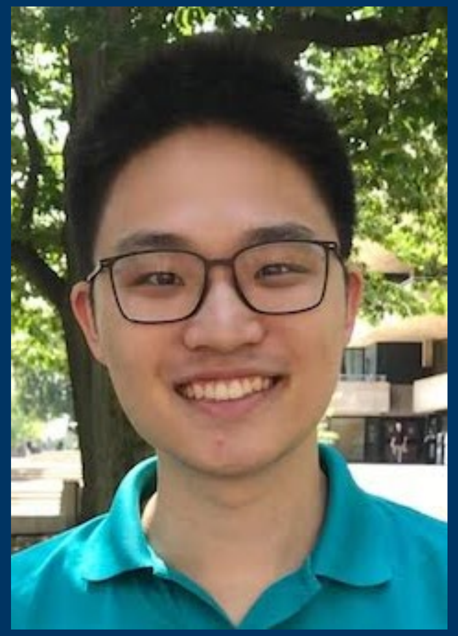
# Mixture Model Auto-Encoders: Deep Clustering through Dictionary Learning

## IEEE ICASSP 2022 Paper #4887



**Alexander Lin**

Harvard University
School of Engineering

**Andrew H. Song**

Harvard Medical School
Brigham and Women’s Hospital

**Demba Ba**

Harvard University
School of Engineering

# Introduction

## The Difficulty of Clustering Natural Signals

# Introduction

## The Difficulty of Clustering Natural Signals




- Unsupervised: No labels

# Introduction

## The Difficulty of Clustering Natural Signals

- Unsupervised: No labels
  - Cannot rely on labeled patterns in training set (like in classification)

# Introduction

## The Difficulty of Clustering Natural Signals

- Unsupervised: No labels
    - Cannot rely on labeled patterns in training set (like in classification)

# Introduction

## The Difficulty of Clustering Natural Signals

- Unsupervised: No labels
  - Cannot rely on labeled patterns in training set (like in classification)
- High-dimensional: Simple metrics (e.g. Euclidean distance) are no longer informative

# Introduction

## The Difficulty of Clustering Natural Signals

- Unsupervised: No labels
  - Cannot rely on labeled patterns in training set (like in classification)
- High-dimensional: Simple metrics (e.g. Euclidean distance) are no longer informative
  - Classical clustering algorithms (e.g. $K$-means) do not work well

# Introduction

## The Difficulty of Clustering Natural Signals


- Unsupervised: No labels
  - Cannot rely on labeled patterns in training set (like in classification)
- High-dimensional: Simple metrics (e.g. Euclidean distance) are no longer informative
  - Classical clustering algorithms (e.g. $K$-means) do not work well

# Modeling High-Dimensional Signals

# Modeling High-Dimensional Signals

**Model-Based**
**Signal Processing**

# Modeling High-Dimensional Signals

**Model-Based Signal Processing**

- Simple models with principled effects (e.g. sparsity, low rank, dictionary learning)
- Incorporate prior knowledge and learns few parameters
- Theoretically understood & interpretable design

# Modeling High-Dimensional Signals

**Model-Based Signal Processing**

**Deep Learning**

- Simple models with principled effects (e.g. sparsity, low rank, dictionary learning)
- Incorporate prior knowledge and learns few parameters
- Theoretically understood & interpretable design

# Modeling High-Dimensional Signals

## Model-Based Signal Processing

- Simple models with principled effects (e.g. sparsity, low rank, dictionary learning)
- Incorporate prior knowledge and learns few parameters
- Theoretically understood & interpretable design

## Deep Learning

- Black-box architectures (e.g. CNN, RNN, Transformer)
- Heavily over-parameterized
- Scalable training on large datasets (e.g. batch processing, GPUs)

# Modeling High-Dimensional Signals

## Model-Based Signal Processing

- Simple models with principled effects (e.g. sparsity, low rank, dictionary learning)
- Incorporate prior knowledge and learns few parameters
- Theoretically understood & interpretable design

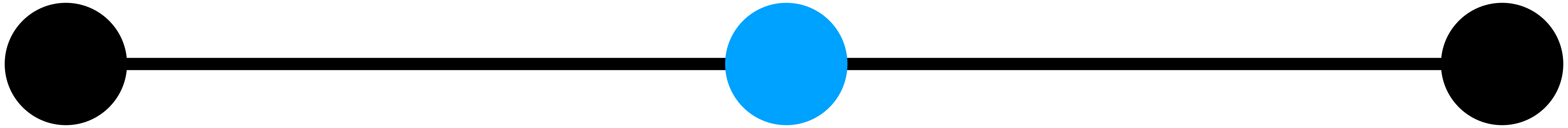
## Model-Based Deep Learning [1]

## Deep Learning

- Black-box architectures (e.g. CNN, RNN, Transformer)
- Heavily over-parameterized
- Scalable training on large datasets (e.g. batch processing, GPUs)

[1] Shlezinger, N., Whang, J., Eldar, Y. C., & Dimakis, A. G. (2020). Model-based deep learning. *arXiv preprint arXiv:2012.08405*.

# Modeling High-Dimensional Signals



### Model-Based Signal Processing

- Simple models with principled effects (e.g. sparsity, low rank, dictionary learning)
- Incorporate prior knowledge and learns few parameters
- Theoretically understood & interpretable design

### Model-Based Deep Learning [1]

- Multi-layer architecture derived from interpretable signal processing model
- Can use model to inject prior knowledge into deep architecture (fewer params)
- Can leverage deep learning technology for scalable training

### Deep Learning

- Black-box architectures (e.g. CNN, RNN, Transformer)
- Heavily over-parameterized
- Scalable training on large datasets (e.g. batch processing, GPUs)

[1] Shlezinger, N., Whang, J., Eldar, Y. C., & Dimakis, A. G. (2020). Model-based deep learning. *arXiv preprint arXiv:2012.08405*.

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images
  - Derived from statistical signal processing models (i.e. *dictionary learning* and *mixture modeling)*

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images
  - Derived from statistical signal processing models (i.e. *dictionary learning* and *mixture modeling)*
  - Trained as a neural network on large datasets

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images
  - Derived from statistical signal processing models (i.e. *dictionary learning* and *mixture modeling)*
  - Trained as a neural network on large datasets
- Achieve superior performance over other state-of-the-art deep learning architectures for the clustering problem, while also providing…

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images
  - Derived from statistical signal processing models (i.e. *dictionary learning* and *mixture modeling)*
  - Trained as a neural network on large datasets
- Achieve superior performance over other state-of-the-art deep learning architectures for the clustering problem, while also providing…
  - Order-of-magnitude fewer parameters (i.e. 50x smaller)

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images
  - Derived from statistical signal processing models (i.e. *dictionary learning* and *mixture modeling)*
  - Trained as a neural network on large datasets
- Achieve superior performance over other state-of-the-art deep learning architectures for the clustering problem, while also providing…
  - Order-of-magnitude fewer parameters (i.e. 50x smaller)
  - Simpler parameter initialization scheme

# Our Work: Model-Based Deep Clustering

## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images
  - Derived from statistical signal processing models (i.e. *dictionary learning* and *mixture modeling)*
  - Trained as a neural network on large datasets
- Achieve superior performance over other state-of-the-art deep learning architectures for the clustering problem, while also providing…
  - Order-of-magnitude fewer parameters (i.e. 50x smaller)
  - Simpler parameter initialization scheme
  - Ability to cluster incomplete/missing data

# Our Work: Model-Based Deep Clustering

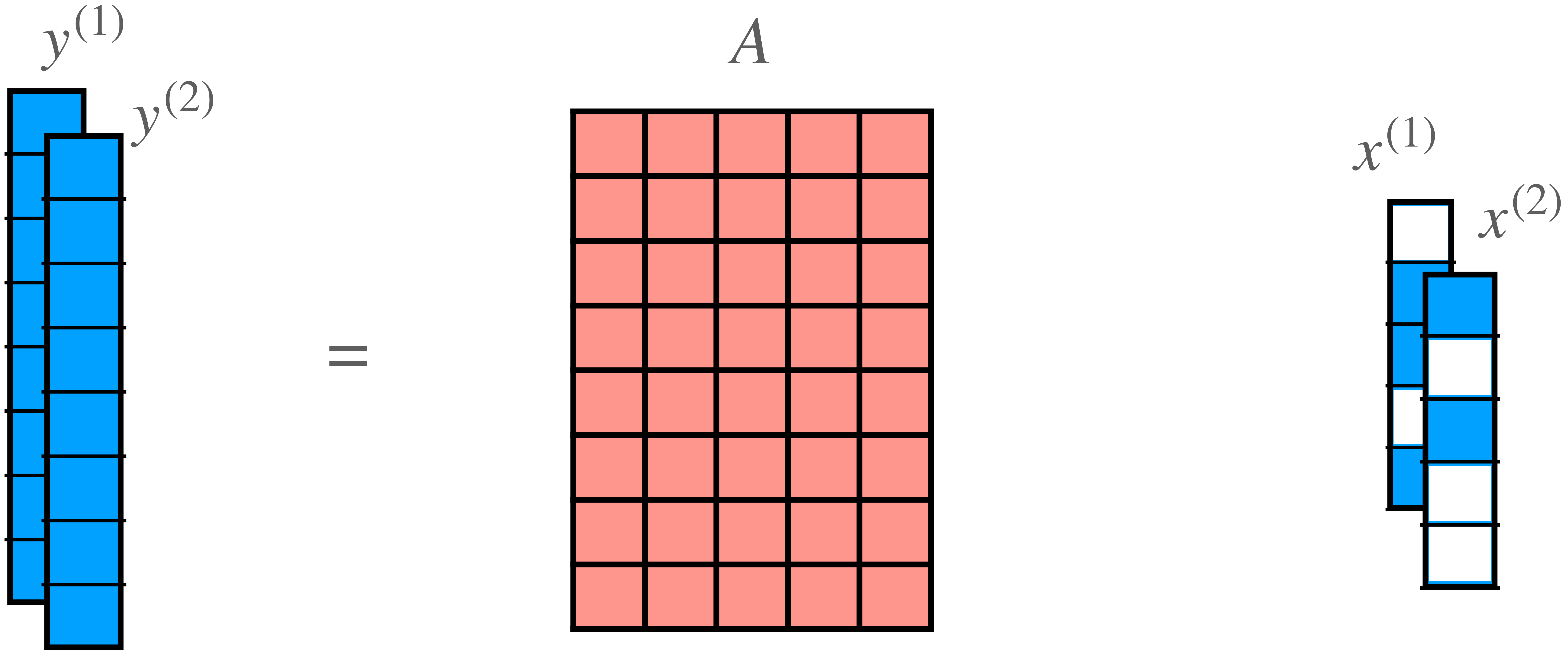
## Applying Model-Based Deep Learning to High-Dim Clustering

- Propose **Mixture Model Auto-Encoders (MixMate)** —a novel deep architecture for clustering signals/images
  - Derived from statistical signal processing models (i.e. *dictionary learning* and *mixture modeling)*
  - Trained as a neural network on large datasets
- Achieve superior performance over other state-of-the-art deep learning architectures for the clustering problem, while also providing…
  - Order-of-magnitude fewer parameters (i.e. 50x smaller)
  - Simpler parameter initialization scheme
  - Ability to cluster incomplete/missing data

Consequences of signal processing model!

# MixMate Model

## Intro to Dictionary Learning



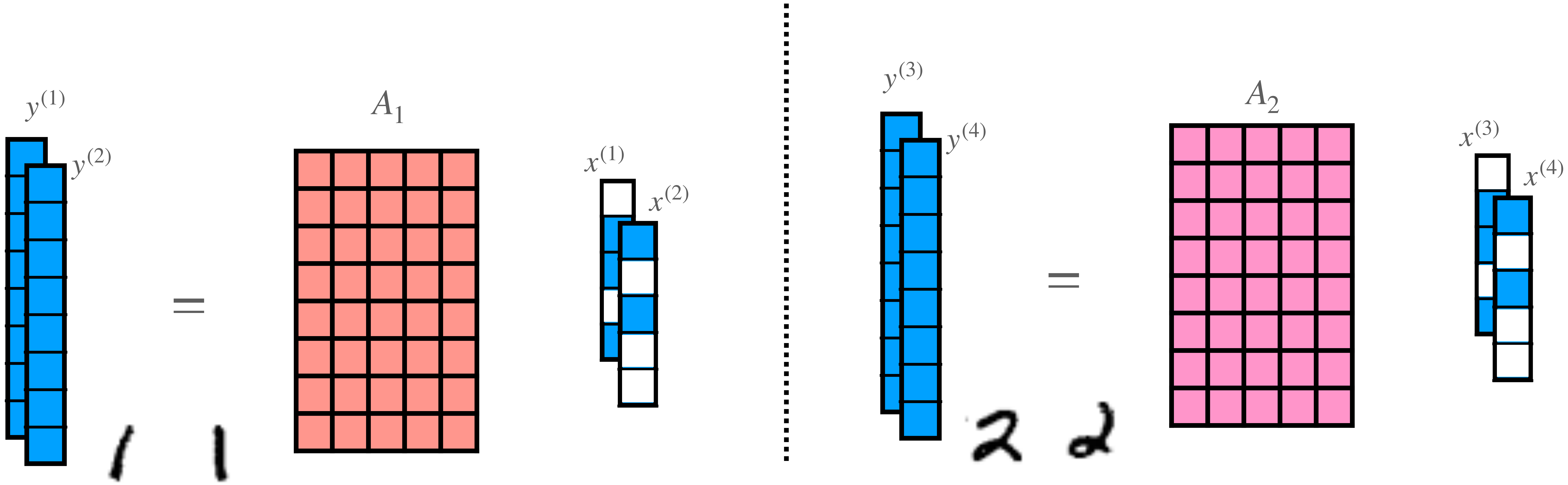

Dataset of Natural Signals
(e.g. images of handwritten digits)

Dictionary
(e.g. collection of strokes)

Sparse Codes
(e.g. how much of each stroke)

# MixMate Model

## Mixture of Dictionary Learning Models




Cluster #1
(e.g. images of
handwritten 1's)

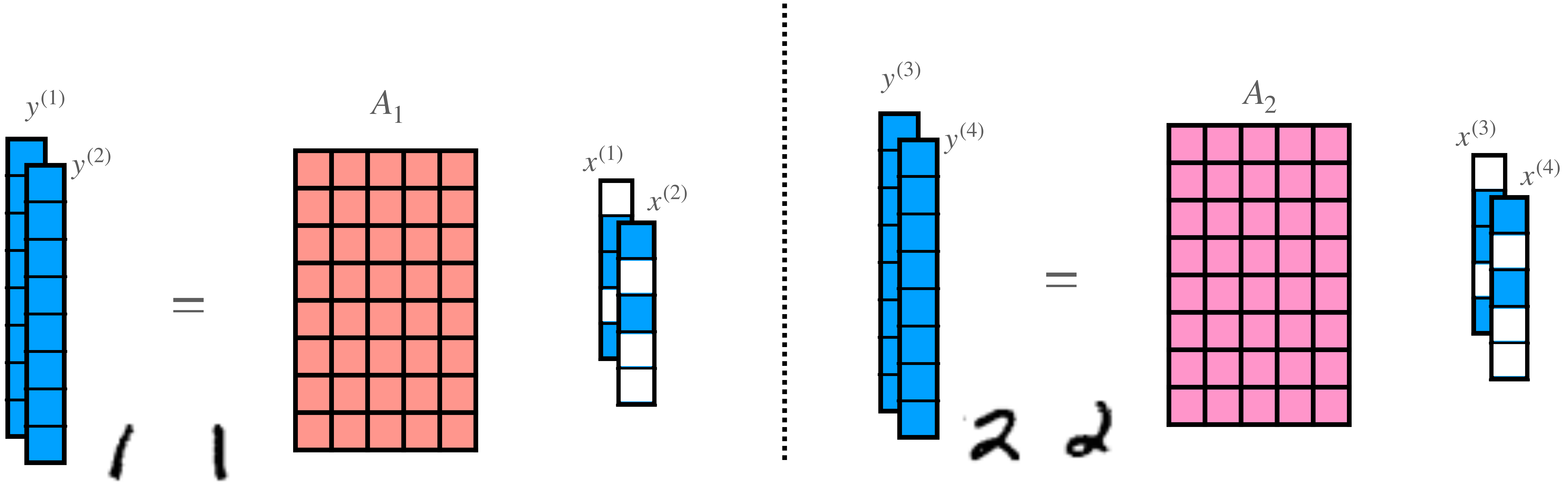
Dictionary #1
(e.g. collection of
strokes for 1's)

Cluster #2
(e.g. images of
handwritten 2's)

Dictionary #2
(e.g. collection of
strokes for 2's)

# MixMate Model

## Mixture of Dictionary Learning Models




Cluster #1 (e.g. images of handwritten 1's)

Dictionary #1 (e.g. collection of strokes for 1's)

Cluster #2 (e.g. images of handwritten 2's)

Dictionary #2 (e.g. collection of strokes for 2's)

<u>Key Insight</u>: Each cluster of images is generated by a *different* dictionary.

# MixMate Model

## Mixture of Dictionary Learning Models

Assume $K$ total clusters.

(Latent) cluster identity: $z \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K)$ $\qquad \sum_{k=1}^{K} \pi_k = 1$

(Latent) sparse code: $x \sim \text{Laplace}(\lambda) \propto \exp(-\lambda ||x||_1)$

(Observed) data: $y|x, z = k \sim \mathcal{N}(A_k x, I) \propto \exp(-||y - A_k x||_2^2)$

Goal: Learn parameters/dictionaries: $A_1, A_2, \ldots, A_K$

Infer latent variables (for each $y$): $z, x$

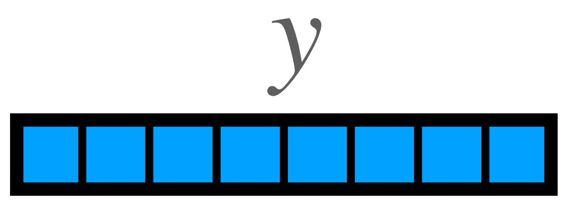
Key Insight: Each cluster of images is generated by a *different* dictionary.

# MixMate Architecture

## Diagram of Inference

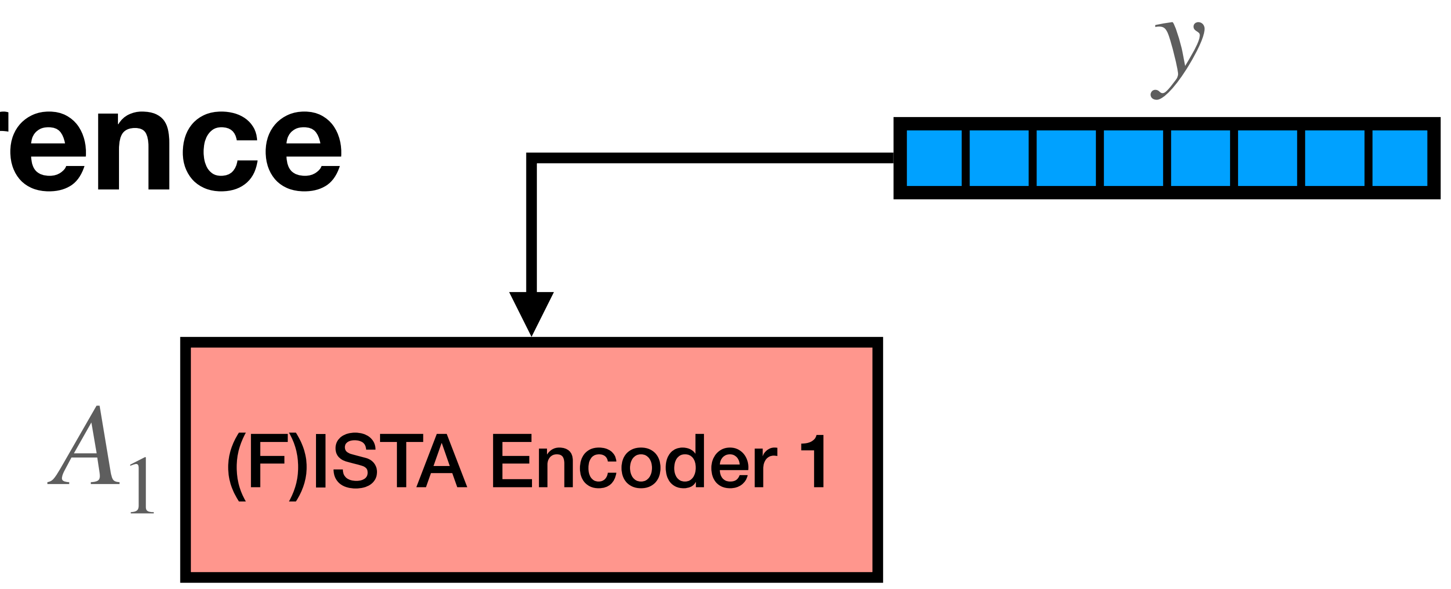
# MixMate Architecture

## Diagram of Inference



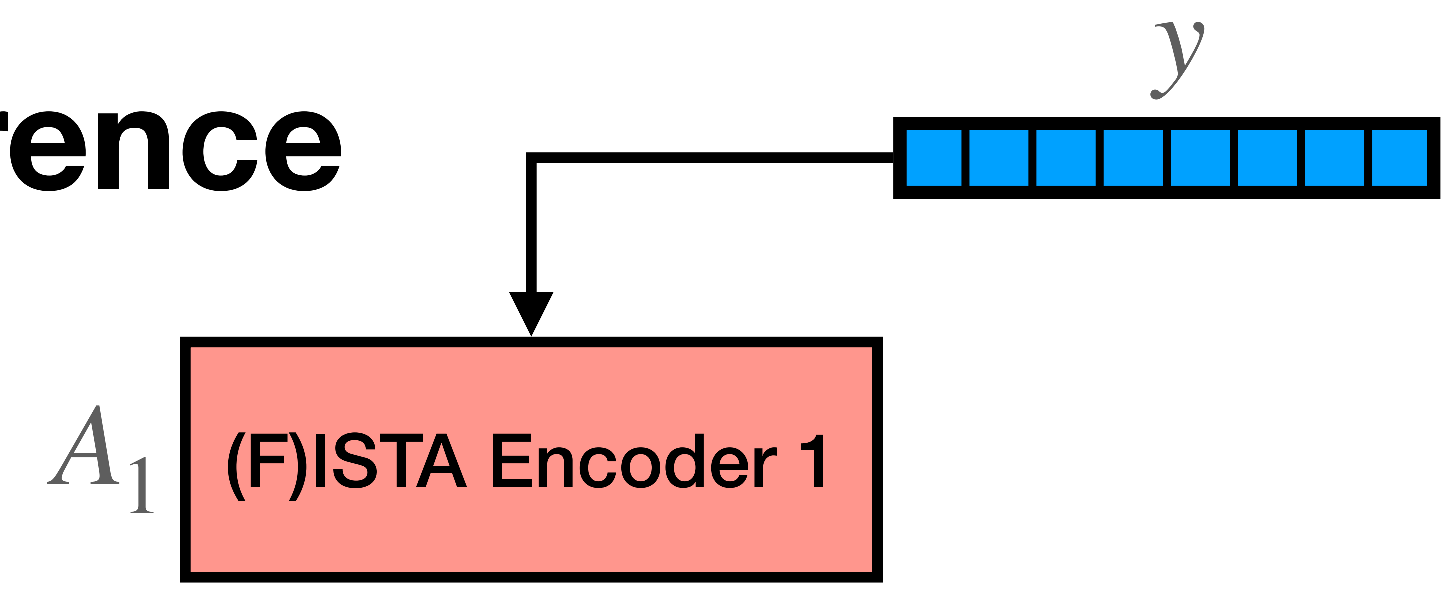
# MixMate Architecture

## Diagram of Inference



# MixMate Architecture

## Diagram of Inference

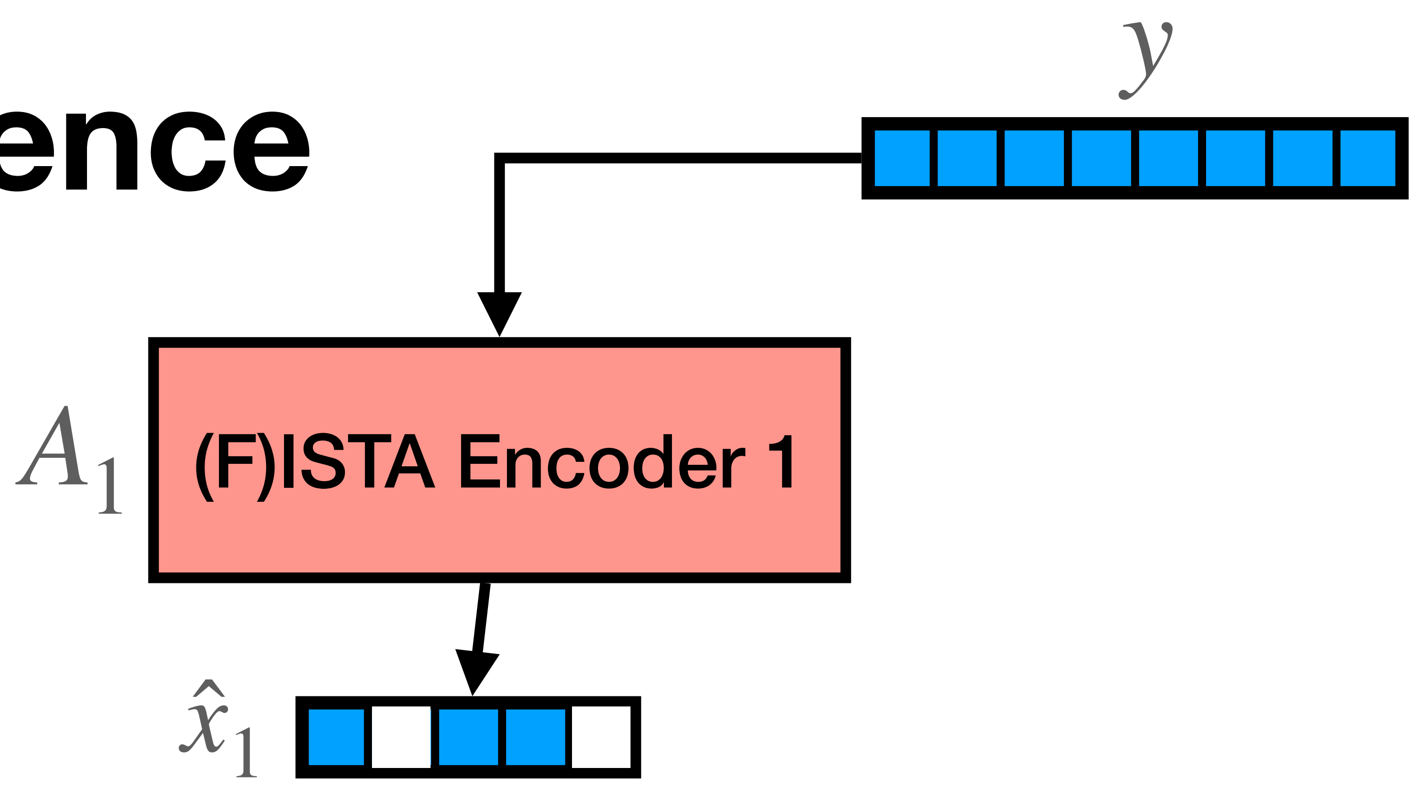



Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

# MixMate Architecture

## Diagram of Inference




Sparse Coding Problem with $A_1$

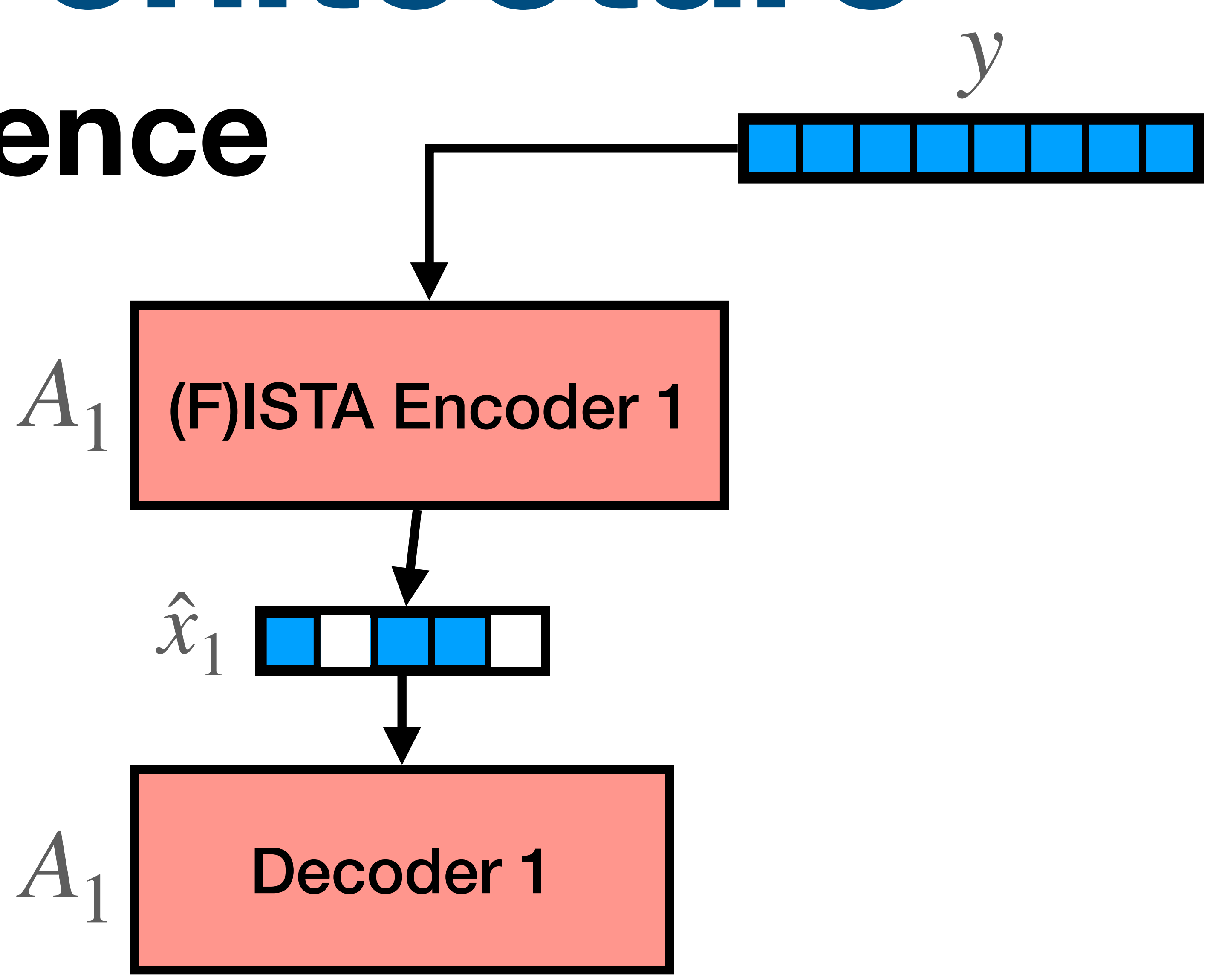
$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

# MixMate Architecture

## Diagram of Inference

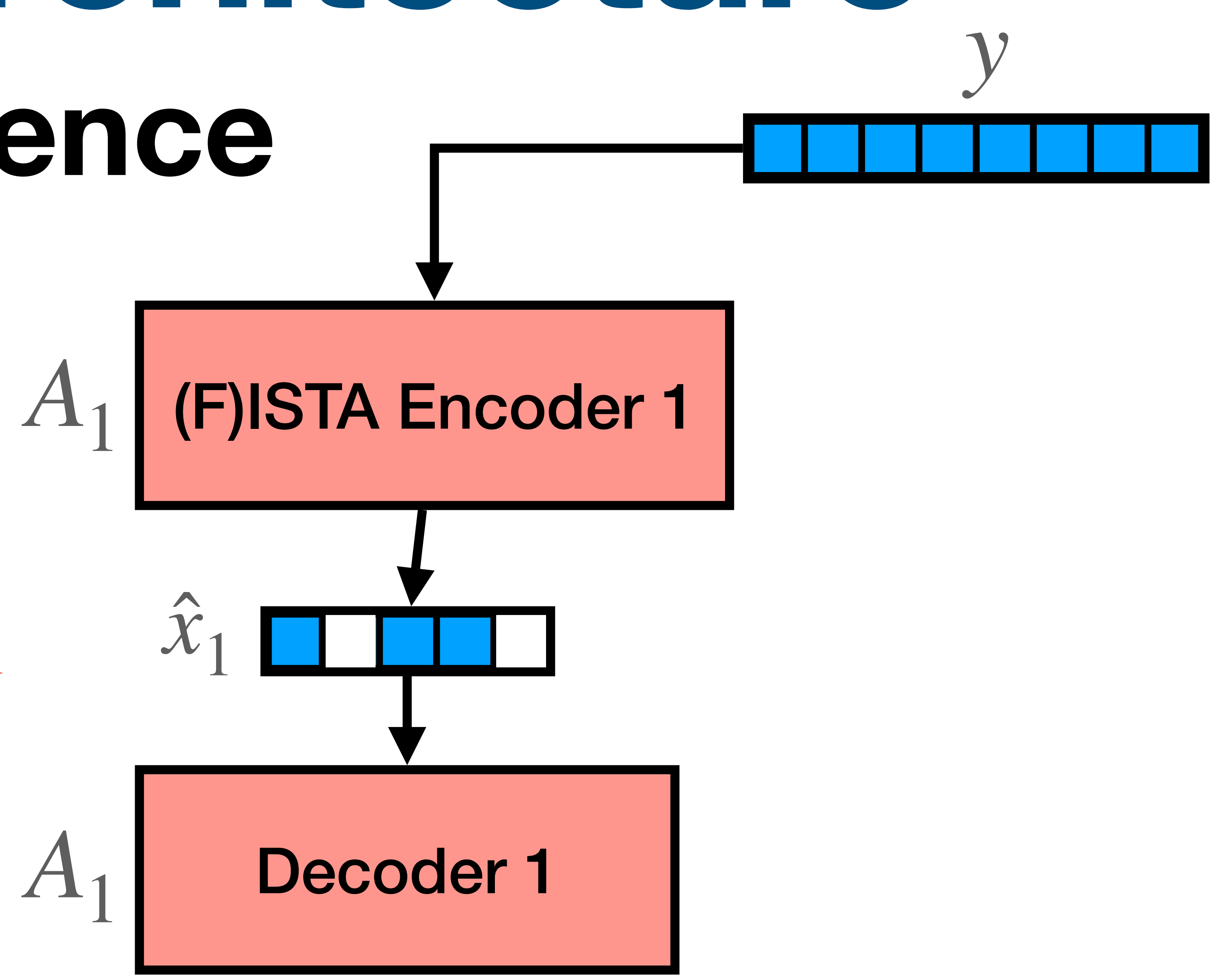
Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$



# MixMate Architecture

## Diagram of Inference




Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

# MixMate Architecture

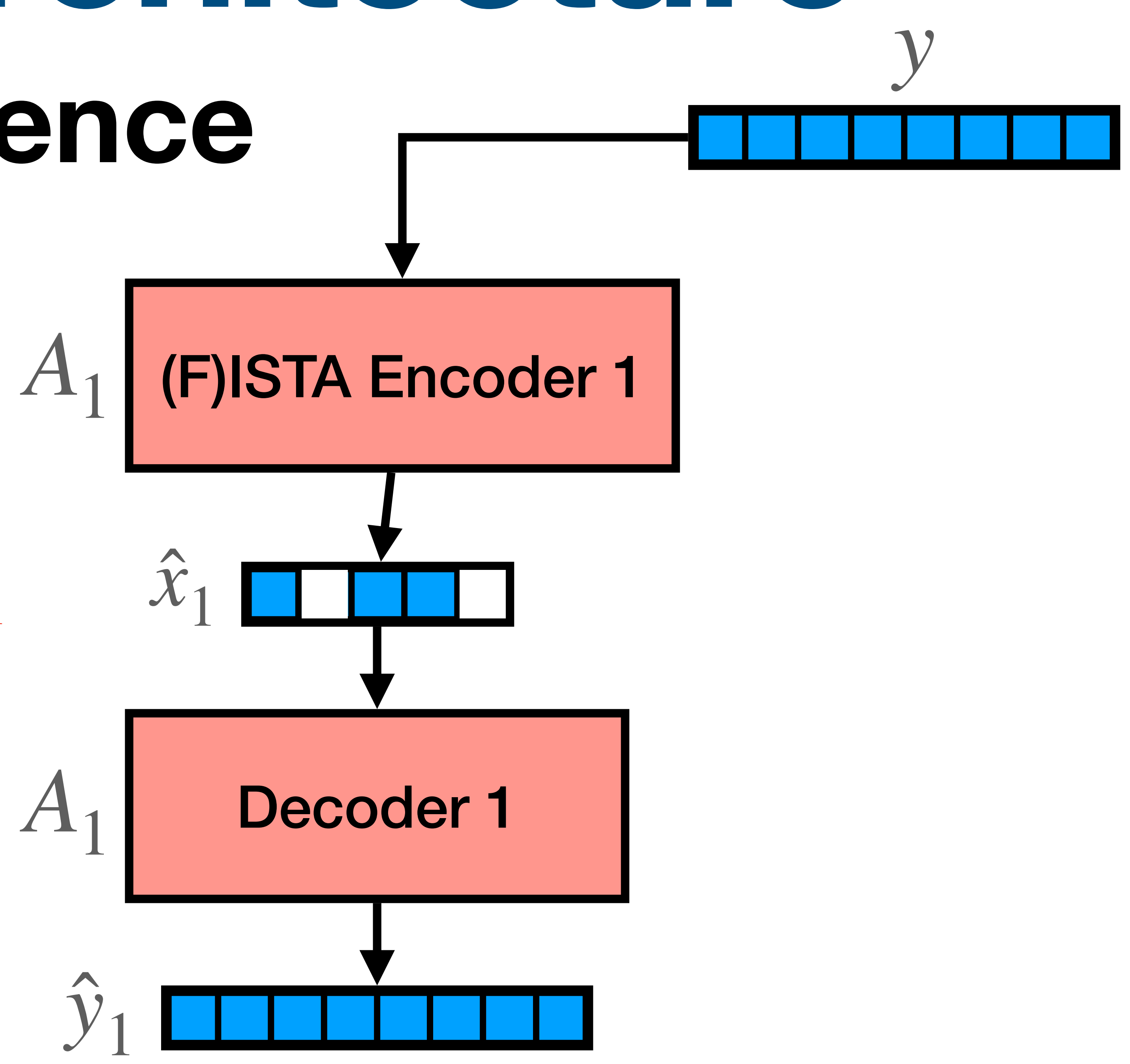
## Diagram of Inference

Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

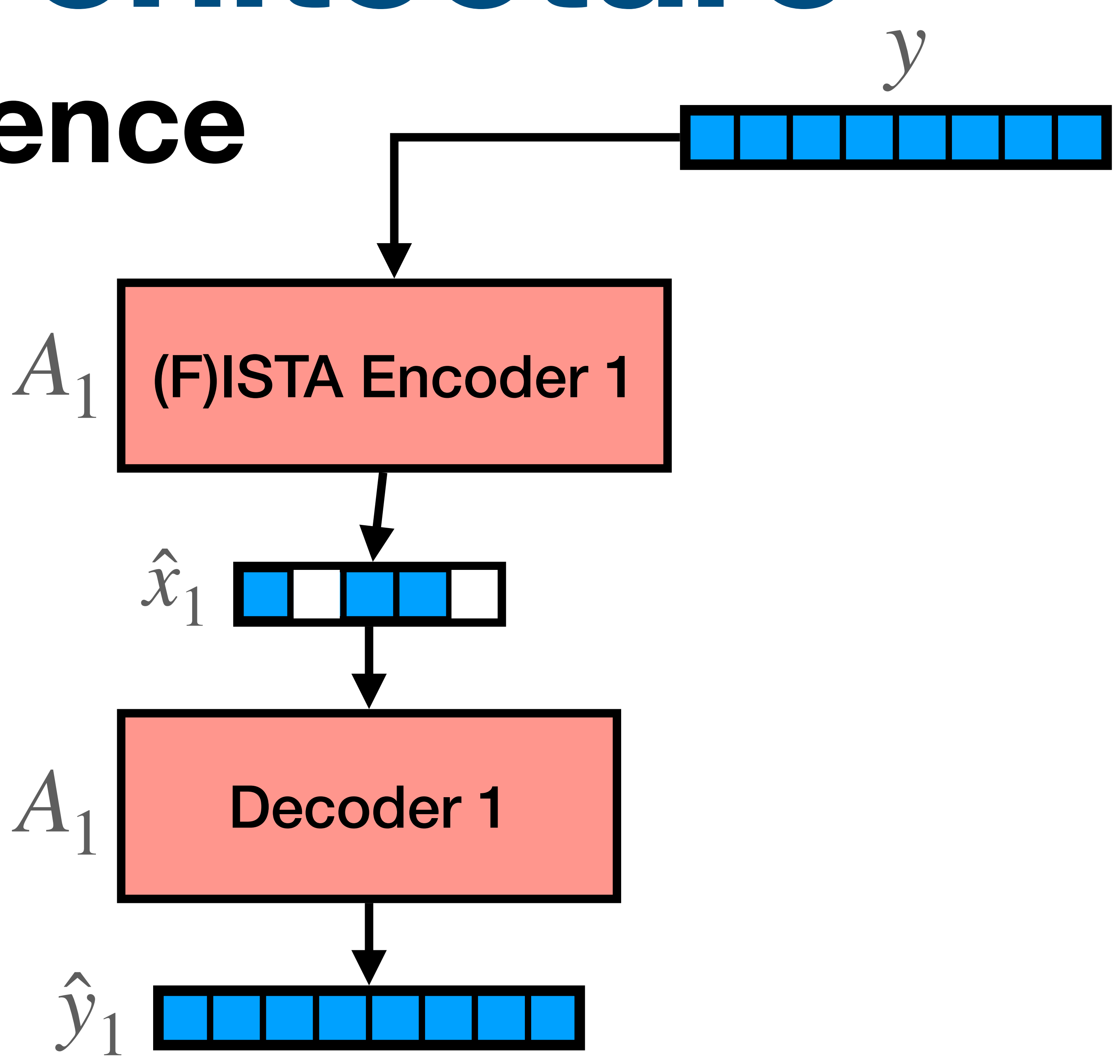
Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$



# MixMate Architecture

## Diagram of Inference




Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$

# MixMate Architecture

## Diagram of Inference

Sparse Coding Problem with $A_1$

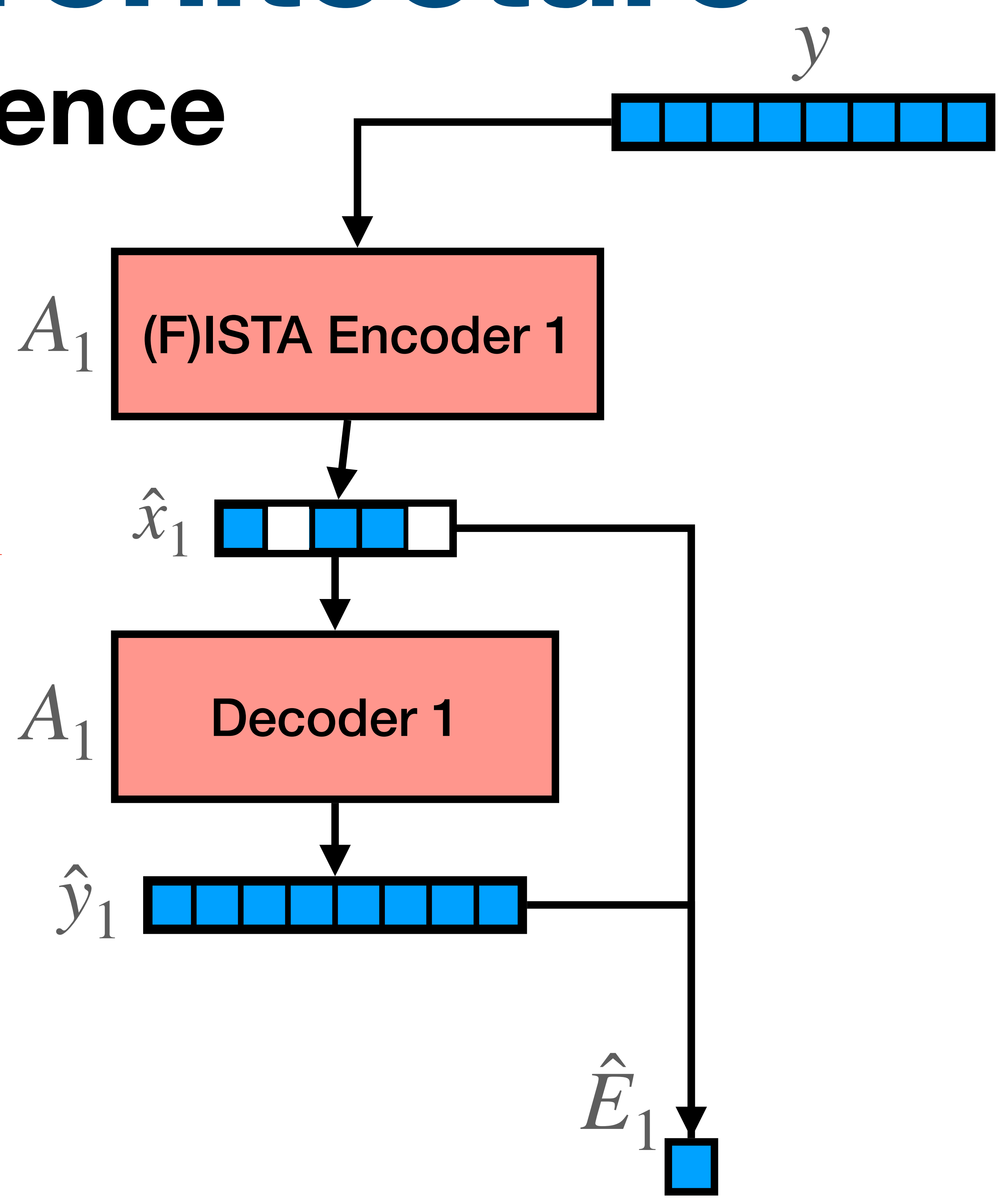
$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$



# MixMate Architecture

## Diagram of Inference

Sparse Coding Problem with $A_1$

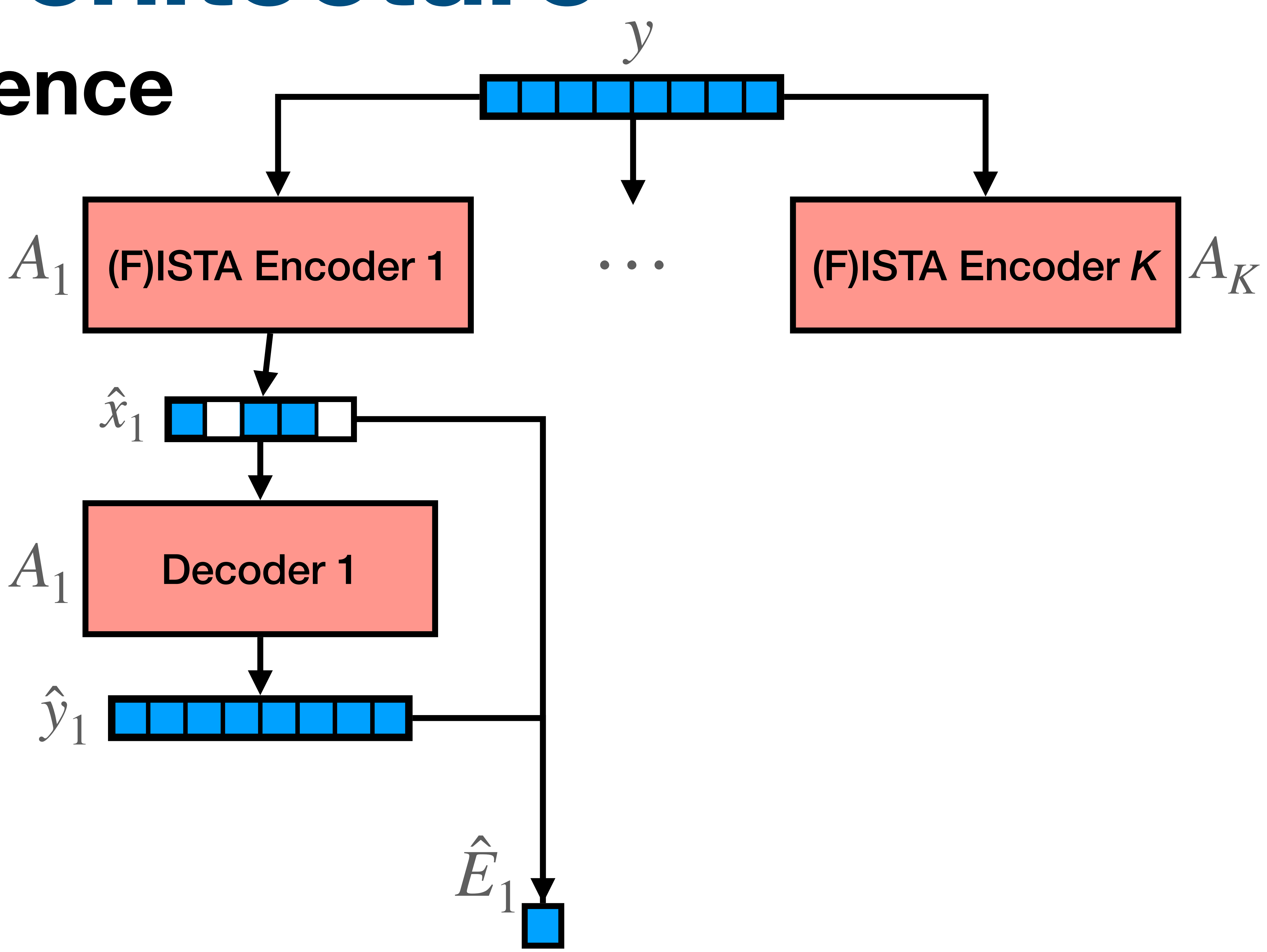
$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$



# MixMate Architecture

## Diagram of Inference

Sparse Coding Problem with $A_1$

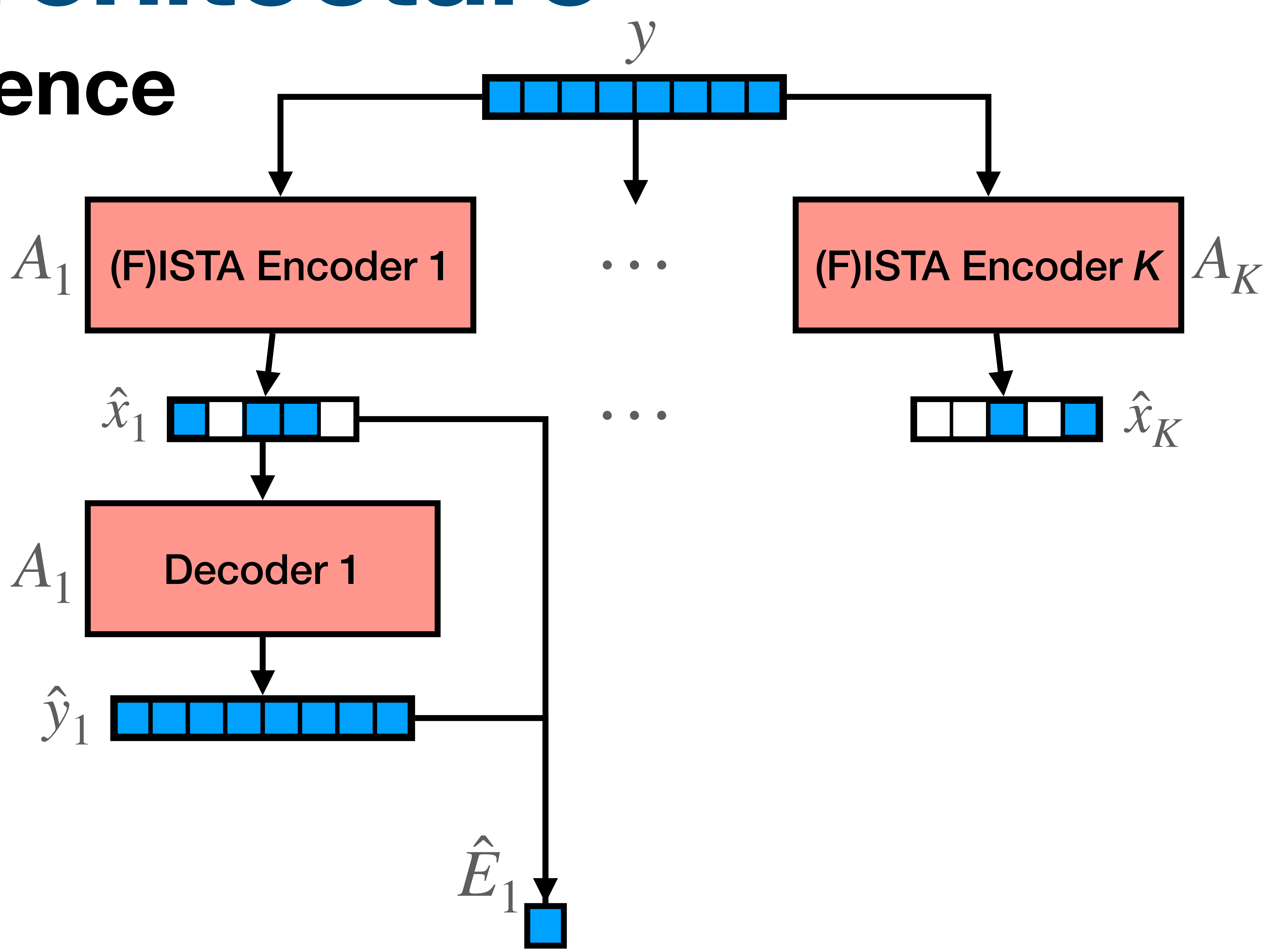
$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$




Sparse Coding Problem with $A_K$

$$\hat{x}_K = \arg\min_x ||y - A_K x||_2^2 + \lambda ||x||_1$$

# MixMate Architecture

## Diagram of Inference

Sparse Coding Problem with $A_1$

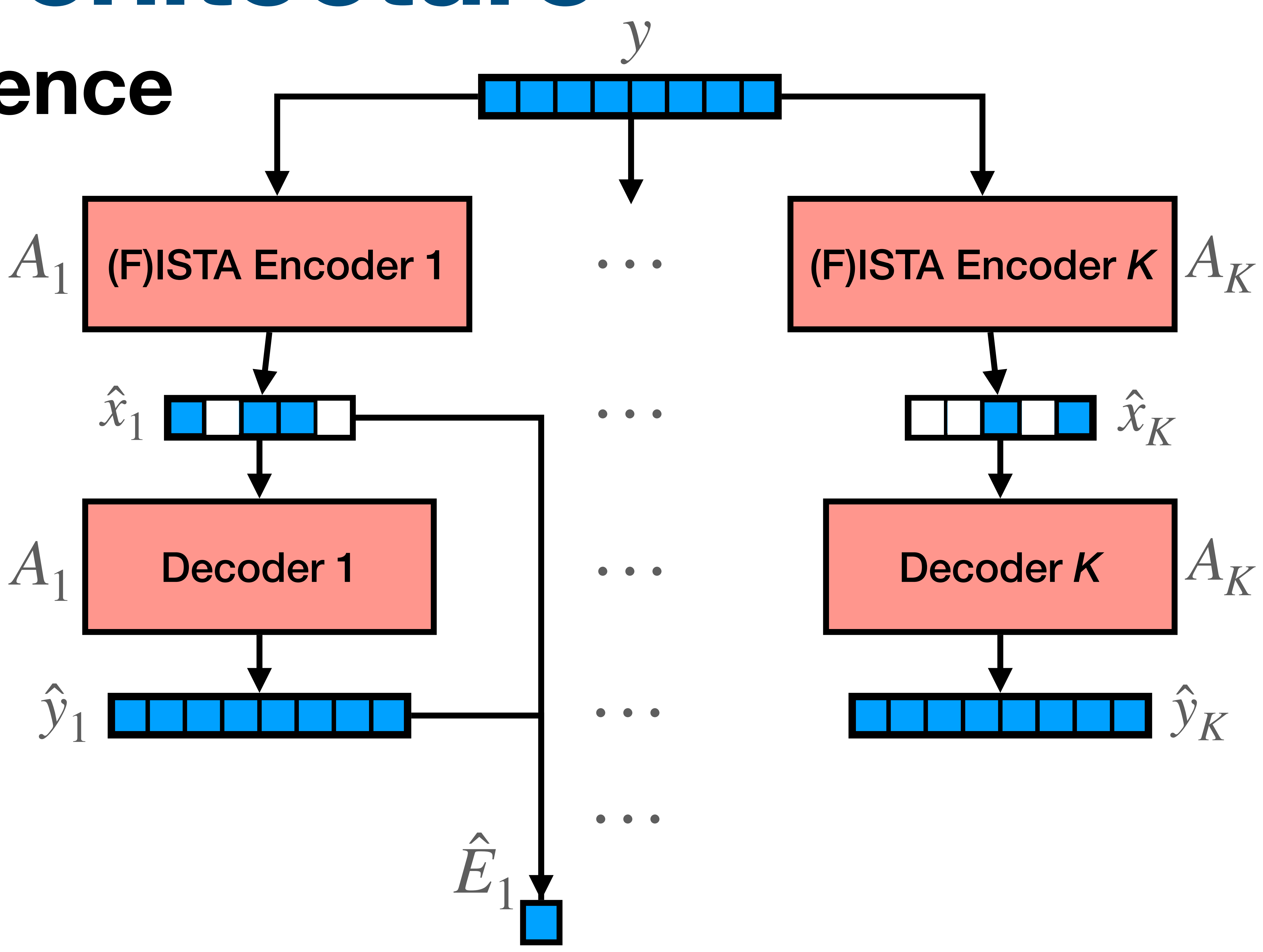
$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$




Sparse Coding Problem with $A_K$

$$\hat{x}_K = \arg\min_x ||y - A_K x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_K$

$$\hat{y}_K = A\hat{x}_K$$

# MixMate Architecture

## Diagram of Inference

Sparse Coding Problem with $A_1$

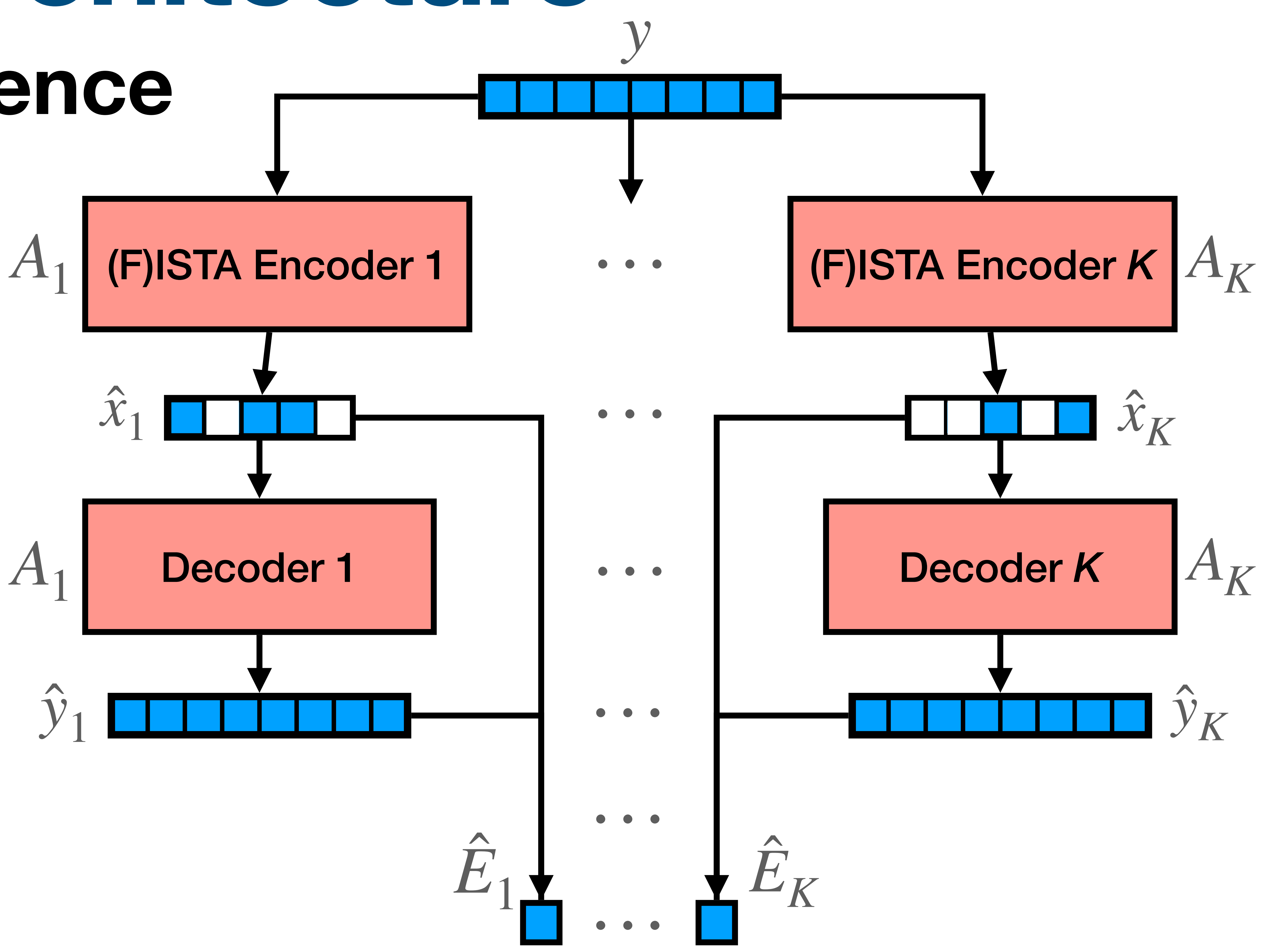
$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$




Sparse Coding Problem with $A_K$

$$\hat{x}_K = \arg\min_x ||y - A_K x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_K$

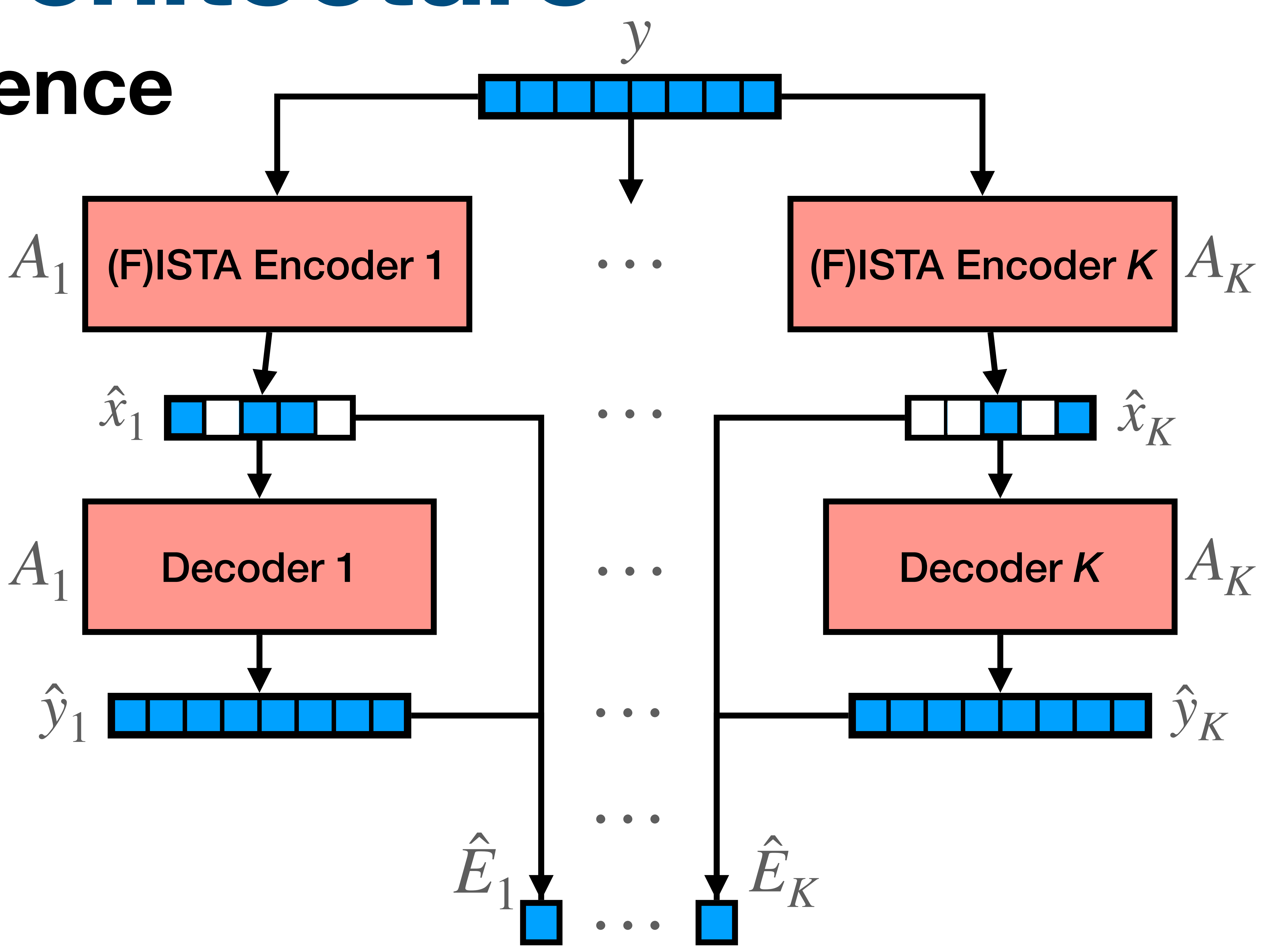
$$\hat{y}_K = A\hat{x}_K$$

Compute energy for $A_K$

$$\hat{E}_K = ||y - \hat{y}_K||_2^2 + \lambda ||\hat{x}_K||_1$$

# MixMate Architecture

## Diagram of Inference




Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$

Compute probability of $z = 1$

$$p(z = 1 \,|\, y) \propto \exp(-E_1)$$

Sparse Coding Problem with $A_K$

$$\hat{x}_K = \arg\min_x ||y - A_K x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_K$

$$\hat{y}_K = A\hat{x}_K$$

Compute energy for $A_K$

$$\hat{E}_K = ||y - \hat{y}_K||_2^2 + \lambda ||\hat{x}_K||_1$$

Compute probability of $z = K$

$$p(z = K \,|\, y) \propto \exp(-E_K)$$

# MixMate Architecture

## Diagram of Inference

Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

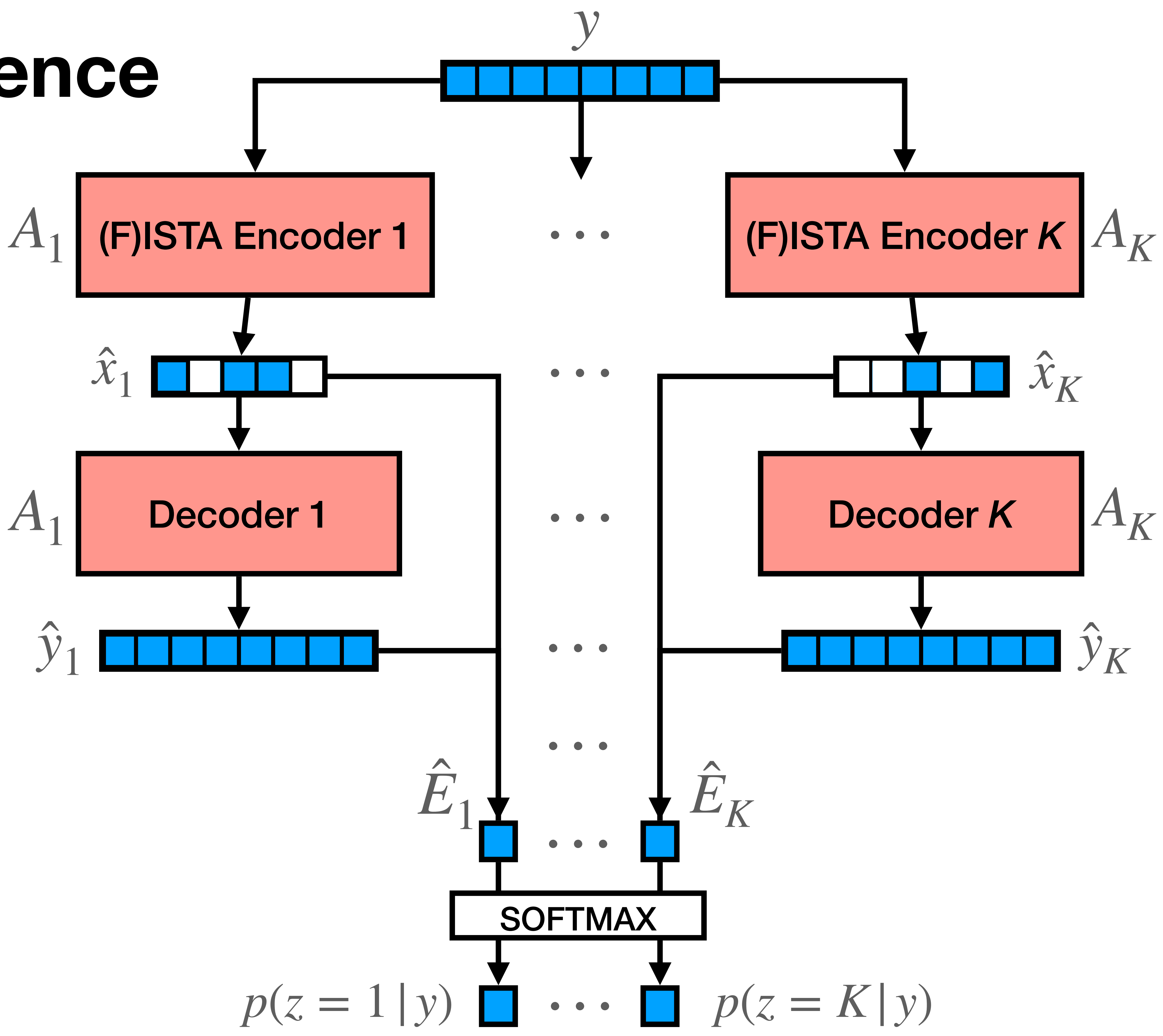
$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$

Compute probability of $z = 1$

$$p(z = 1 | y) \propto \exp(-E_1)$$




Sparse Coding Problem with $A_K$

$$\hat{x}_K = \arg\min_x ||y - A_K x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_K$

$$\hat{y}_K = A\hat{x}_K$$

Compute energy for $A_K$

$$\hat{E}_K = ||y - \hat{y}_K||_2^2 + \lambda ||\hat{x}_K||_1$$

Compute probability of $z = K$

$$p(z = K | y) \propto \exp(-E_K)$$

# MixMate Architecture

## Diagram of Inference

Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

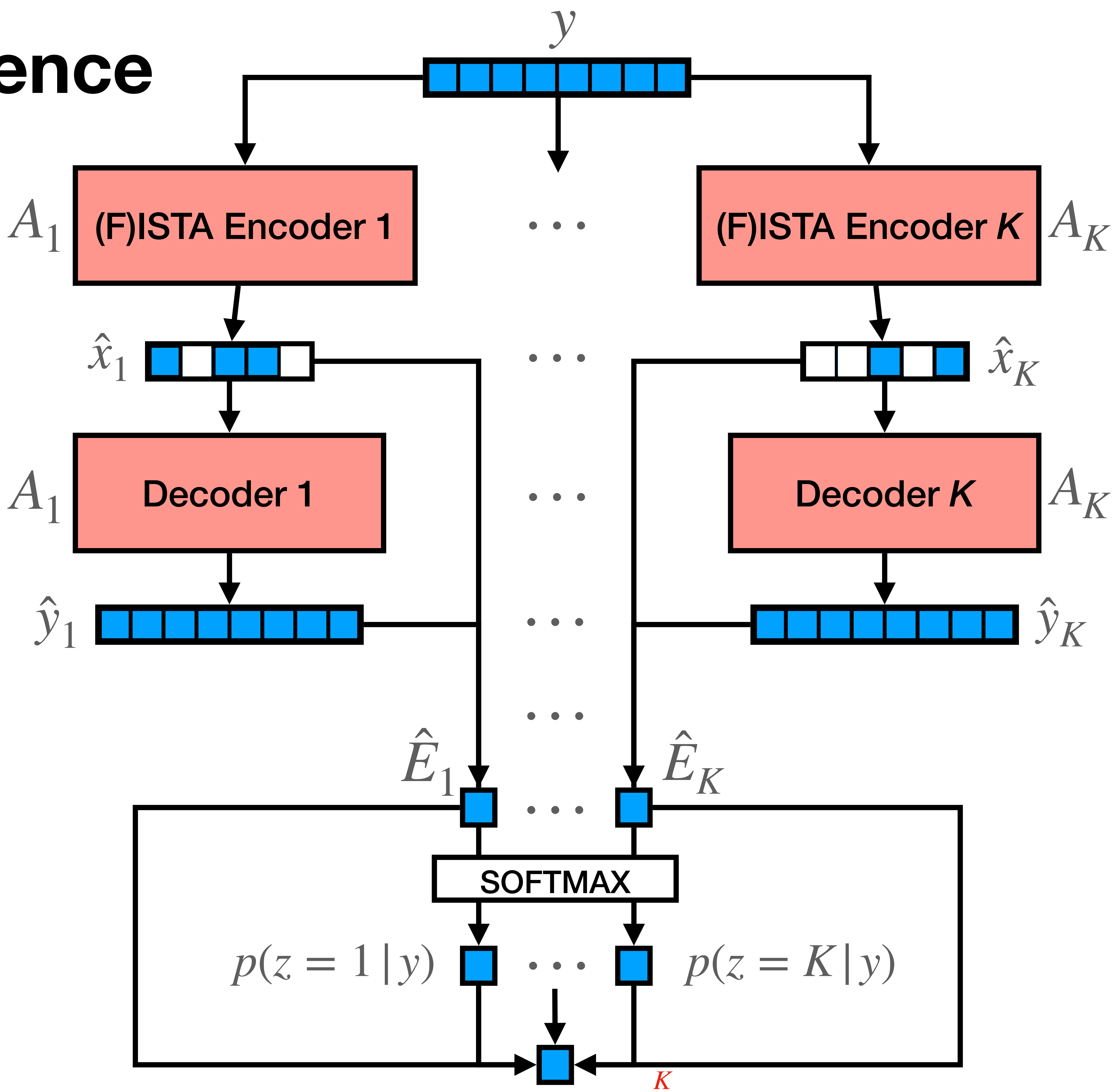
$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$

Compute probability of $z = 1$

$$p(z = 1 | y) \propto \exp(-E_1)$$




Attention Loss$(y) = \sum_{k=1}^{K} p(z = k | y) \cdot \hat{E}_k$

Sparse Coding Problem with $A_K$

$$\hat{x}_K = \arg\min_x ||y - A_K x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_K$

$$\hat{y}_K = A\hat{x}_K$$

Compute energy for $A_K$

$$\hat{E}_K = ||y - \hat{y}_K||_2^2 + \lambda ||\hat{x}_K||_1$$

Compute probability of $z = K$

$$p(z = K | y) \propto \exp(-E_K)$$

# MixMate Architecture

## Diagram of Inference

Back-propagate through entire architecture to train dictionaries

Sparse Coding Problem with $A_1$

$$\hat{x}_1 = \arg\min_x ||y - A_1 x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_1$

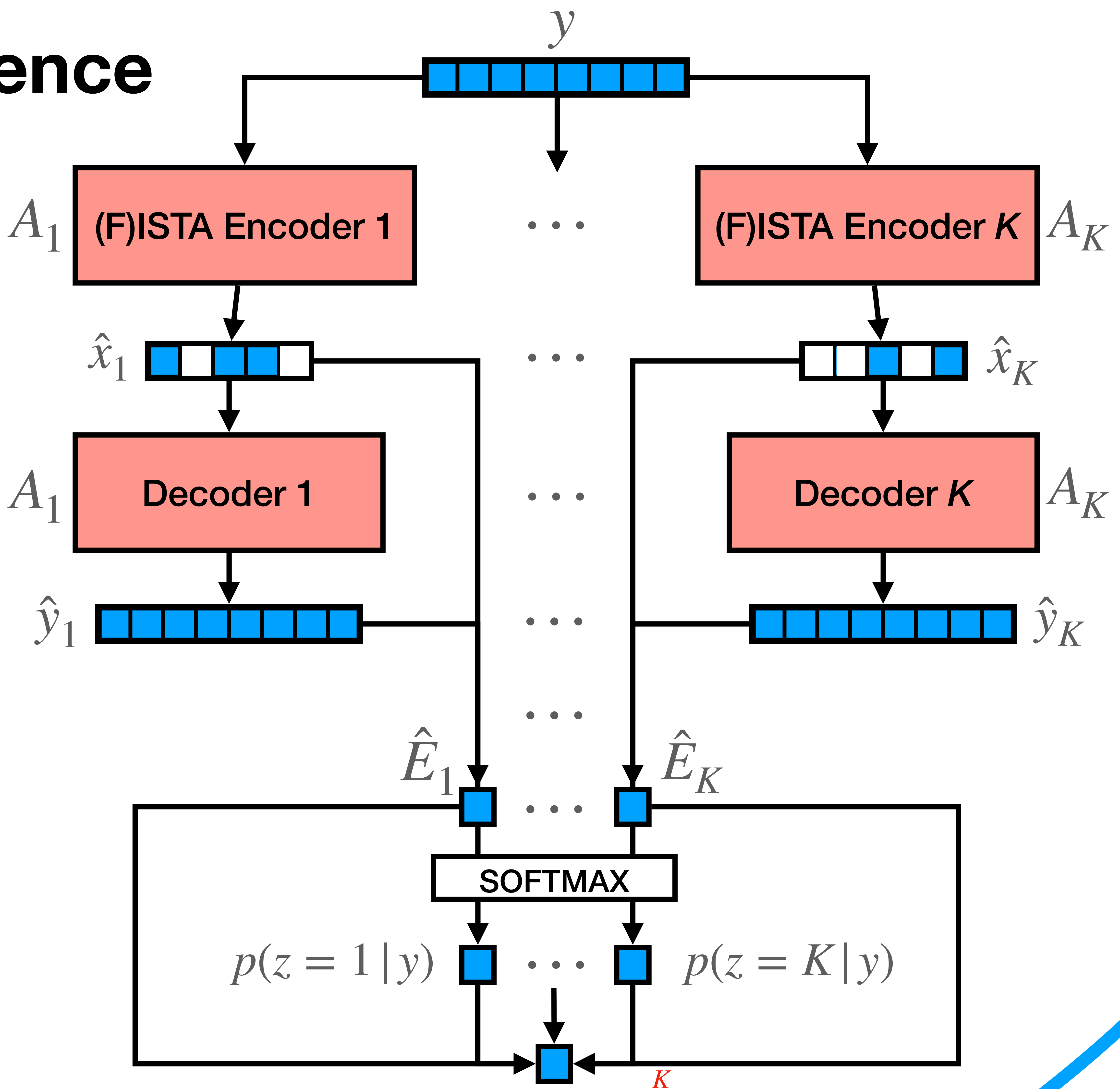
$$\hat{y}_1 = A\hat{x}_1$$

Compute energy for $A_1$

$$\hat{E}_1 = ||y - \hat{y}_1||_2^2 + \lambda ||\hat{x}_1||_1$$

Compute probability of $z = 1$

$$p(z = 1 | y) \propto \exp(-E_1)$$




Attention Loss$(y) = \sum_{k=1}^{K} p(z = k | y) \cdot \hat{E}_k$

Sparse Coding Problem with $A_K$

$$\hat{x}_K = \arg\min_x ||y - A_K x||_2^2 + \lambda ||x||_1$$

Reconstruct Data with $A_K$

$$\hat{y}_K = A\hat{x}_K$$

Compute energy for $A_K$

$$\hat{E}_K = ||y - \hat{y}_K||_2^2 + \lambda ||\hat{x}_K||_1$$

Compute probability of $z = K$

$$p(z = K | y) \propto \exp(-E_K)$$

# Results: Image Clustering Datasets

| | DEC* | DCN* | DAMIC* | $K$-DAE* | MixMate INIT | MixMate TRAIN |
|---|---|---|---|---|---|---|
| **MNIST** | | | | | | |
| NMI | 0.80 | 0.81 | **0.86** | **0.86** | 0.75 | **0.86** $\pm$ 0.03 |
| ARI | 0.75 | 0.75 | 0.82 | 0.82 | 0.72 | **0.85** $\pm$ 0.04 |
| ACC | 0.84 | 0.83 | 0.88 | 0.88 | 0.84 | **0.92** $\pm$ 0.04 |
| Params | 2.1 M | 2.1 M | 22.1 M | 21.4 M | 0.4 M | |
| **Fashion** | | | | | | |
| NMI | 0.54 | 0.55 | 0.65 | 0.65 | 0.60 | **0.68** $\pm$ 0.02 |
| ARI | 0.40 | 0.42 | 0.48 | 0.48 | 0.44 | **0.52** $\pm$ 0.01 |
| ACC | 0.51 | 0.50 | 0.60 | 0.60 | 0.57 | **0.63** $\pm$ 0.01 |
| Params | N/A | N/A | N/A | N/A | 0.4 M | |
| **USPS** | | | | | | |
| NMI | 0.77 | 0.68 | 0.78 | 0.80 | 0.79 | **0.82** $\pm$ 0.01 |
| ARI | N/A | N/A | 0.70 | 0.71 | 0.73 | **0.76** $\pm$ 0.02 |
| ACC | 0.76 | 0.69 | 0.75 | 0.77 | 0.79 | **0.81** $\pm$ 0.03 |
| Params | N/A | N/A | N/A | N/A | 0.08 M | |

# Results: Image Clustering Datasets

Clustering metrics (higher=better, 1.0 is best)

| | DEC* | DCN* | DAMIC* | $K$-DAE* | MixMate INIT | MixMate TRAIN |
|---|---|---|---|---|---|---|
| **MNIST** | | | | | | |
| NMI | 0.80 | 0.81 | **0.86** | **0.86** | 0.75 | **0.86** $\pm$ 0.03 |
| ARI | 0.75 | 0.75 | 0.82 | 0.82 | 0.72 | **0.85** $\pm$ 0.04 |
| ACC | 0.84 | 0.83 | 0.88 | 0.88 | 0.84 | **0.92** $\pm$ 0.04 |
| Params | 2.1 M | 2.1 M | 22.1 M | 21.4 M | 0.4 M | |
| **Fashion** | | | | | | |
| NMI | 0.54 | 0.55 | 0.65 | 0.65 | 0.60 | **0.68** $\pm$ 0.02 |
| ARI | 0.40 | 0.42 | 0.48 | 0.48 | 0.44 | **0.52** $\pm$ 0.01 |
| ACC | 0.51 | 0.50 | 0.60 | 0.60 | 0.57 | **0.63** $\pm$ 0.01 |
| Params | N/A | N/A | N/A | N/A | 0.4 M | |
| **USPS** | | | | | | |
| NMI | 0.77 | 0.68 | 0.78 | 0.80 | 0.79 | **0.82** $\pm$ 0.01 |
| ARI | N/A | N/A | 0.70 | 0.71 | 0.73 | **0.76** $\pm$ 0.02 |
| ACC | 0.76 | 0.69 | 0.75 | 0.77 | 0.79 | **0.81** $\pm$ 0.03 |
| Params | N/A | N/A | N/A | N/A | 0.08 M | |

# Results: Image Clustering Datasets

Other state-of-the-art deep clustering networks [1]

Clustering metrics (higher=better, 1.0 is best)

| | DEC* | DCN* | DAMIC* | $K$-DAE* | MixMate | |
|---|---|---|---|---|---|---|
| | | | | | INIT | TRAIN |
| **MNIST** | | | | | | |
| NMI | 0.80 | 0.81 | **0.86** | **0.86** | 0.75 | **0.86** $\pm$ 0.03 |
| ARI | 0.75 | 0.75 | 0.82 | 0.82 | 0.72 | **0.85** $\pm$ 0.04 |
| ACC | 0.84 | 0.83 | 0.88 | 0.88 | 0.84 | **0.92** $\pm$ 0.04 |
| Params | 2.1 M | 2.1 M | 22.1 M | 21.4 M | 0.4 M | |
| **Fashion** | | | | | | |
| NMI | 0.54 | 0.55 | 0.65 | 0.65 | 0.60 | **0.68** $\pm$ 0.02 |
| ARI | 0.40 | 0.42 | 0.48 | 0.48 | 0.44 | **0.52** $\pm$ 0.01 |
| ACC | 0.51 | 0.50 | 0.60 | 0.60 | 0.57 | **0.63** $\pm$ 0.01 |
| Params | N/A | N/A | N/A | N/A | 0.4 M | |
| **USPS** | | | | | | |
| NMI | 0.77 | 0.68 | 0.78 | 0.80 | 0.79 | **0.82** $\pm$ 0.01 |
| ARI | N/A | N/A | 0.70 | 0.71 | 0.73 | **0.76** $\pm$ 0.02 |
| ACC | 0.76 | 0.69 | 0.75 | 0.77 | 0.79 | **0.81** $\pm$ 0.03 |
| Params | N/A | N/A | N/A | N/A | 0.08 M | |

[1] Opochinsky, Y., Chazan, S. E., Gannot, S., & Goldberger, J. (2020, May). K-autoencoders deep clustering. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4037-4041). IEEE.

# Results: Image Clustering Datasets

Other state-of-the-art deep clustering networks [1]

Clustering metrics (higher=better, 1.0 is best)

| | DEC* | DCN* | DAMIC* | $K$-DAE* | MixMate | |
|---|---|---|---|---|---|---|
| | | | | | INIT | TRAIN |
| **MNIST** | | | | | | |
| NMI | 0.80 | 0.81 | **0.86** | **0.86** | 0.75 | **0.86** $\pm$ 0.03 |
| ARI | 0.75 | 0.75 | 0.82 | 0.82 | 0.72 | **0.85** $\pm$ 0.04 |
| ACC | 0.84 | 0.83 | 0.88 | 0.88 | 0.84 | **0.92** $\pm$ 0.04 |
| Params | 2.1 M | 2.1 M | 22.1 M | 21.4 M | | 0.4 M |
| **Fashion** | | | | | | |
| NMI | 0.54 | 0.55 | 0.65 | 0.65 | 0.60 | **0.68** $\pm$ 0.02 |
| ARI | 0.40 | 0.42 | 0.48 | 0.48 | 0.44 | **0.52** $\pm$ 0.01 |
| ACC | 0.51 | 0.50 | 0.60 | 0.60 | 0.57 | **0.63** $\pm$ 0.01 |
| Params | N/A | N/A | N/A | N/A | | 0.4 M |
| **USPS** | | | | | | |
| NMI | 0.77 | 0.68 | 0.78 | 0.80 | 0.79 | **0.82** $\pm$ 0.01 |
| ARI | N/A | N/A | 0.70 | 0.71 | 0.73 | **0.76** $\pm$ 0.02 |
| ACC | 0.76 | 0.69 | 0.75 | 0.77 | 0.79 | **0.81** $\pm$ 0.03 |
| Params | N/A | N/A | N/A | N/A | | 0.08 M |

MixMate obtains best performance on all metrics for all datasets…

[1] Opochinsky, Y., Chazan, S. E., Gannot, S., & Goldberger, J. (2020, May). K-autoencoders deep clustering. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4037-4041). IEEE.

# Results: Image Clustering Datasets

Other state-of-the-art deep clustering networks [1]

| | DEC* | DCN* | DAMIC* | $K$-DAE* | MixMate | |
|---|---|---|---|---|---|---|
| | | | | | INIT | TRAIN |
| **MNIST** | | | | | | |
| NMI | 0.80 | 0.81 | **0.86** | **0.86** | 0.75 | **0.86** $\pm$ 0.03 |
| ARI | 0.75 | 0.75 | 0.82 | 0.82 | 0.72 | **0.85** $\pm$ 0.04 |
| ACC | 0.84 | 0.83 | 0.88 | 0.88 | 0.84 | **0.92** $\pm$ 0.04 |
| Params | 2.1 M | 2.1 M | 22.1 M | 21.4 M | | 0.4 M |
| **Fashion** | | | | | | |
| NMI | 0.54 | 0.55 | 0.65 | 0.65 | 0.60 | **0.68** $\pm$ 0.02 |
| ARI | 0.40 | 0.42 | 0.48 | 0.48 | 0.44 | **0.52** $\pm$ 0.01 |
| ACC | 0.51 | 0.50 | 0.60 | 0.60 | 0.57 | **0.63** $\pm$ 0.01 |
| Params | N/A | N/A | N/A | N/A | | 0.4 M |
| **USPS** | | | | | | |
| NMI | 0.77 | 0.68 | 0.78 | 0.80 | 0.79 | **0.82** $\pm$ 0.01 |
| ARI | N/A | N/A | 0.70 | 0.71 | 0.73 | **0.76** $\pm$ 0.02 |
| ACC | 0.76 | 0.69 | 0.75 | 0.77 | 0.79 | **0.81** $\pm$ 0.03 |
| Params | N/A | N/A | N/A | N/A | | 0.08 M |

Clustering metrics (higher=better, 1.0 is best)

MixMate obtains best performance on all metrics for all datasets…

…with the fewest number of parameters (up to 50x fewer)!

[1] Opochinsky, Y., Chazan, S. E., Gannot, S., & Goldberger, J. (2020, May). K-autoencoders deep clustering. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4037-4041). IEEE.

# Results: Image Clustering Datasets

Other state-of-the-art deep clustering networks [1]

| | DEC* | DCN* | DAMIC* | $K$-DAE* | MixMate | |
|---|---|---|---|---|---|---|
| | | | | | INIT | TRAIN |
| **MNIST** | | | | | | |
| NMI | 0.80 | 0.81 | **0.86** | **0.86** | 0.75 | **0.86** $\pm$ 0.03 |
| ARI | 0.75 | 0.75 | 0.82 | 0.82 | 0.72 | **0.85** $\pm$ 0.04 |
| ACC | 0.84 | 0.83 | 0.88 | 0.88 | 0.84 | **0.92** $\pm$ 0.04 |
| Params | 2.1 M | 2.1 M | 22.1 M | 21.4 M | 0.4 M | |
| **Fashion** | | | | | | |
| NMI | 0.54 | 0.55 | 0.65 | 0.65 | 0.60 | **0.68** $\pm$ 0.02 |
| ARI | 0.40 | 0.42 | 0.48 | 0.48 | 0.44 | **0.52** $\pm$ 0.01 |
| ACC | 0.51 | 0.50 | 0.60 | 0.60 | 0.57 | **0.63** $\pm$ 0.01 |
| Params | N/A | N/A | N/A | N/A | 0.4 M | |
| **USPS** | | | | | | |
| NMI | 0.77 | 0.68 | 0.78 | 0.80 | 0.79 | **0.82** $\pm$ 0.01 |
| ARI | N/A | N/A | 0.70 | 0.71 | 0.73 | **0.76** $\pm$ 0.02 |
| ACC | 0.76 | 0.69 | 0.75 | 0.77 | 0.79 | **0.81** $\pm$ 0.03 |
| Params | N/A | N/A | N/A | N/A | 0.08 M | |

Clustering metrics (higher=better, 1.0 is best)

MixMate obtains best performance on all metrics for all datasets…

…with the fewest number of parameters (up to 50x fewer)!

These numbers don't change even if 90% of images have 25% of pixels missing → MixMate is robust to incomplete data!

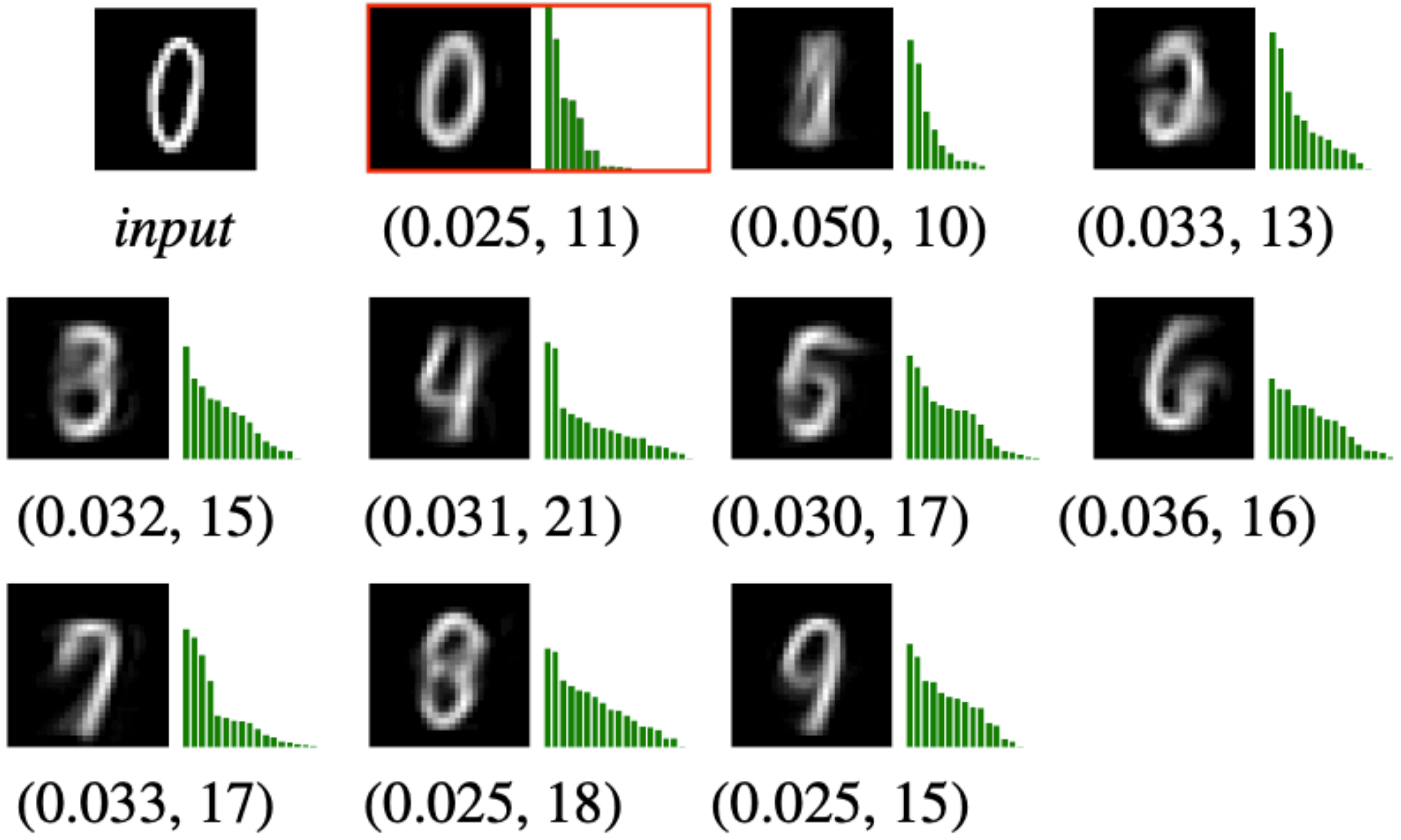
[1] Opochinsky, Y., Chazan, S. E., Gannot, S., & Goldberger, J. (2020, May). K-autoencoders deep clustering. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4037-4041). IEEE.

# Why does MixMate work so well?

## Latent sparsity really helps!

Each of the 10 auto-encoder's output is labeled with (recon error, L0 norm of code)

Clustering depends on *both* reconstruction error and latent sparsity.

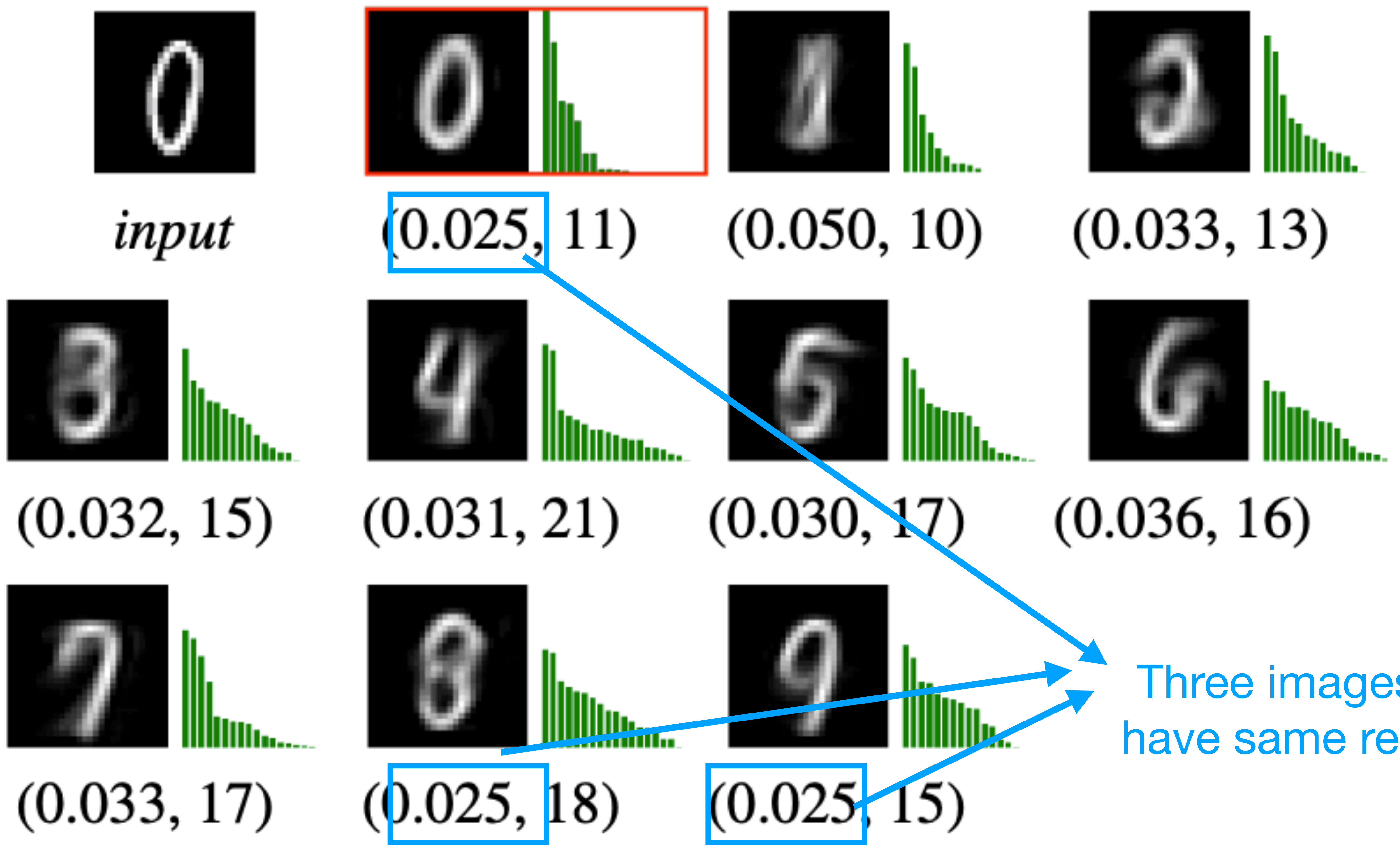


# Why does MixMate work so well?

## Latent sparsity really helps!

Each of the 10 auto-encoder's output is labeled with (recon error, L0 norm of code)

Clustering depends on *both* reconstruction error and latent sparsity.



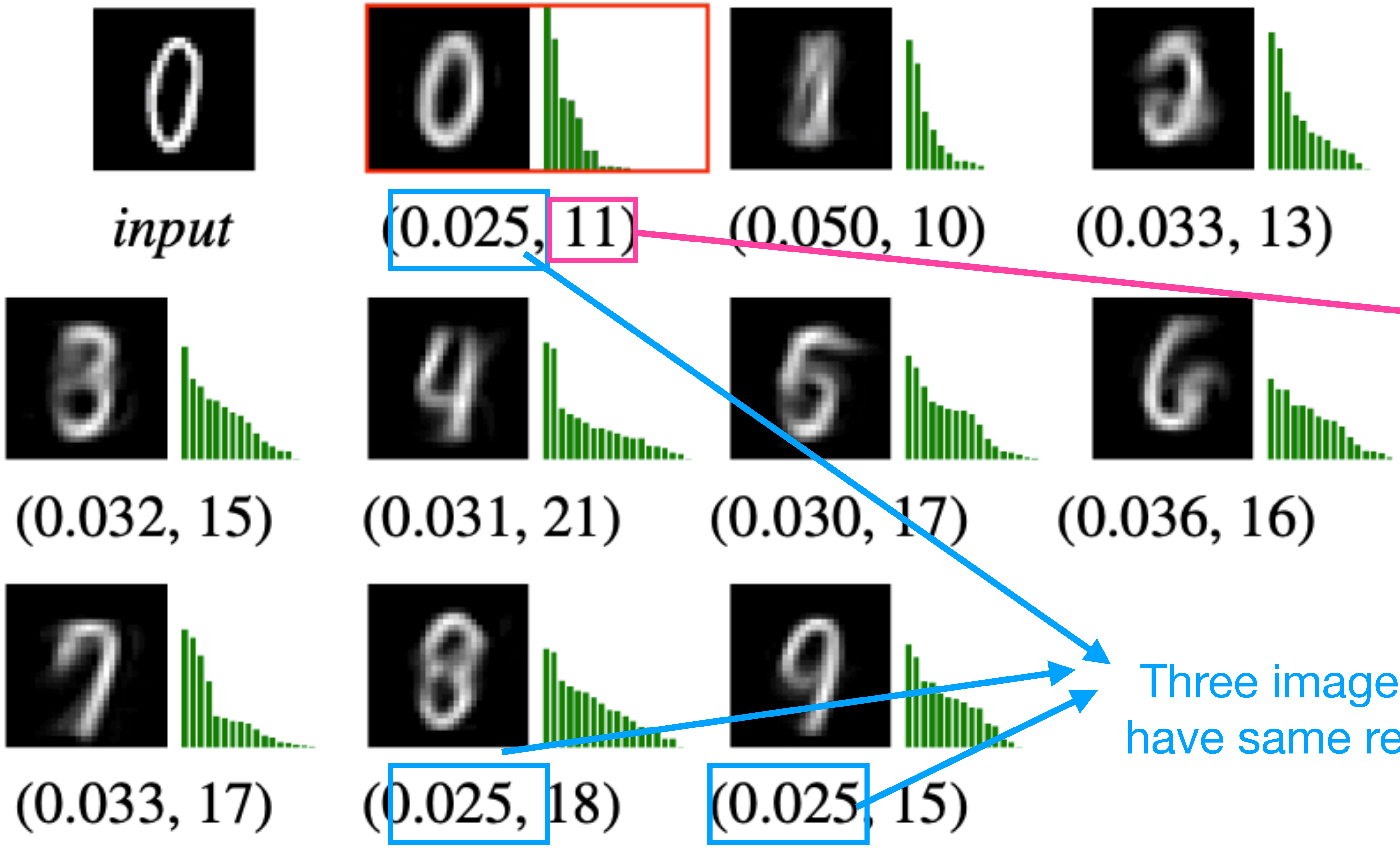

Three images (0, 8, 9) have same recon error!

# Why does MixMate work so well?

## Latent sparsity really helps!

Each of the 10 auto-encoder's output is labeled with (recon error, L0 norm of code)

Clustering depends on *both* reconstruction error and latent sparsity.




Cluster 0 has the sparsest code → data is clustered correctly as cluster 0!

Three images (0, 8, 9) have same recon error!

# For more information…

- Paper: https://ieeexplore.ieee.org/document/9747848
  - Lin, A., Song, A. H., & Ba, D. (2022, May). **Mixture Model Auto-Encoders: Deep Clustering through Dictionary Learning.** In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3368-3372). IEEE.
  - More information on theory behind our MixMate architecture, initialization scheme, tuning the sparsity level, etc.
- Code: https://github.com/al5250/mixmate