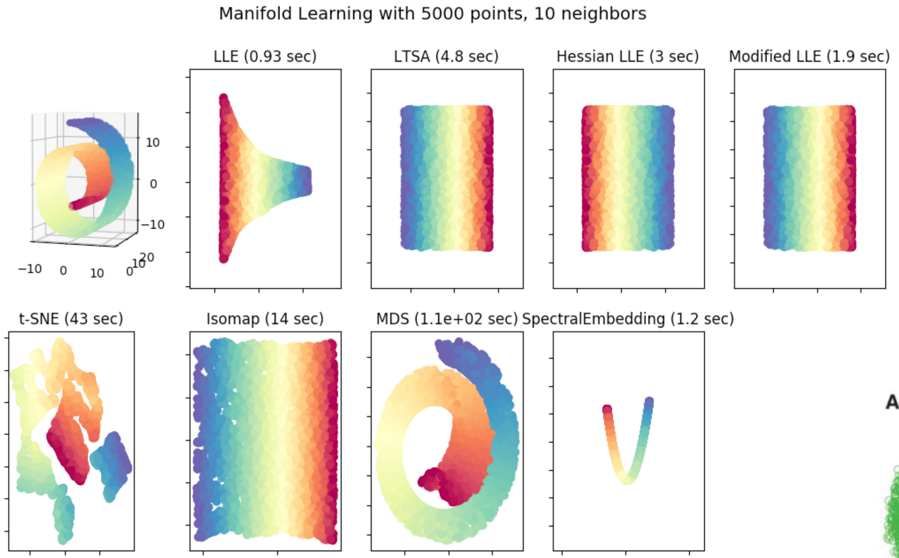
- Principal component analysis (**PCA**)
- Kernel PCA
- Independent component analysis (ICA)
- Isomap
- Local linear embedding (LLE)
- **Auto-encoder** (AE)
- …

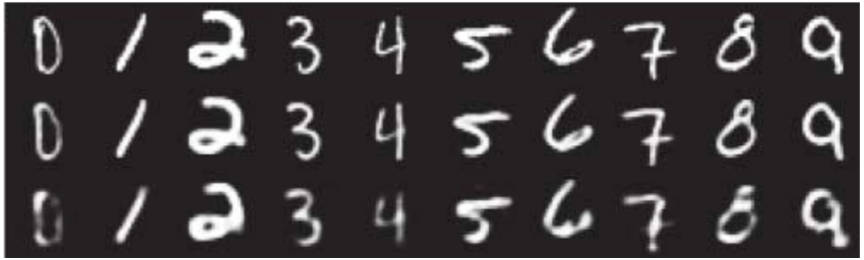


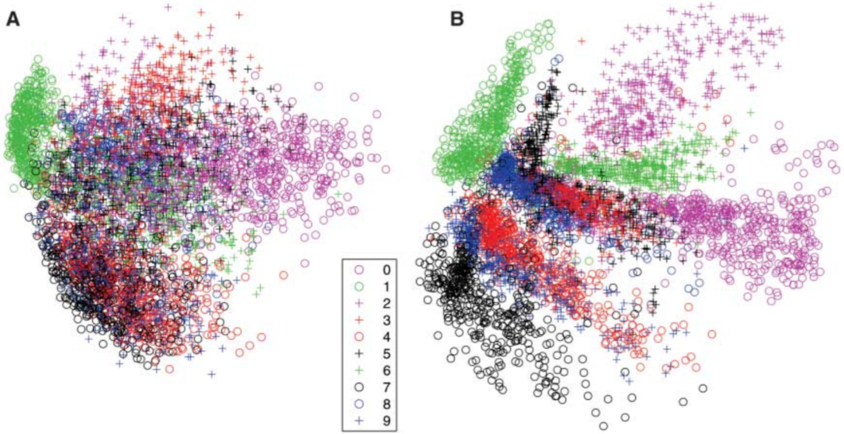


Reduced-Dimension Feature

- High-dimensional data may be well-described by lower-dim latent variables





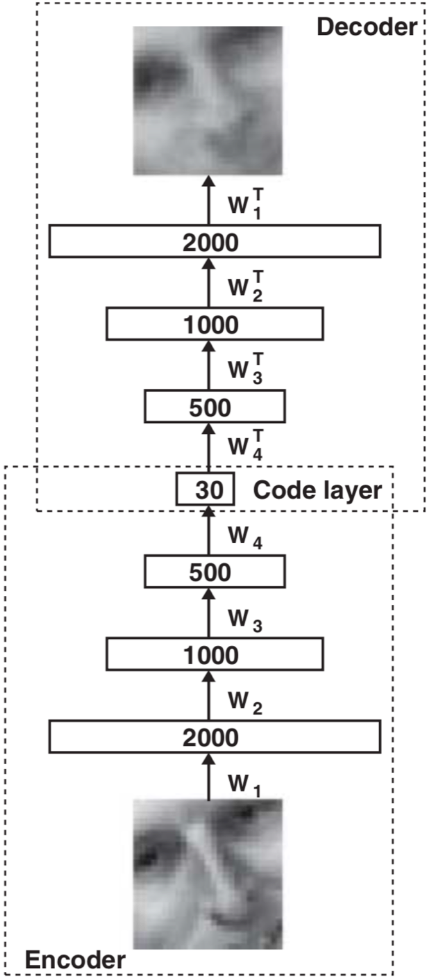


Auto-Encoder (AE): Bottleneck Network

- **Bottleneck** neural network architecture: *M<N*
- Encoder and decoder networks are jointly trained such that latent variables can regenerate original data with smallest distortion

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{\mathbf{x}\sim\Pr(\mathbf{x})} \Big[\mathcal{L}\big(\mathbf{x}, g_\phi(f_\theta(\mathbf{x}))\big)\Big]$$

Original data	$\mathbf{x} \in \mathbb{R}^N$
Latent variable	$\mathbf{z} \in \mathbb{R}^M$
Encoder network	$\mathbf{z} = f_\theta(\mathbf{x})$
Decoder network	$\mathbf{x}' = g_\phi(\mathbf{z})$
Loss function (e.g. MSE)	$\mathcal{L}(\mathbf{x}, \mathbf{x}')$



AE as Nonlinear PCA (NLPCA)

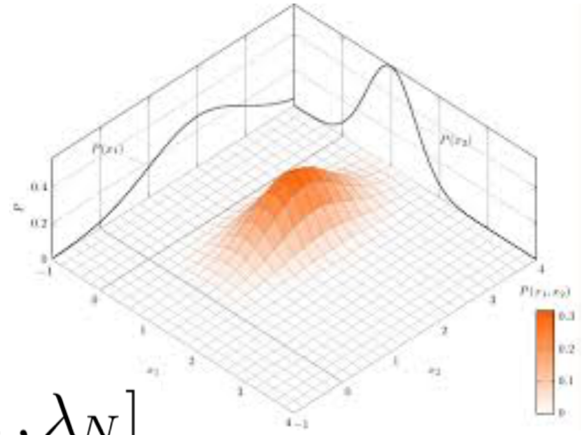
- AE is often called **NLPCA** due to analogy
- Without nonlinear activations, an optimal AE model coincides with PCA for Gaussian data under MSE distortion (Karhunen-Loeve)

Encoder Affine Transform $$f_\theta(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b}$$

Decoder Affine Transform $$g_\phi(\mathbf{z}) = \mathbf{W}'\mathbf{z} + \mathbf{b}'$$

Consider Gaussian data $$\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$$

Covariance EVD $$\mathbf{C} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\mathrm{T}}$$



$$\mathbf{\Lambda} = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_N]$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$$

Eigen projection gives minimum MSE

$$\mathbf{W} = \mathbf{I}_{M,N}\mathbf{\Phi}^{\mathrm{T}}$$
$$\mathbf{b} = -\mathbf{W}\mathbf{m}$$
$$\mathbf{W}' = \mathbf{\Phi}\mathbf{I}_{N,M}$$
$$\mathbf{b}' = \mathbf{m}$$

$$\bar{\mathcal{L}}_M = \mathbb{E}_{\mathbf{x}}\left[\left\|\mathbf{W}'(\mathbf{W}\mathbf{x} + \mathbf{b}) + \mathbf{b}' - \mathbf{x}\right\|^2\right] = \sum_{n=M+1}^{N} \lambda_n$$

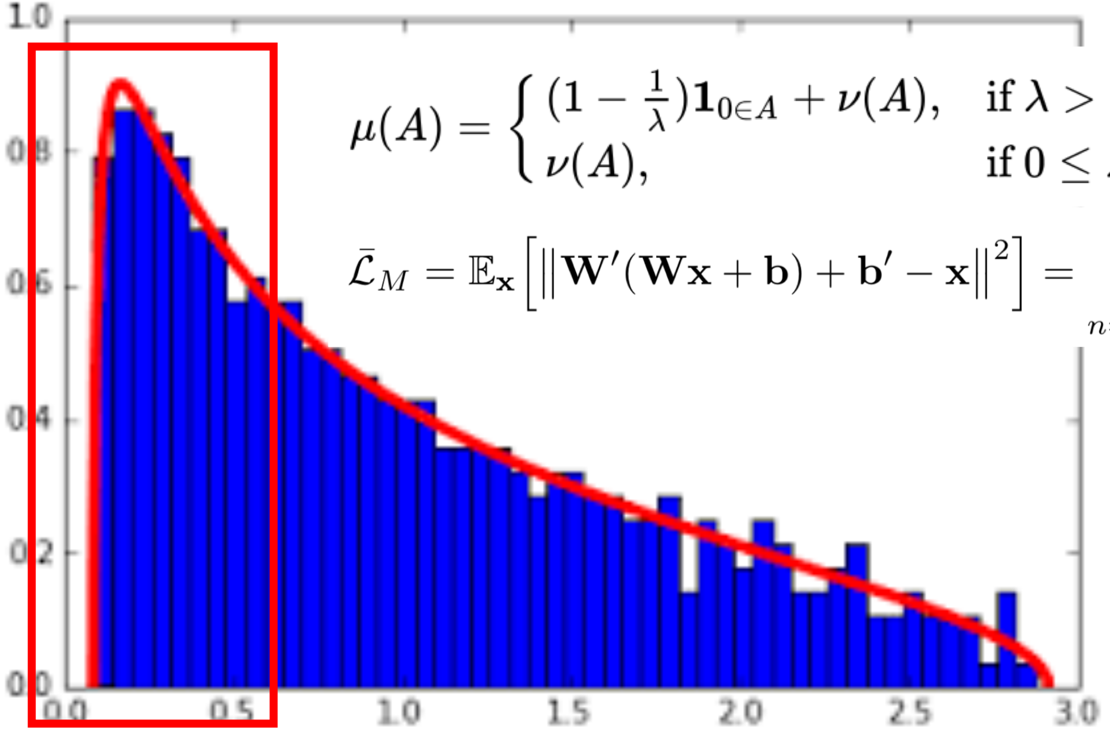
PCA: Eigen-Spectrum

- Random matrix theorem:
  If covariance matrix follows i.i.d. Gaussian Gram matrix, eigenvalue distribution follows Marchenko-Pastur distribution

$$\mu(A) = \begin{cases} (1 - \frac{1}{\lambda})\mathbf{1}_{0\in A} + \nu(A), & \text{if } \lambda > 1 \\ \nu(A), & \text{if } 0 \le \lambda \le 1, \end{cases}$$

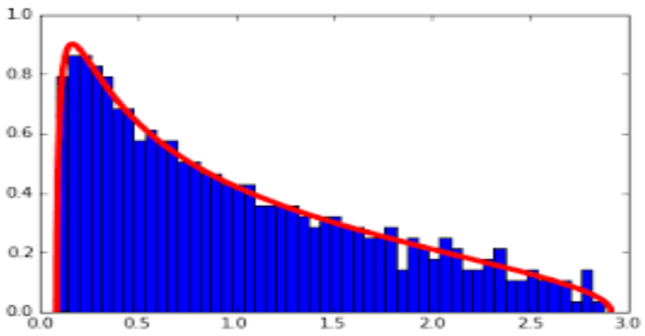
$$\bar{\mathcal{L}}_M = \mathbb{E}_{\mathbf{x}}\left[\left\|\mathbf{W}'(\mathbf{W}\mathbf{x} + \mathbf{b}) + \mathbf{b}' - \mathbf{x}\right\|^2\right] = \sum_{n=M+1}^{N} \lambda_n$$

Cumulative is well approximated by exponential



Rateless Property

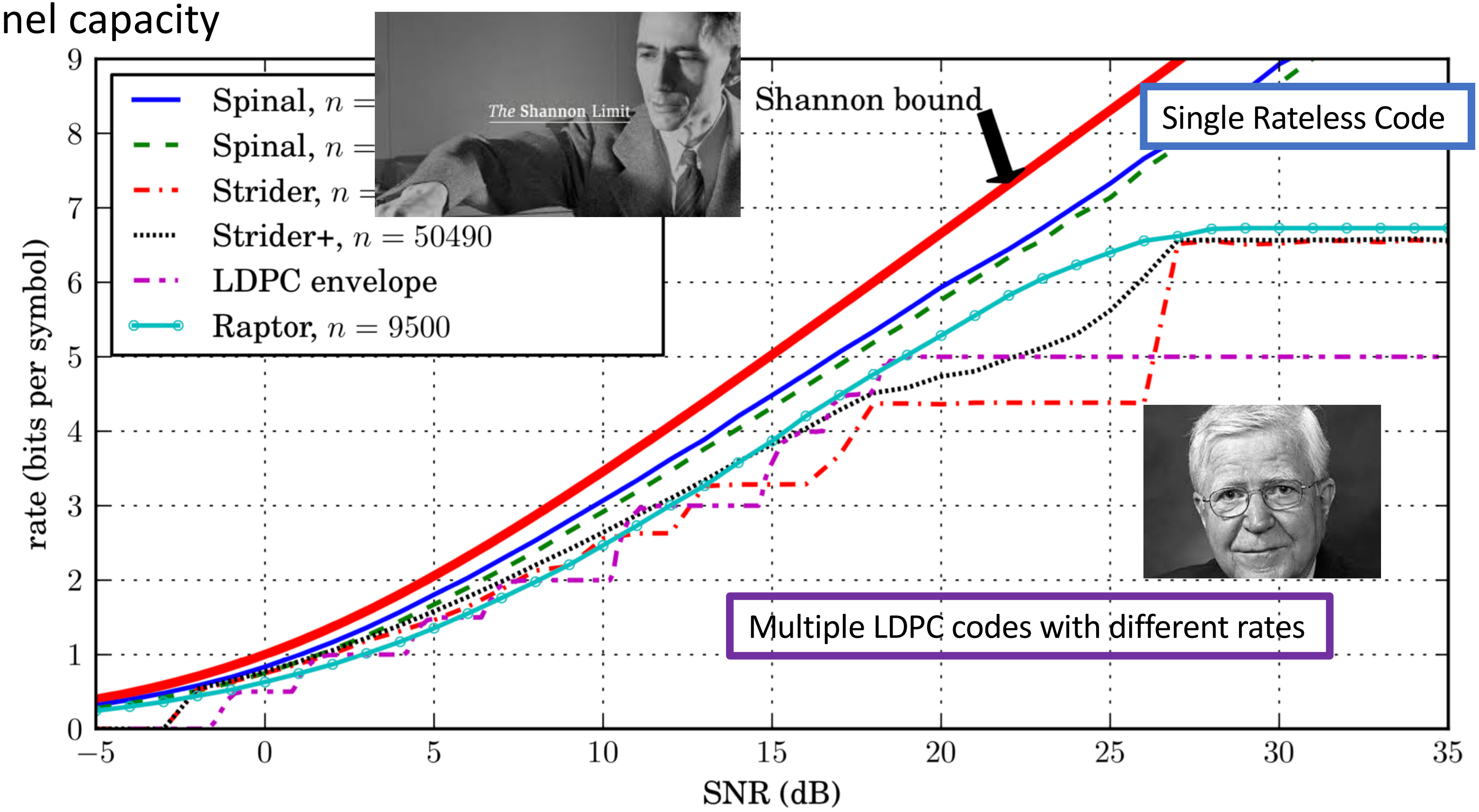
- PCA universally achieves best MSE for all dimensionality $1<M<N$ under Gaussian datasets
- The downstream users can freely change the dimensionality by discarding the least-principal components or appending the most-principal components without changing encoder and decoder models
- The MSE is gracefully improved by increasing the compression rate M/N
- We do not need to pre-determine the dimensionality when training the model
- This **rateless** property can resolve the issue:



How many latent variables do we need for training the AE model?

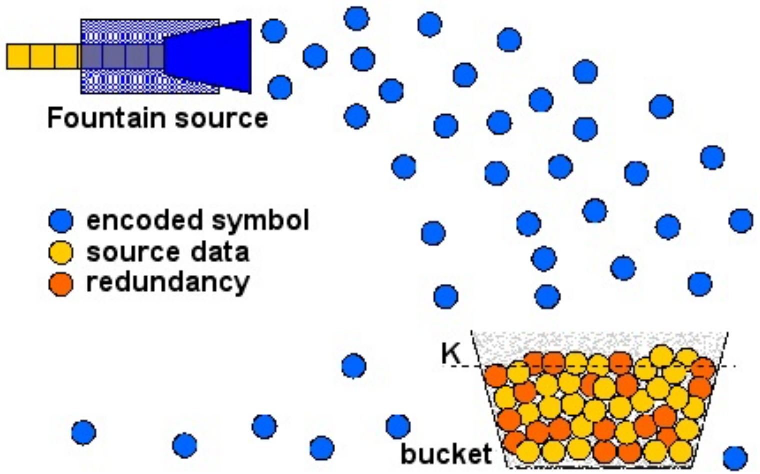
Coding Theory: Rateless Channel Codes

- Capacity approaching codes need to pre-determine code rates under the knowledge of channel capacity



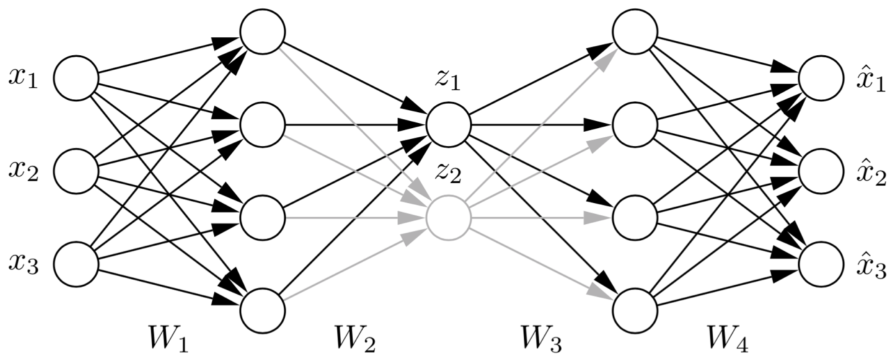
Rateless Fountain Codes

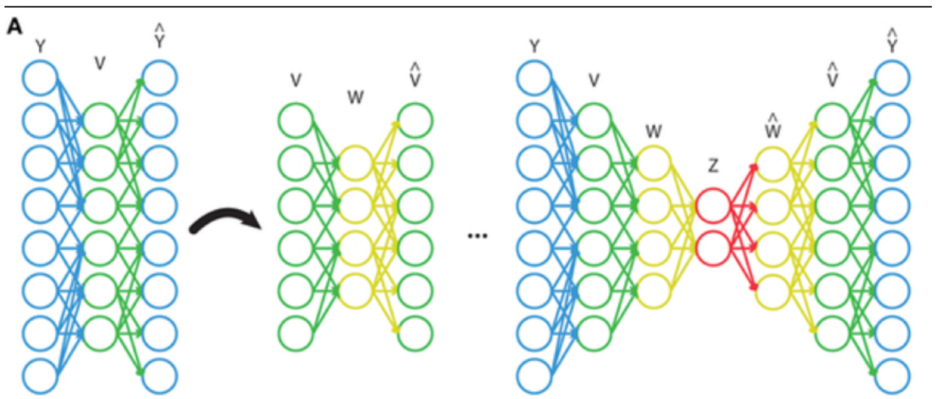
- Continue sending more redundant parity until the user satisfies
  - Luby-Transform (LT) codes [2002], Online codes [2002], Raptor codes [2006], Tornado codes [2004]
- We do not pre-determine the code rates
- Rateless codes are capacity-achievable
- We introduce “**rateless**” AE which does not have to determine the dimensionality beforehand



Dimensionality Flexibility

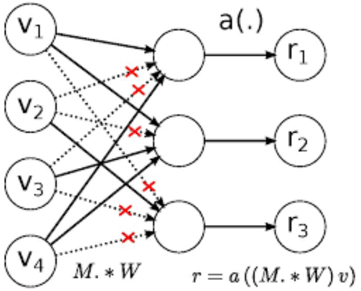
- For PCA, principal components are sorted in significance, thus scalable
  - For AE, latent variables are equally important, thus not adaptable
- Once AE is learned with pre-determined dims, it requires another learning to reduce or expand dims
  - **Hierarchical AE** (hAE) to append dim for residual reconstruction
  - **Stacked AE** (sAE) to further reduce dimensionality
- Conditional update for progressive learning usually does not work best and often finite-tuning is required while flexibility is compromised
- We propose a very simple **dropout** mechanism to realize ratelessness

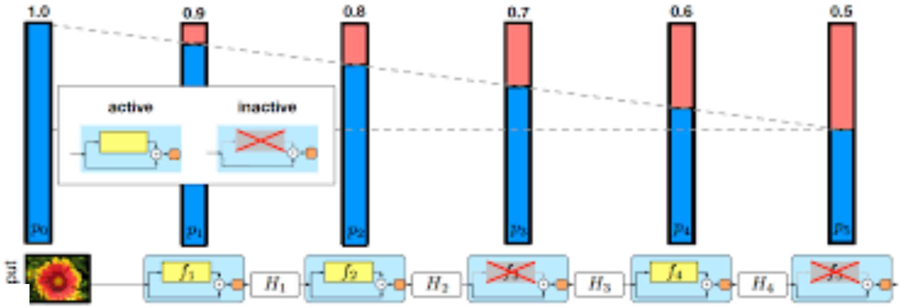


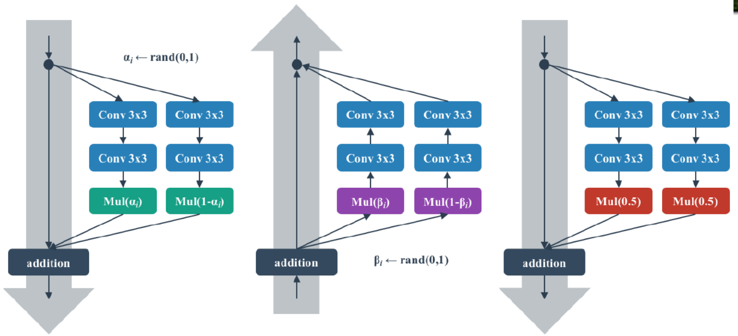


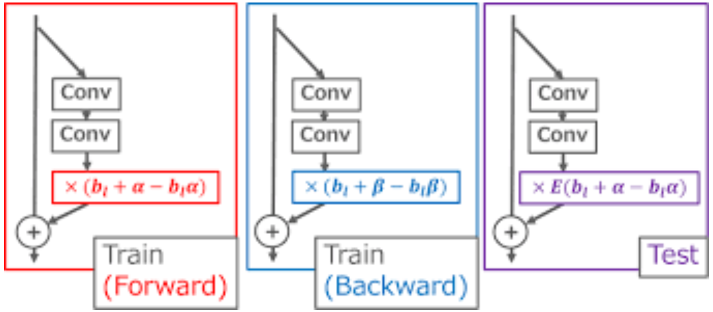
Dropout Regularization

- Dropout is an effective method to prevent **over-training** by regularizing over-parameterized networks
- It can be viewed as **Bayesian** approximation [Gal2016]
- There are many different regularization techniques: DropConnect, DropBlock, StochasticDepth, DropPath, ShakeDrop, SpatialDrop, ZoneOut, Shake-Shake, etc.



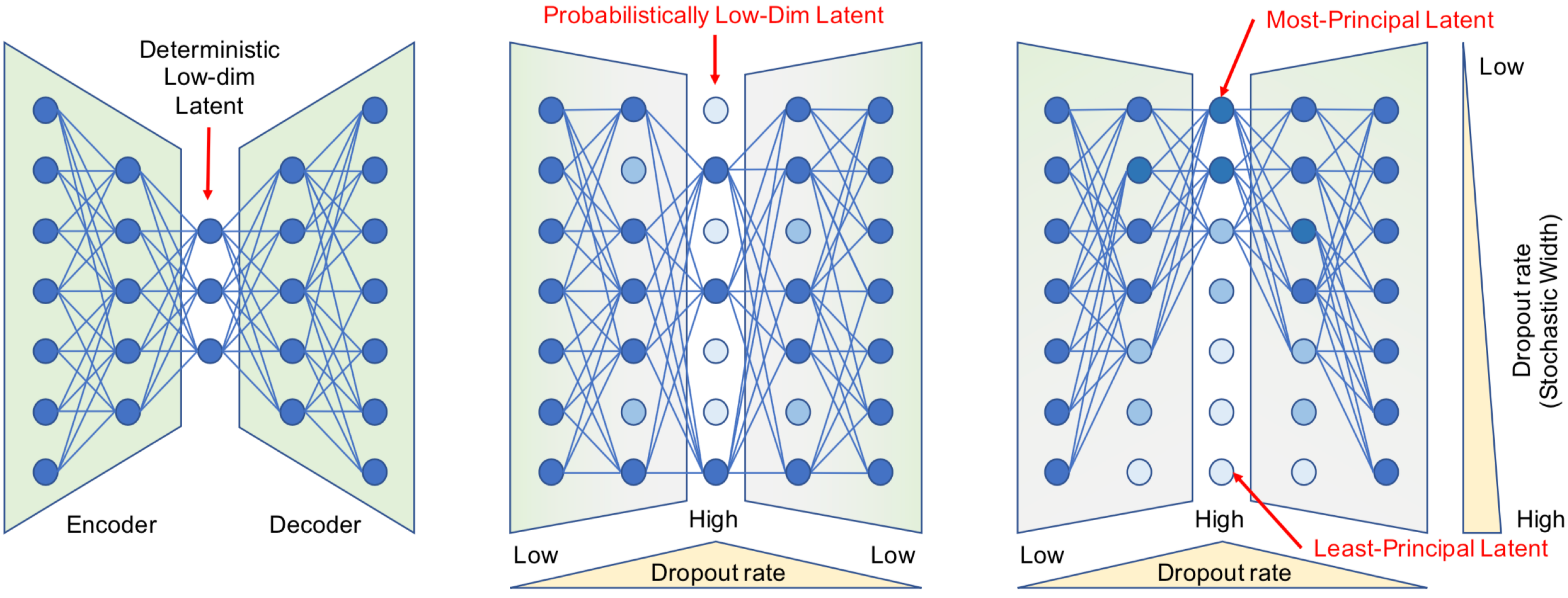






Proposal: Stochastic Bottleneck

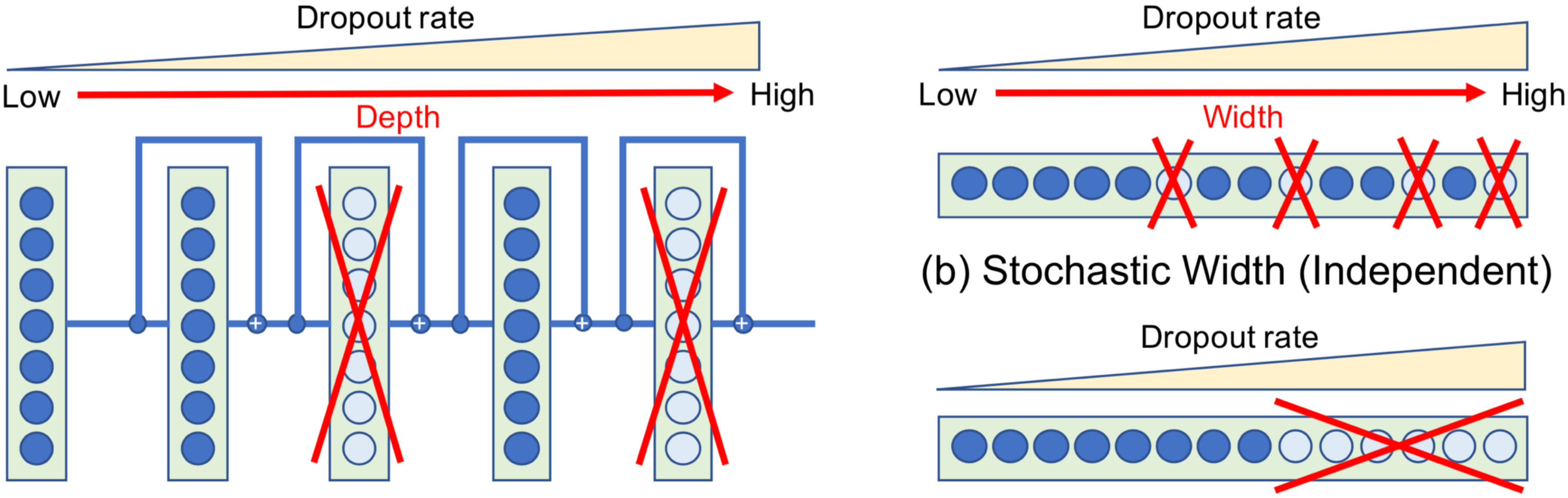
- Simple idea: Non-uniform dropout mechanism



(a) Conventional Bottleneck AE　(b) Sparse AE　(c) Stochastic Bottleneck AE

Stochastic Width: Tail Drop

- Non-uniform dropout has been used in StochasticDepth for ResNet
- Not only depth direction, we use width direction to concentrate important feature in upper neurons



(a) Stochastic Depth

(b) Stochastic Width (Independent)

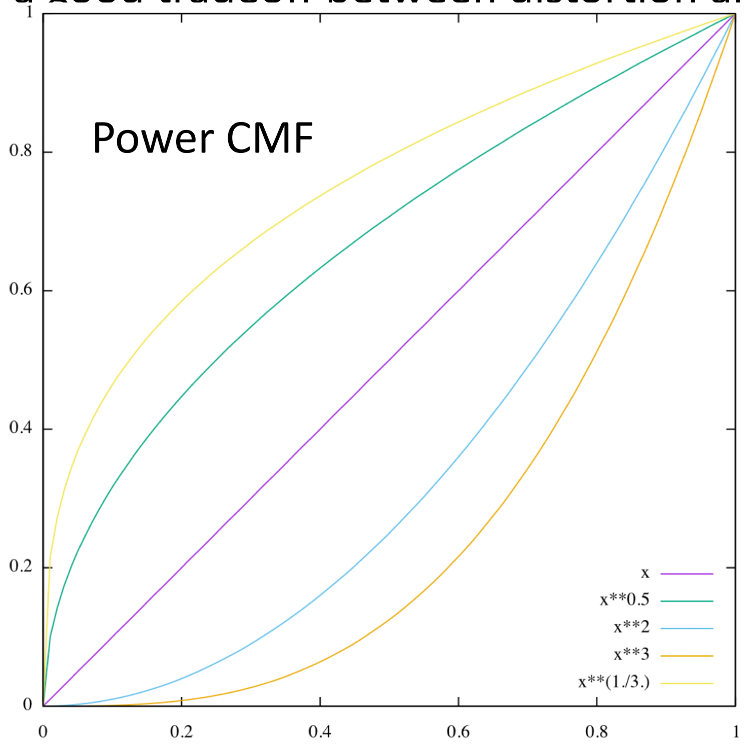
(c) Stochastic Width (Tail Drop)

Drop random burst length of tail

Tail Drop Distribution

- We tested various eigenspectrum model: Poisson, Laplacian, exponential, sigmoid, Lorentzian, polynomial, and Wigner distribution
- Power cumulative mass function (CMF) showed a good tradeoff between distortion and compression rate.
- Best power order parameter is chosen dependent on datasets

$$\Pr(D < \tau M) = \tau^{\beta}$$



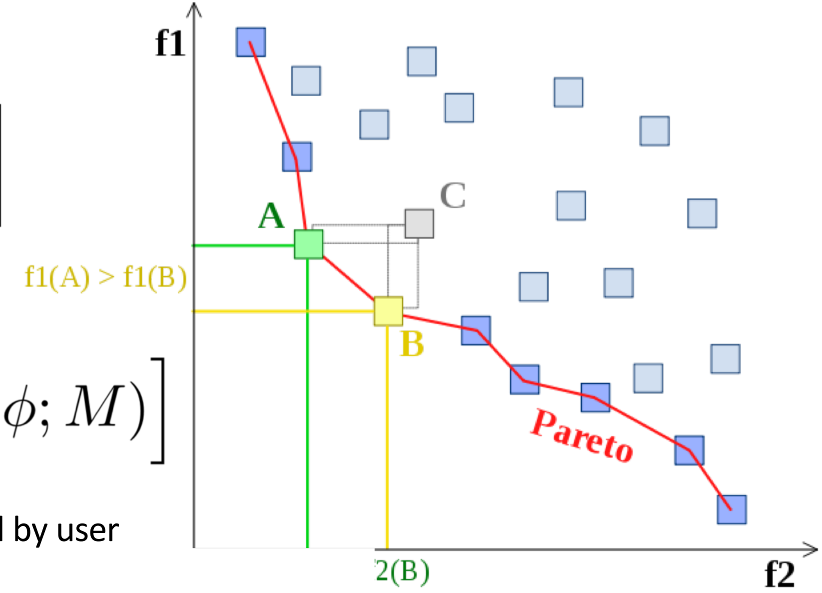
Multi-Objective Learning

- Rateless objective is multi-task learning

Single: $$\min_{\theta,\phi} \mathop{\mathbb{E}}_{\mathbf{x}\sim\Pr(\mathbf{x})} \left[\mathcal{L}(\mathbf{x}, g_\phi(f_\theta(\mathbf{x})))\right]$$

Multi: $$\min_{\theta,\phi} \left[\bar{\mathcal{L}}(\theta,\phi;1), \bar{\mathcal{L}}(\theta,\phi;2), \ldots, \bar{\mathcal{L}}(\theta,\phi;M)\right]$$

$\bar{\mathcal{L}}(\theta,\phi;L)$:Expected loss when the first *L* latent variables retained by user



Simple multi-objective optimization with weighted sum:

$$\min_{\theta,\phi} \sum_{L=1}^{M} \omega_L \bar{\mathcal{L}}(\theta,\phi;L)$$

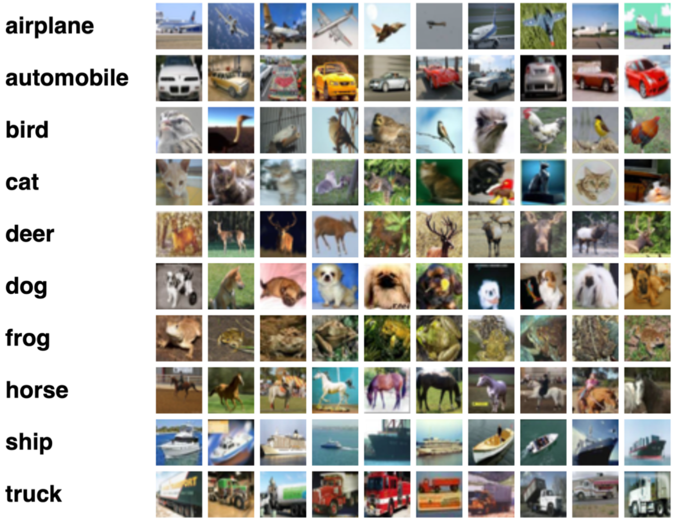
$$\Pr(D = M - L) = \omega_L$$

e.g.) balanced weights: $\omega_L \simeq 1/\bar{\mathcal{L}}^\star(\theta,\phi;L)$

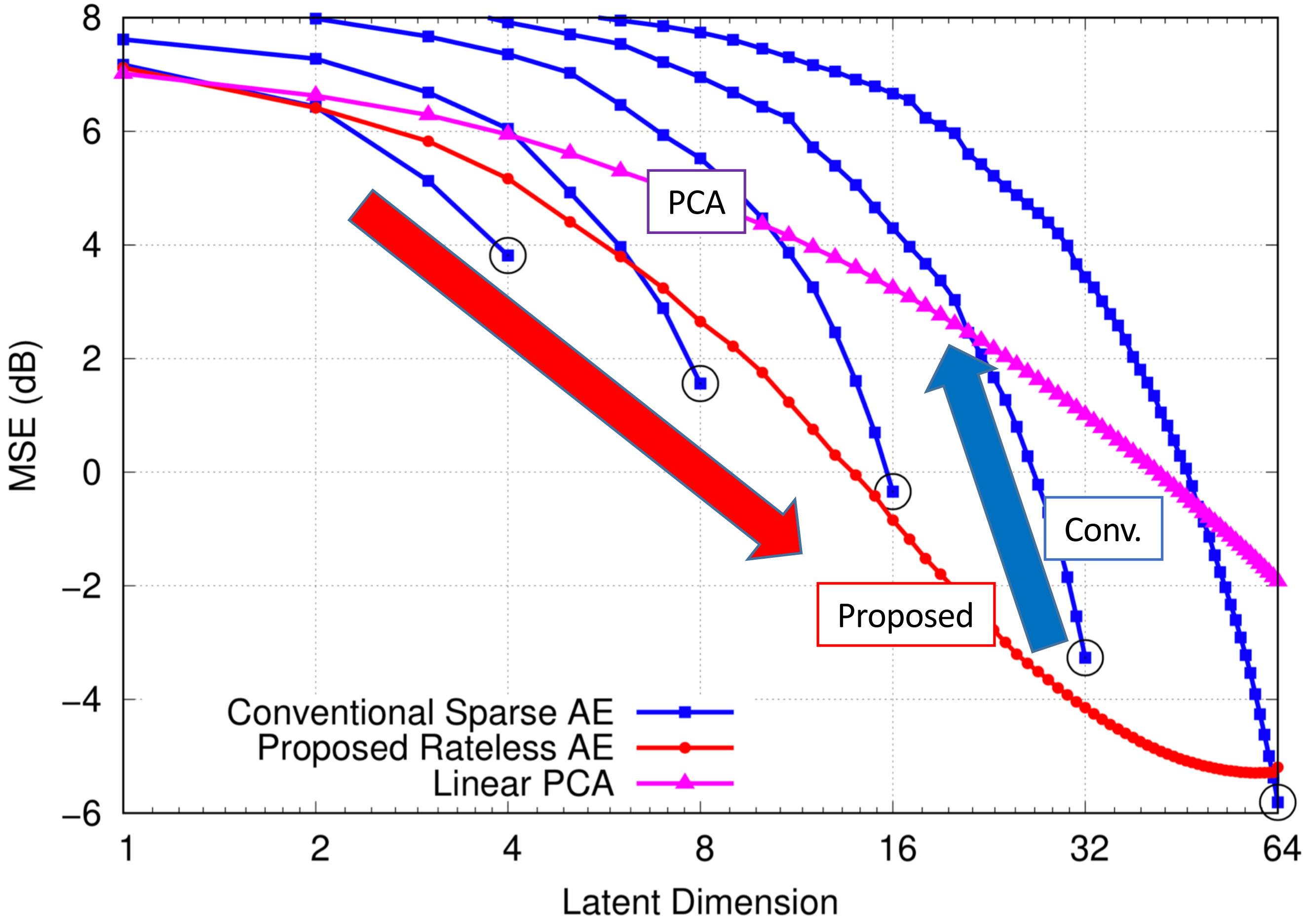
Weighted metric method, …

Toy Experiments

- AE architecture
  - 3 layers 1024 or 2048 nodes
  - Adam (0.001)
  - Mini-batch 100
  - Max 500 epochs
  - Power CMF TailDrop
- Datasets
  - **MNIST**
  - **CIFAR-10**
  - FMNIST
  - KMNIST
  - SVHN
  - CIFAR-100

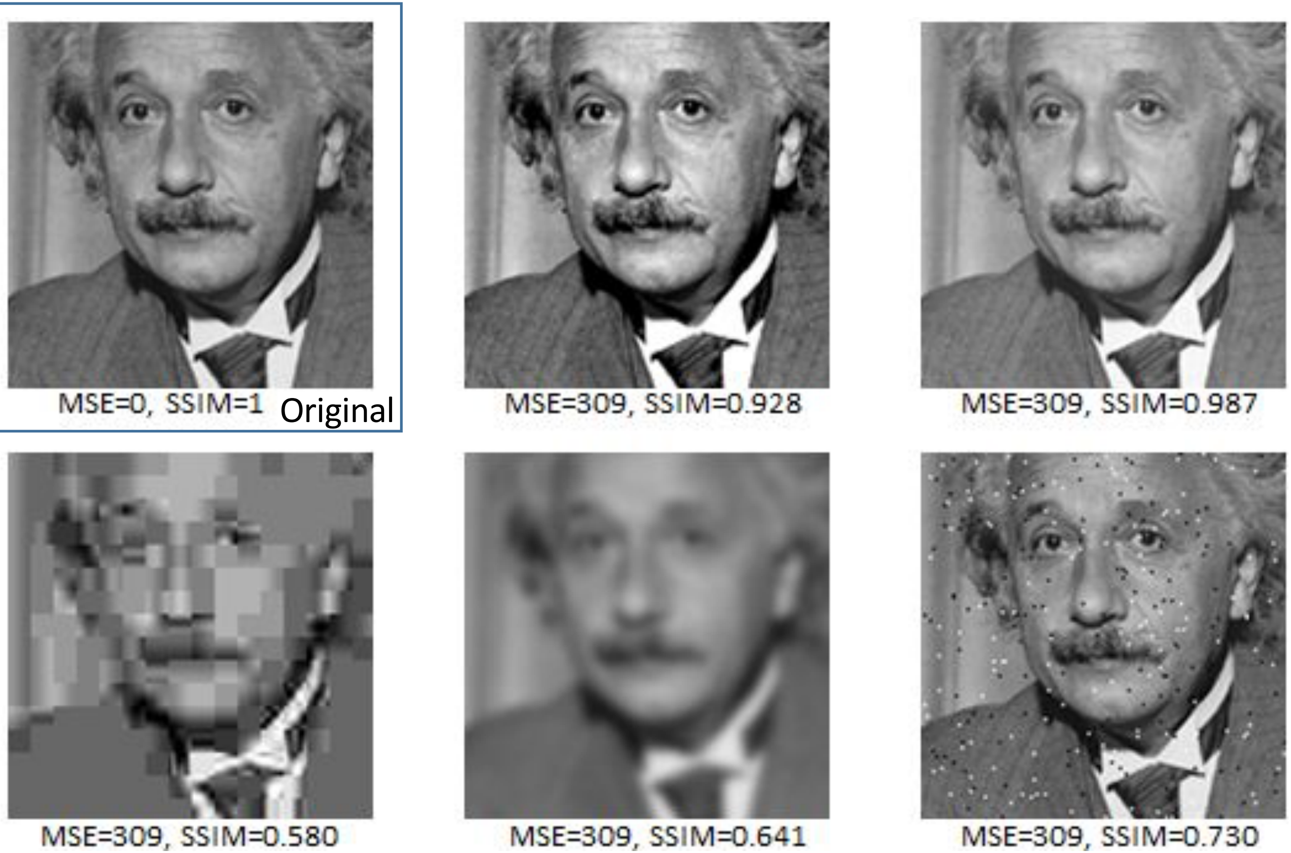


MSE Distortion Measure (MNIST)

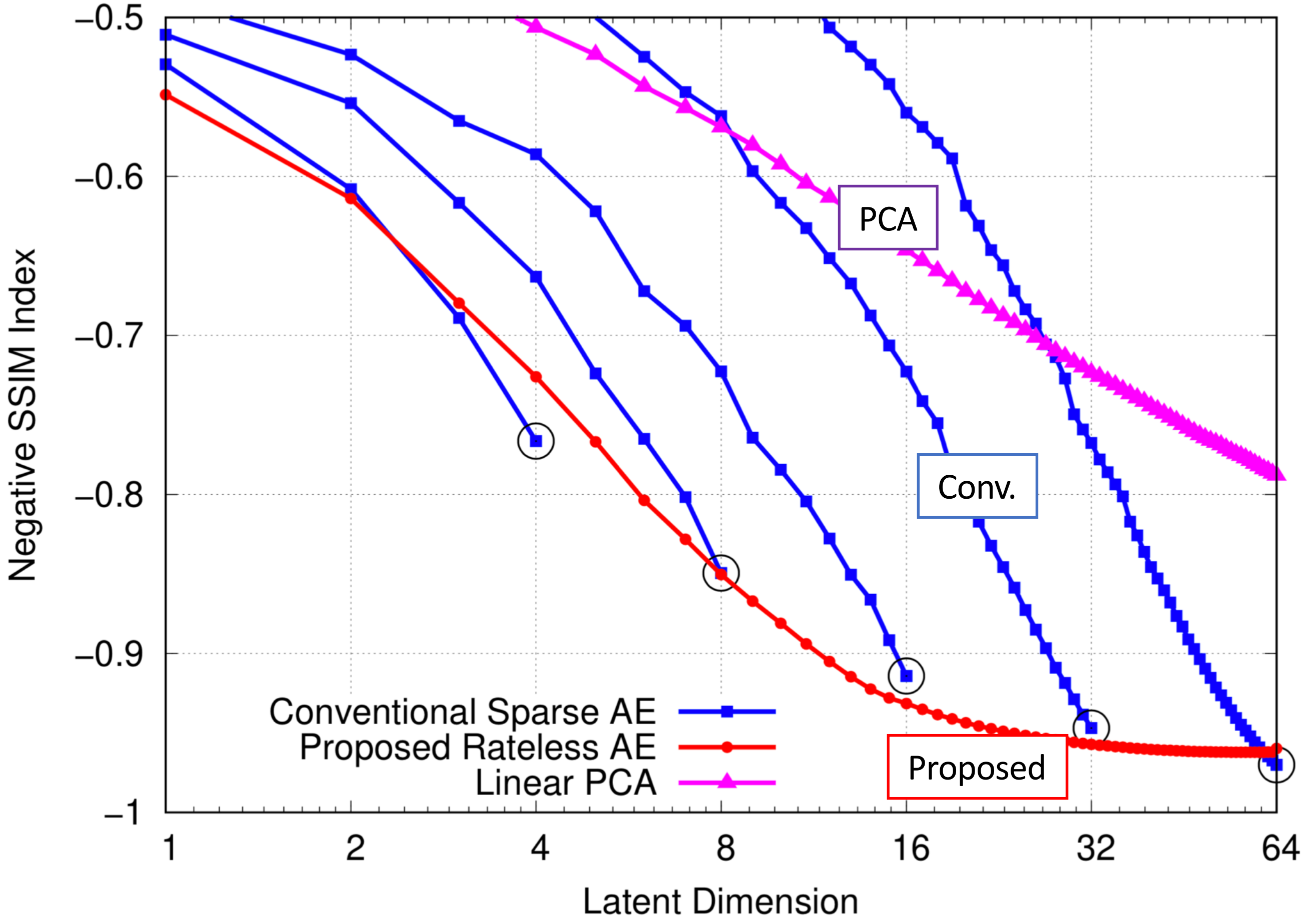


MSE vs. Structural Similarity (SSIM)

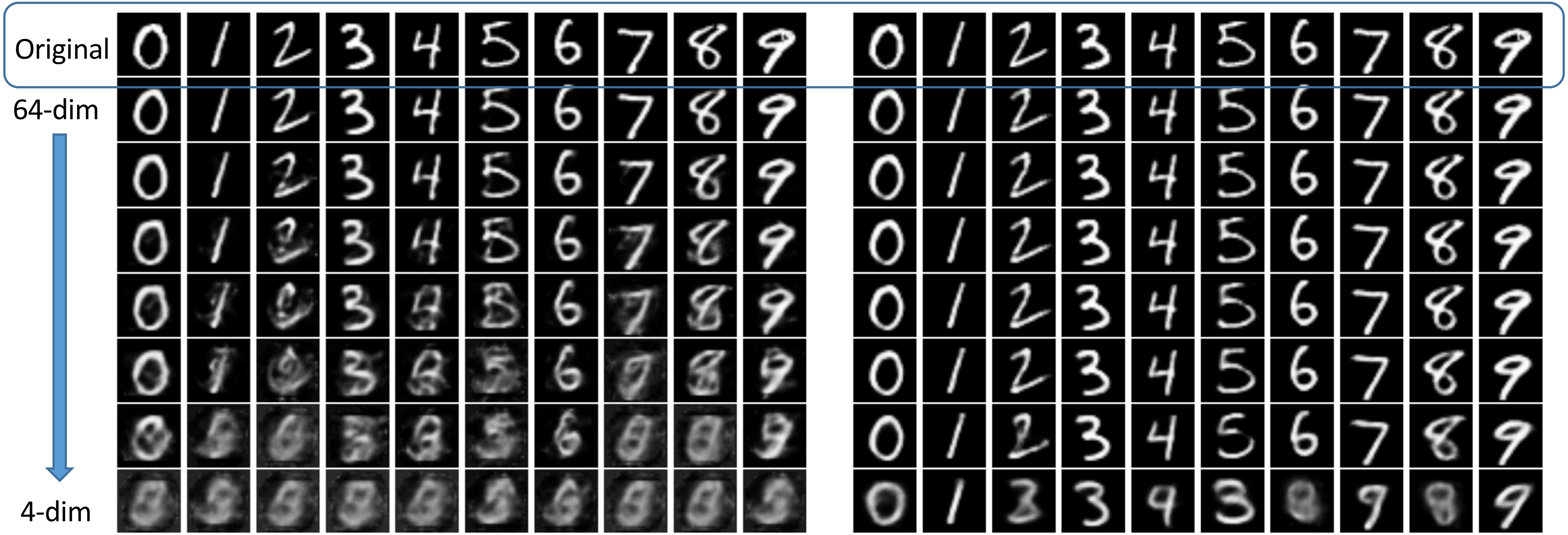
- MSE does not fully tell perceptual distortion



SSIM Distortion Measure (MNIST)



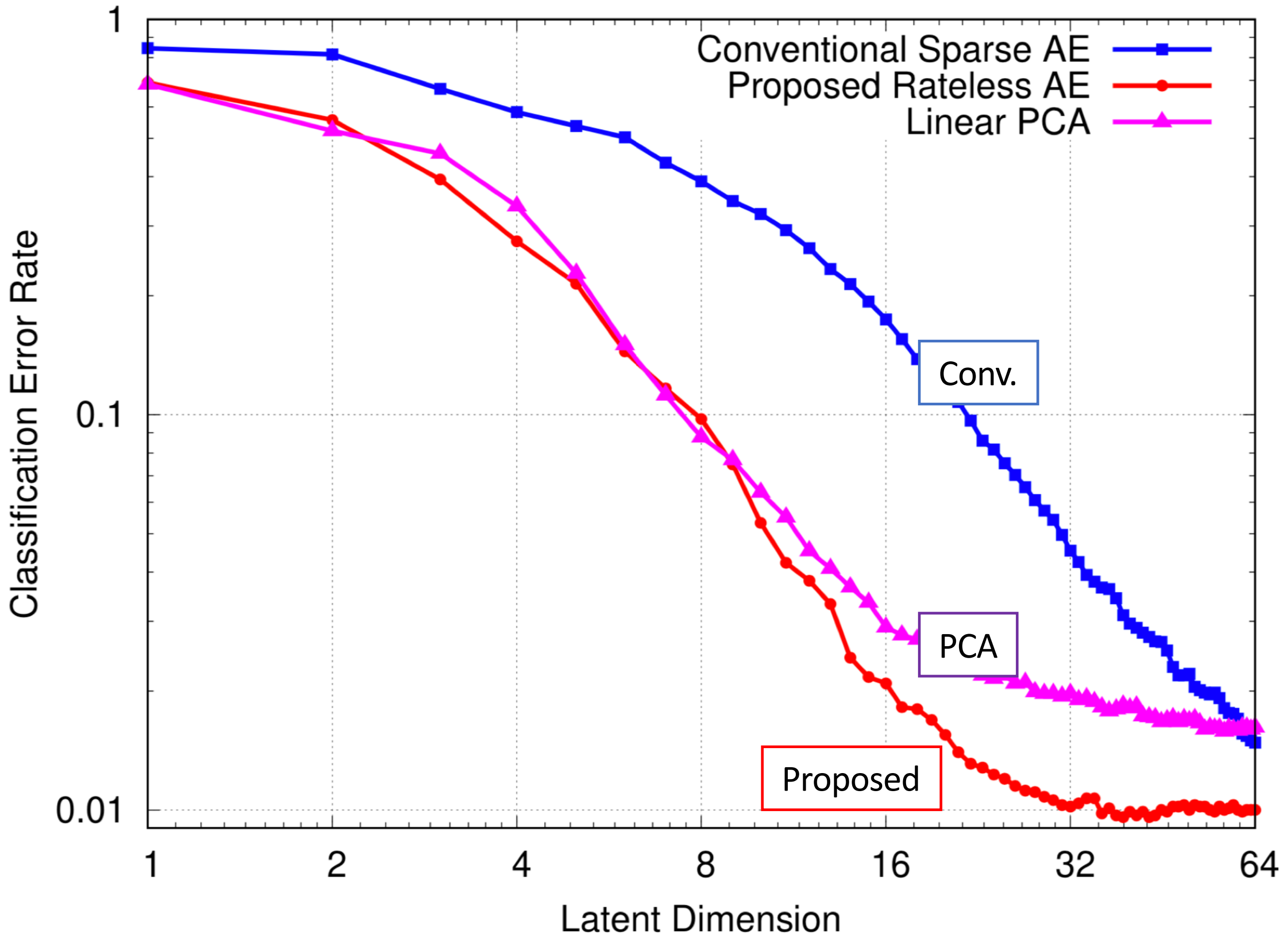
Reconstructed Image Snapshots



Conventional AE

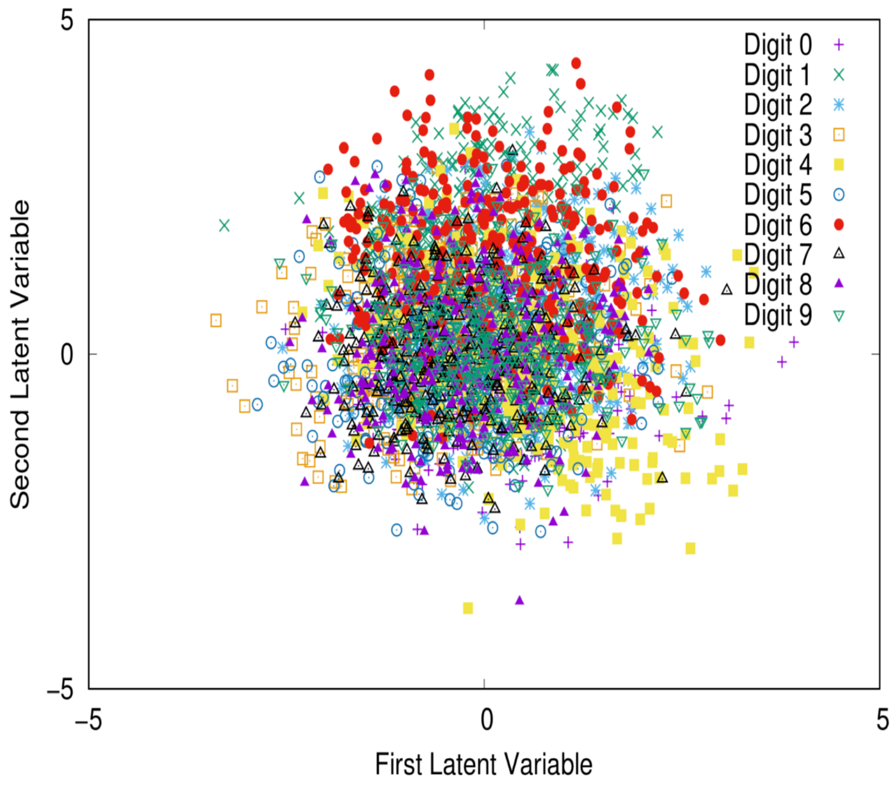
Proposed AE

Support Vector Machine (SVM) Classification (MNIST)

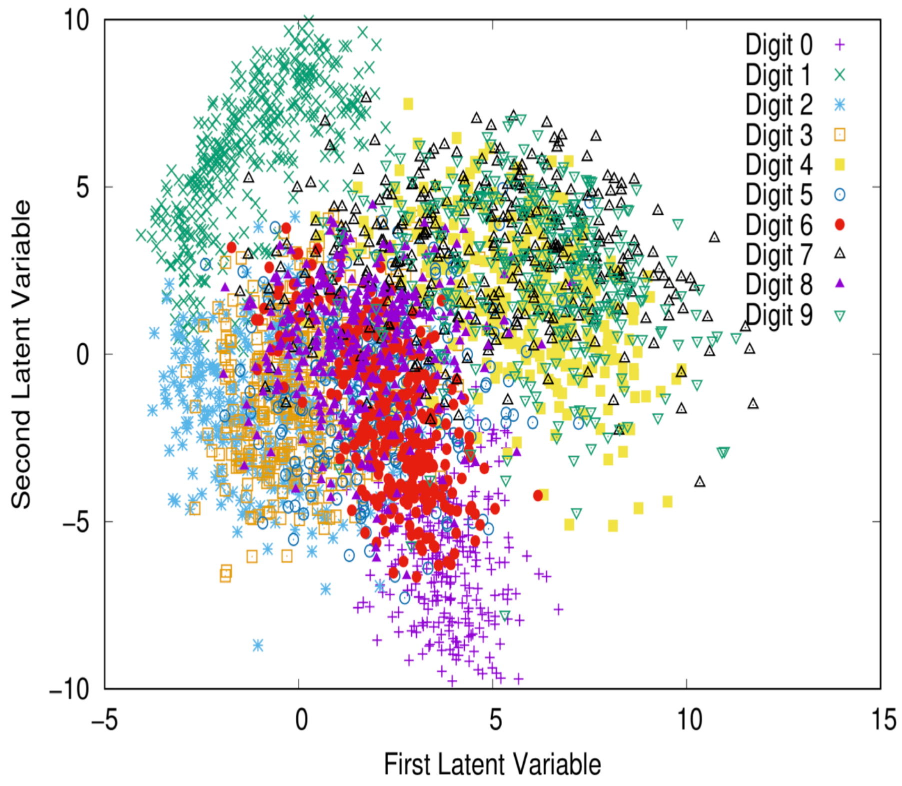


Latent Space Geometry

- The first 2 latent variables



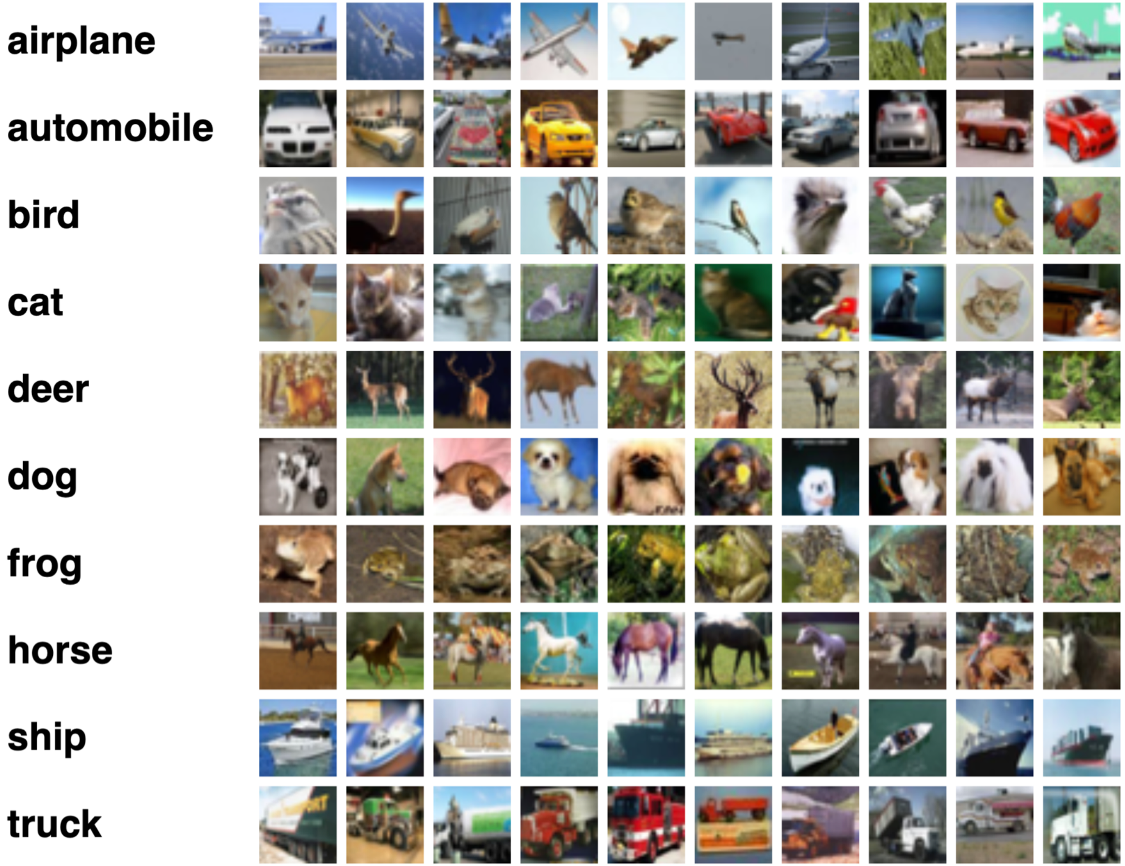
Conventional AE



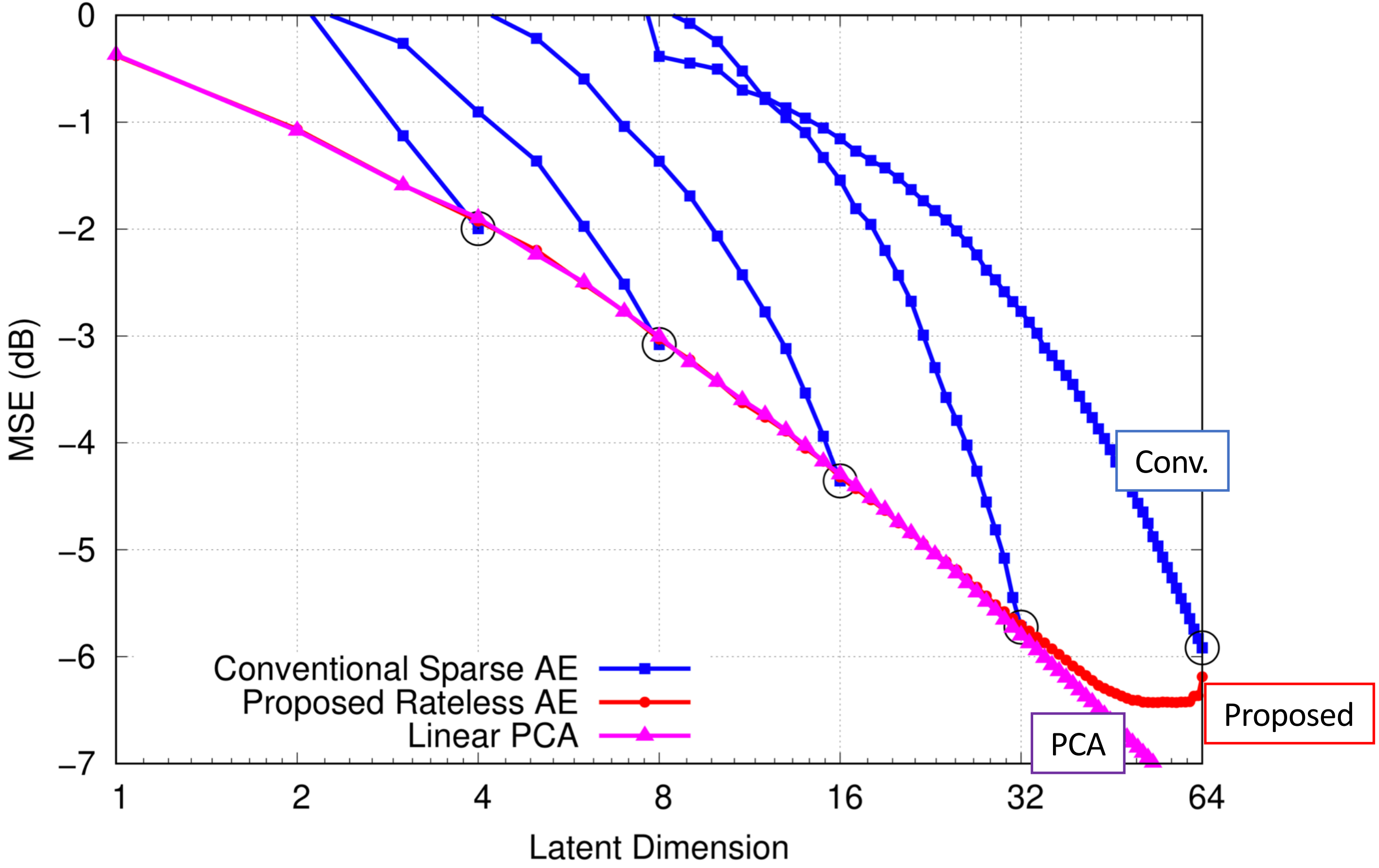
Proposed AE

CIFAR-10

- 32x32 color images
- 10-class natural photos
- 50,000 training
- 10,000 test



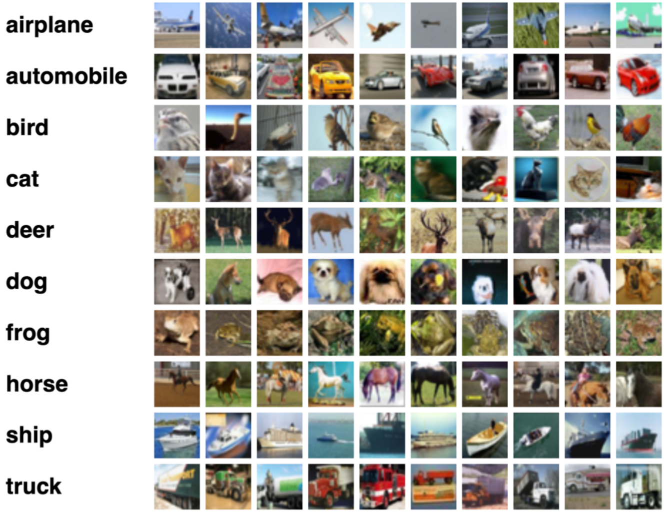
MSE Distortion (CIFAR-10)



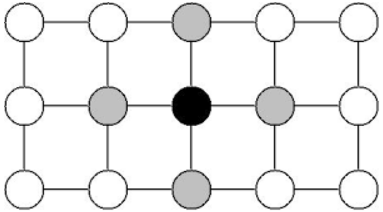
MNIST vs. CIFAR-10

- MNIST data is gray-scale image, but nearly binary (white or black) whose statistics are far from Gaussian distribution
- CIFAR-10 uses color natural photos. Such photos are well modeled by Gauss-Markov random field (GMRF)
- Hence, PCA surprisingly performs well for CIFAR-10 if we consider MSE distortion
- However, SSIM and accuracy measure …

MNIST: Bernoulli like

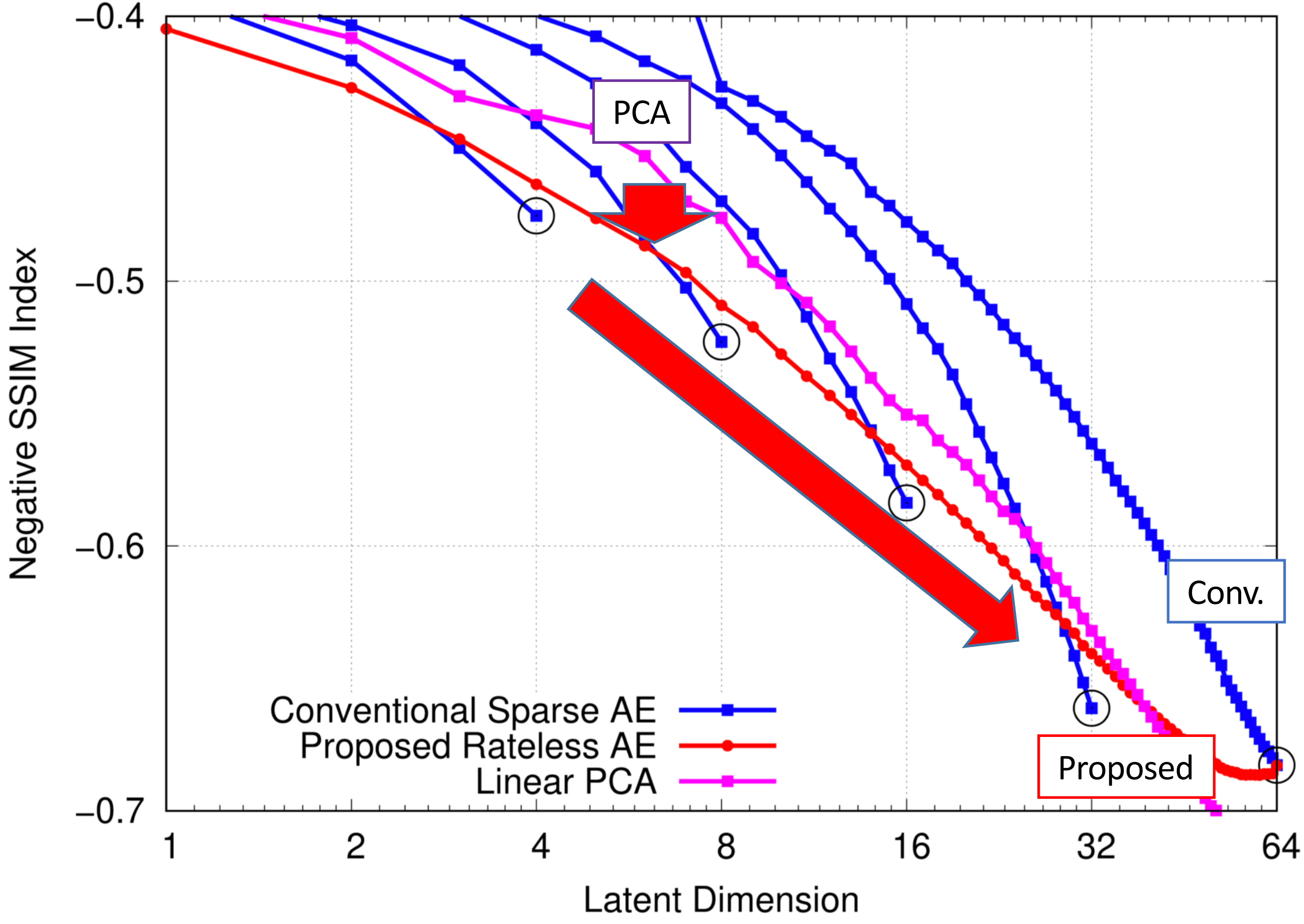


CIFAR-10: Gauss like

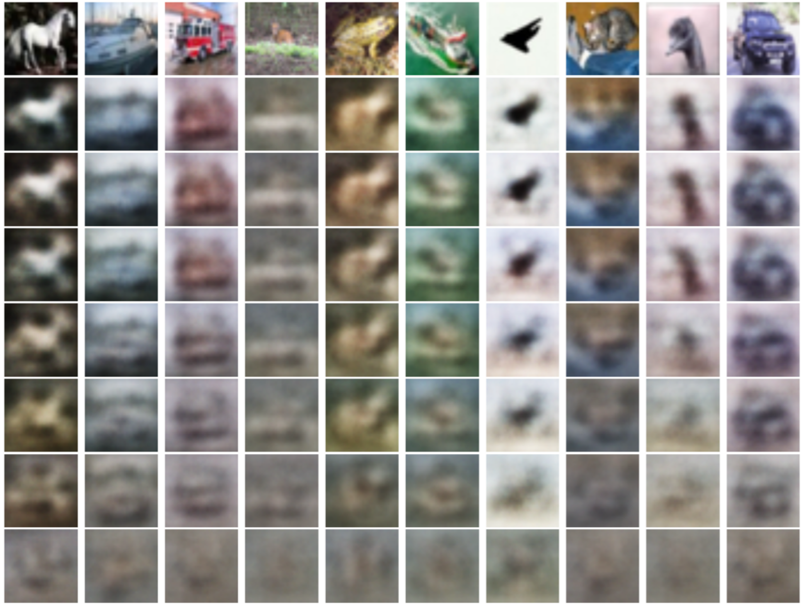


Gauss-Markov

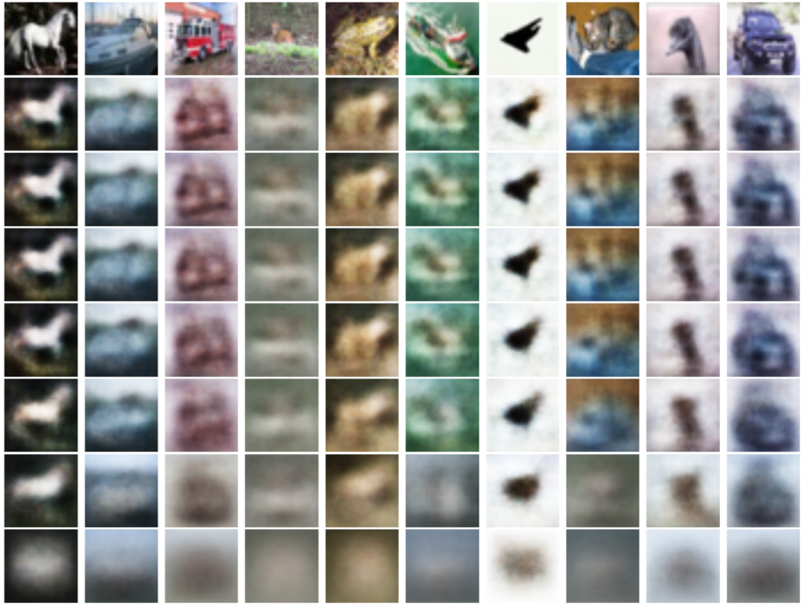
SSIM Measure (CIFAR-10)



Reconstructed Image Snapshots (CIFAR-10)



Conventional AE



Proposed AE

Dimensionality L		64	54	44	34	24	14	4
MSE (dB)	Conv. AE	−5.92	−4.96	−3.96	−2.97	−1.91	−0.96	0.92
	Prop. AE	**−6.19**	**−6.43**	**−6.30**	**−5.82**	**−5.11**	**−4.05**	**−1.92**
SSIM Index	Conv. AE	0.64	0.61	0.57	0.53	0.48	0.44	0.37
	Prop. AE	**0.66**	**0.67**	**0.67**	**0.64**	**0.60**	**0.54**	**0.44**
SVM Acc.	Conv. AE	**0.47**	0.47	0.46	0.44	0.40	0.32	0.20
	Prop. AE	**0.47**	**0.48**	**0.47**	**0.48**	**0.46**	**0.42**	**0.29**

Example Use Case

Rateless:
- Single unified AE model regardless of dimensionality for application invariant

Conventional:
- What purpose?
- Dimensionality?
- Which AE models?

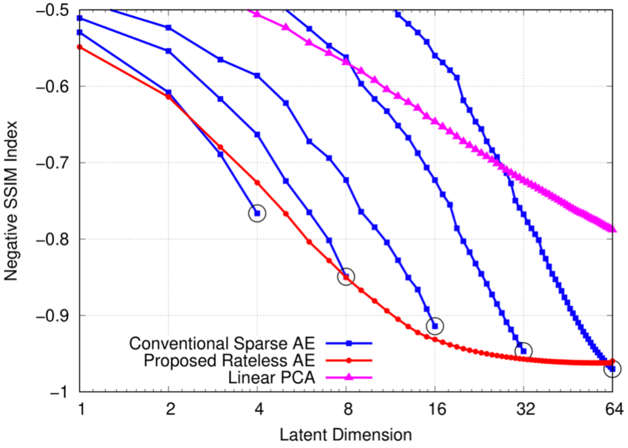
Moderate dimension
we need to diagnose!

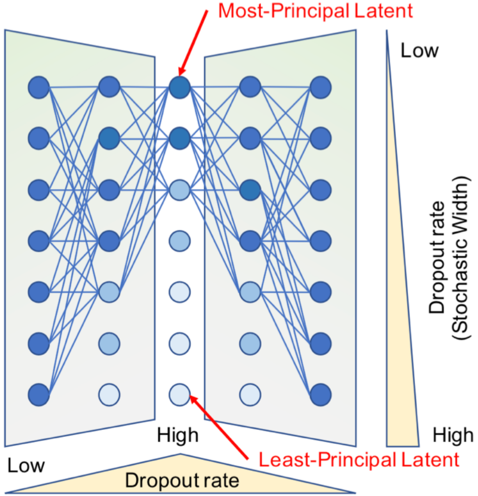
All dimensions
we need to analyze!

We do not care many
but final results

Summary

- We introduced a new **rateless** concept in auto-encoder design
- We proposed **Stochastic Bottleneck** architecture
  - Non-identical dropout rates for **Stochastic Width** and Depth
- New regularization called **TailDrop** was investigated
- Proposed AE offers an excellent trade-off between distortion and compression rates
  - Benefits in MSE, SSIM, and SVM accuracy were confirmed
- Demonstrated the benefit for various benchmark datasets





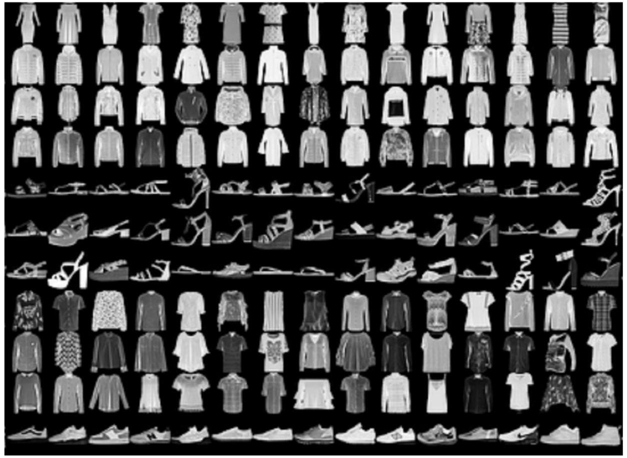
(c) Stochastic Bottleneck AE

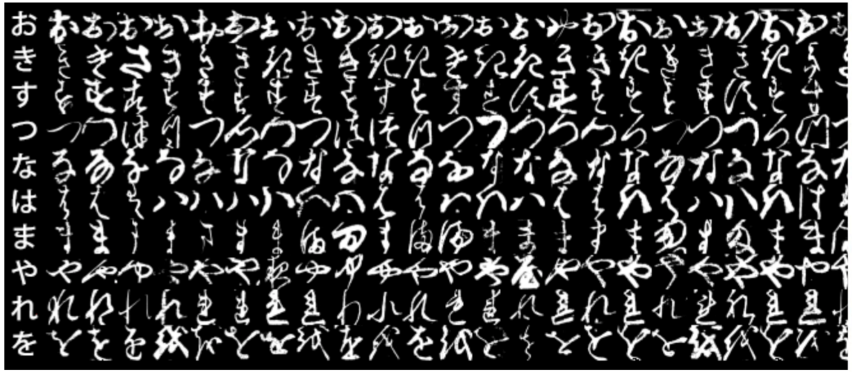
- Questions?
  - koike@merl.com
  - More results in arXiv 2005.02870

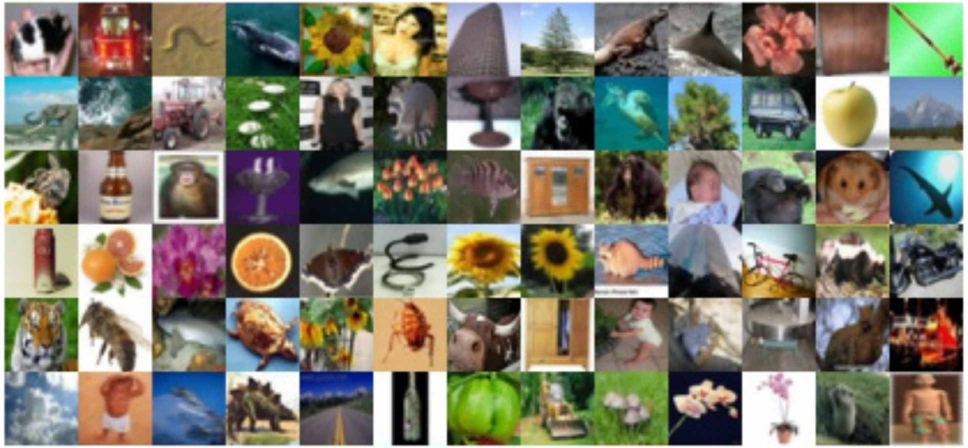
MITSUBISHI ELECTRIC
Changes for the Better

More Results in ArXiv

- Datasets
  - MNIST
  - CIFAR-10
  - **FMNIST**
  - **KMNIST**
  - **SVHN**
  - **CIFAR-100**

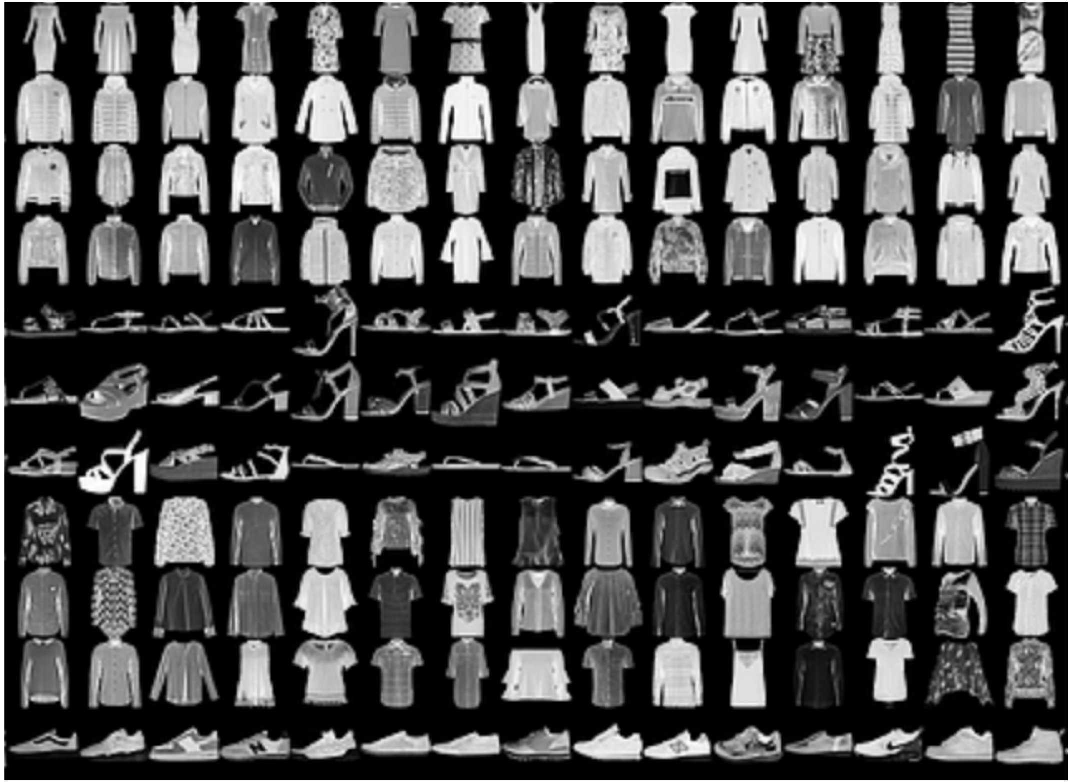




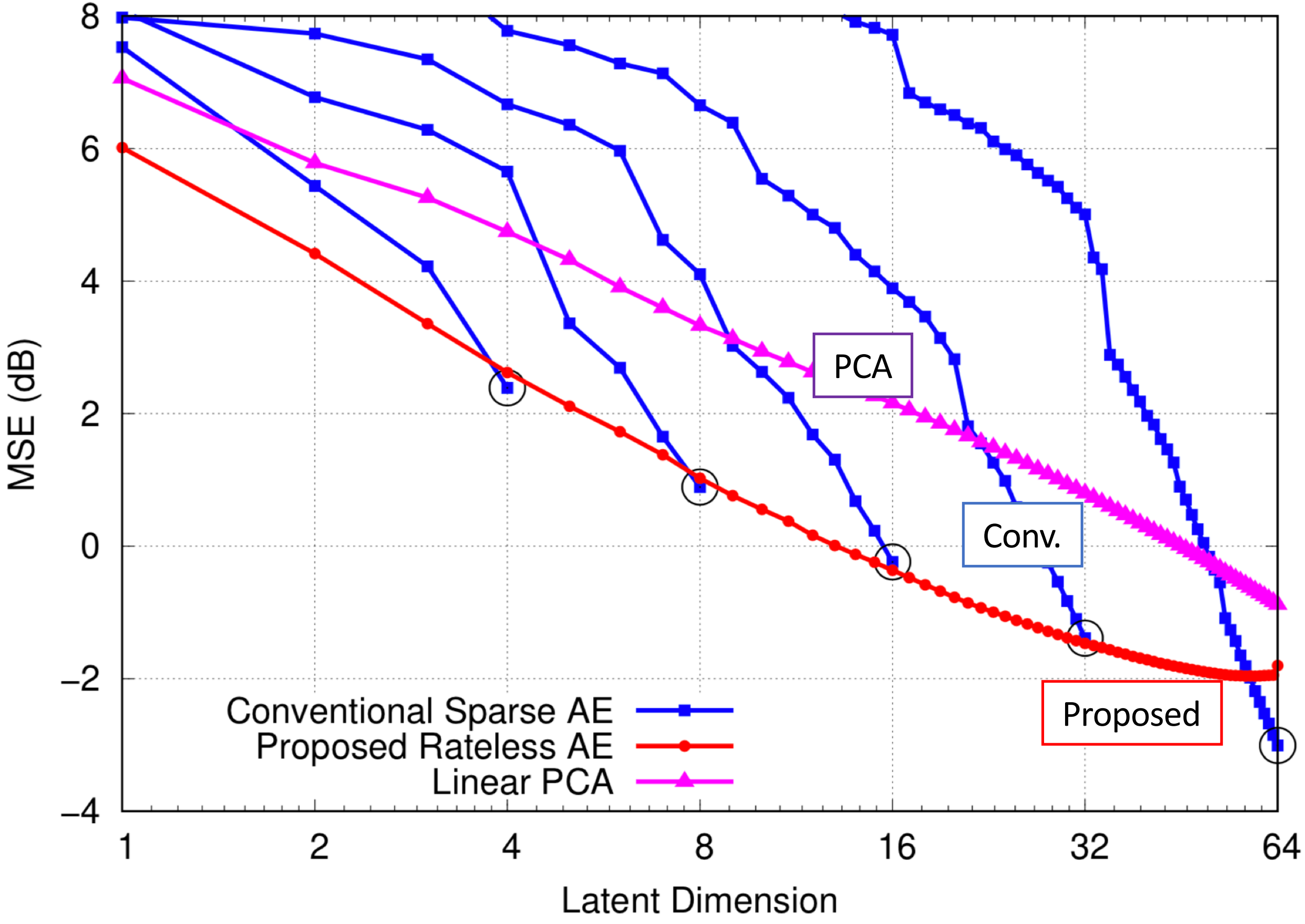


Fashion MNIST (FMNIST)

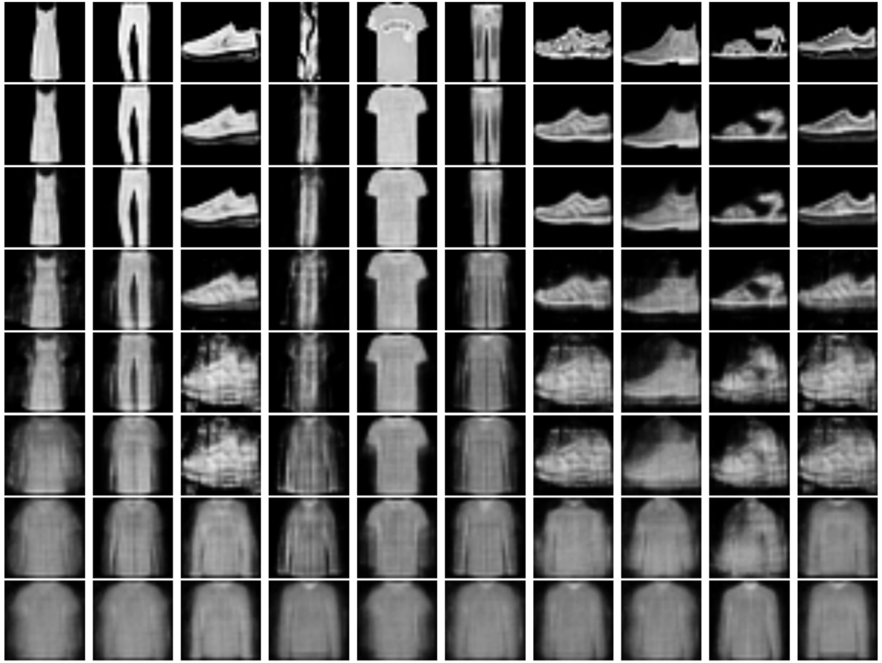
- 28x28 gray-scale images
- 10-class fashion photos
- 60,000 train
- 10,000 test



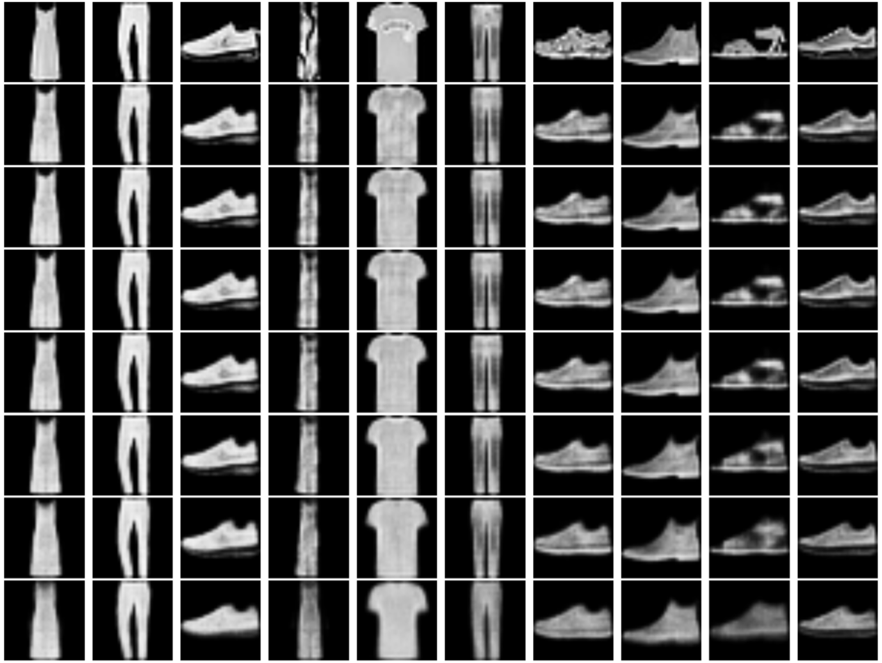
MSE Measure (FMNIST)



Snapshots (FMNIST)



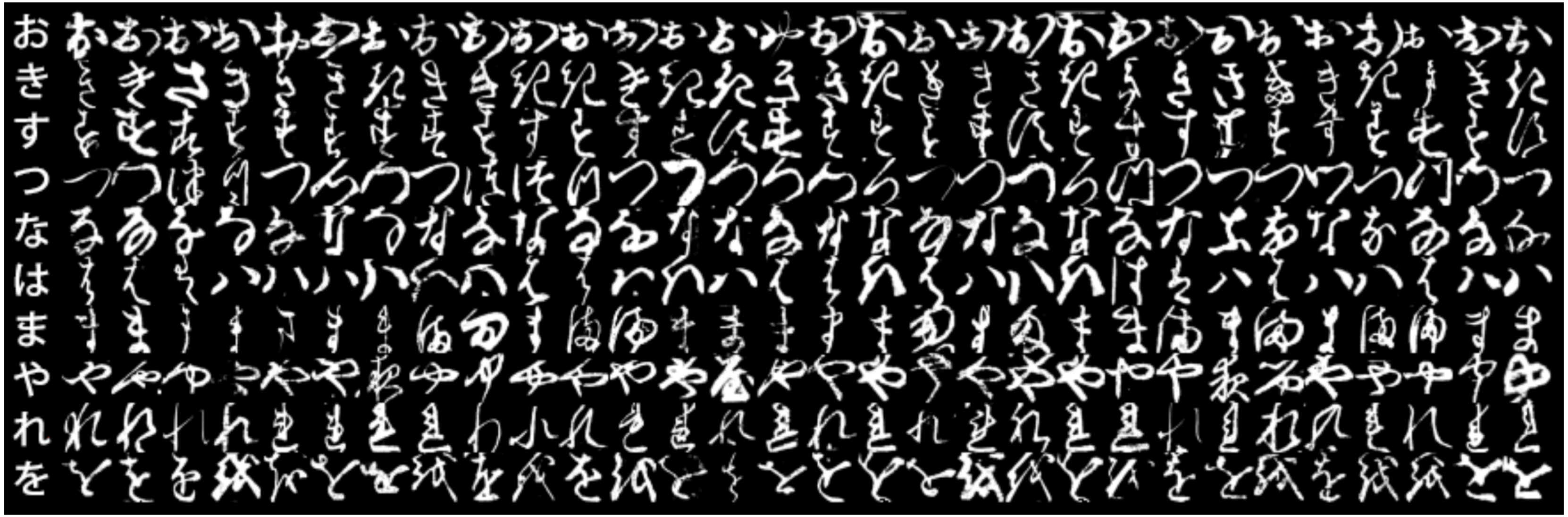
Conventional AE



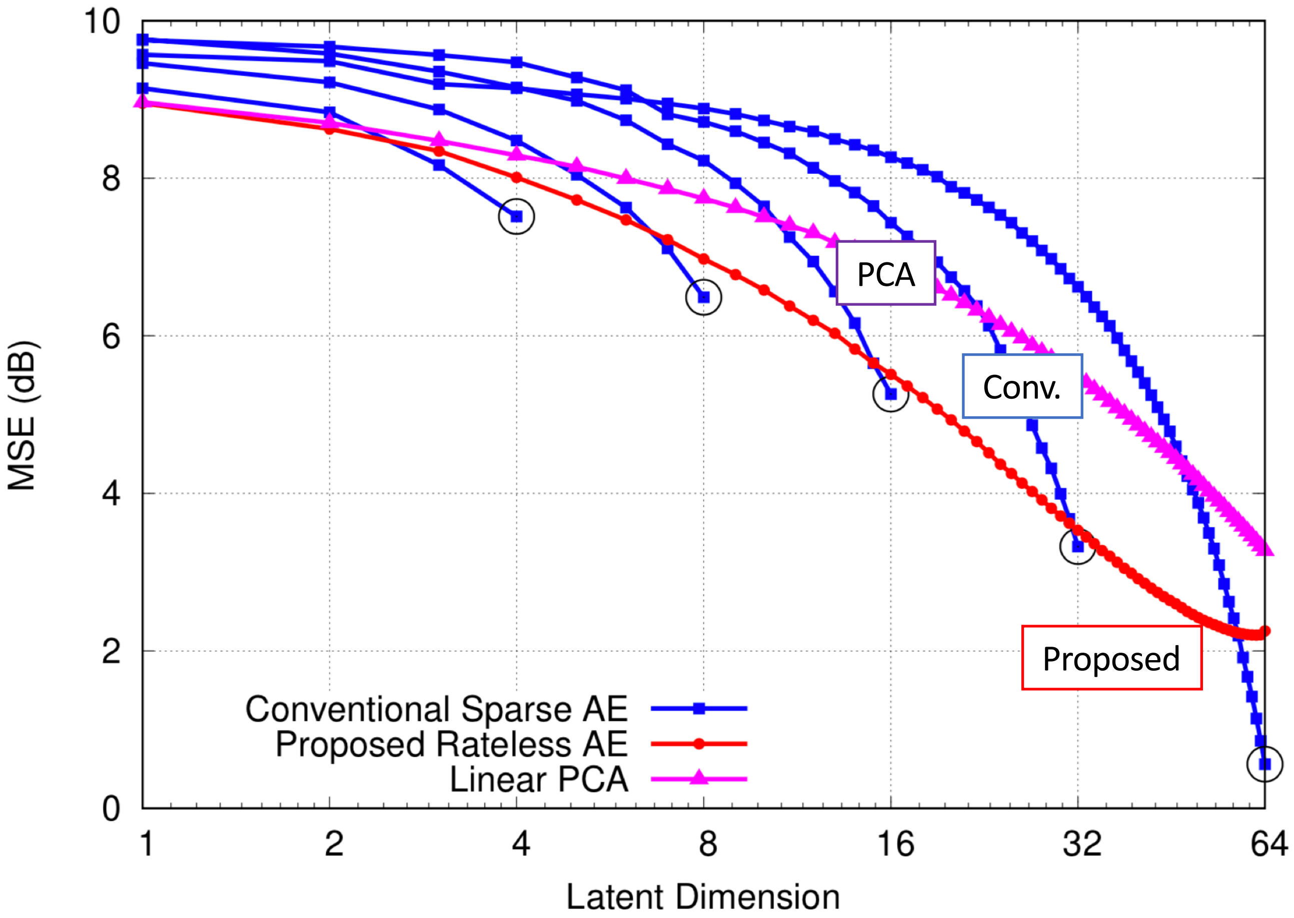
Proposed AE

Kuzushiji MNIST (KMNIST)

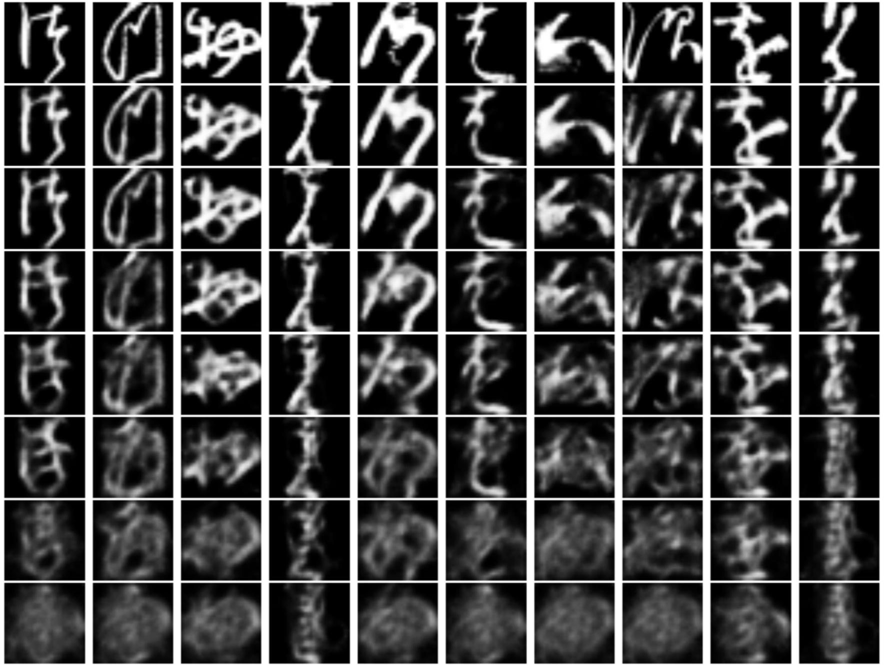
- 28x28 gray-scale images
- 10-class ancient Japanese letters
- 60,000 training data
- 10,000 test data



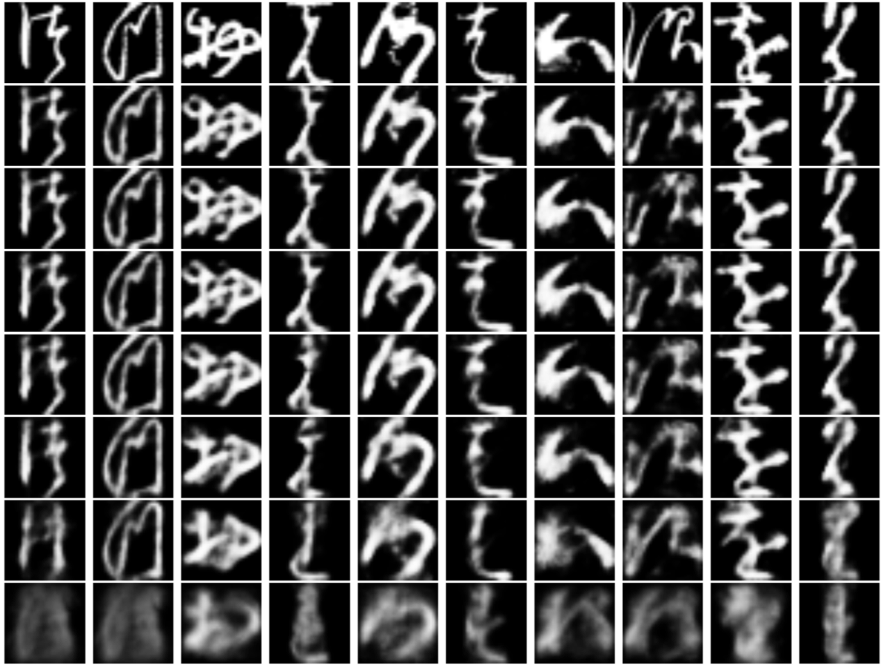
MSE Measure (KMNIST)



Snapshots (KMNIST)



Conventional AE

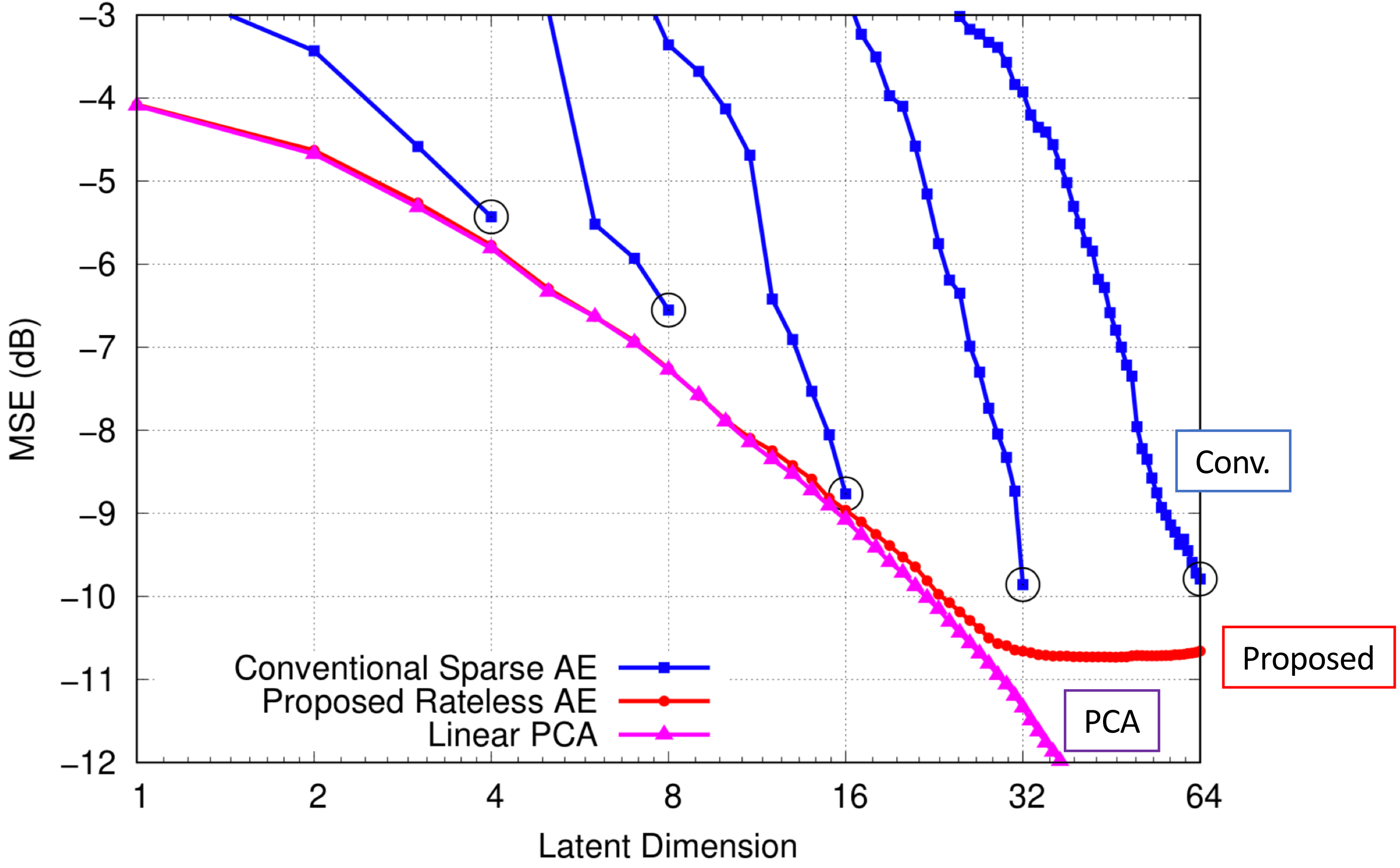


Proposed AE

Street-View House Numbers (SVHN)

- 32x32 color images
- 10-class cropped digits
- 73,257 training
- 26,032 test

MSE Measure (SVHN)



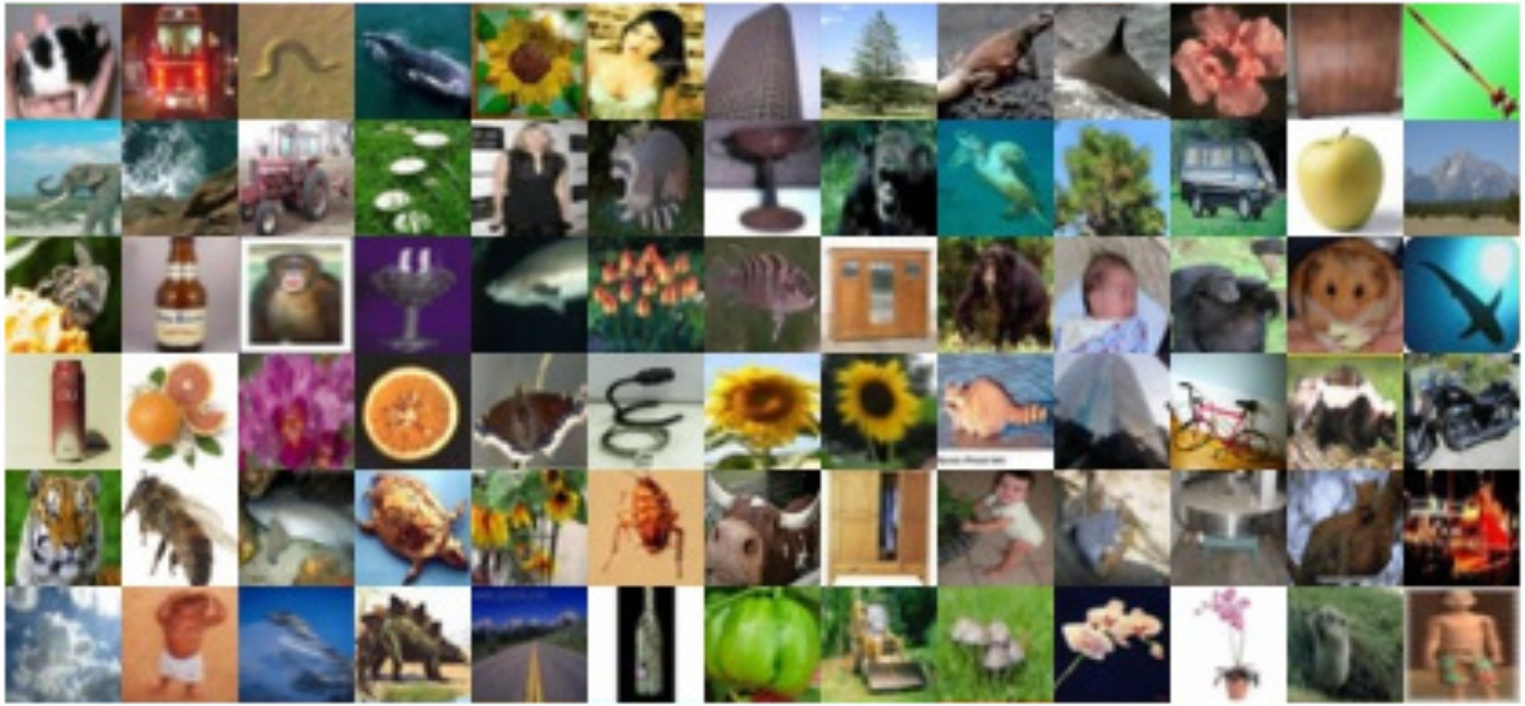
Snapshots (SVHN)

Conventional AE

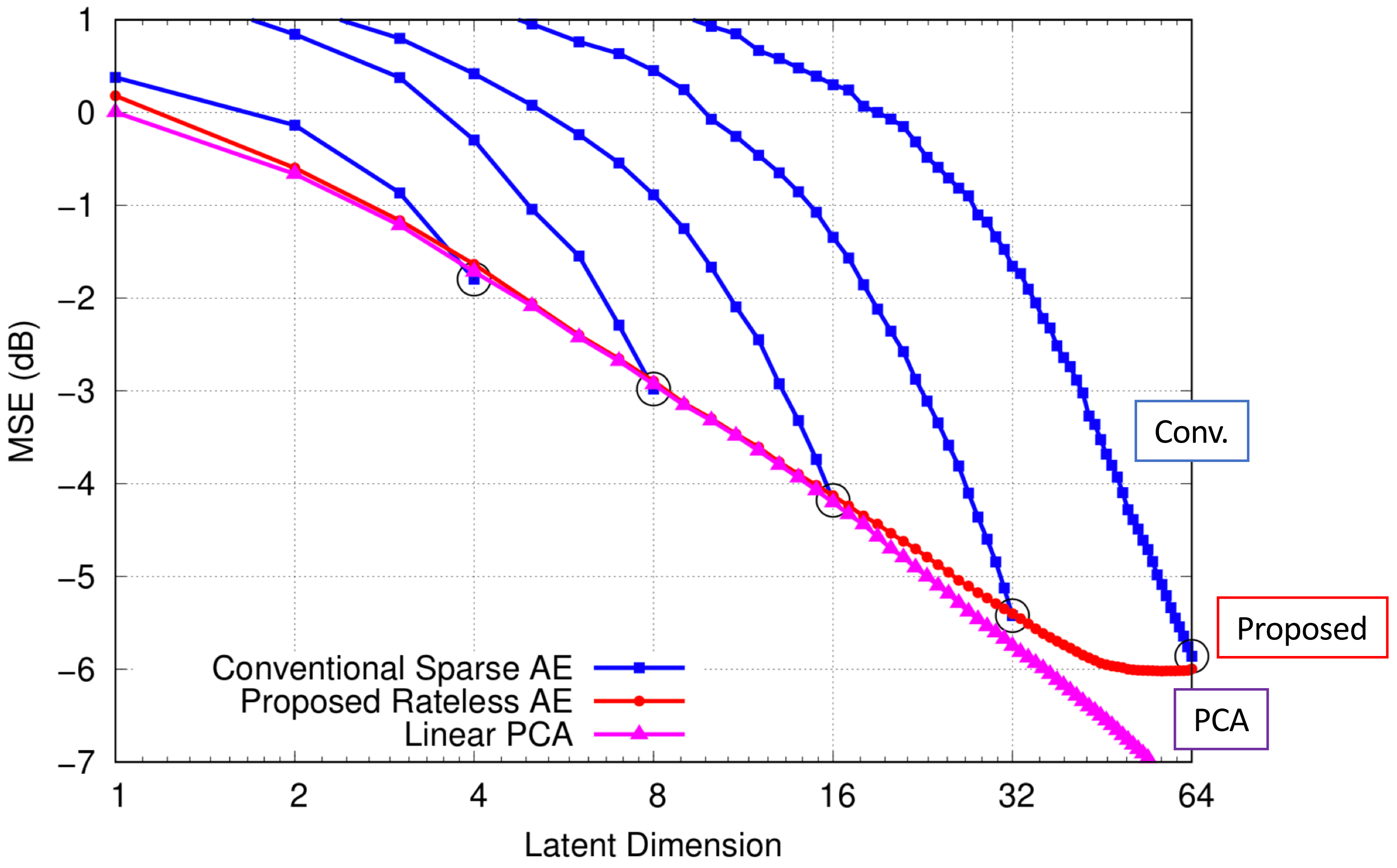
Proposed AE

CIFAR-100

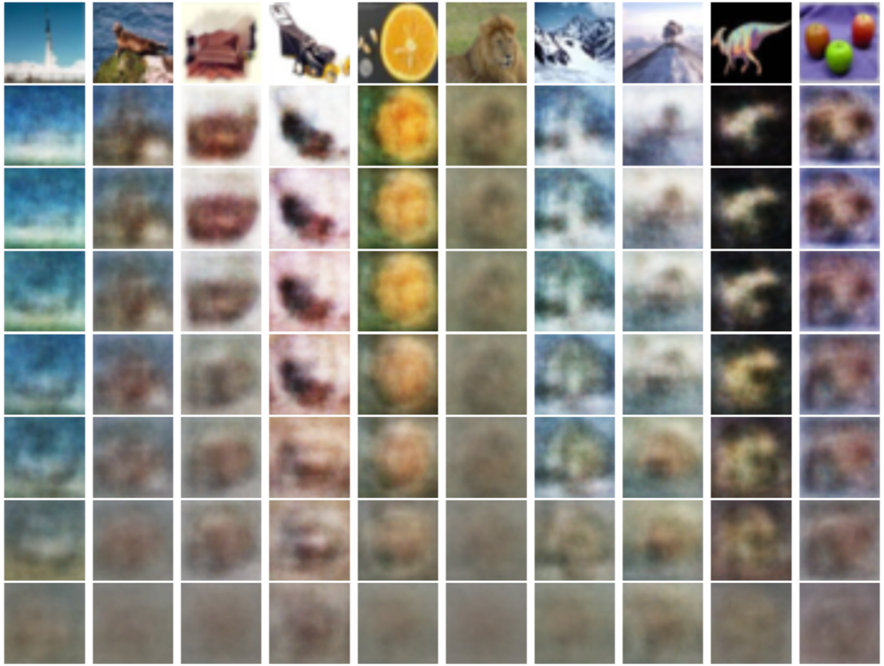
- 32x32 color images
- 100-class natural photos (20 super-classes)
- 50,000 training data, 10,000 test data



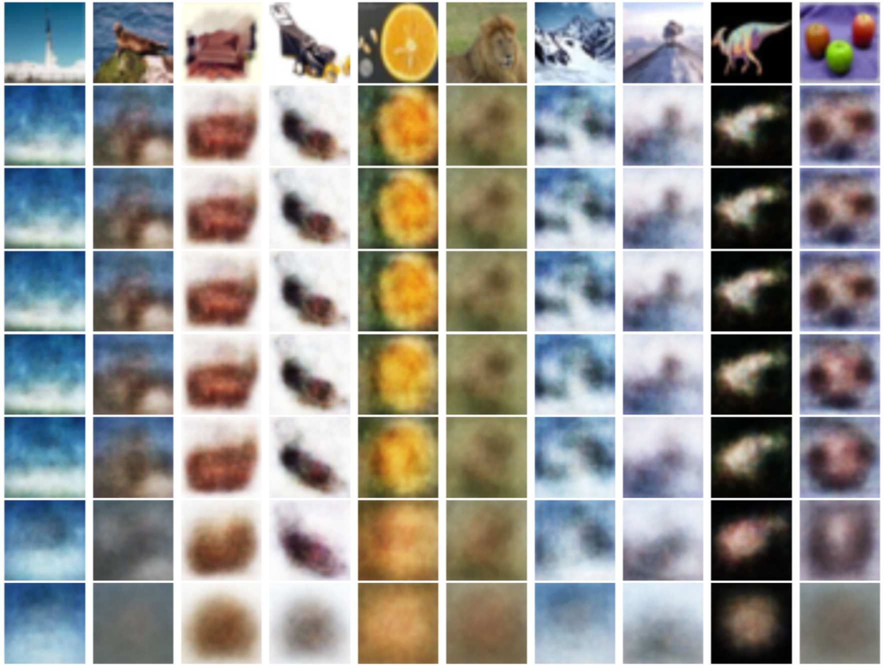
MSE Measure (CIFAR-100)



Snapshots (CIFAR-100)



Conventional AE



Proposed AE

MITSUBISHI
ELECTRIC
Changes for the Better