CS 4824/ECE 4424: Autoencoder

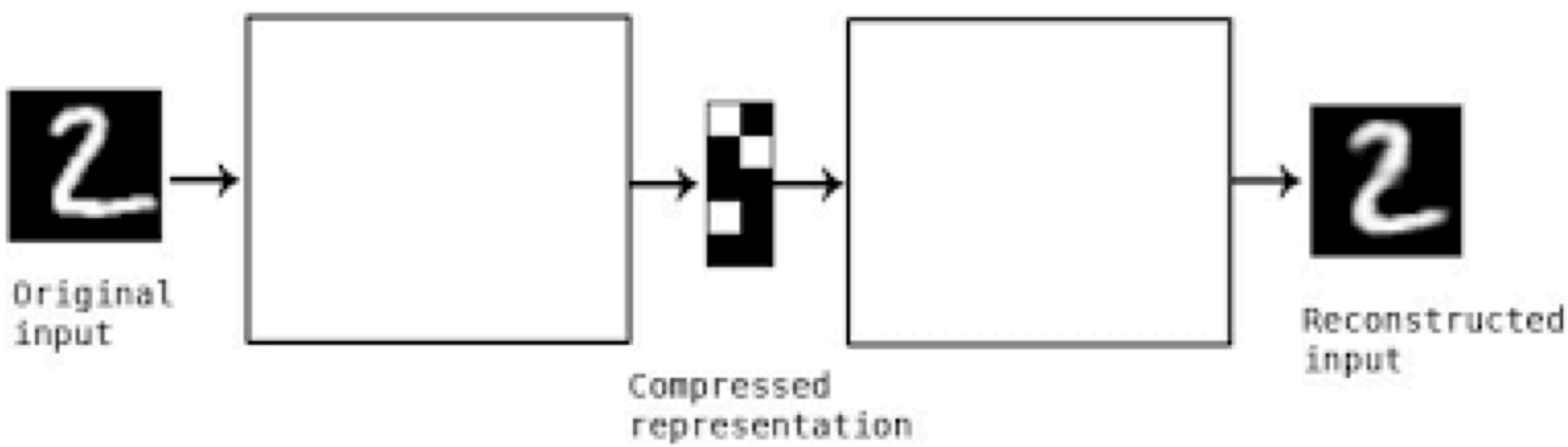
Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

Autoencoder

- Special type of feed forward network for
  - Compression
  - Denoising
  - Sparse representation
  - Data generation

Autoencoder

- Encoder: $f\,(\quad)$
- Decoder: $g\,(\quad)$
- Autoencoder: $g\,(f\,(\mathbf{x})) = \mathbf{x}$



Linear Autoencoder

- f and g are linear
  - Matrix representation: $\mathbf{W}_f$ and $\mathbf{W}_g$

- Schematic

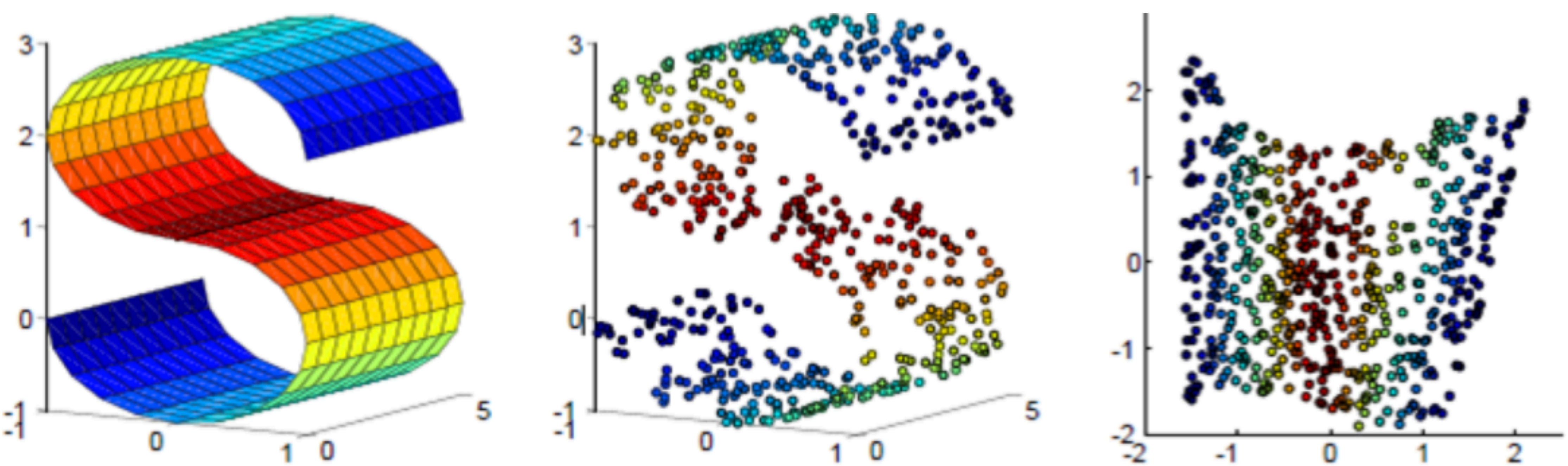
Linear Autoencoder

- **Objective**: find weights $\mathbf{W}_f$ and $\mathbf{W}_g$ that minimizes the reconstruction error
  - $\arg\min_{\mathbf{W}} \frac{1}{2} \sum_n \|\mathbf{W}_g \mathbf{W}_f \mathbf{x}_n - \mathbf{x}_n\|_2^2$
- Algorithm: Backpropagation
  - Gradient descent
- Hidden nodes: compressed representation

Nonlinear Autoencoder

- *f* and *g* are nonlinear functions
  - $\arg\min_{\mathbf{W}} \frac{1}{2} \sum_{n} \| g(f(\mathbf{x}_n; \mathbf{W}_f); \mathbf{W}_g) - \mathbf{x}_n \|_2^2$

- Hidden nodes: nonlinear manifold



Deep Autoencoder

Deep Autoencoder

- f and g often consist of multiple layers
- In theory, one hidden layer in f and g is sufficient to represent any possible compression
- Multiple hidden layers in f and g is often better

Sparse Representations

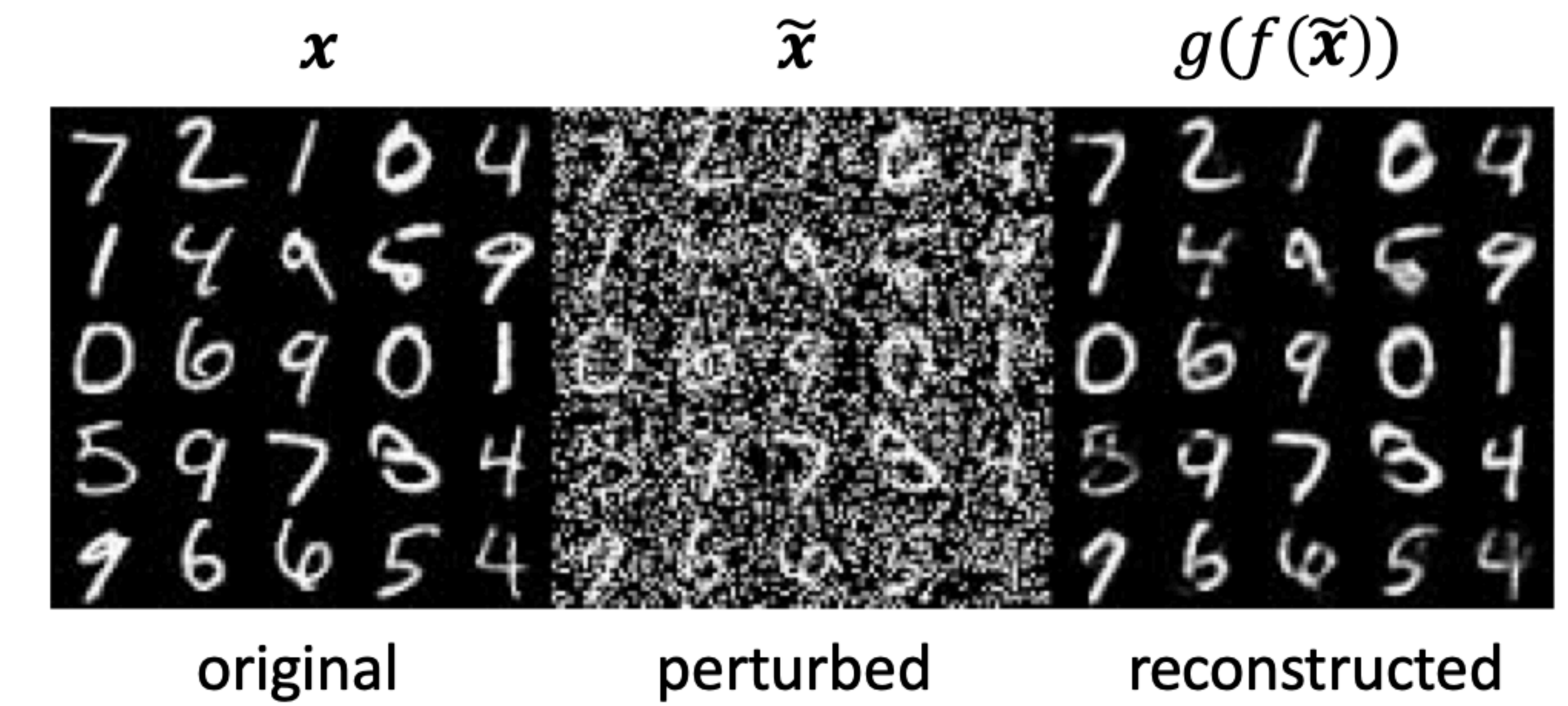
- When more hidden nodes than inputs, use regularization to constrain autoencoder

- Example: force hidden nodes to be sparse
  - $$\arg\min_{\mathbf{W}} \frac{1}{2} \sum_n \|g(f(\mathbf{x}_n; \mathbf{W}_f); \mathbf{W}_g) - \mathbf{x}_n\|_2^2 + c\,nnz(f(\mathbf{x}_n; \mathbf{W}_f))$$
  - Where $cnnz(f(\mathbf{x}_n; \mathbf{W}_f))$ is the number of non-zero entries produced by f

- Approximate objective: L1 regularization
  - $$\arg\min_{\mathbf{W}} \frac{1}{2} \sum_n \|g(f(\mathbf{x}_n; \mathbf{W}_f); \mathbf{W}_g) - \mathbf{x}_n\|_2^2 + c\|f(\mathbf{x}_n; \mathbf{W}_f)\|_1$$

Denoising Autoencoder

- Consider noisy version $\tilde{\mathbf{x}}$ of the input $\mathbf{x}$
- Data denoising:
  - $$\arg\min_{\mathbf{W}} \frac{1}{2} \sum_n \|g(f(\tilde{\mathbf{x}}_n; \mathbf{W}_f); \mathbf{W}_g) - \mathbf{x}_n\|_2^2 + c\|f(\tilde{\mathbf{x}}_n; \mathbf{W}_f)\|_1$$



Probabilistic Autoencoder

- Let *f* and *g* represent conditional distributions
  - $f : Pr(\mathbf{h} \,|\, \mathbf{x}; \mathbf{W}_f)$ and $g : Pr(\tilde{\mathbf{x}} \,|\, \mathbf{h}; \mathbf{W_g})$
  - by using sigmoid, softmax or linear units at the hidden and output layers
- Schematic

Probabilistic Autoencoder

Generative Model

- Sample $\mathbf{h}$ from some distribution $\Pr(\mathbf{h})$
- sample $\mathbf{x}$ from the decoder: $Pr(\mathbf{h} \mid \mathbf{x}; \mathbf{W}_g)$