REUMass Amherst 2015 Data Science Bootcamp

Day 4: Unsupervised Learning

Prof. Ben Marlin
marlin@cs.umass.edu

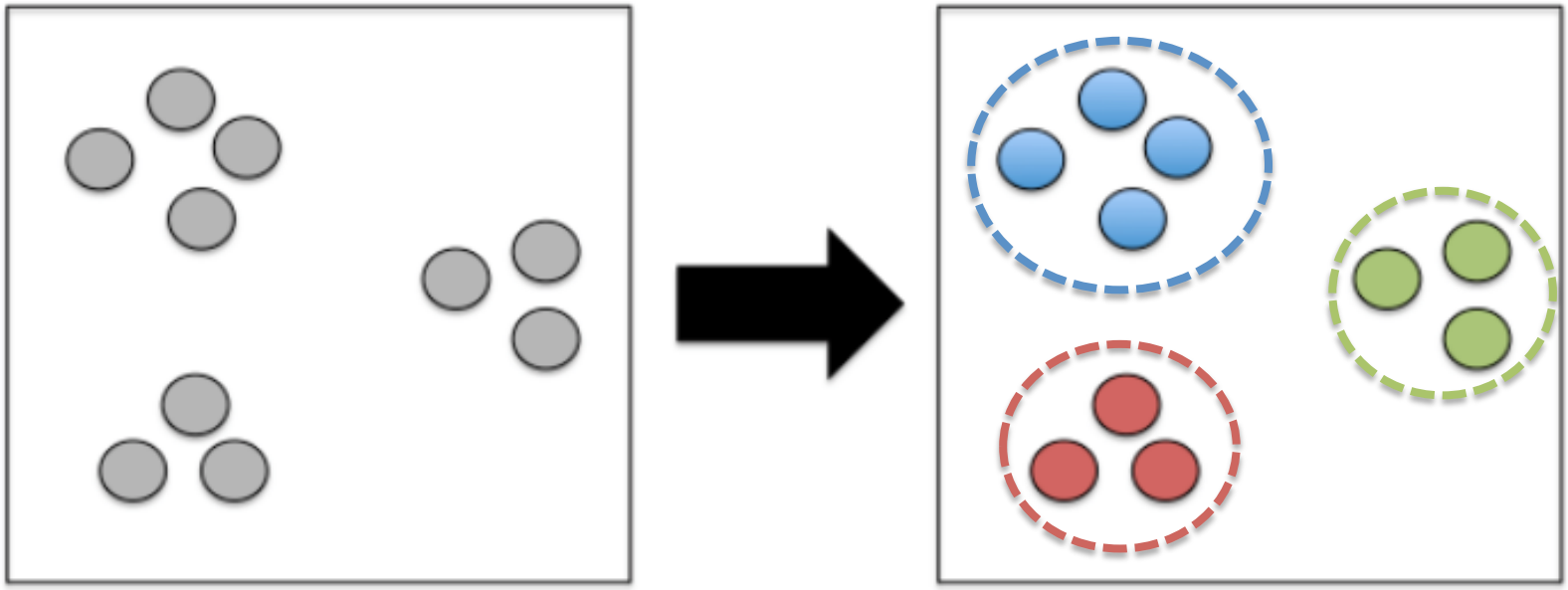
Plan for Day 4:

- Clustering
- Dimensionality Reduction

Clustering

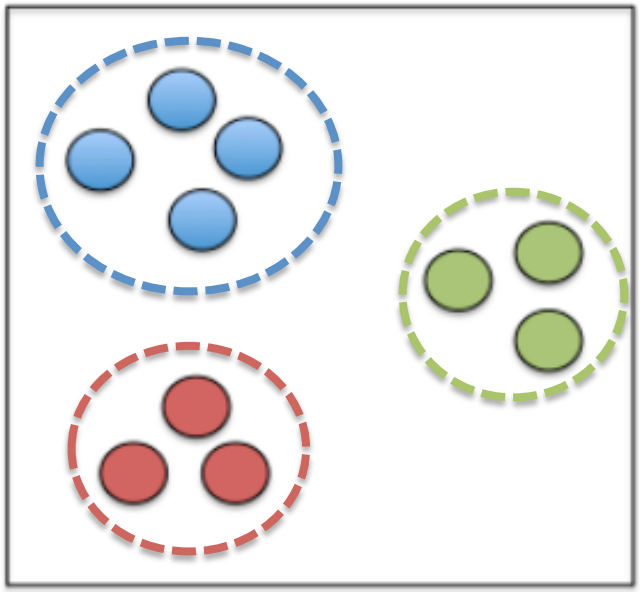
Definition: The Clustering Task

Given a collection of data cases $\mathbf{x}_i \in \mathbb{R}^D$, partition the data cases into groups such that the data cases within each partition are more similar to each other than they are to data cases in other partitions.

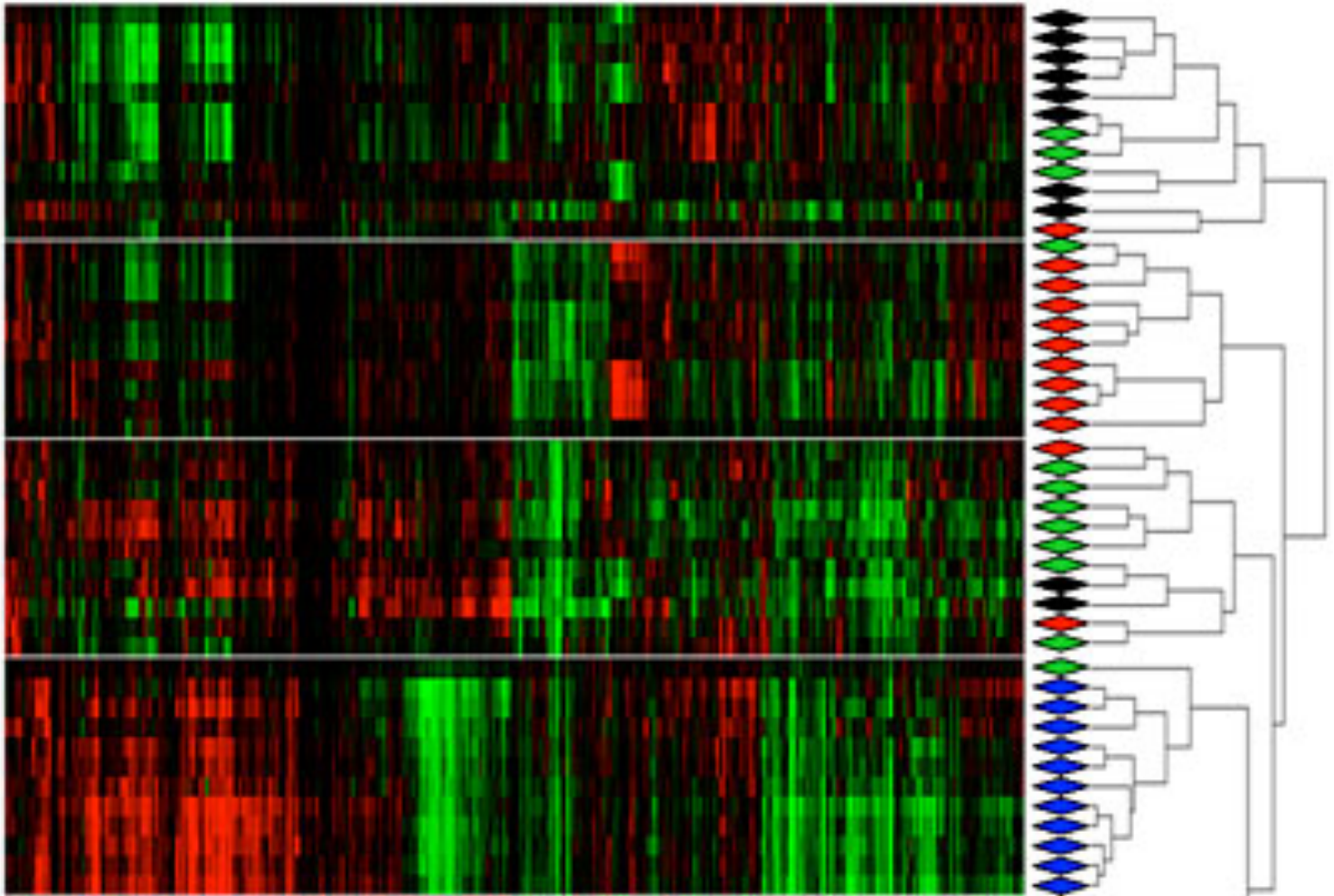


Definition of a Partitioning

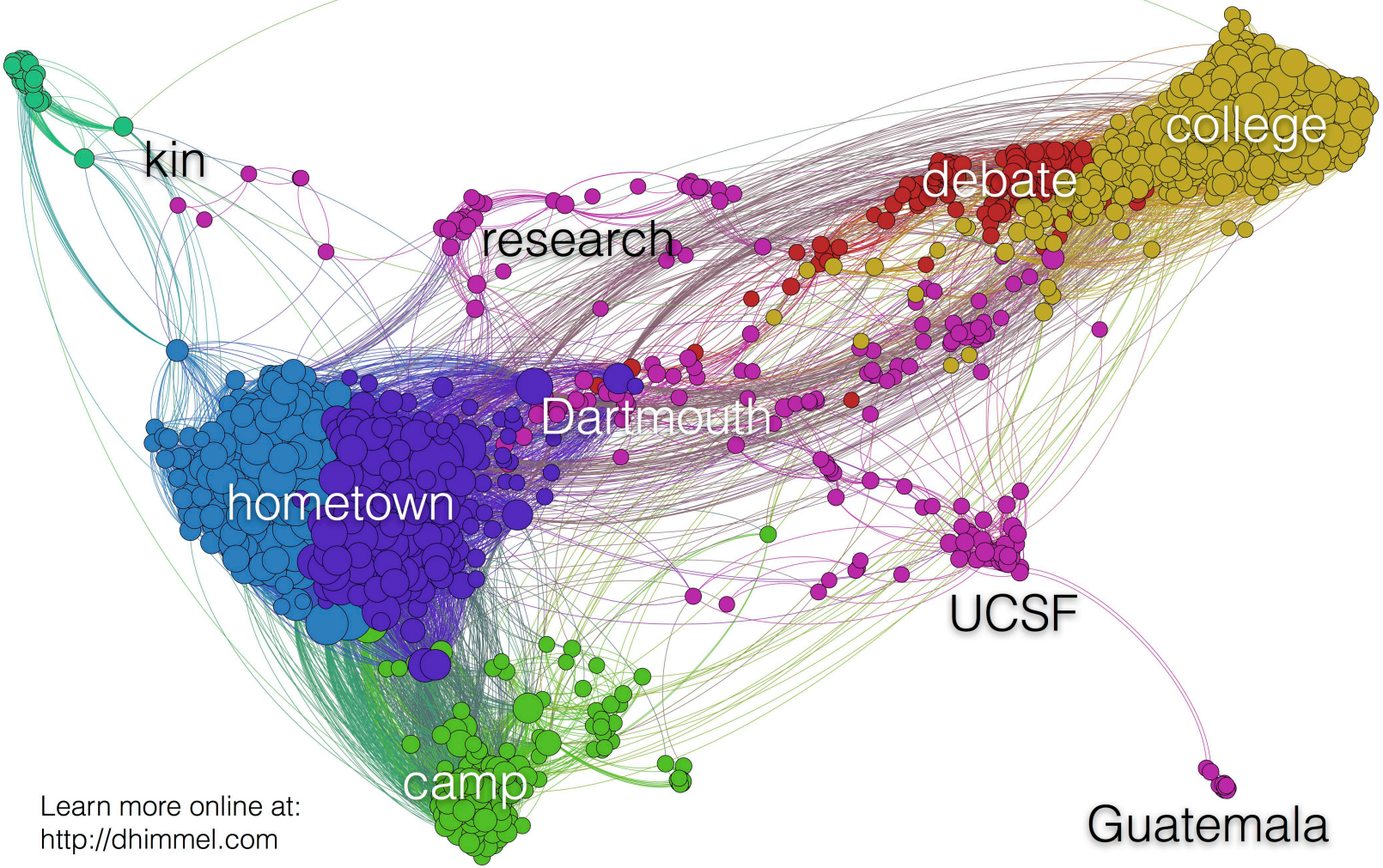
- Suppose we have N data cases $\mathcal{D} = \{\mathbf{x}_i\}_{i=1:N}$.
- A clustering of the N cases into K clusters is a partitioning of $\mathcal{D}$ into K mutually disjoint subsets $\mathcal{C} = \{C_1, \ldots, C_K\}$ such that $C_1 \cup \ldots \cup C_K = \mathcal{D}$.



Example: Gene Expression Data



Example: Online Community Detection



Learn more online at:
http://dhimmel.com

Example: Super Pixels



The K-Means Algorithm

Suppose we let z_i indicate which cluster $\mathbf{x}_i$ belongs to and $\mu_k \in \mathbb{R}^D$ be the cluster centroid/prototype for cluster k. The two main steps of the algorithm can then be expressed as follows:

1. $$z_i = \arg\min_k ||\mu_k - \mathbf{x}_i||_2^2$$

2. $$\mu_k = \frac{\sum_{i=1}^N [z_i = k]\mathbf{x}_i}{\sum_{i=1}^N [z_i = k]}$$

The K-Means Algorithm

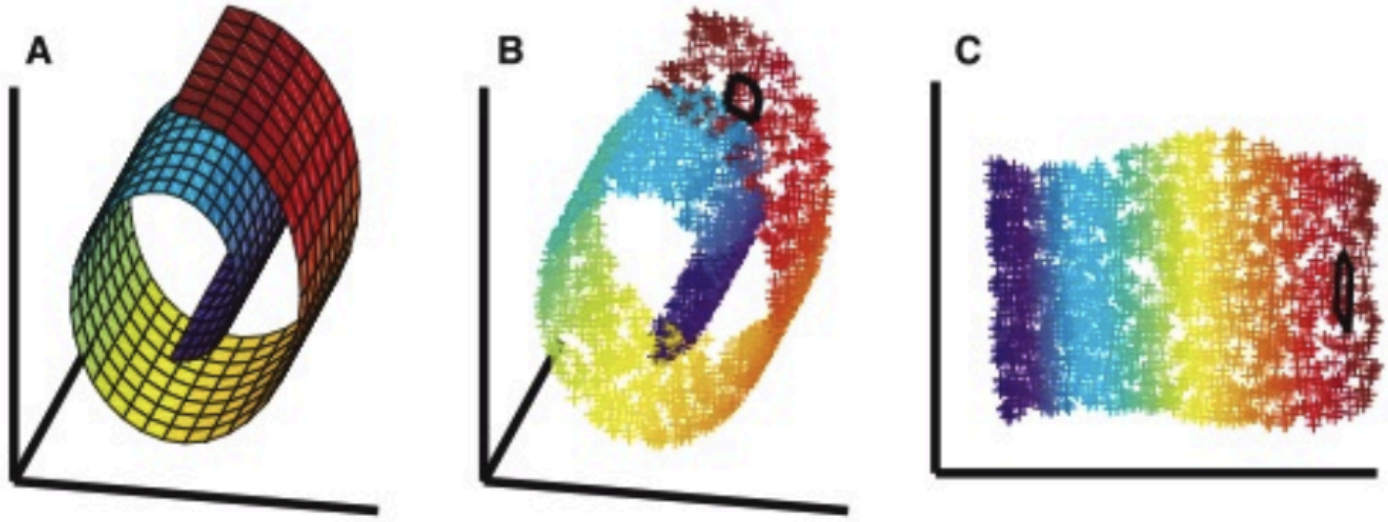
- The K-Means algorithm attempts to minimize the sum of the within-cluster variation over all clusters (also called the within-cluster sum of squares):

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{\mathbf{x}_i, \mathbf{x}_j \in C_k} ||\mathbf{x}_i - \mathbf{x}_j||_2^2$$

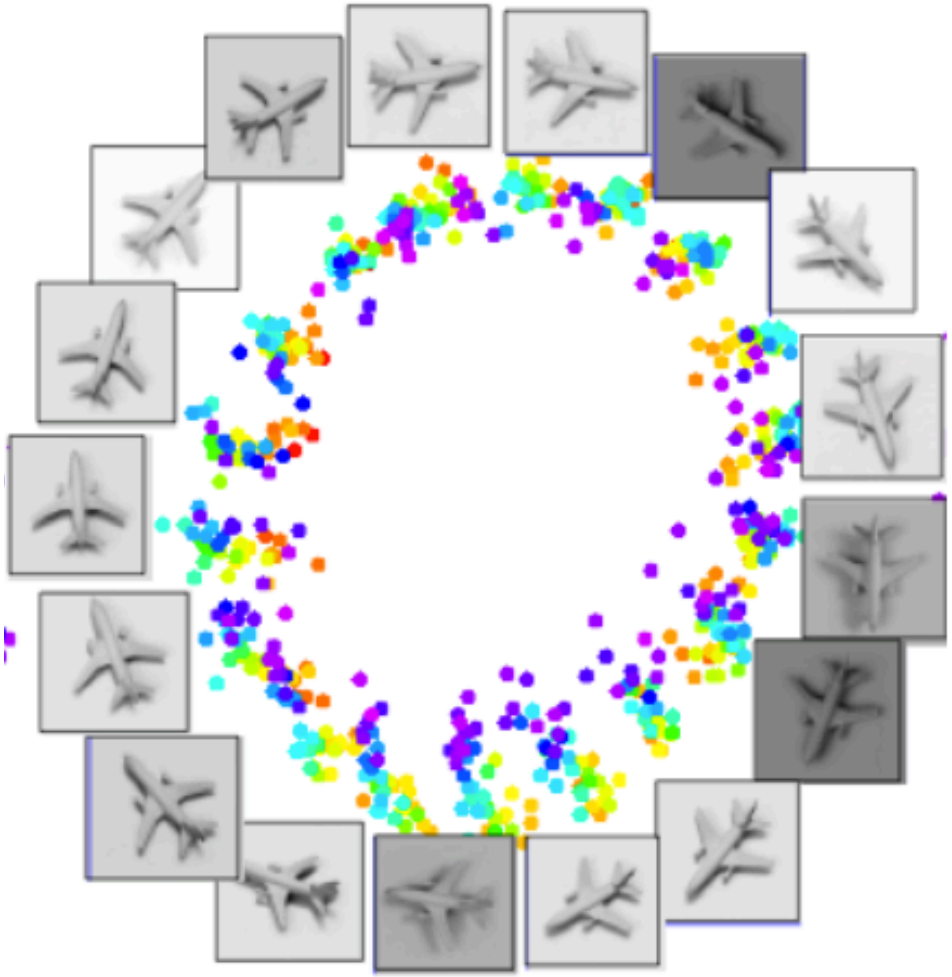
Dimensionality Reduction

Definition: The Dimensionality Reduction Task

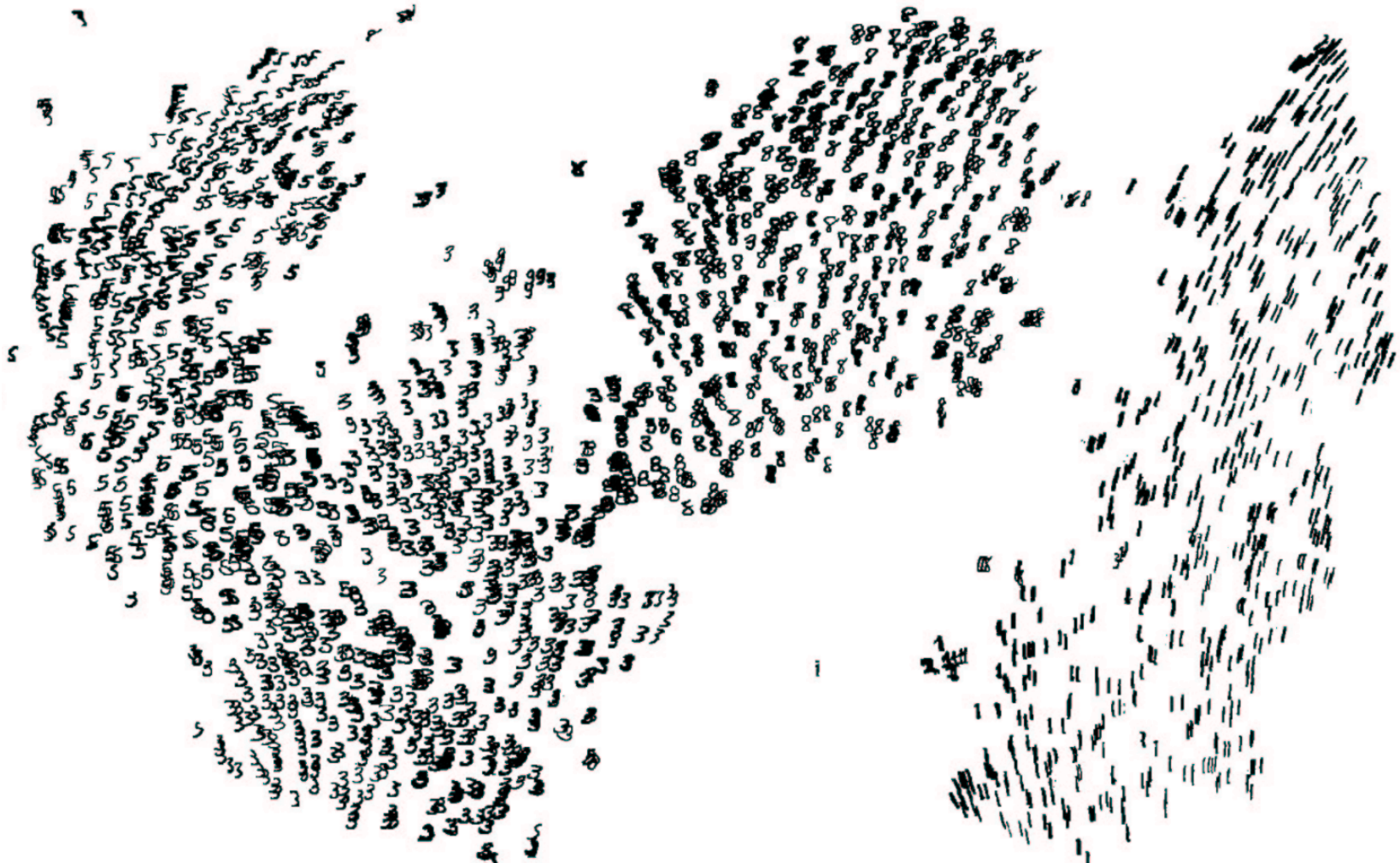
Given a collection of feature vectors $\mathbf{x}_i \in \mathbb{R}^D$, map the feature vectors into a lower dimensional space $\mathbf{z}_i \in \mathbb{R}^K$ where $K < D$ while preserving certain properties of the data.



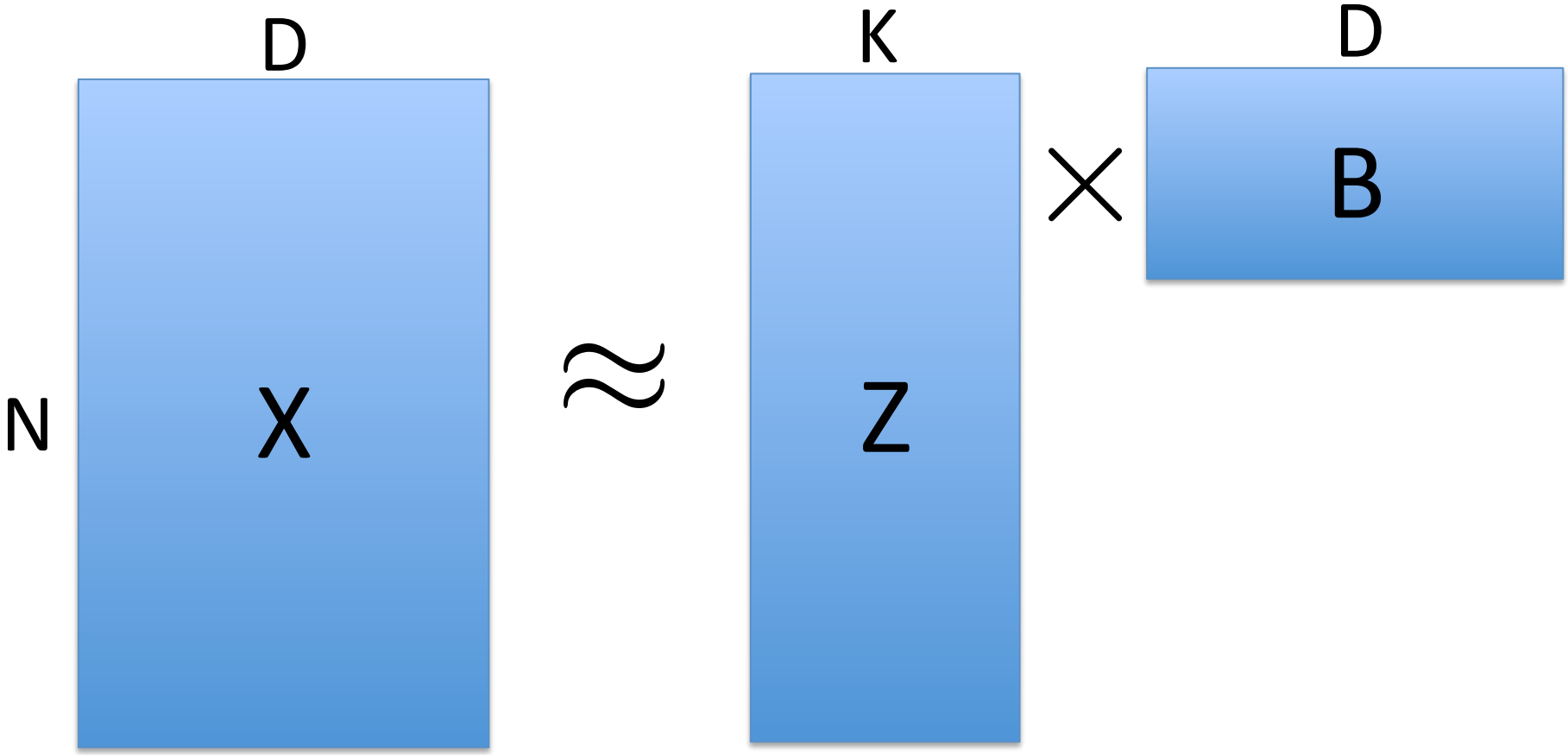
Example: Image Manifolds



Example: Digits



Linear Dimensionality Reduction



Linear Dimensionality Reduction

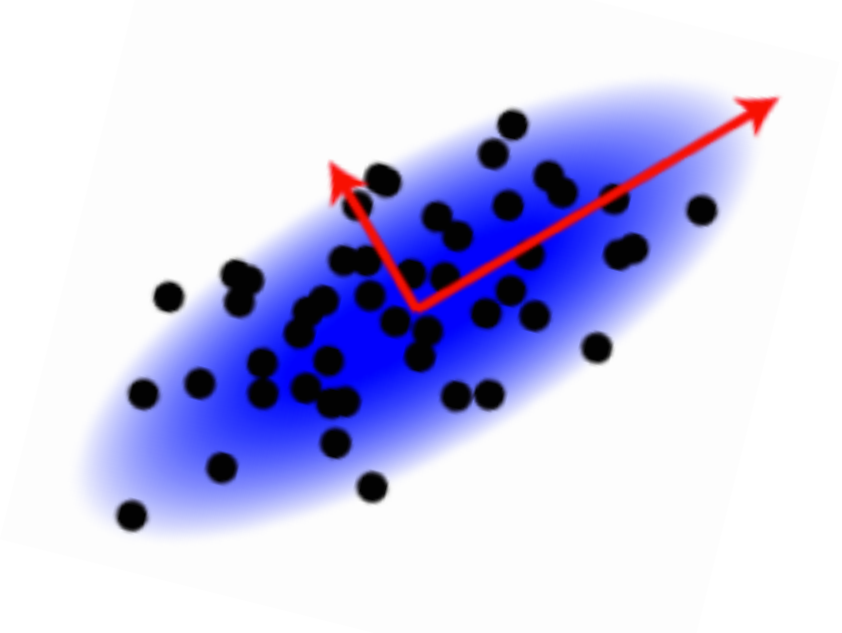
- One possible learning criteria is to minimize the sum of squared errors when reconstructing $\mathbf{X}$ from $\mathbf{Z}$ and $\mathbf{B}$. This leads to:

$$\arg\min_{\mathbf{Z},\mathbf{B}} ||\mathbf{X} - \mathbf{ZB}||_F$$

where $||\mathbf{A}||_F$ is the Frobenius norm of matrix $\mathbf{A}$ (the sum of the squares of all matrix entries).

Principal Components Analysis

Under the assumption that the matrix B is orthonormal, we obtain a classical method called *Principal Components Analysis* where the basis elements correspond to directions of maximum variation in the data.



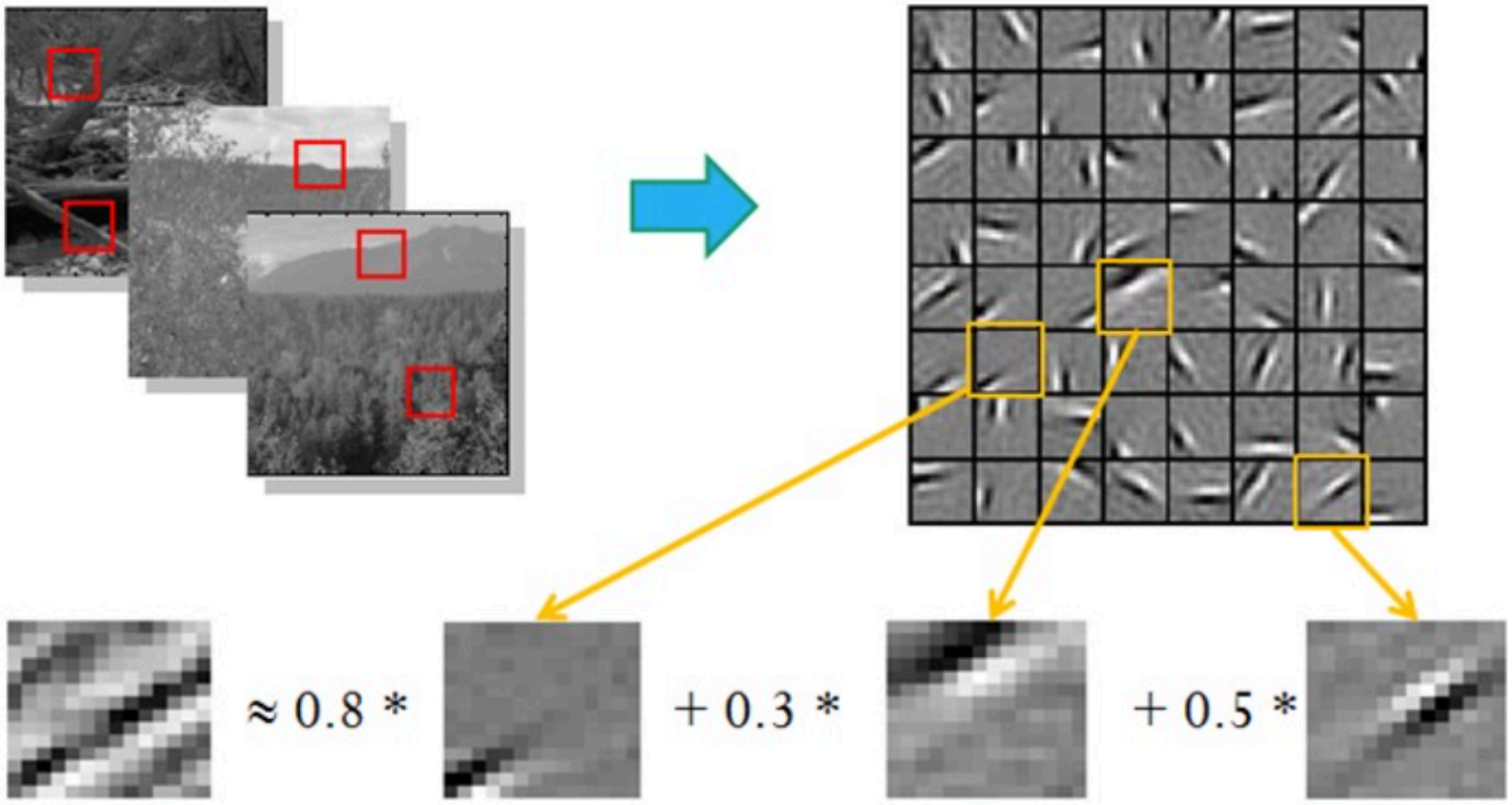
Sparse Coding

Under the additional constraint that the rows of Z are sparse, we obtain a method called *Sparse Coding*:

$$\min_{\mathbf{Z},\mathbf{B}} ||\mathbf{X} - \mathbf{ZB}||_F - \lambda||\mathbf{Z}||_1$$

$$\text{such that } ||B_k||_2 = 1 \text{ for all } k$$

Sparse Coding



Multi-Dimensional Scaling

- MDS is a non-linear dimensionality reduction method that is explicitly designed to minimize the distortion in the pairwise distances between points when projecting them into a low dimensional embedding.
- Least-squares MDS learns the embeddings $\mathbf{z}_i$ by minimizing the following objective function, known as the *stress* function:

$$\min_{\mathbf{z}_1,\ldots,\mathbf{z}_N} \sum_{i<j} (d_{ij} - ||\mathbf{z}_i - \mathbf{z}_j||_2)^2$$

ISOMAP

- Isometric feature mapping (Isomap) is a non-linear dimensionality reduction method that is designed to minimize the distortion in geodesic distances on a manifold when projecting them into a low dimensional embedding.