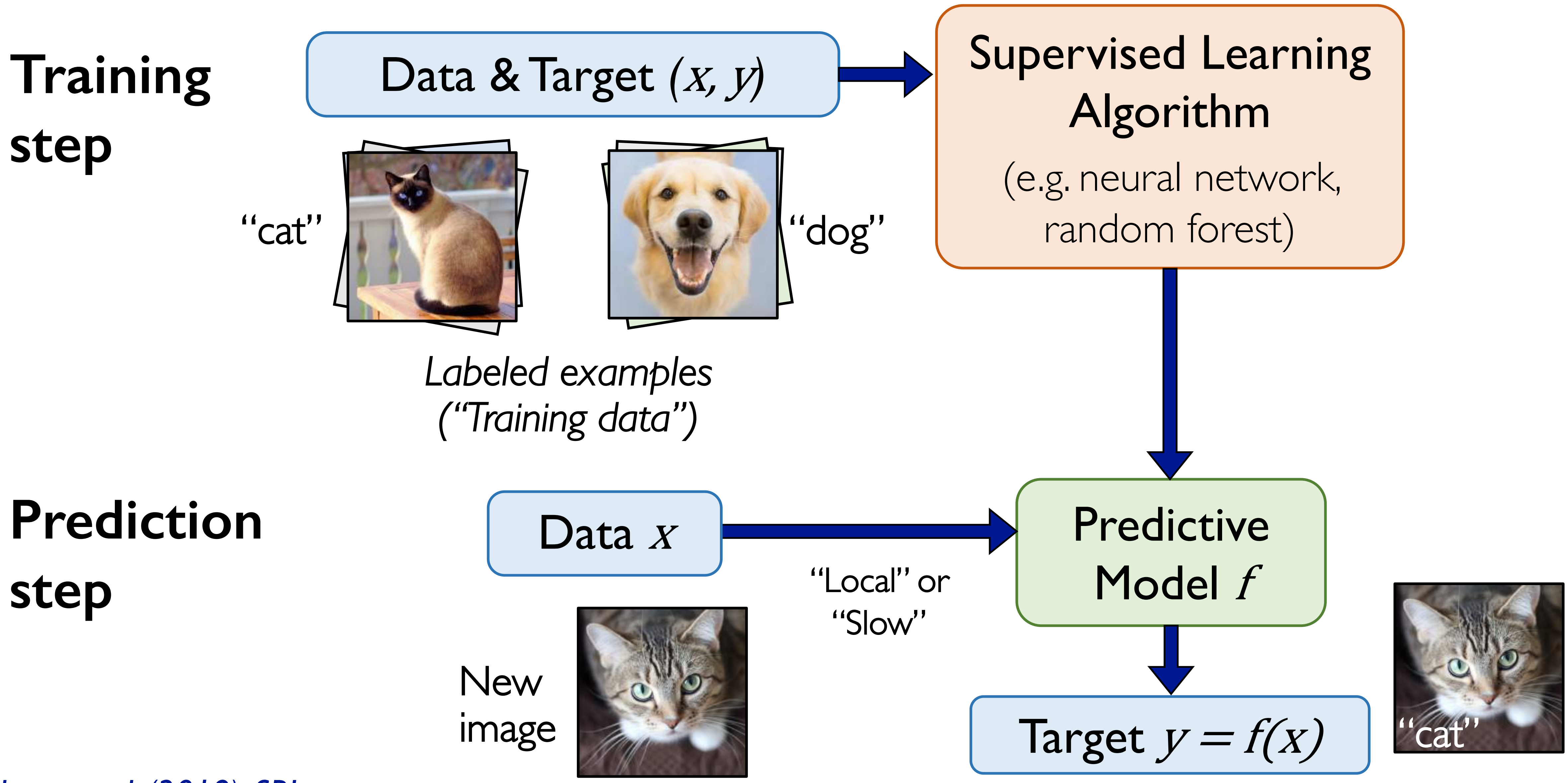
Acknowledgment: Alex Ioannidis (Stanford University) contributed slides to this presentation

Unsupervised Learning
for Geoscience Applications

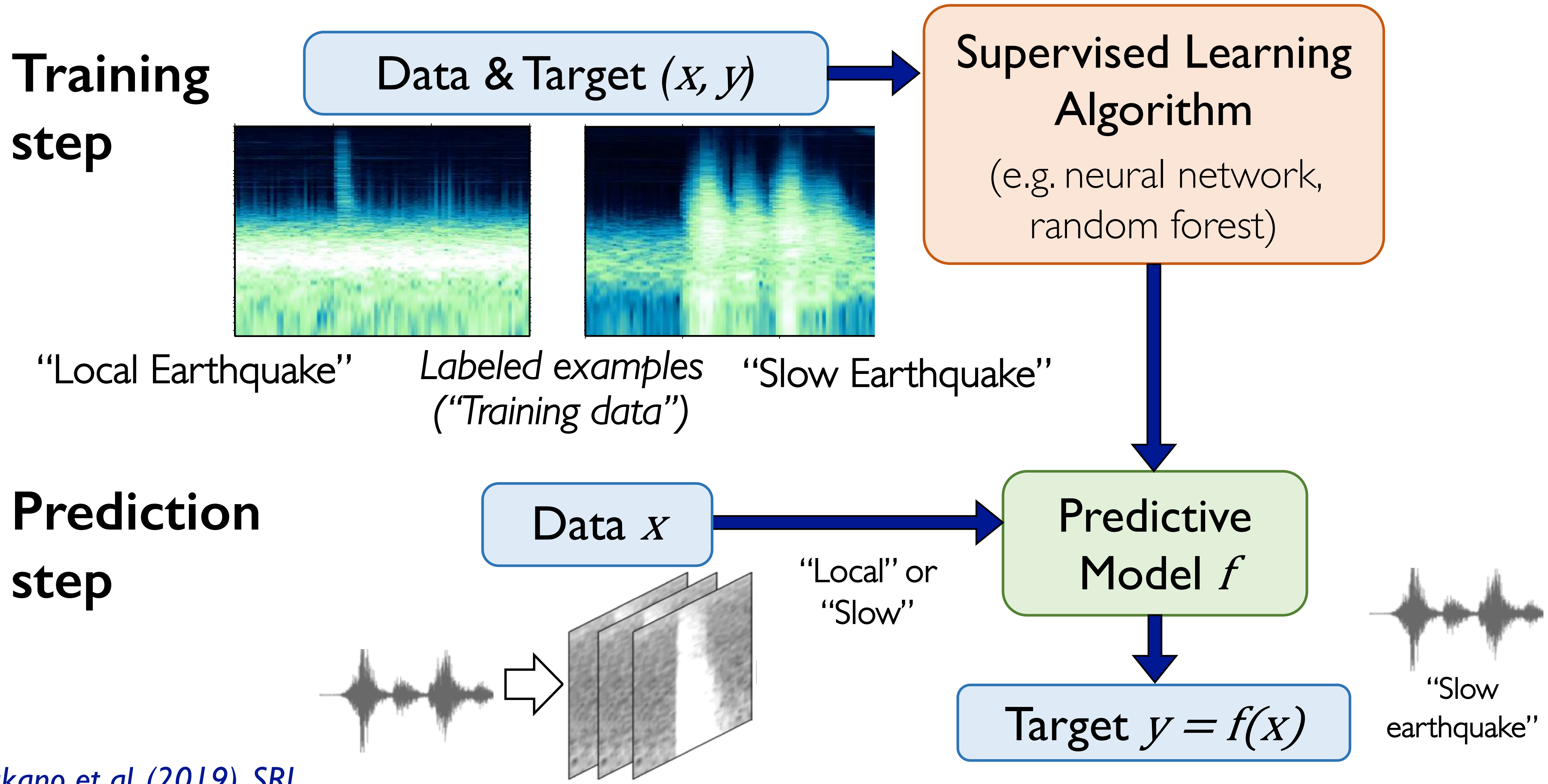
Karianne J. Bergen
Data Science Initiative Postdoctoral Fellow
Harvard University

Supervised Learning: *Building models from examples*



Nakano et al. (2019), SRL

Supervised Learning: *Building models from examples*



Nakano et al. (2019), SRL

Unsupervised Learning: *Finding patterns in data*

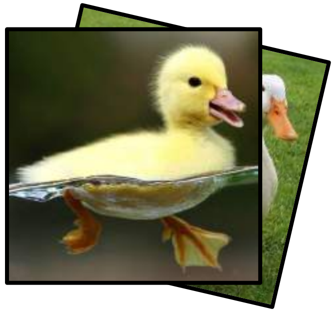
Data X → Unsupervised Learning Algorithm (e.g. PCA, k-means) → Structure in X

Images only
(no labels)

Common patterns or features in images

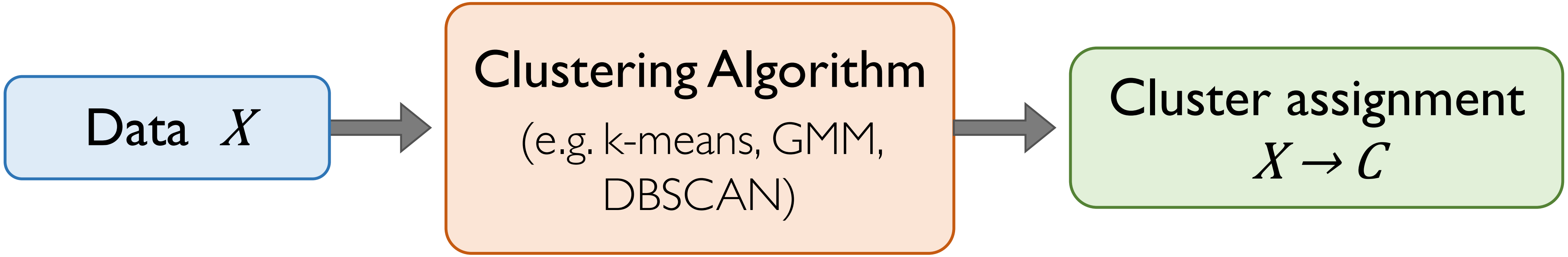
OR

Groups of similar images



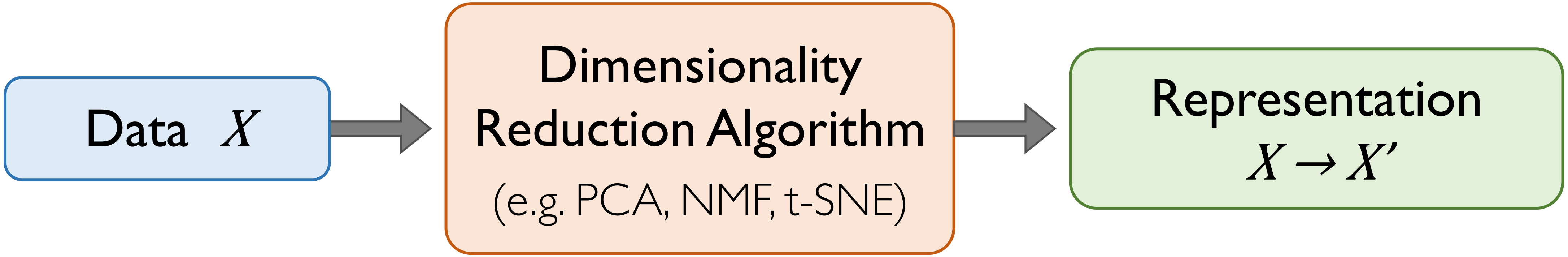
Clustering

Identifying homogenous subgroups of samples



Dimensionality Reduction / Feature Learning

Finding a new (low-dimensional) representation to characterize the data



Unsupervised learning

When is it used ?

- Exploratory Data Analysis & Visualization
- Preprocessing for Supervised Learning
- Learning without Labels

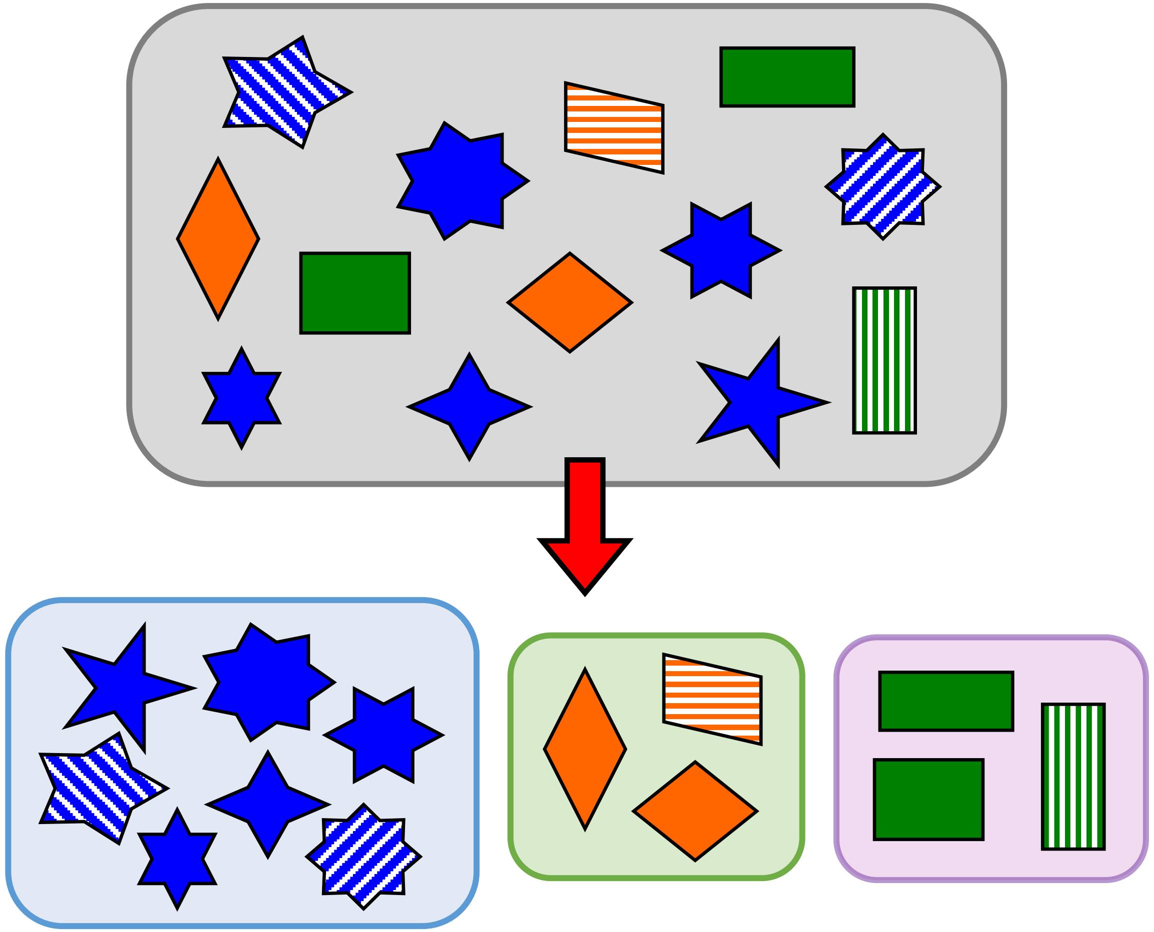
Unsupervised learning
Why is it challenging ?

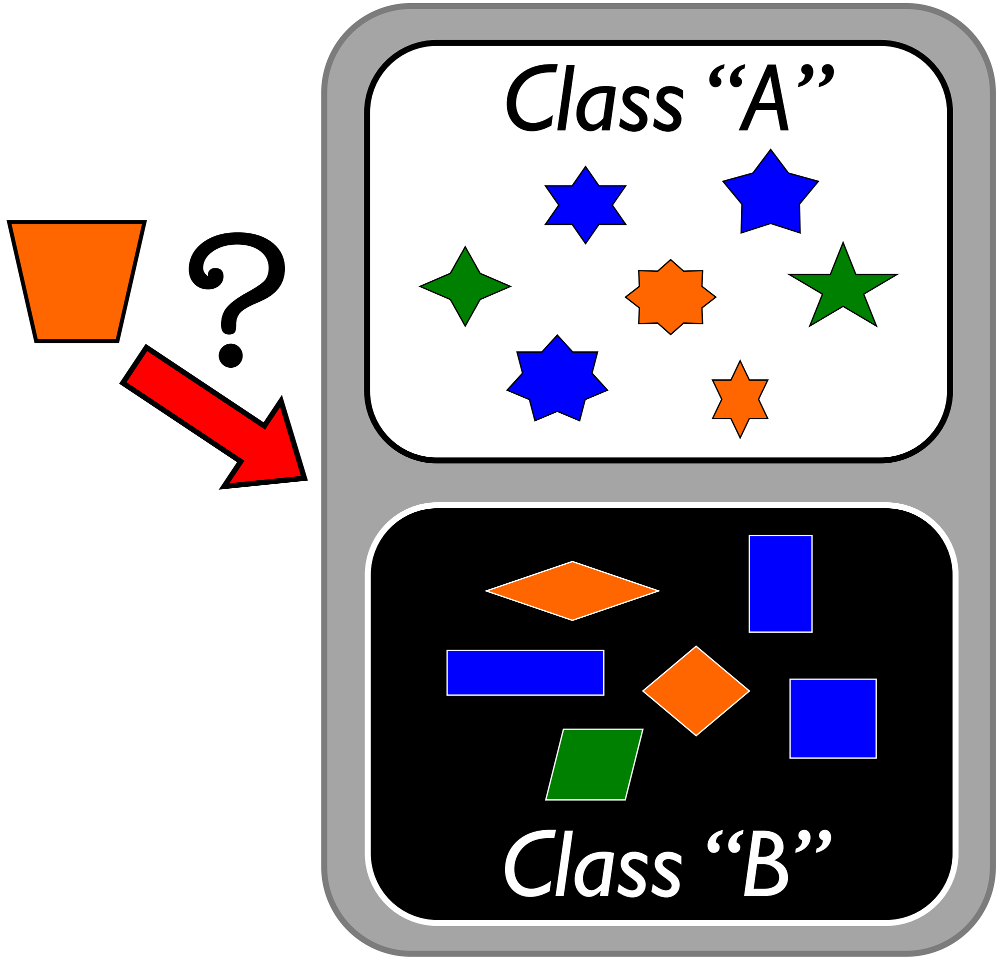
Cluster Analysis

K-means & Hierarchical clustering

Clustering: identifies subgroups within data –

common within-group characteristics, differences across groups





Groupings determined from the data itself, unlike classification

Types of clustering algorithms

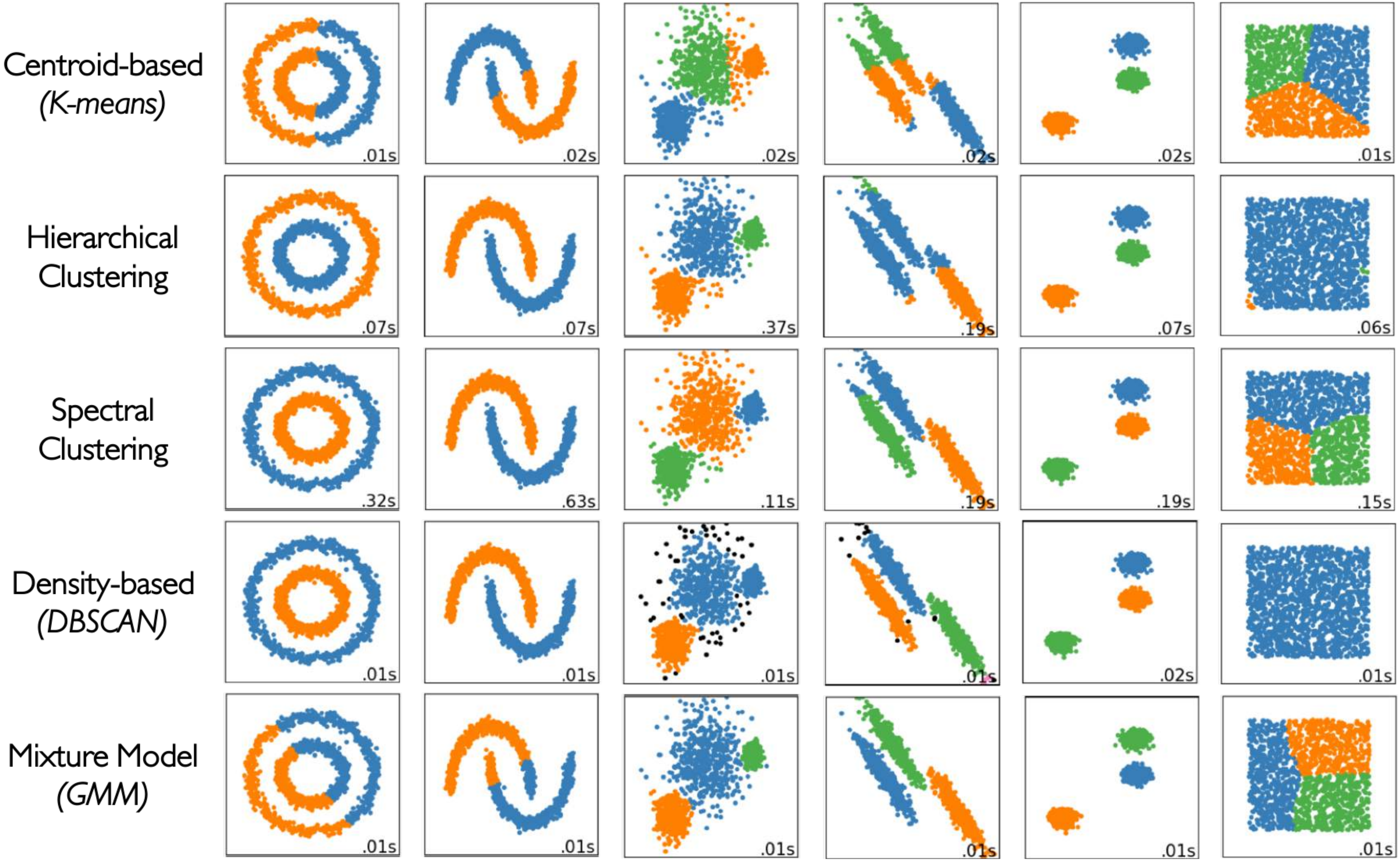
Hierarchical (Agglomerative) Clustering

Centroid-based Clustering

Spectral Methods

Density-based Clustering

Mixture Models



Centroid-based
(K-means)
.01s
.02s
.02s
.02s
.02s
.01s
Hierarchical
Clustering
.07s
.07s
.37s
.19s
.07s
.06s
Spectral
Clustering
.32s
.63s
.11s
.19s
.19s
.15s
Density-based
(DBSCAN)
.01s
.01s
.01s
.01s
.02s
.01s
Mixture Model
(GMM)
.01s
.01s
.01s
.01s
.01s
.01s

Types of clustering algorithms

Hierarchical (Agglomerative) Clustering

Centroid-based Clustering

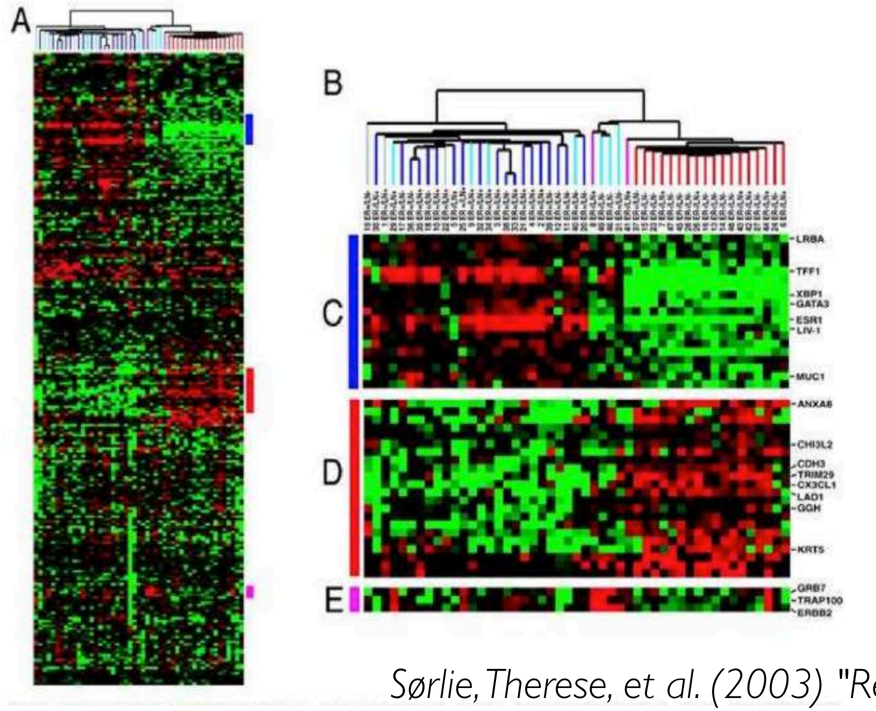
Spectral Methods

Density-based Clustering

Mixture Models

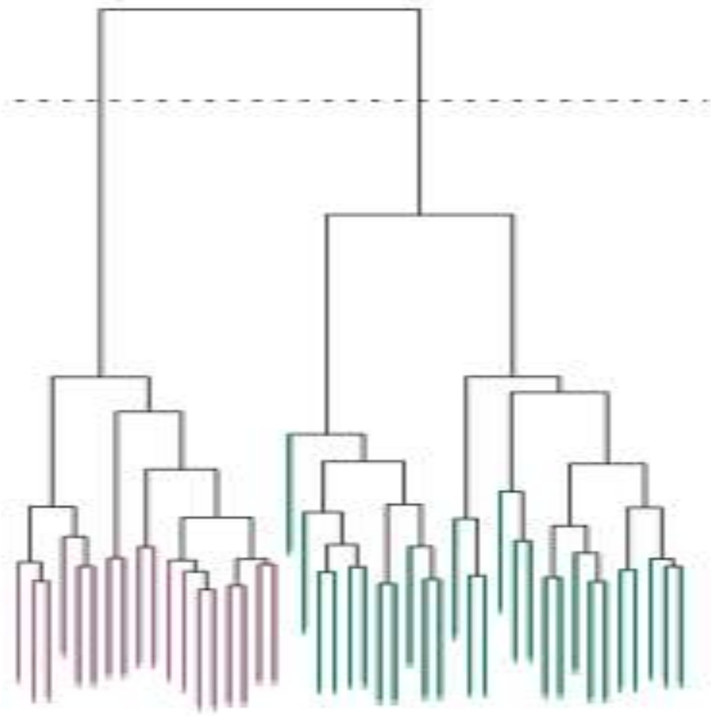
Hierarchical Clustering

- Merges clusters/observations that are “closest” together
- Represented as a hierarchy rather than a partition of data



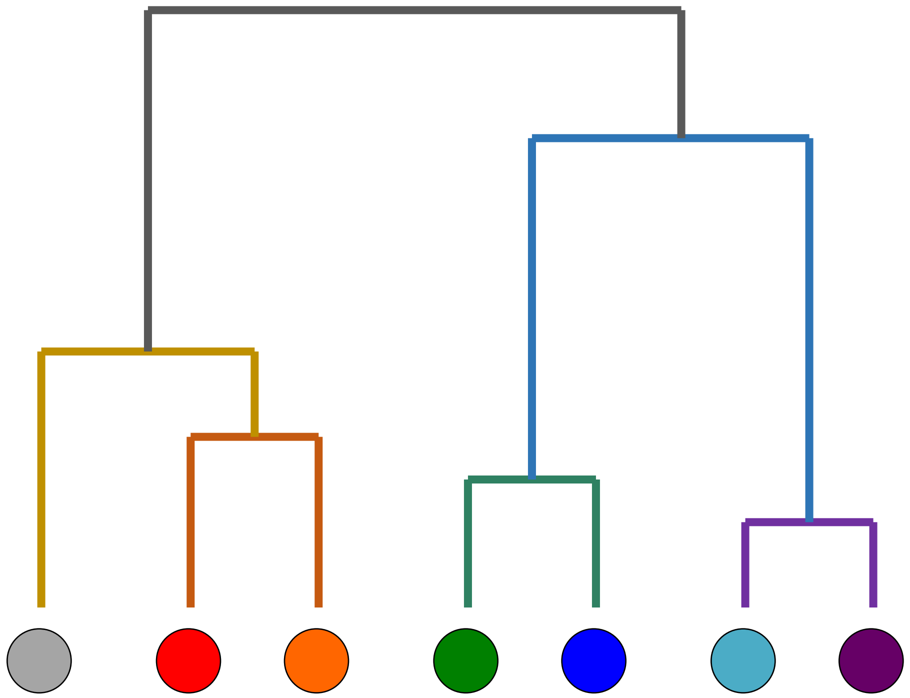
Sørlie, Therese, et al. (2003) "Repeated observation of breast tumor subtypes in independent gene expression data sets," PNAS.

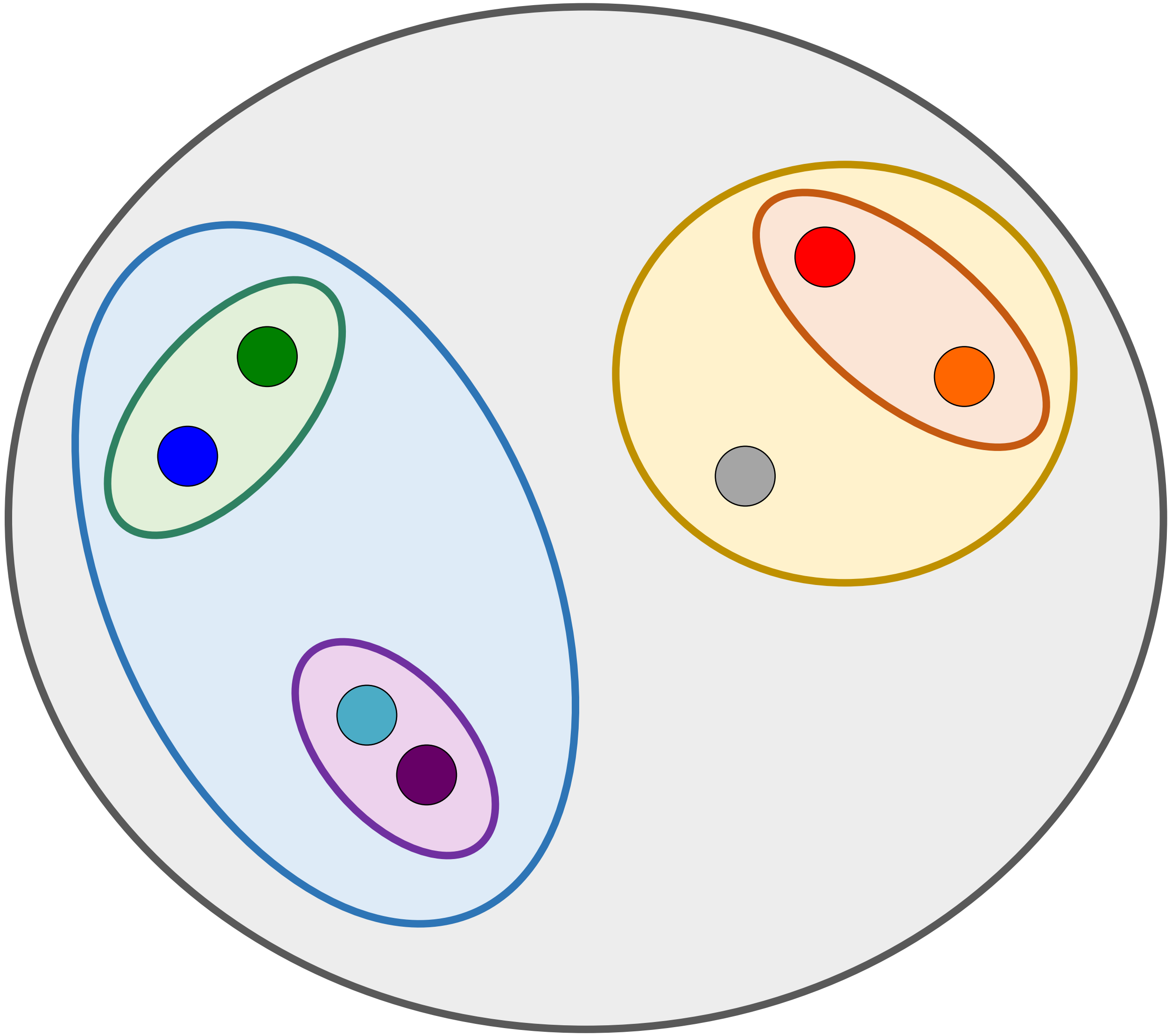
Figure 10.9 , ISL 2013



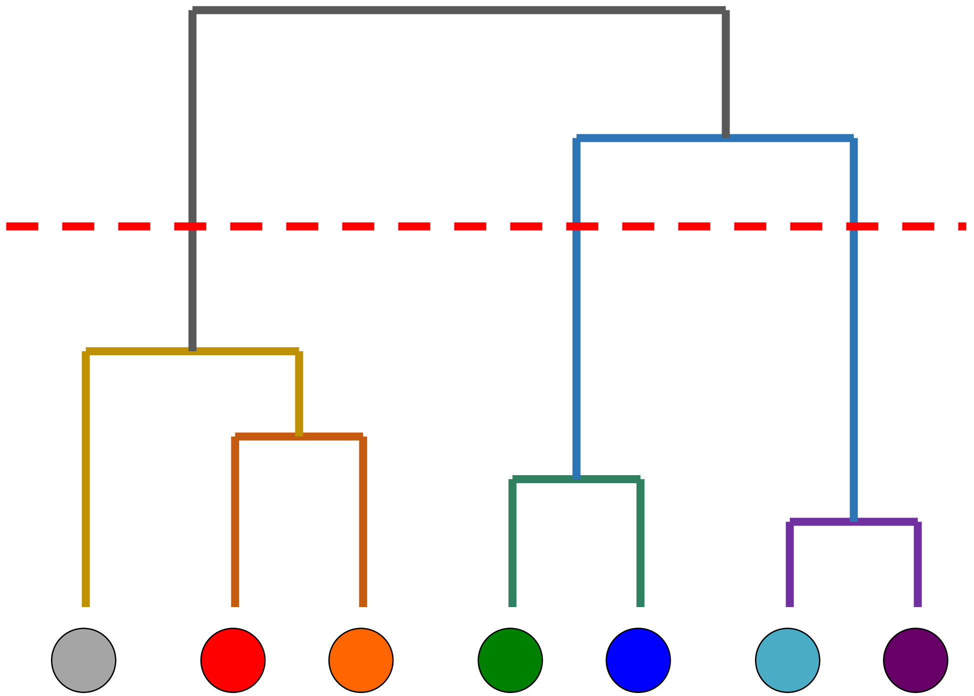
Dendrogram

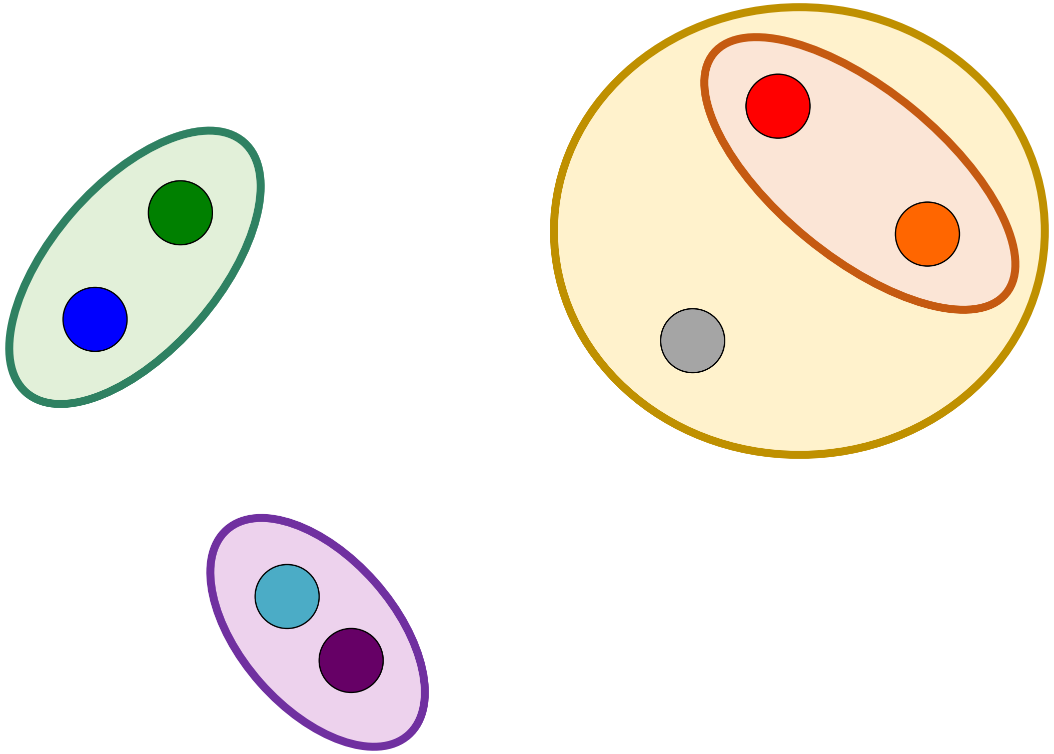
Hierarchical Clustering



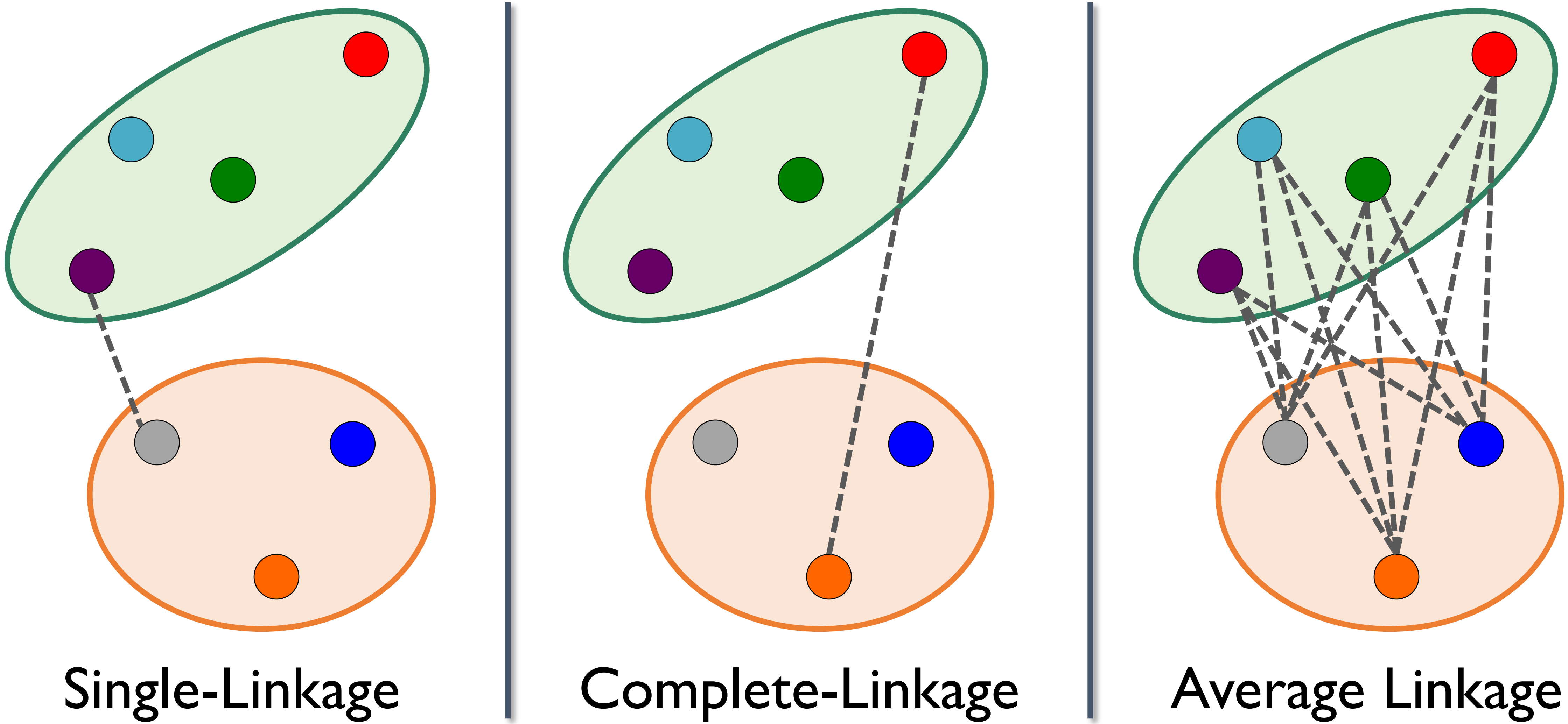


Hierarchical Clustering





Hierarchical Clustering is a family of clustering methods.



What does it mean for two clusters to be "close"?

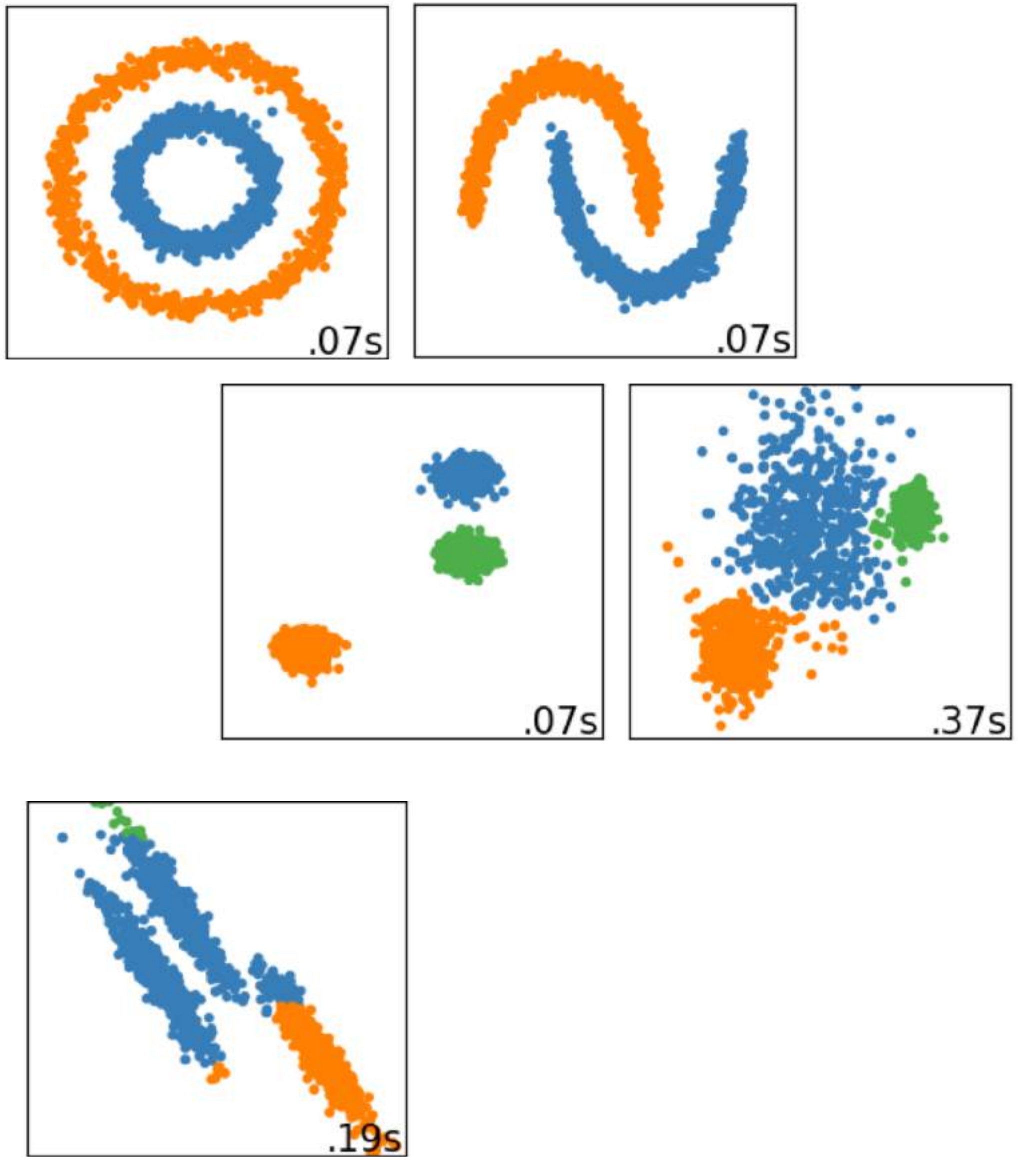
Hierarchical Clustering

Advantages

- Don't need to know # of clusters
- Can find non-spherical clusters

Disadvantages

- Doesn't scale to large data sets
- # clusters can be difficult to determine
- Can be sensitive to noise/outliers



Types of clustering algorithms

Hierarchical (Agglomerative) Clustering

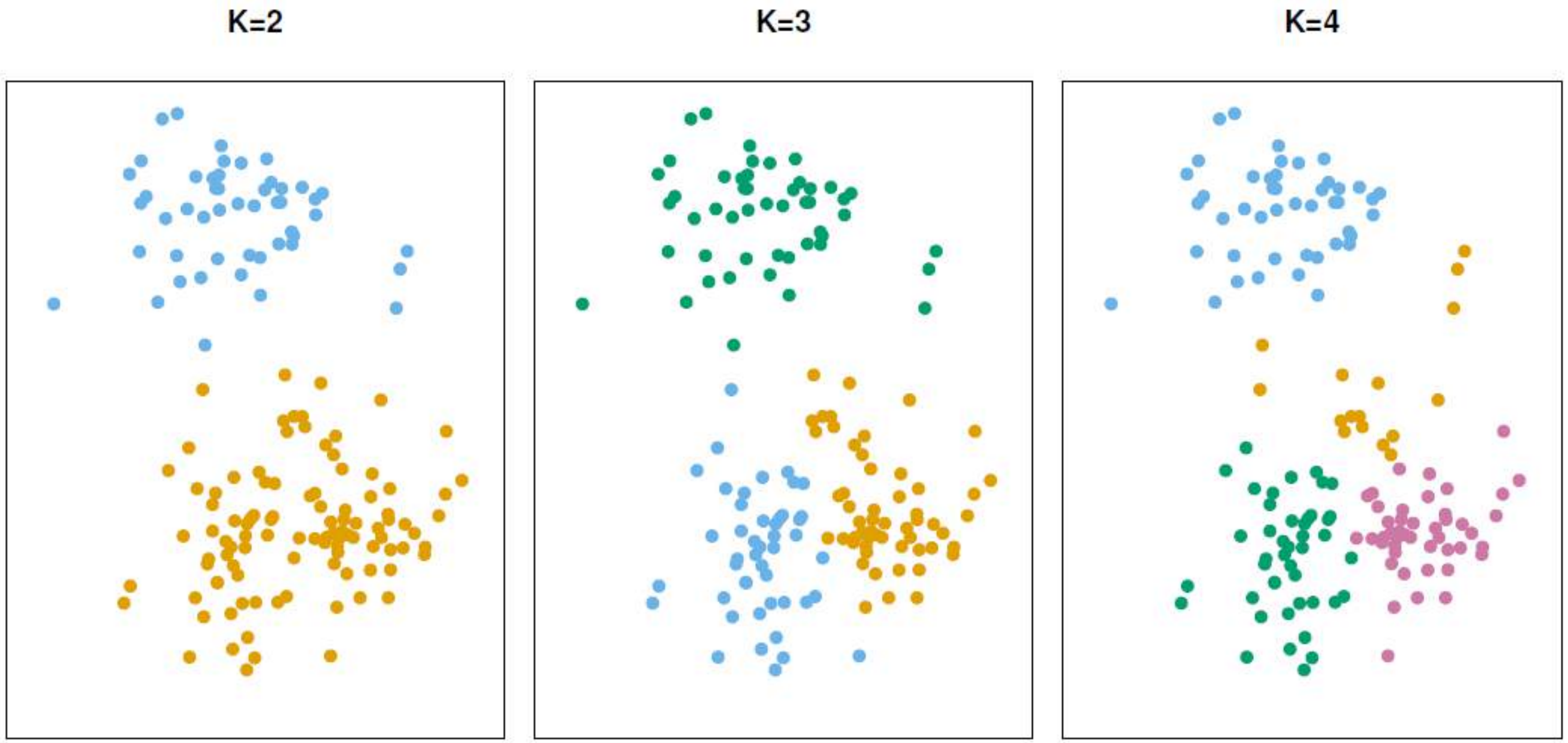
Centroid-based Clustering

Spectral Methods

Density-based Clustering

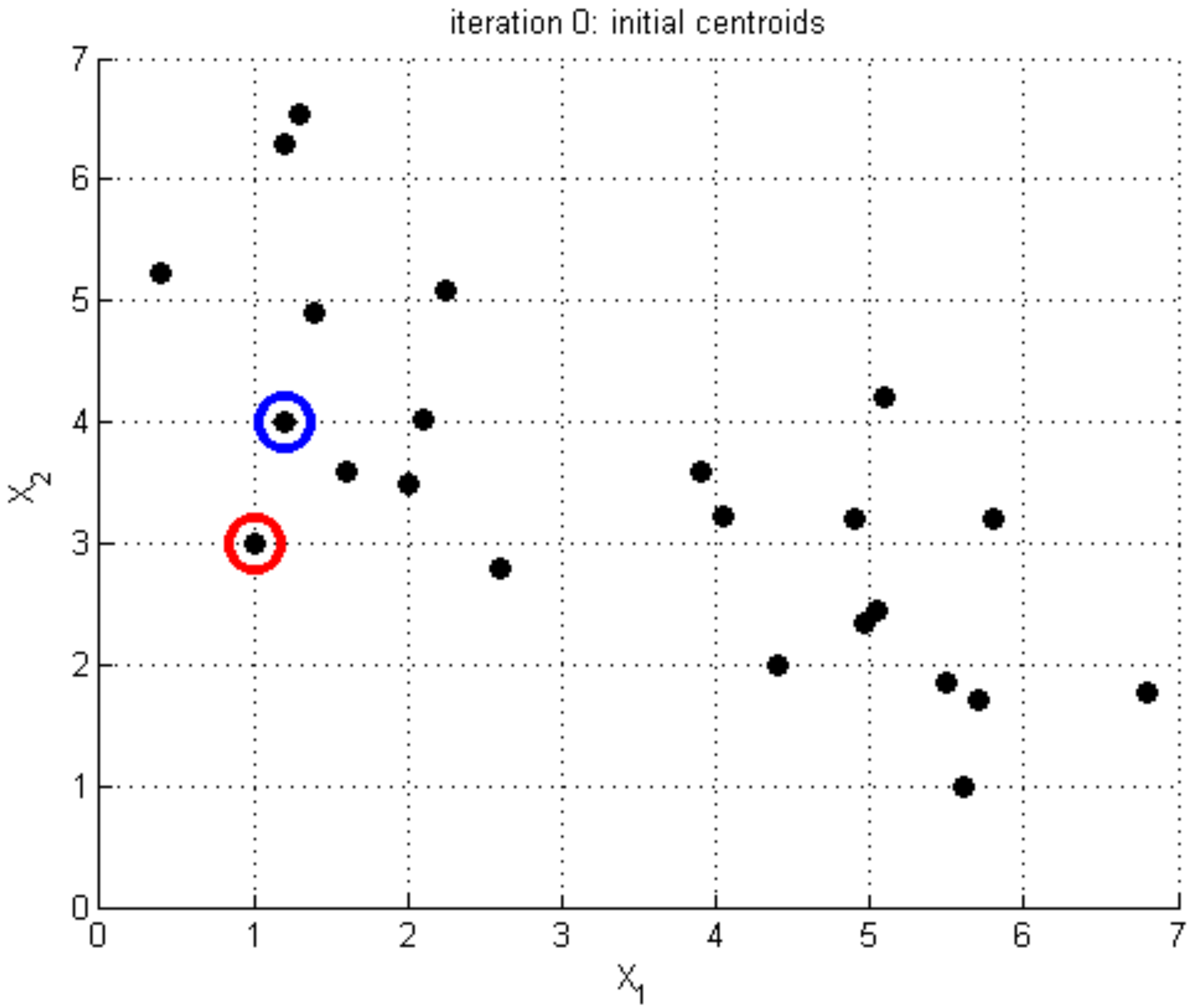
Mixture Models

K-means Clustering



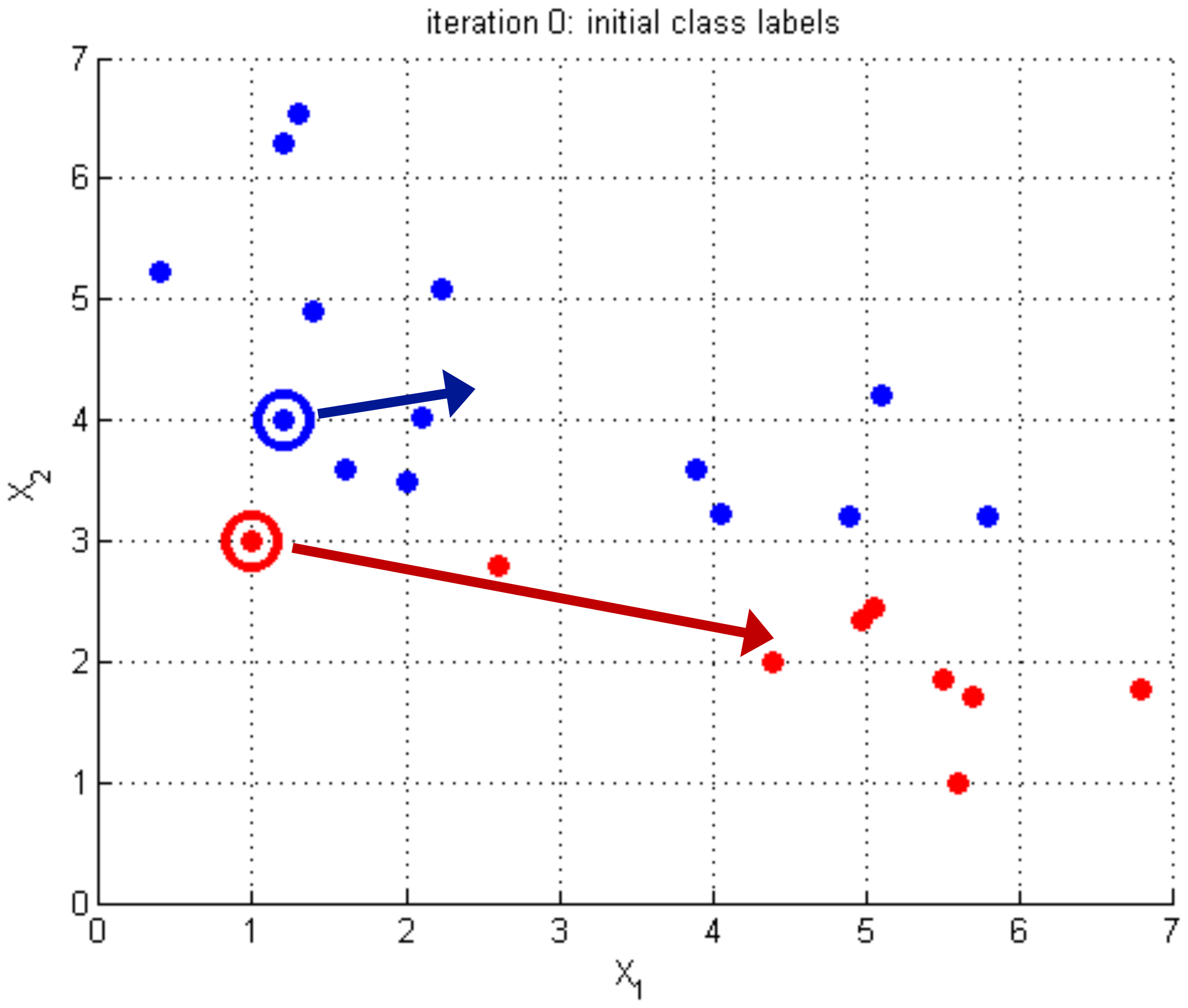
- Groups data into K distinct clusters
- Cluster defined by a centroid vector (mean of samples in cluster), each observation assigned to single cluster (nearest centroid)

K-means Algorithm



Pick initial centroids

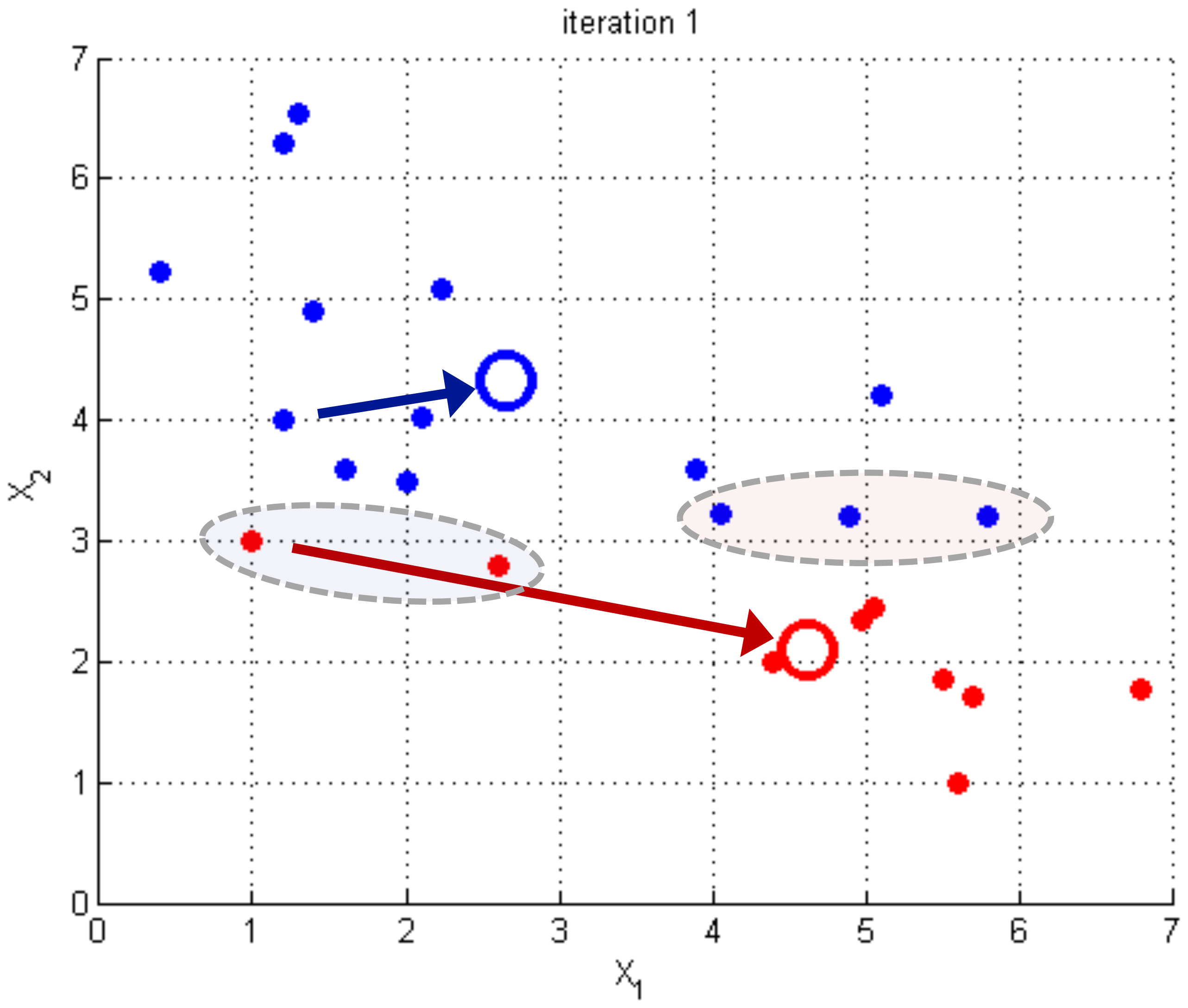
K-means Algorithm



Pick initial centroids

Assign initial clusters

K-means Algorithm

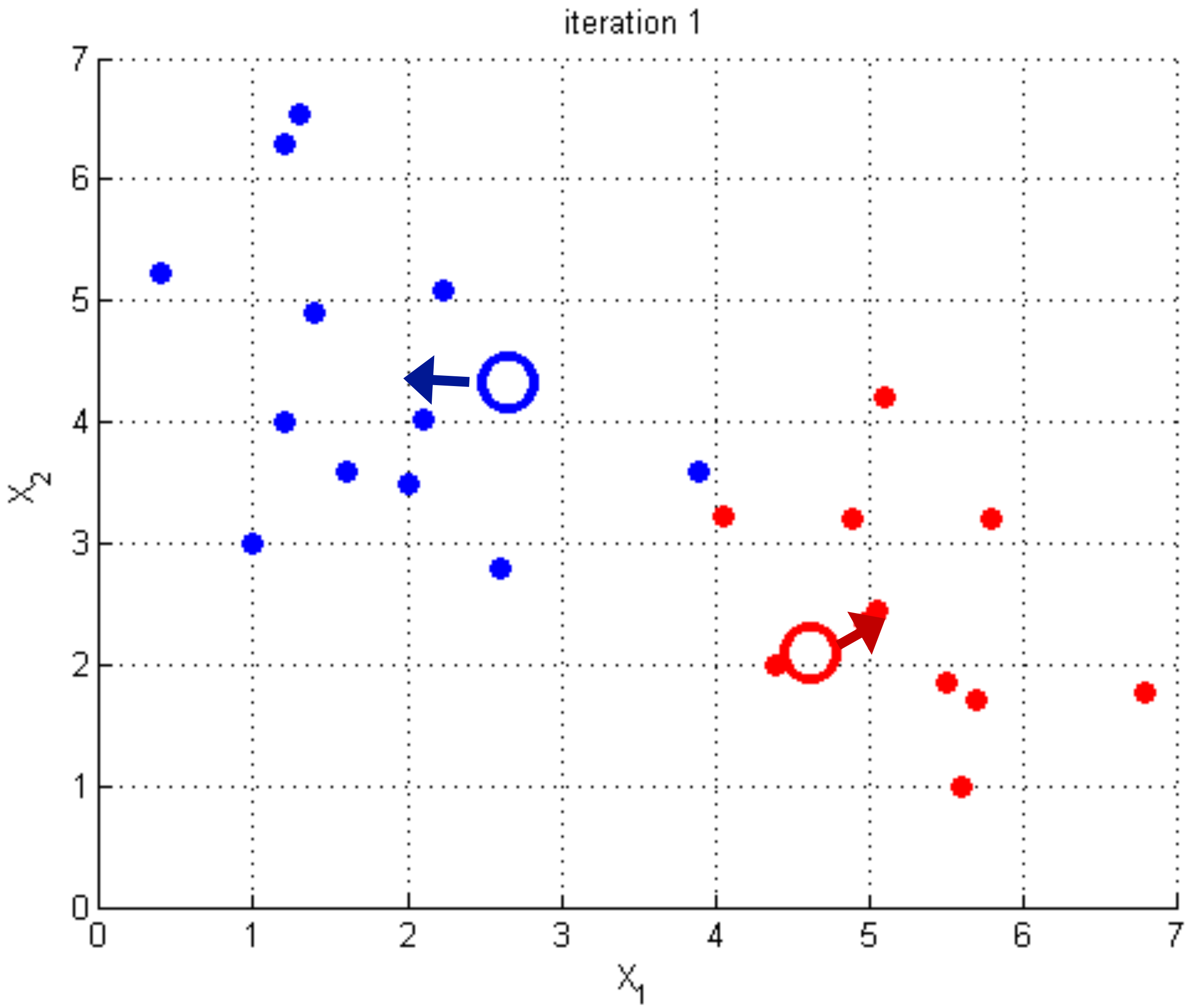


Pick initial centroids

Assign initial clusters

Update centroids

K-means Algorithm



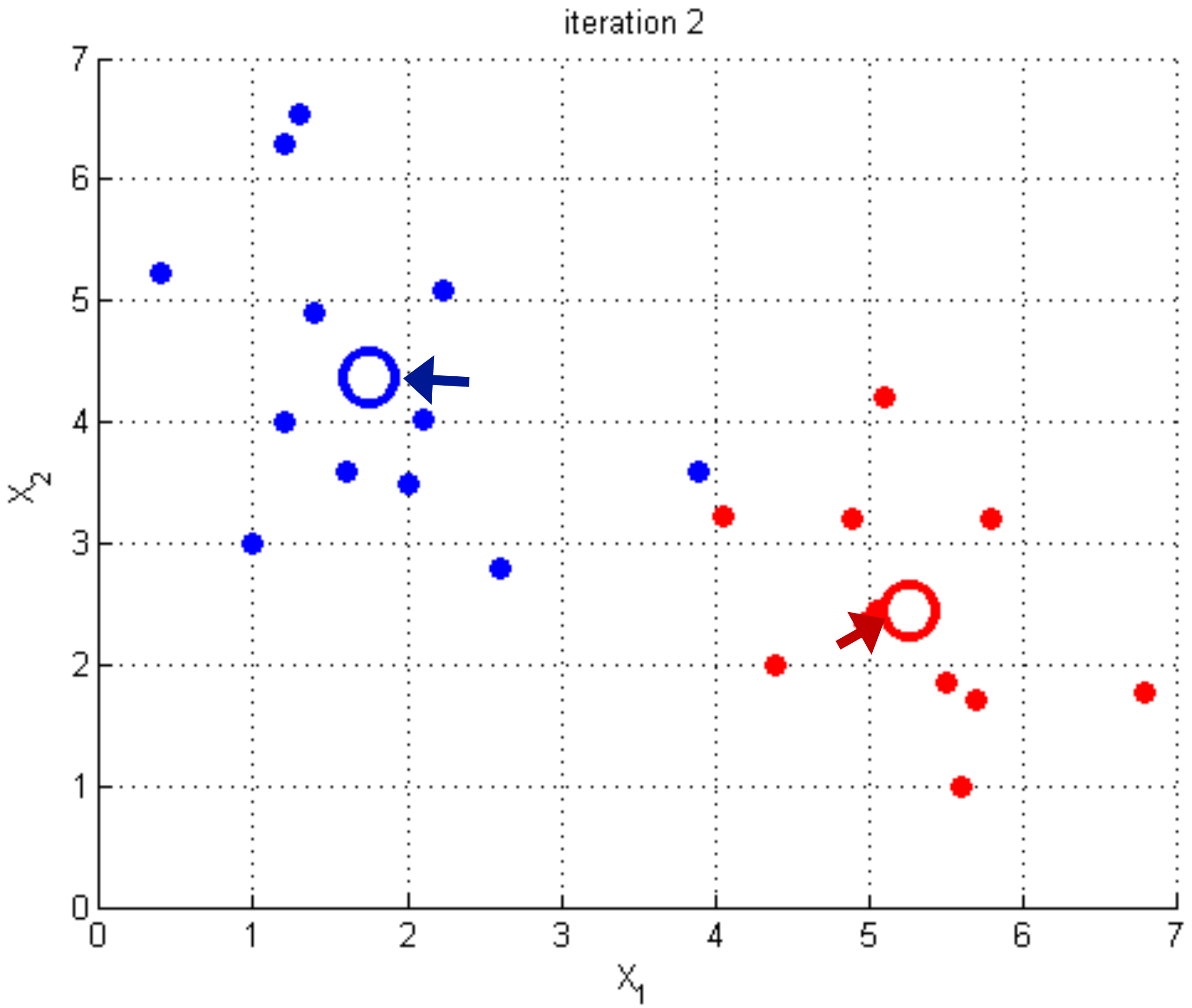
Pick initial centroids

Assign initial clusters

Update centroids

Reassign clusters

K-means Algorithm



Pick initial centroids

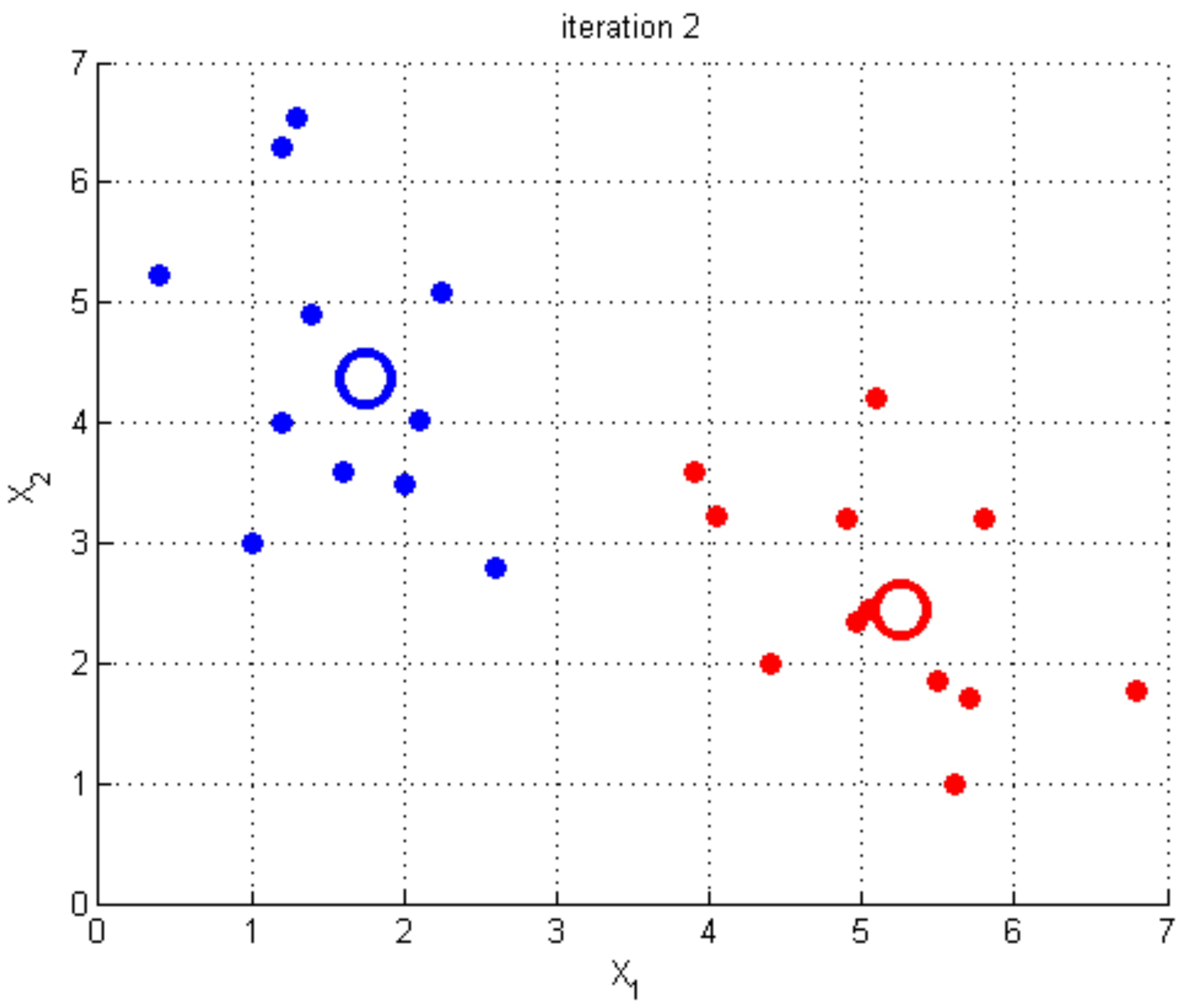
Assign initial clusters

Update centroids

Reassign clusters

Update centroids

K-means Algorithm



Pick initial centroids

Assign initial clusters

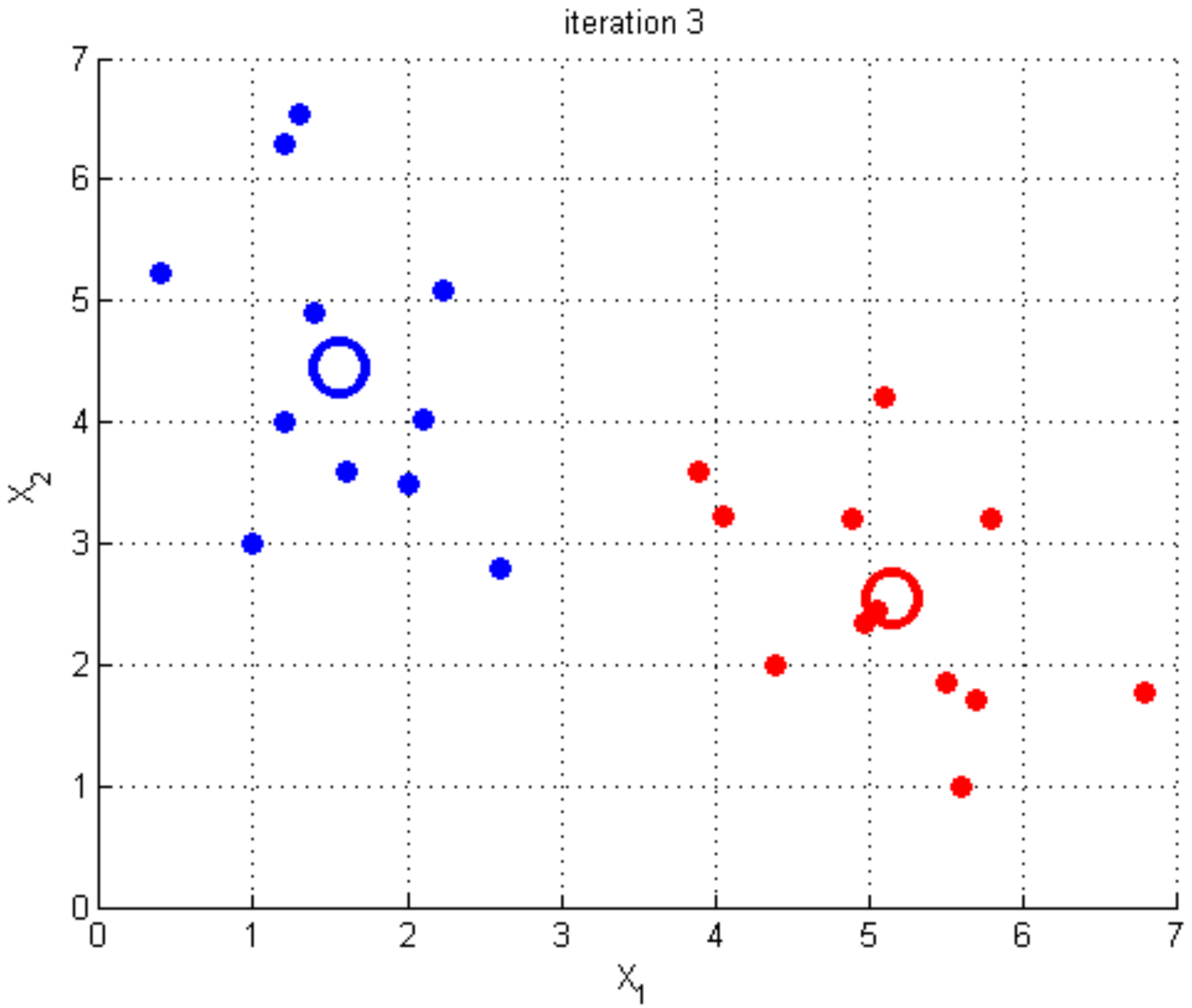
Update centroids

Reassign clusters

Update centroids

Reassign clusters

K-means Algorithm



Pick initial centroids

Assign initial clusters

Update centroids

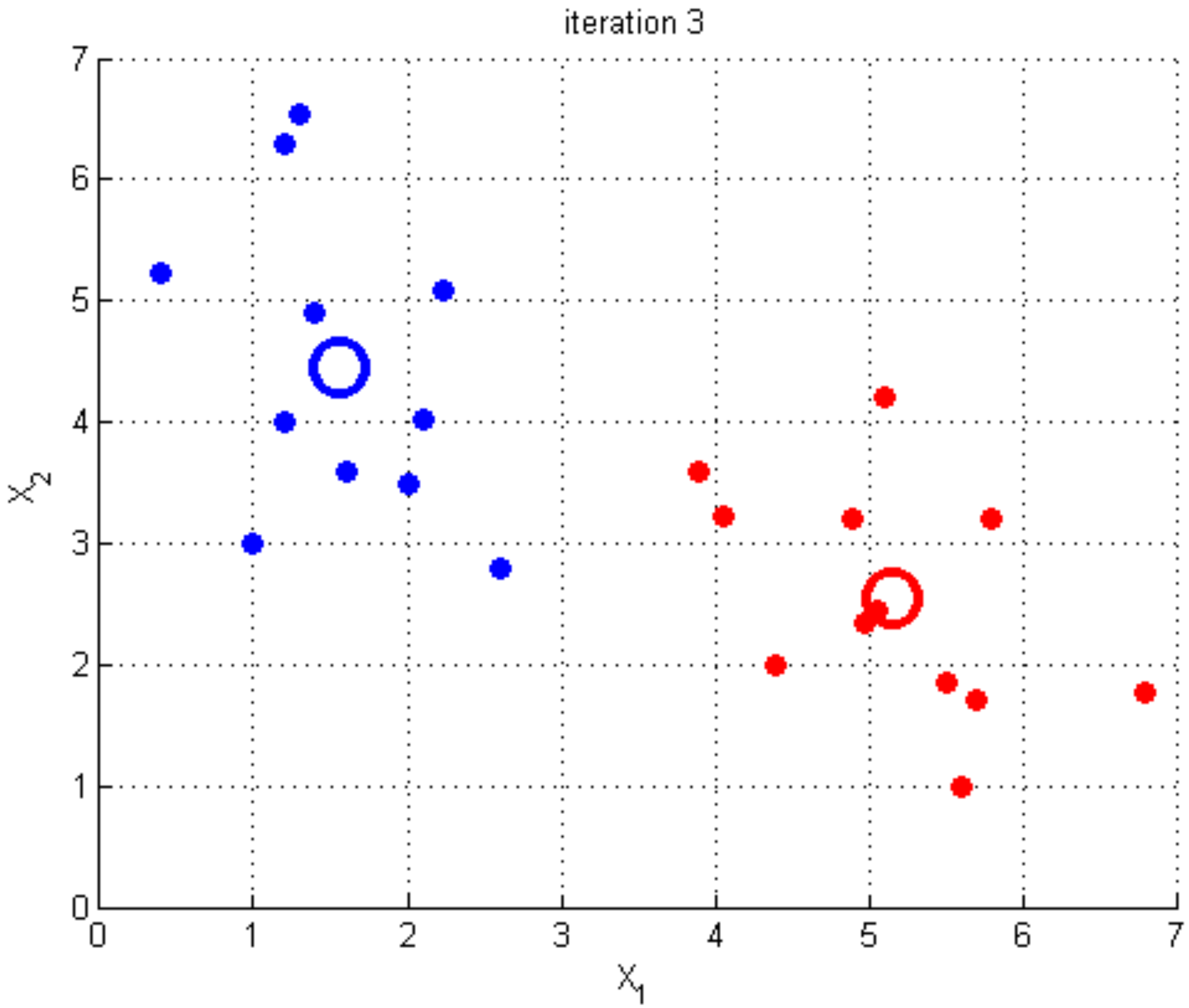
Reassign clusters

Update centroids

Reassign clusters

Update centroids

K-means Algorithm



Pick initial centroids

Assign initial clusters

Update centroids

Reassign clusters

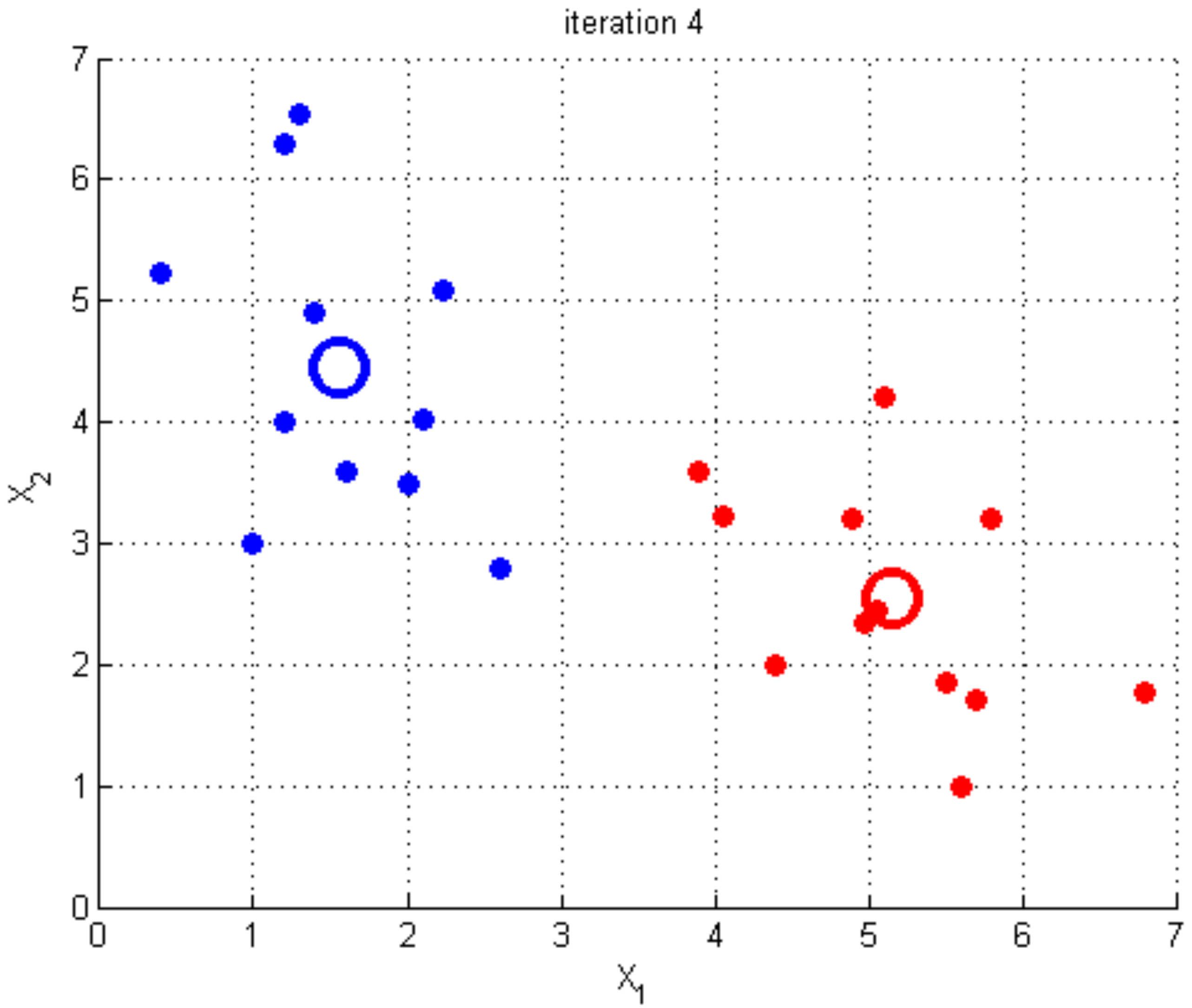
Update centroids

Reassign clusters

Update centroids

Reassign clusters

K-means Algorithm



Pick initial centroids

Assign initial clusters

Update centroids

Reassign clusters

Update centroids

Reassign clusters

Update centroids

Reassign clusters

Converged

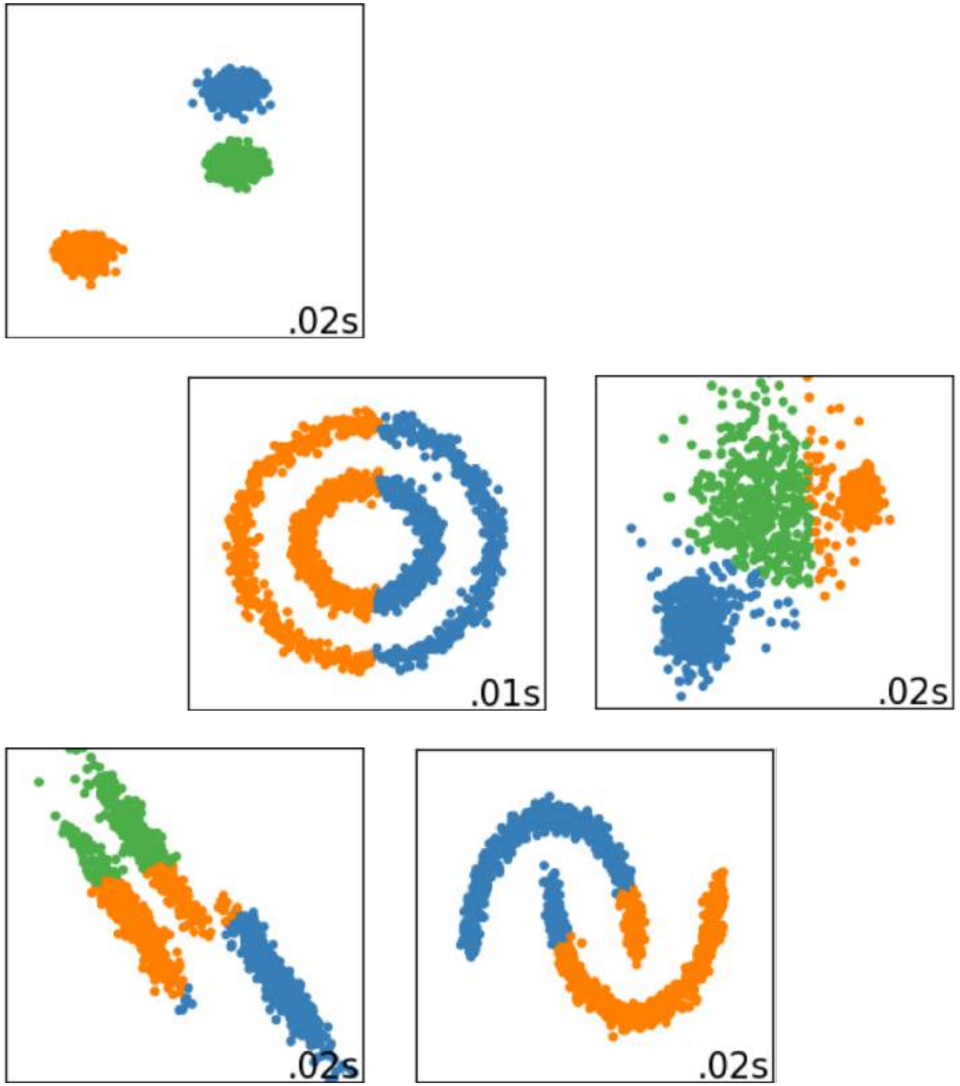
K-means clustering

Advantages

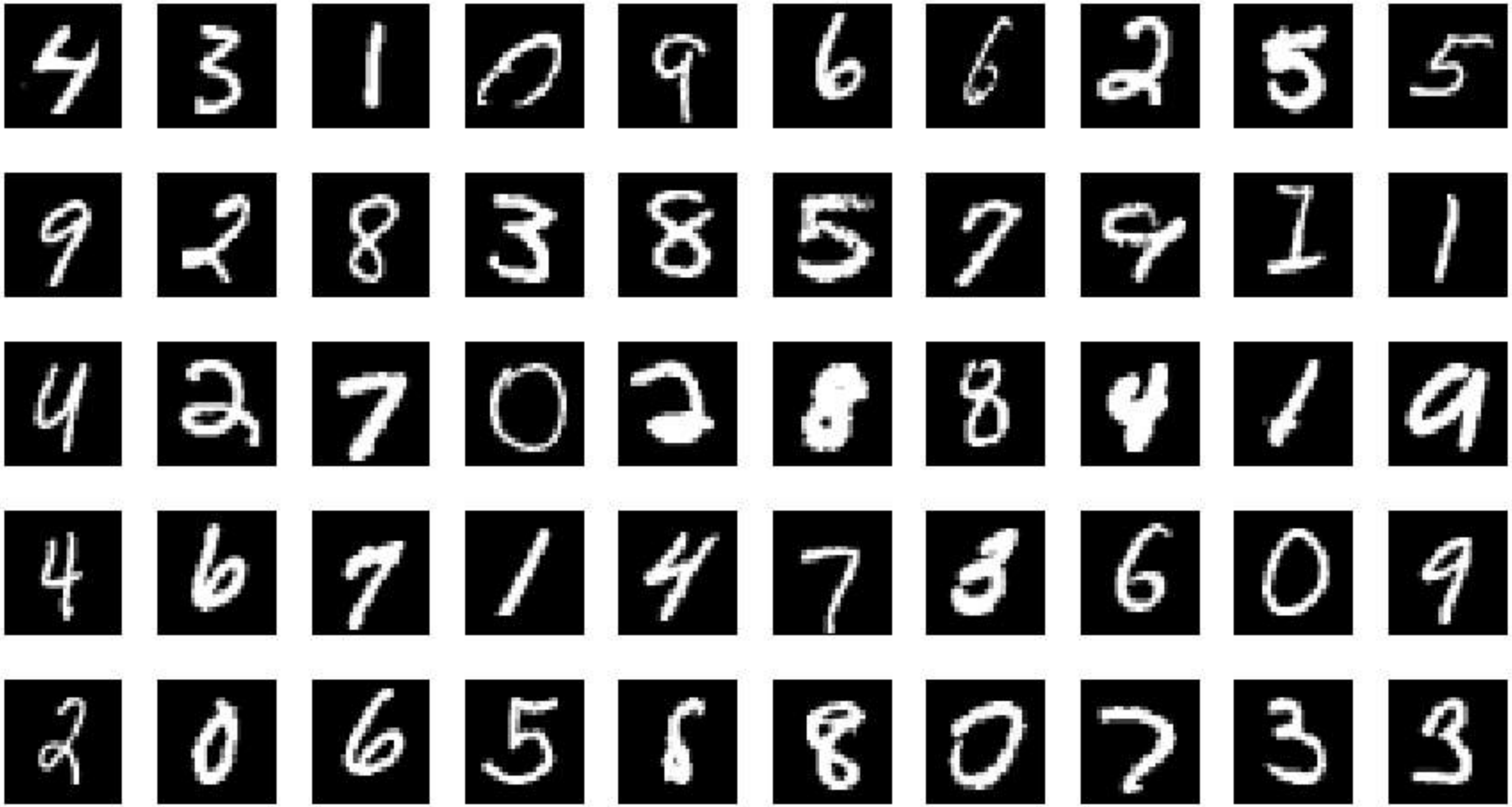
- Easy to implement
- Converges quickly (few iterations)
- Scales better than hierarchical clustering

Disadvantages

- # clusters must be specified
- (Hyper-)spherical, similar-sized clusters
- Sensitive to outliers in data
- Sensitive to initialization of centroids



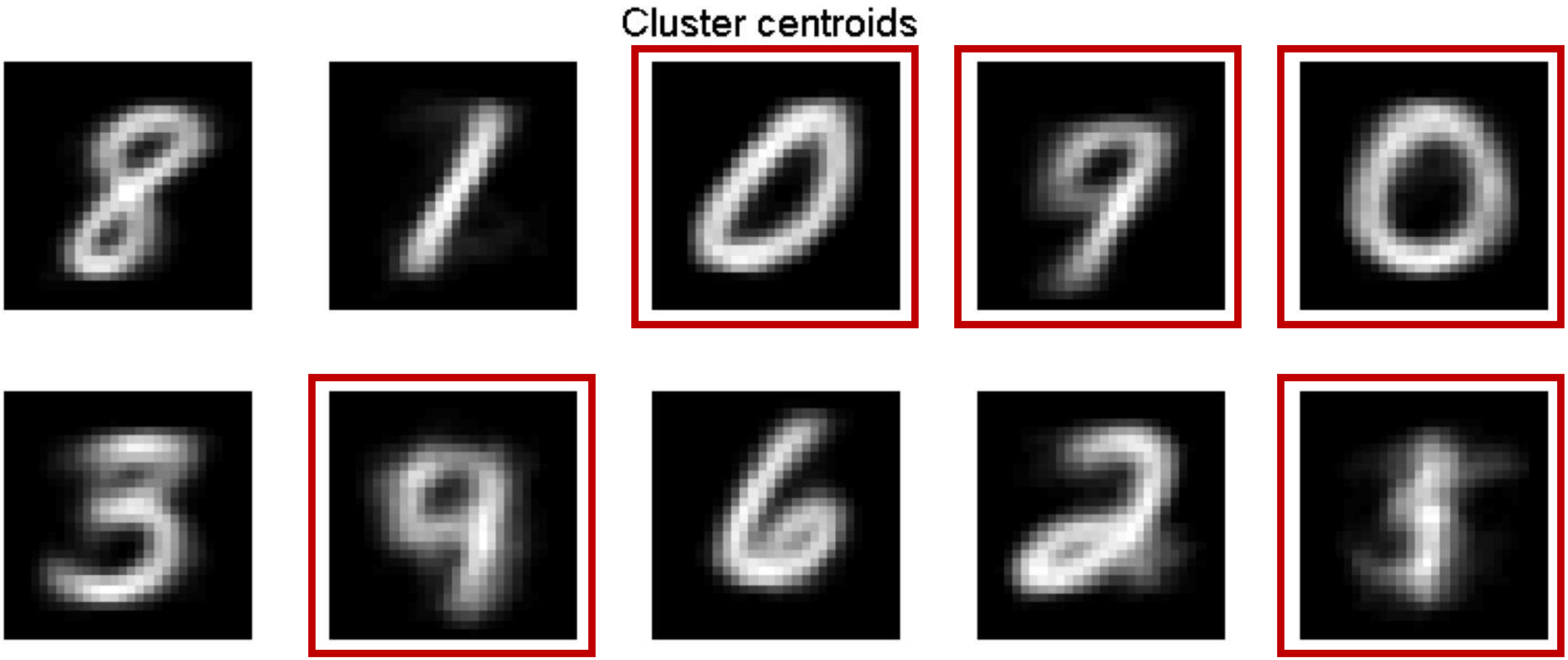
Handwritten digit clustering



MNIST dataset: http://yann.lecun.com/exdb/mnist/

Handwritten digits: cluster centroids

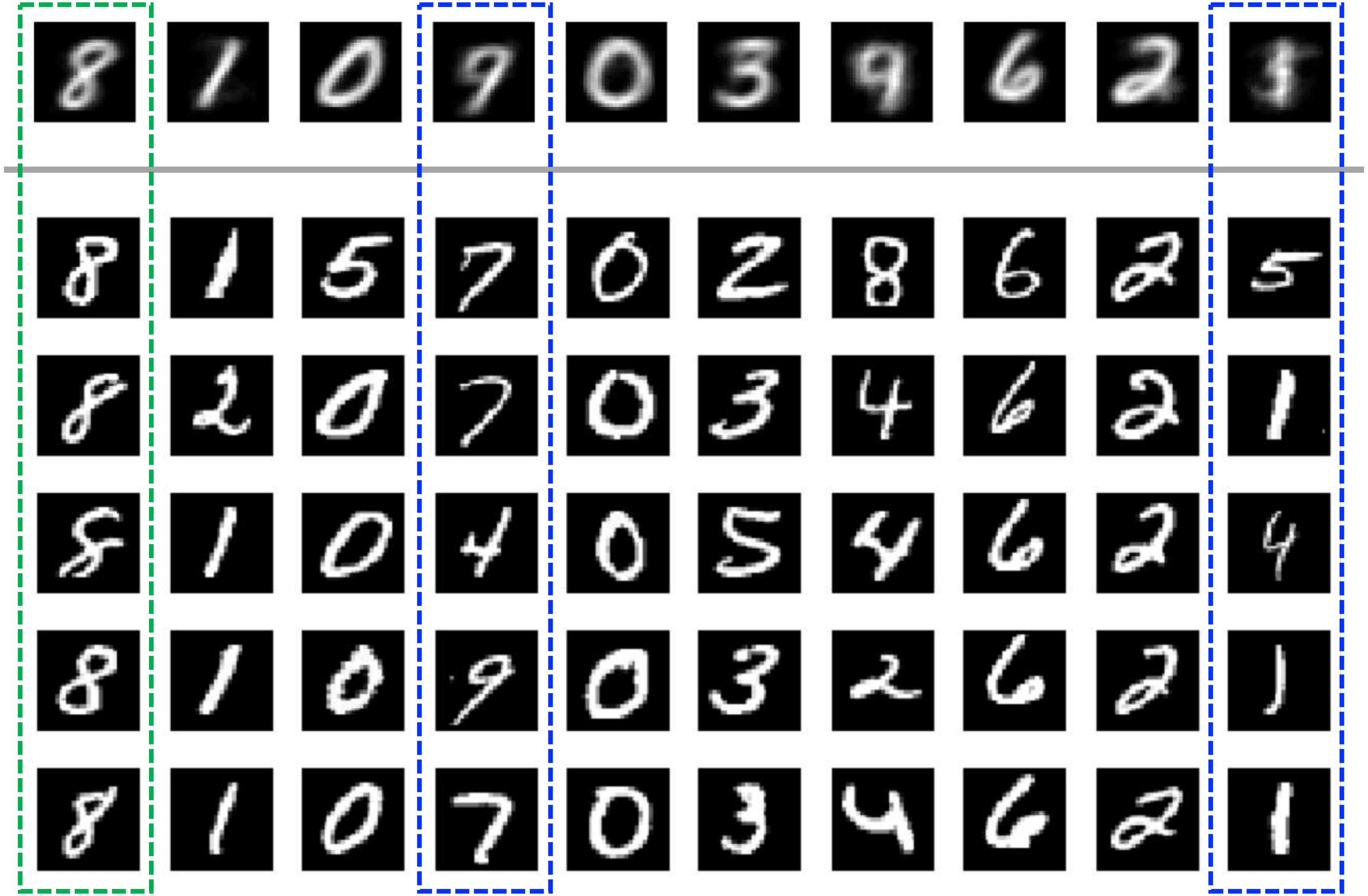
Apply K-means to find K=10 clusters:



Handwritten digits: visualizing clusters

Cluster centroid:

Sample of digits assigned to cluster:



Types of clustering algorithms

Hierarchical (Agglomerative) Clustering

Centroid-based Clustering

Spectral Methods

Density-based Clustering

Mixture Models

Spectral Methods

Encodes local neighborhoods in similarity graphs – clustering using graph cuts

Density-based Clustering

Identify high-density regions in feature space separated by low-density regions

Mixture Models

Each cluster represented by parametric distribution – probabilistic (soft) clusters

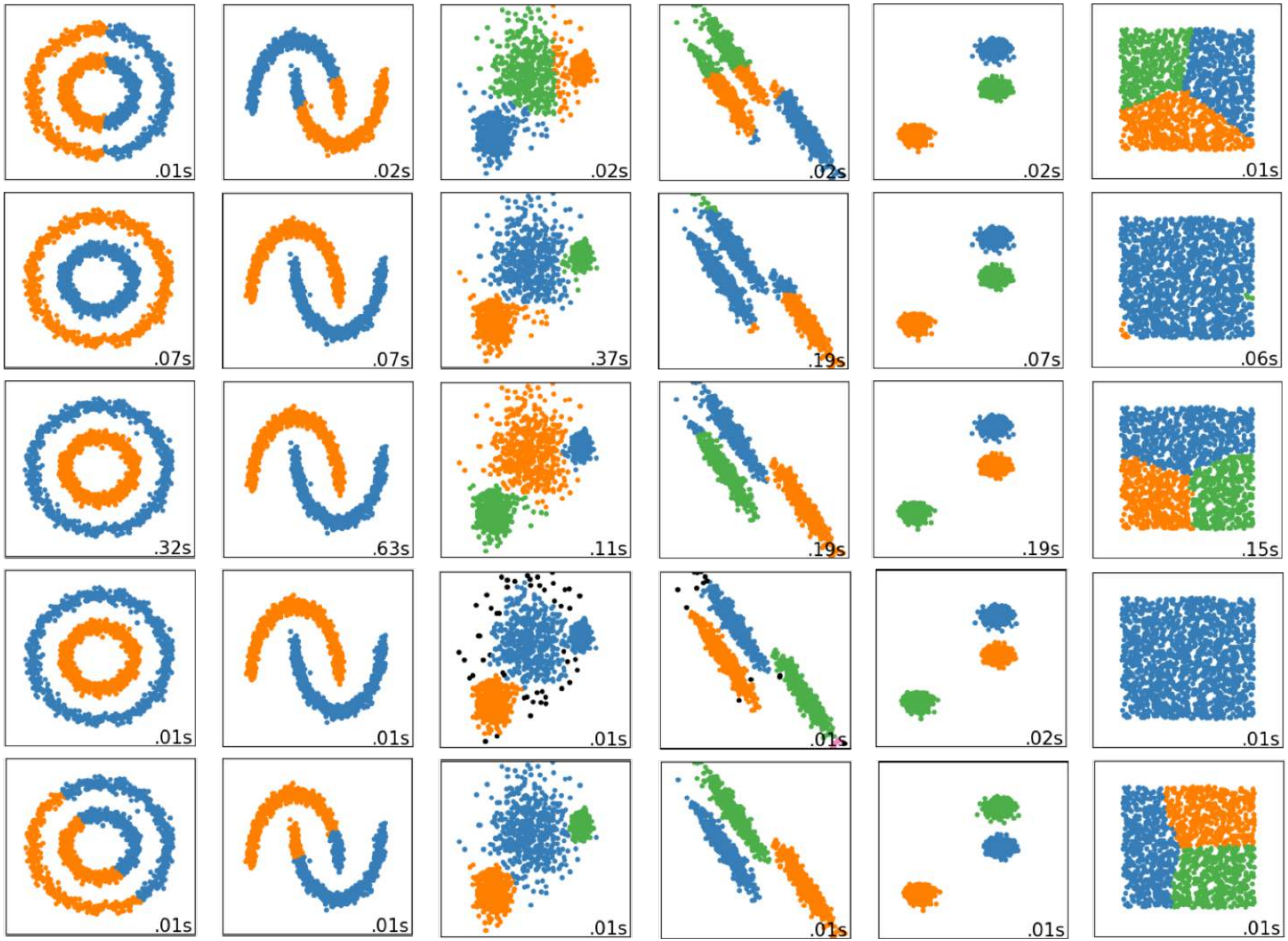
Centroid-based
(K-means)

Hierarchical
Clustering

Spectral
Clustering

Density-based
(DBSCAN)

Mixture Model
(GMM)

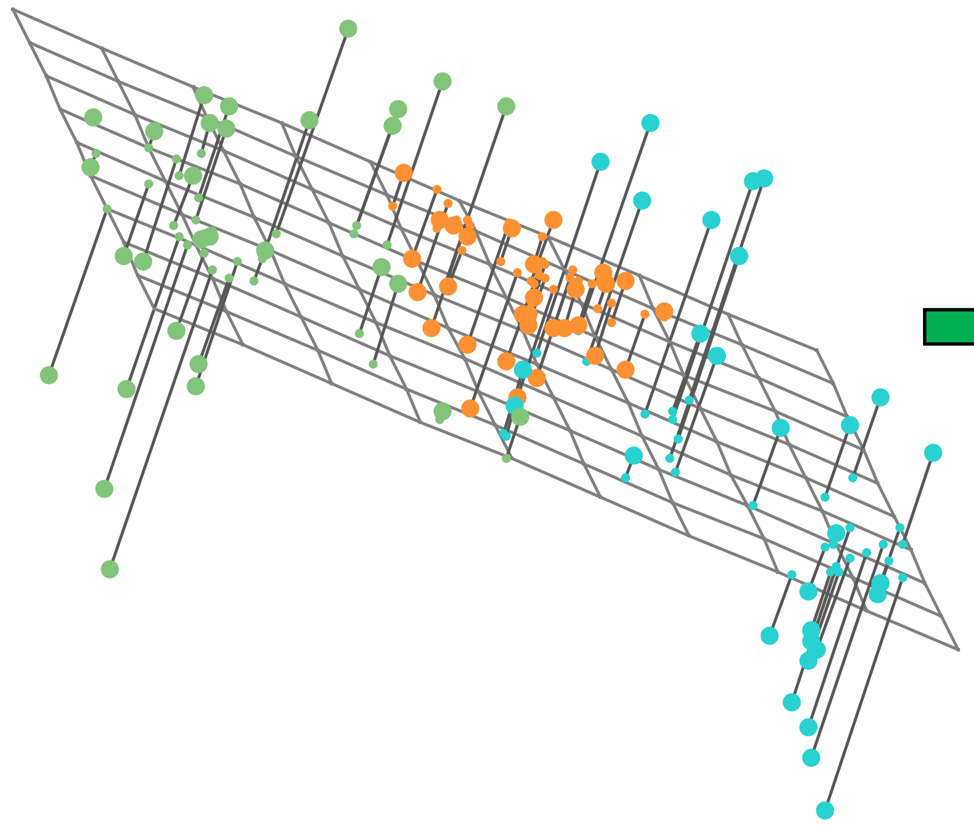


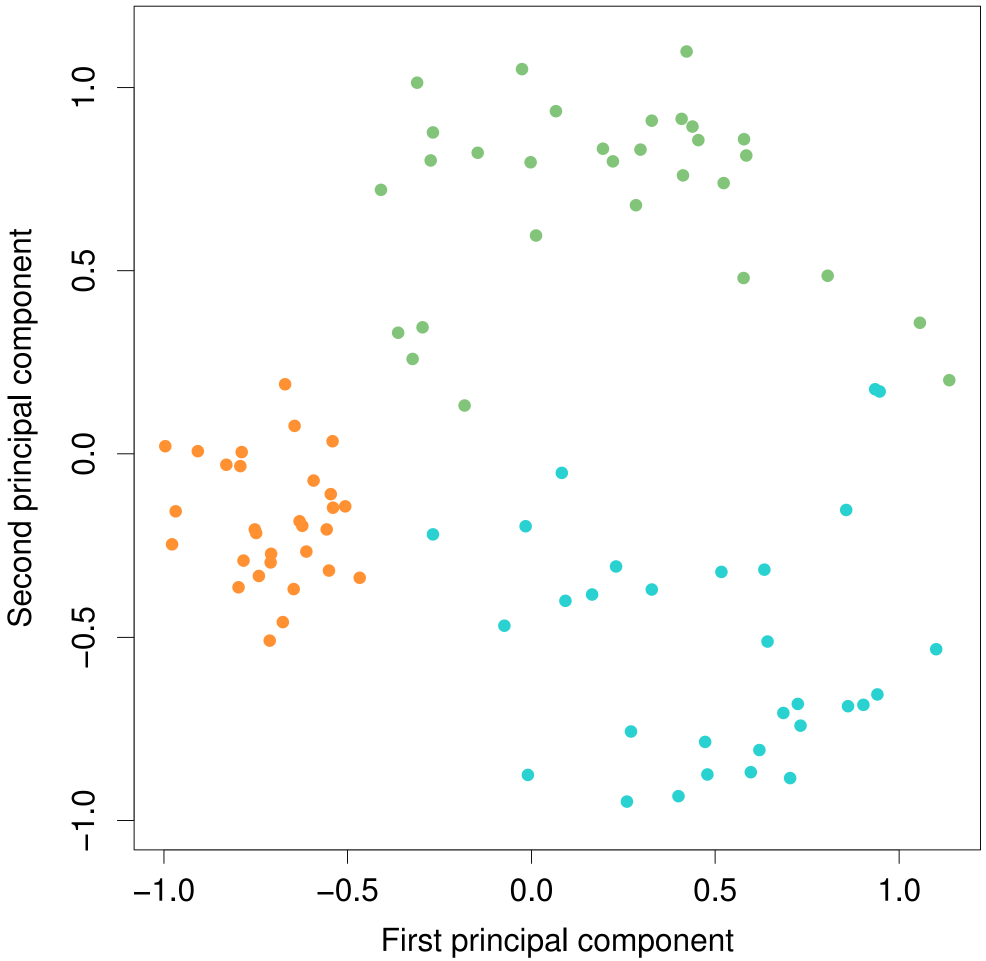
Dimensionality Reduction / Feature Learning

Linear Methods

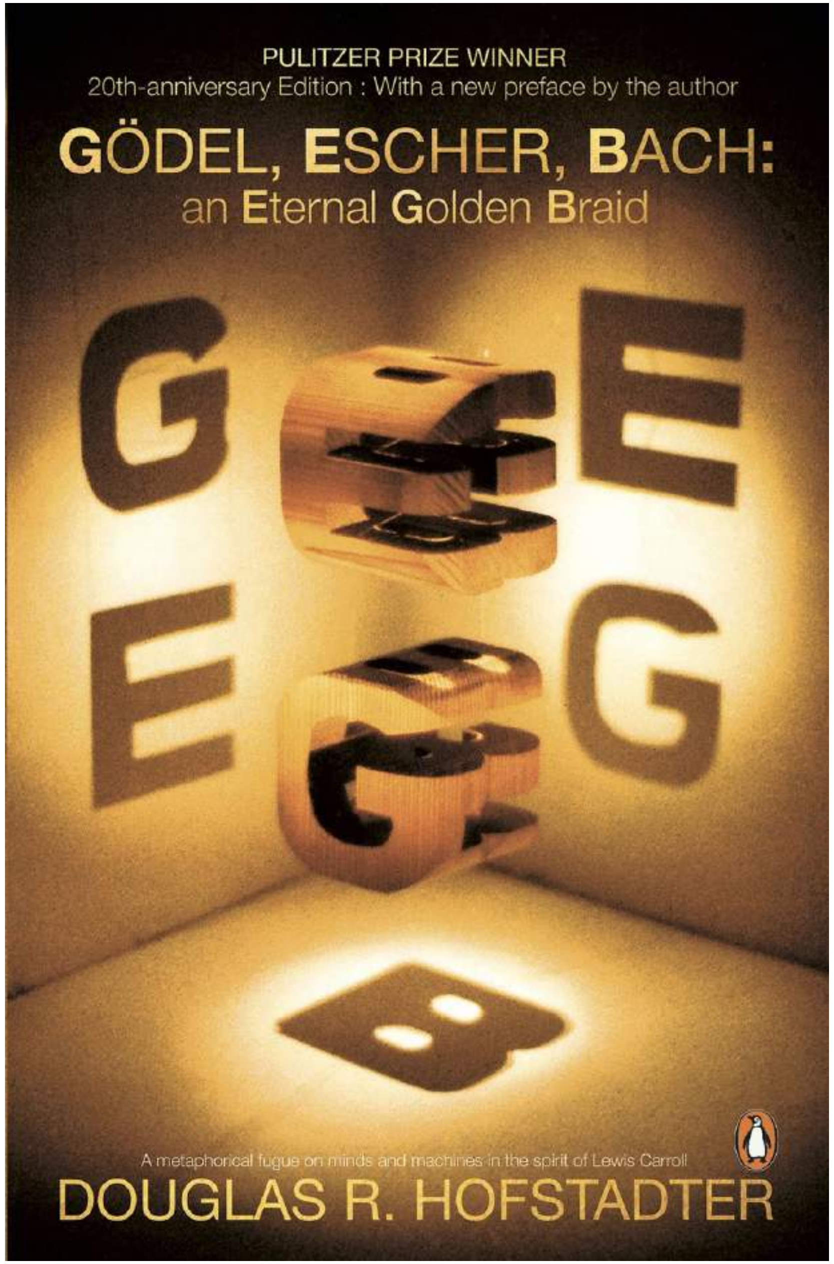
Dimensionality Reduction

Goal: *Find a linear transformation to lower-dimensional feature space that preserves the key characteristics of the original (high-dimensional) data.*





Projections



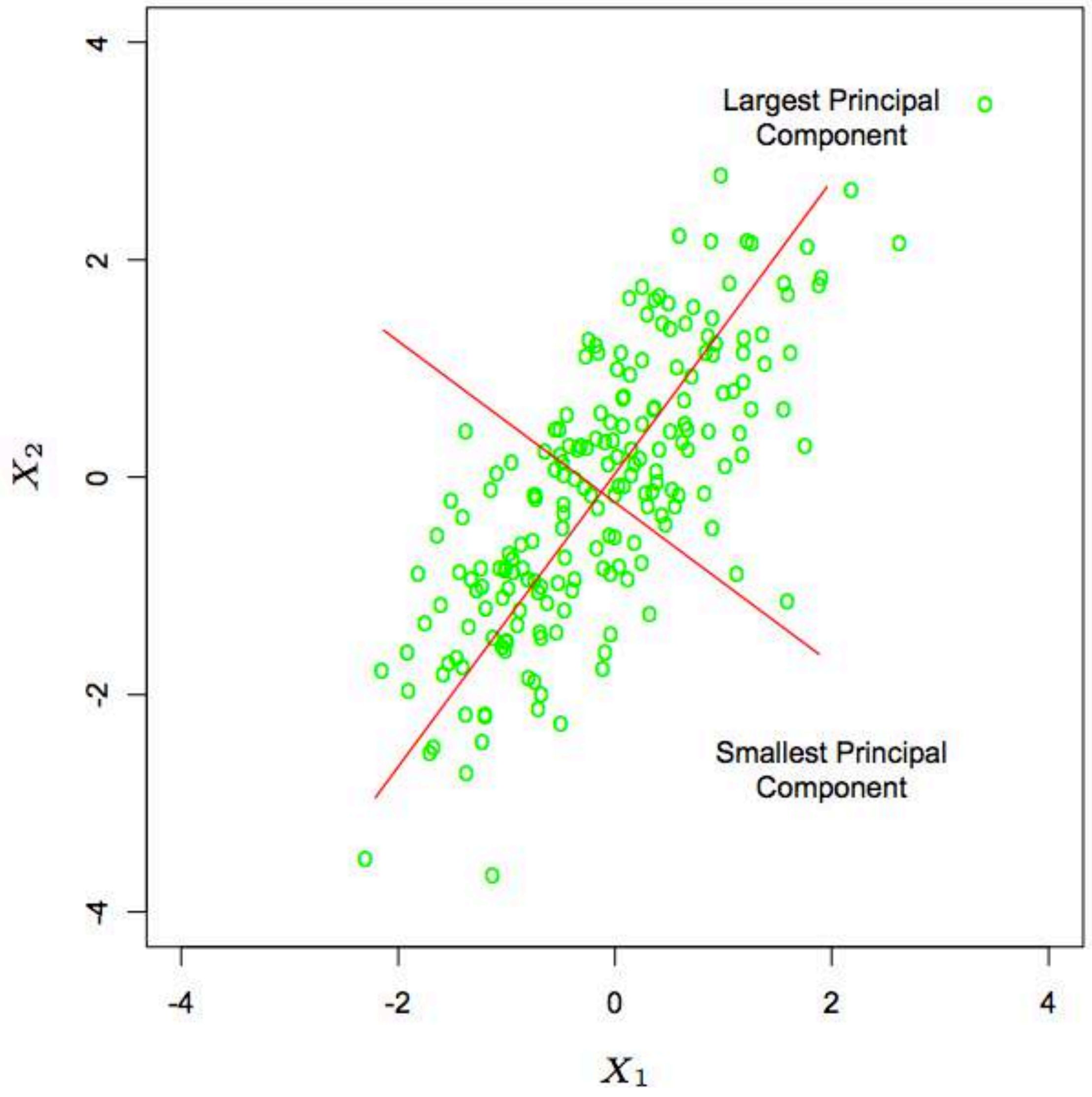
PCA: Maximal Variance Projection

What is principal component analysis?

Projection to lower dimensional feature space that captures the most variance in the data (orthogonal directions).

Principal components are linear combinations of original features.

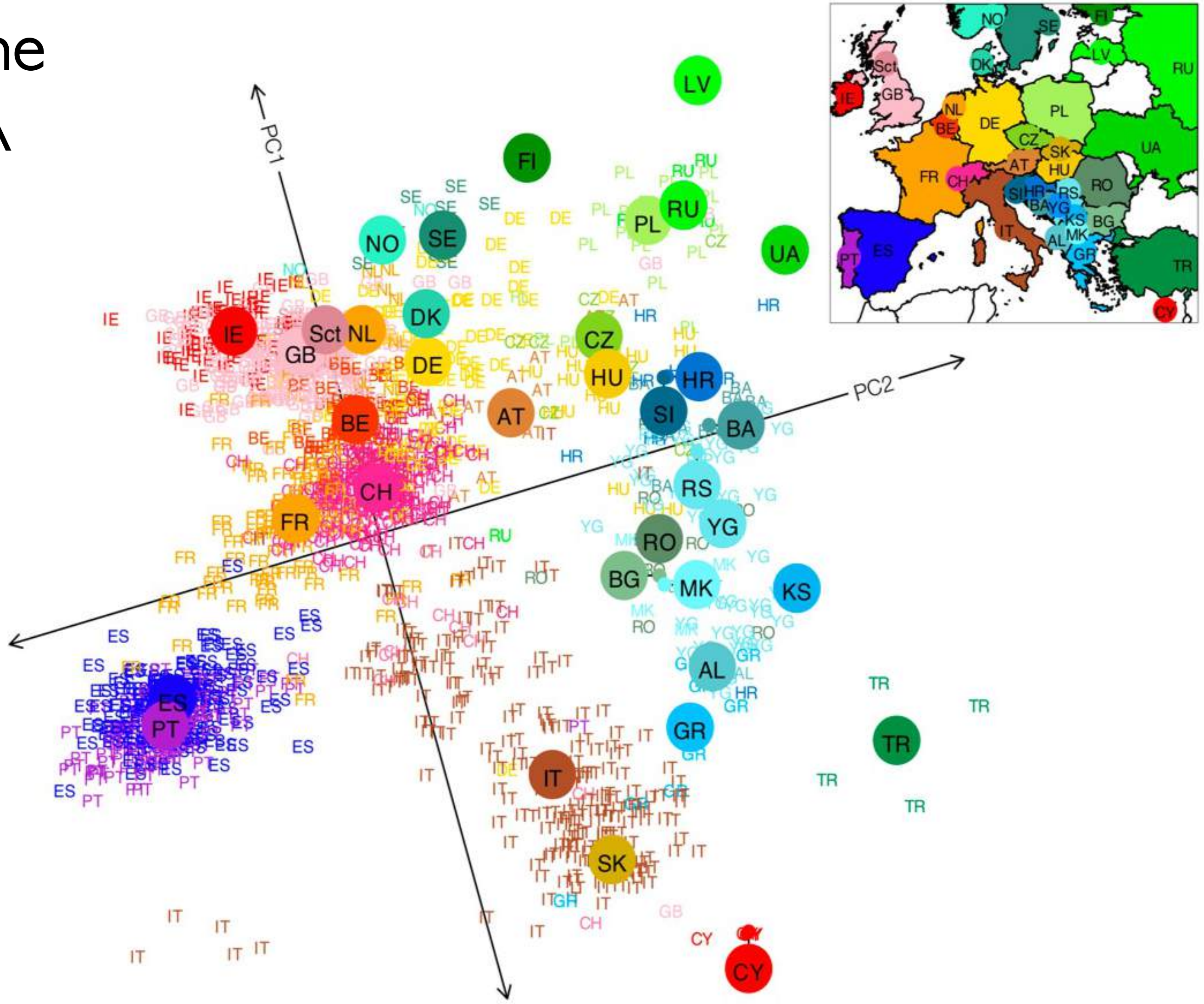
Principal components are eigenvectors of covariance matrix.



Example: PCA for high-dimensional data

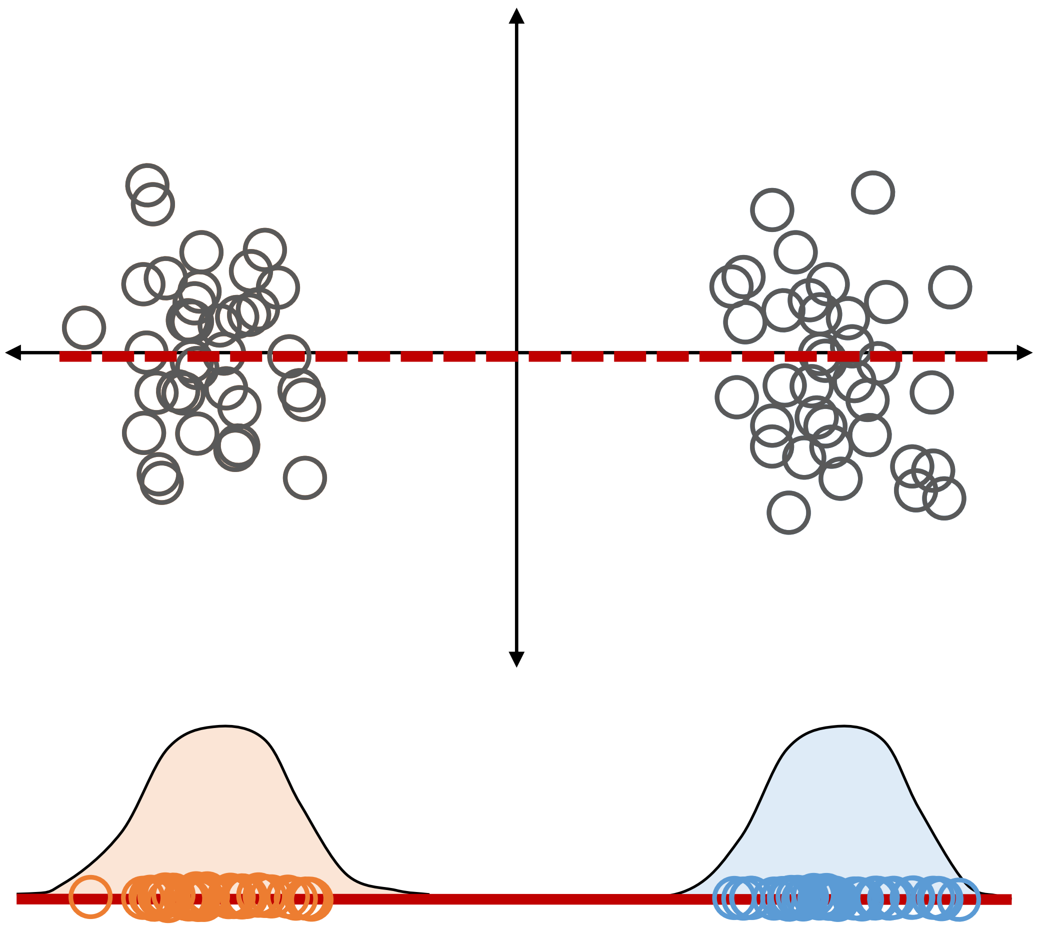
500,000 DNA sites in human genome projected to 2 dimensions with PCA

Principal components correspond to geography → ancestry

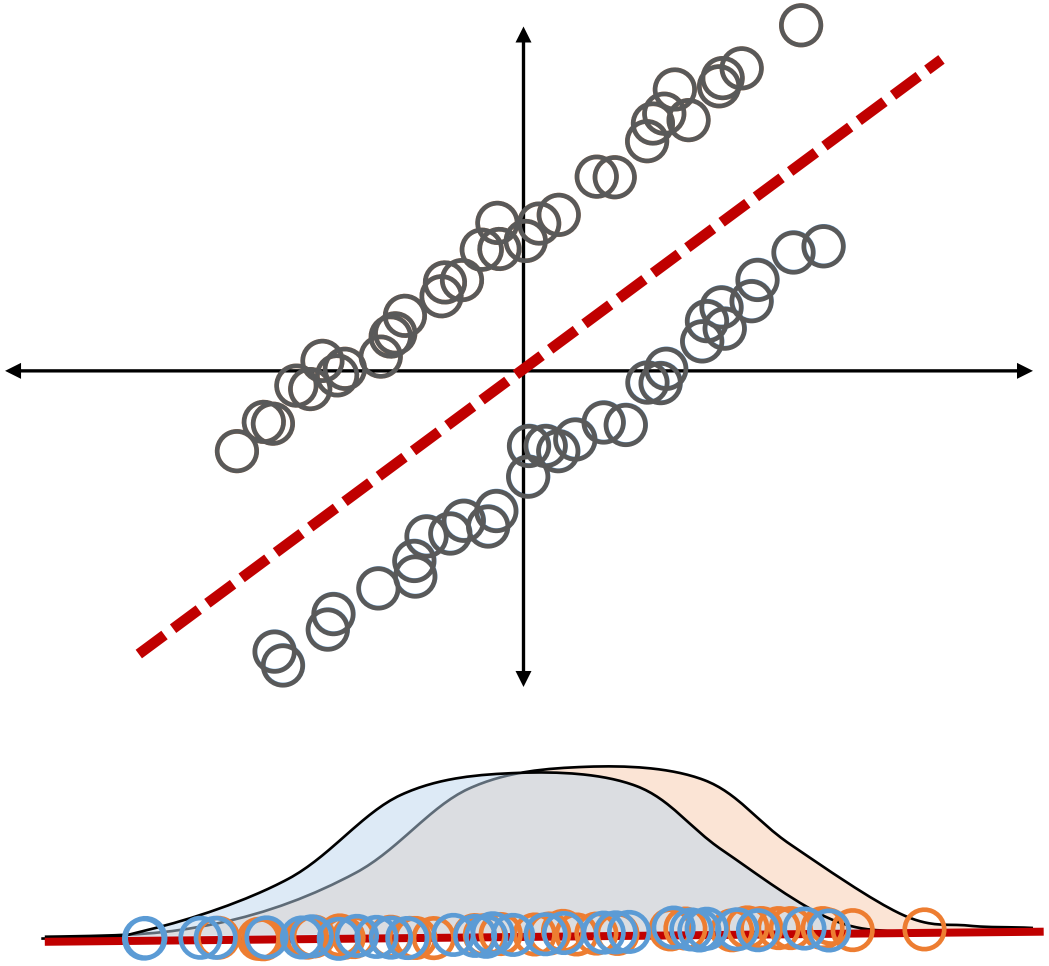


Novembre et al. (2008), Nature

PCA does not always give the “best” projection.



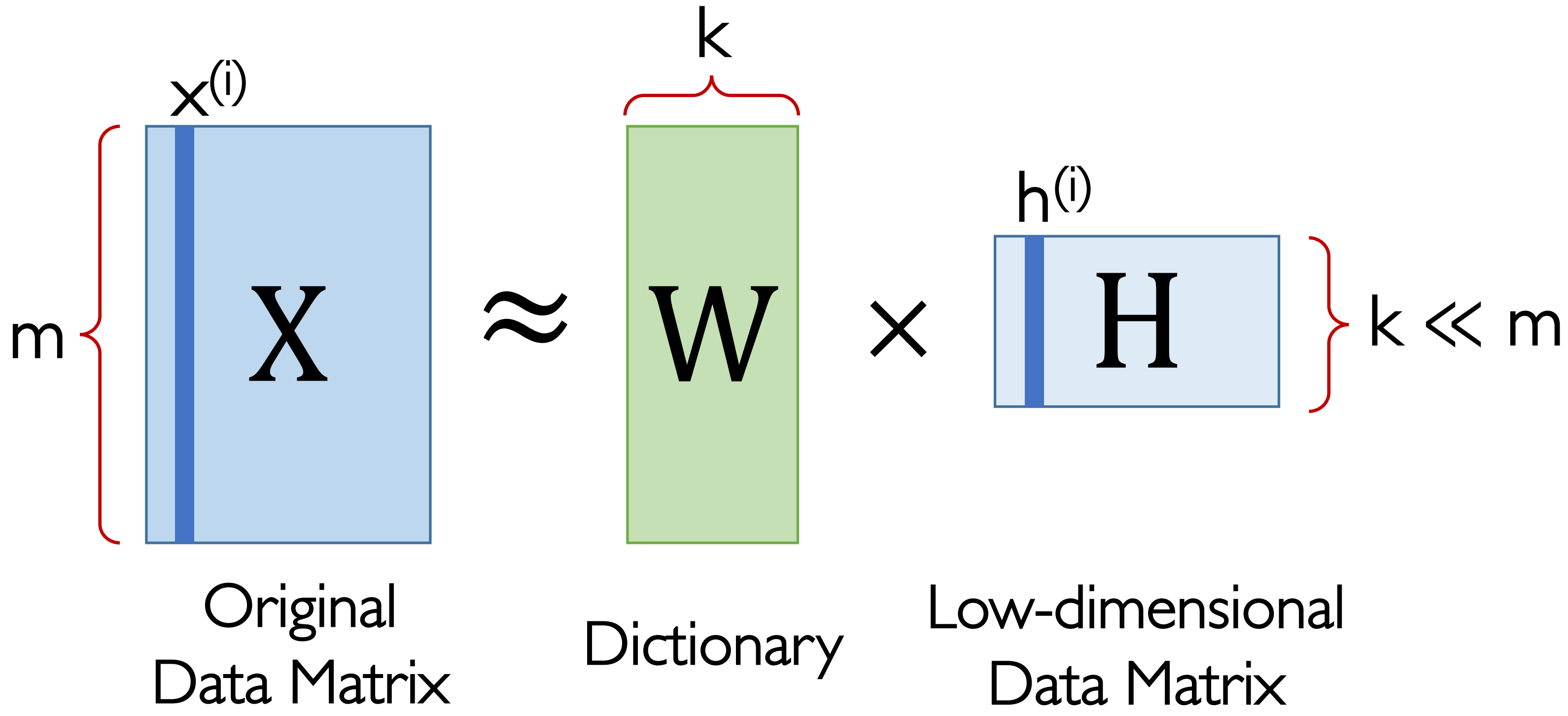
First PC finds clusters



First PC misses clusters

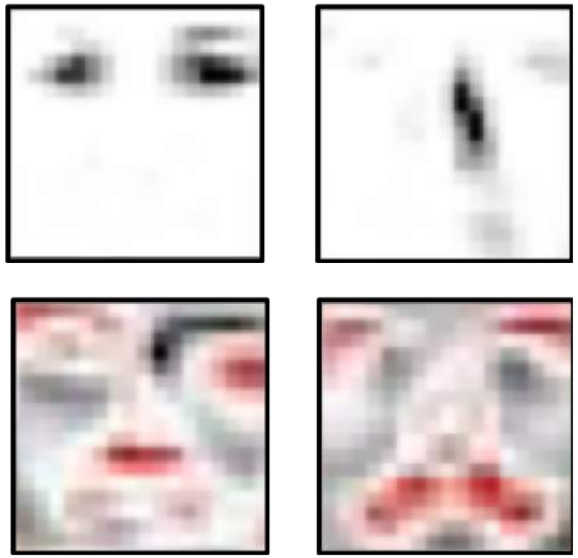
Non-negative Matrix Factorization (NMF)

Data approximated by positive linear combination of k vectors containing only non-negative values → k-dimensional representation

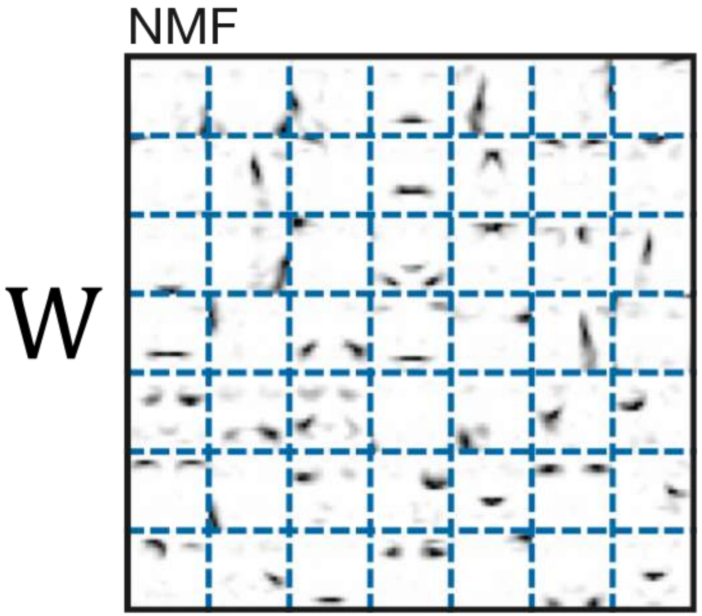


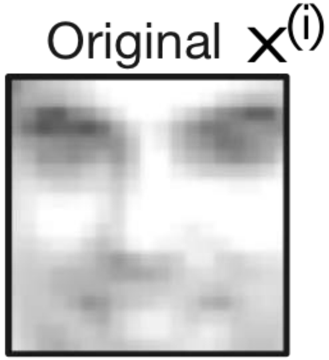
$$X, W, H \geq 0$$

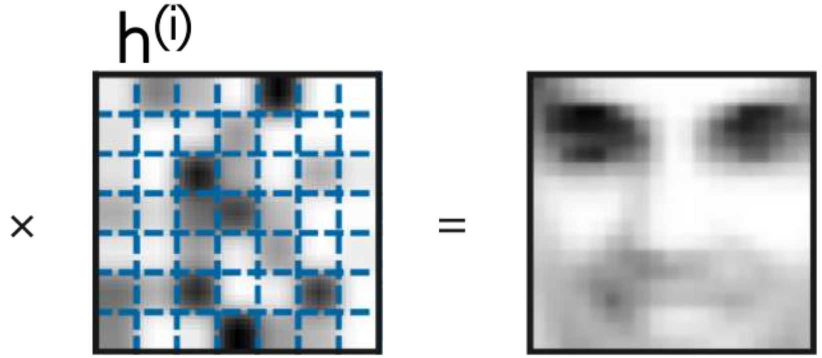
Non-negative constraint → sparsity, interpretability



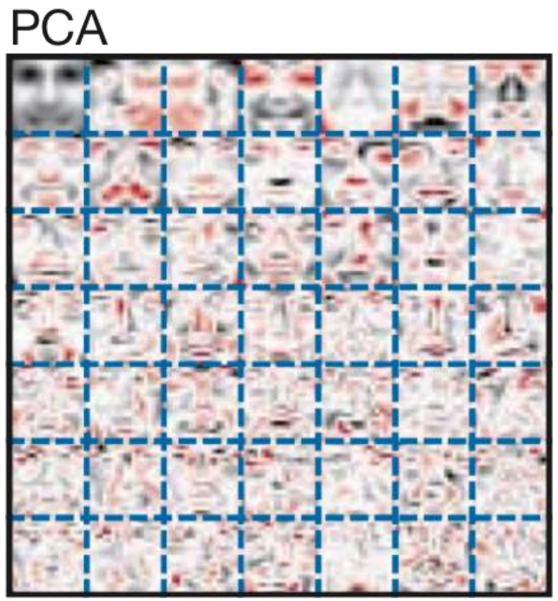
Parts-based representation (*sparse*)



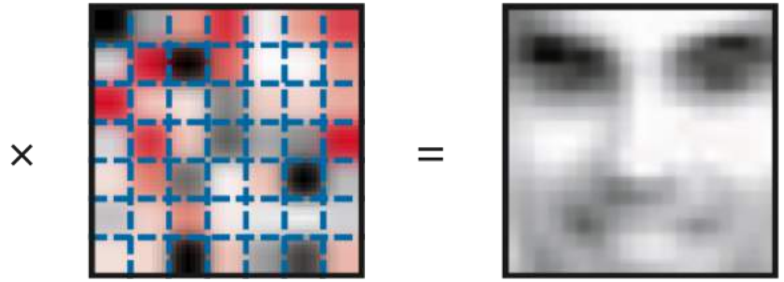




Each face approximated as sum of facial elements



PCs dense, less interpretable

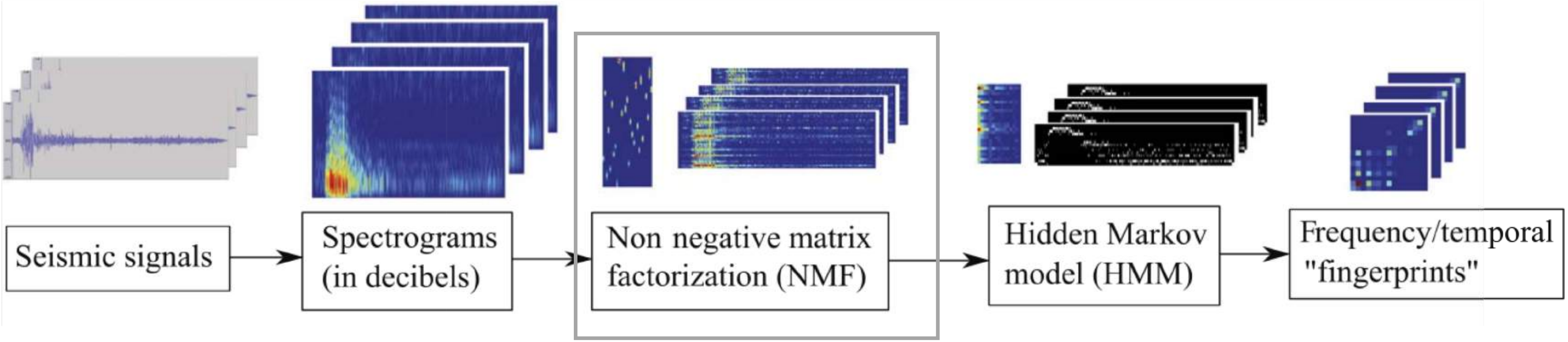


Lee & Seung (1999), Nature

Holtzman et al. (2018)

Geoscience Example 1:

NMF and K-means to characterize seismic source properties

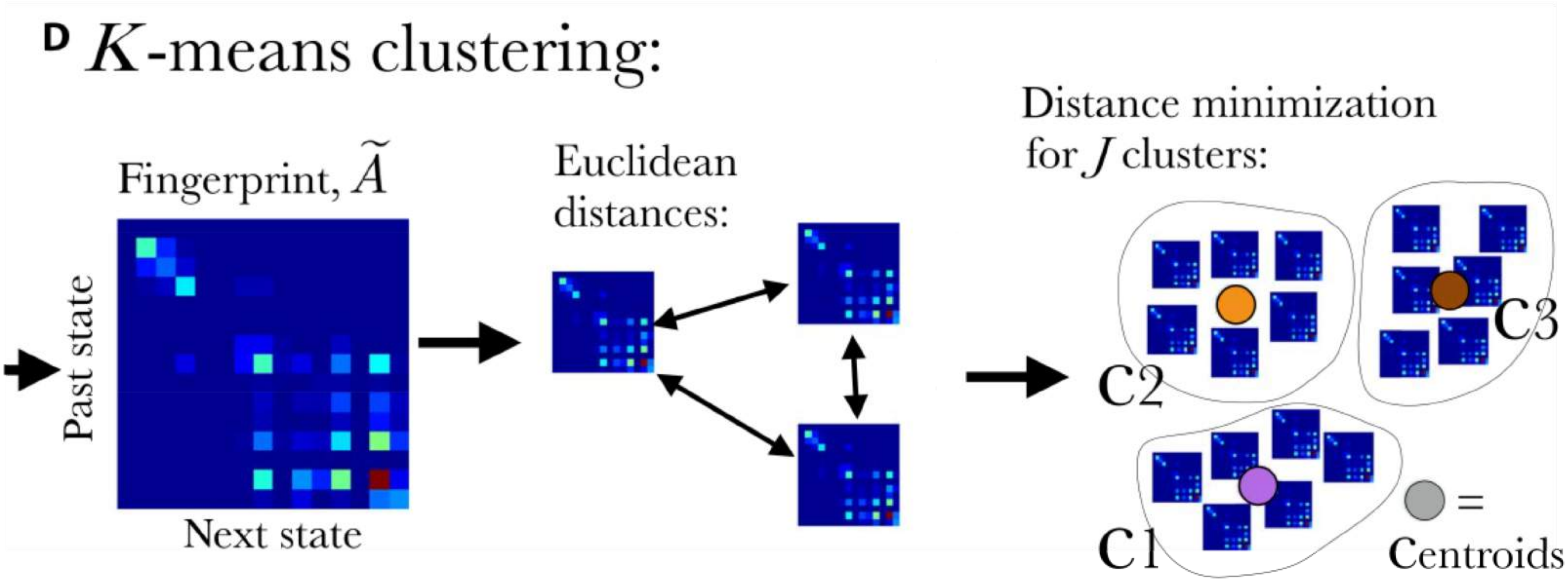


1) Learn feature representation with NMF and Hidden Markov Model
2) Cluster 46,000 earthquakes in Geysers geothermal field

Geoscience Example 1:

Holtzman et al. (2018)

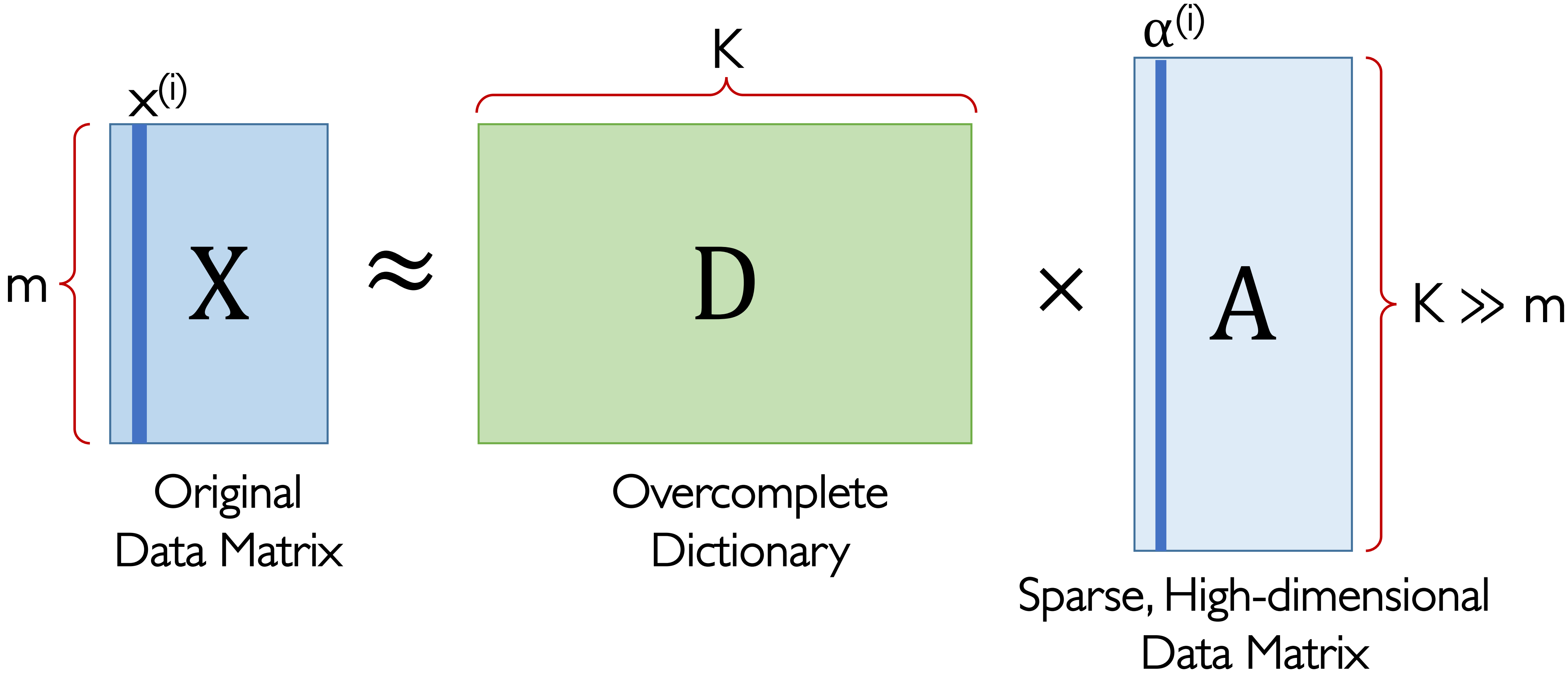
NMF and K-means to characterize seismic source properties



1) Learn feature representation with NMF and Hidden Markov Model
2) Cluster 46,000 earthquakes in Geysers geothermal field

Dictionary Learning & Sparse Coding

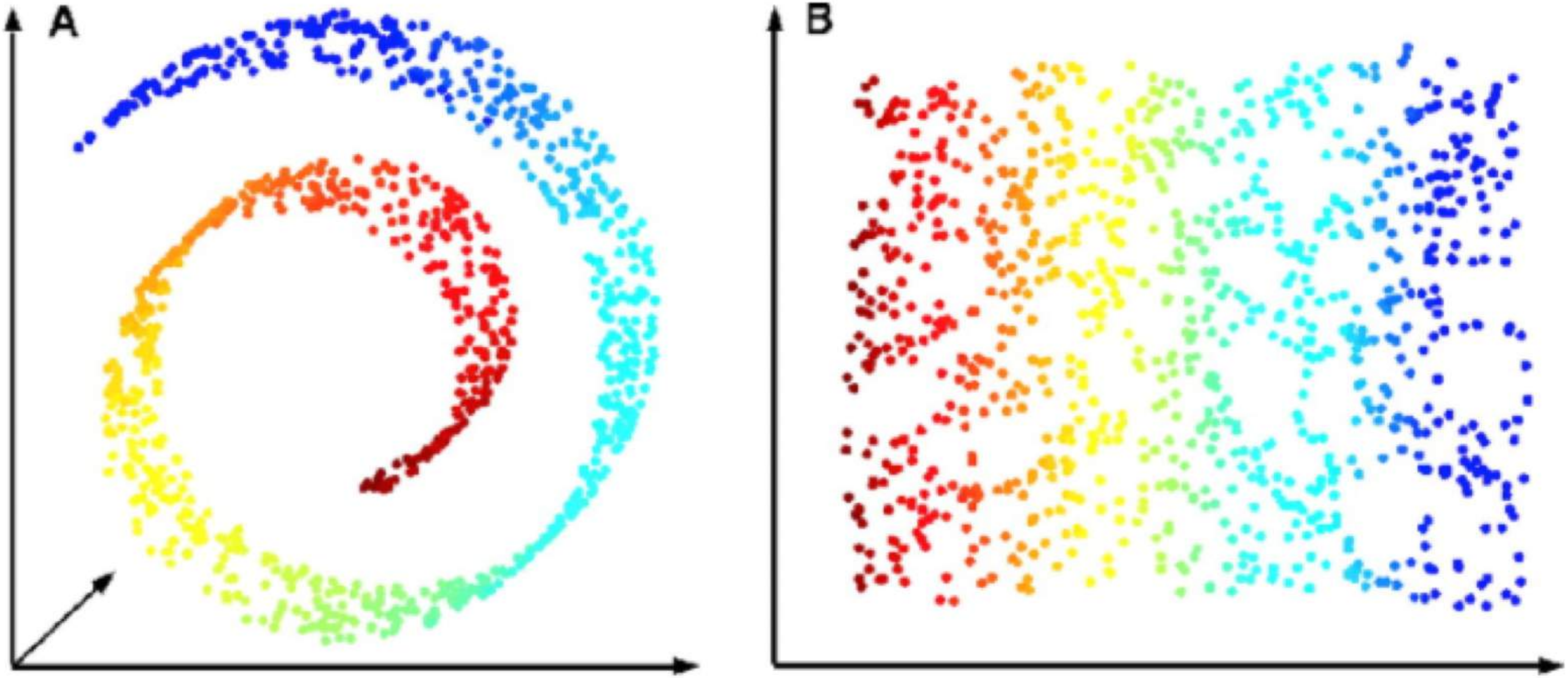
- Method for feature learning / representation learning – learns a sparse representation of the data
- Overcomplete basis → data sparse in a higher dimensional feature space



Dimensionality Reduction & Manifold Learning

Non-linear Methods

Assumption: data live on a non-linear, low-dimensional manifold.

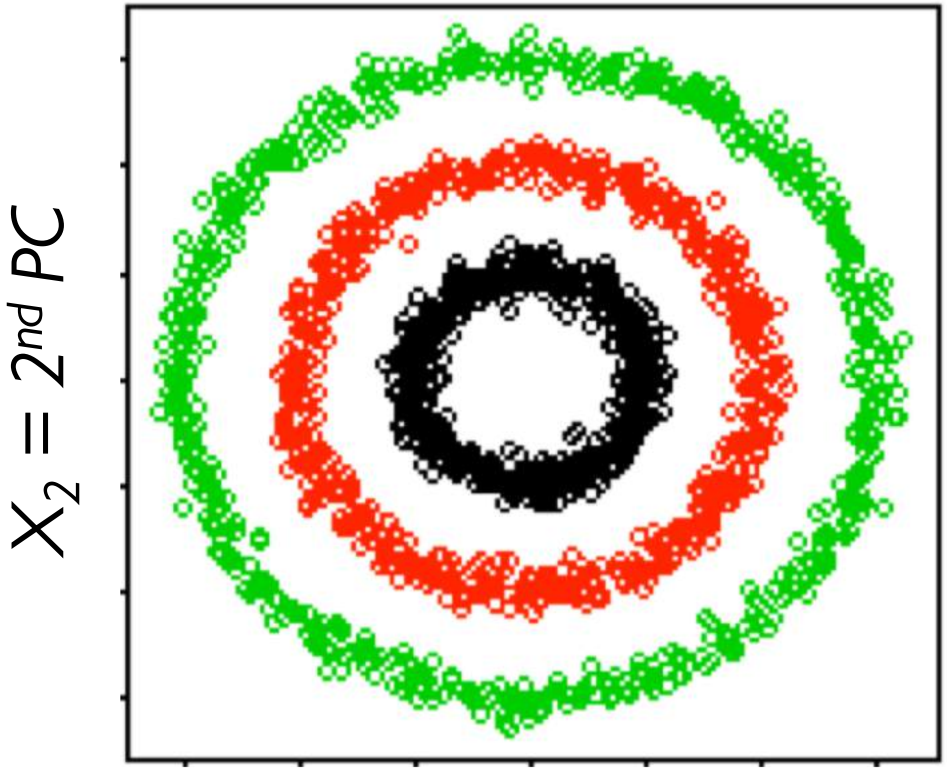


Dimensionality reduction by (linear) projection onto a 2D plane will not preserve structure (color progression).

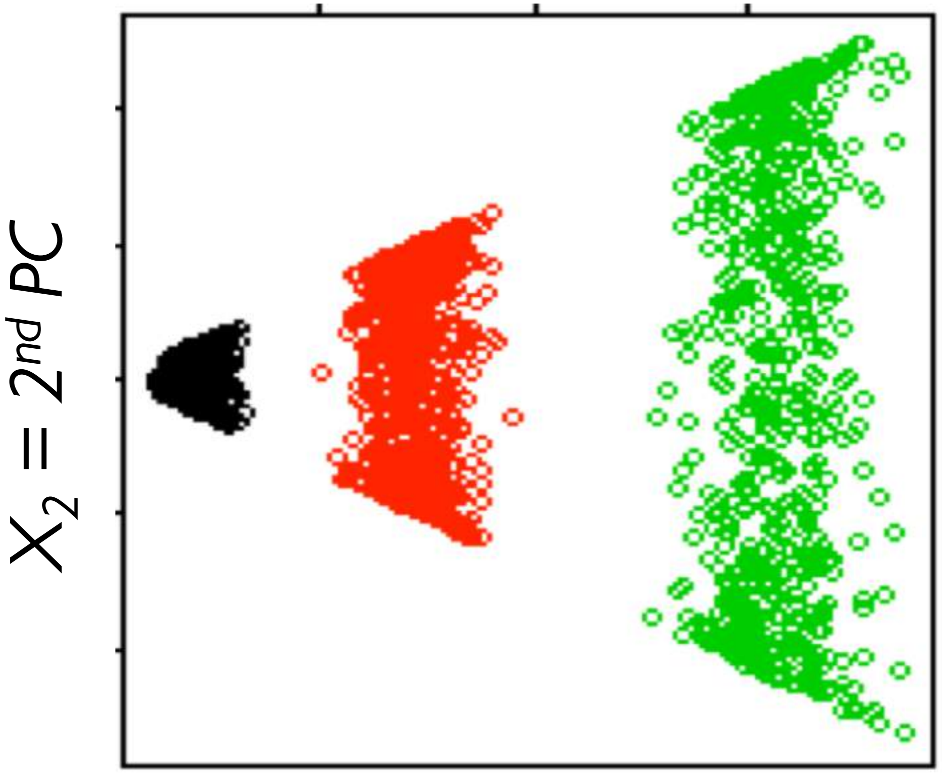
Kernel PCA

Applies PCA to (implicit) higher-dimensional representation of data.

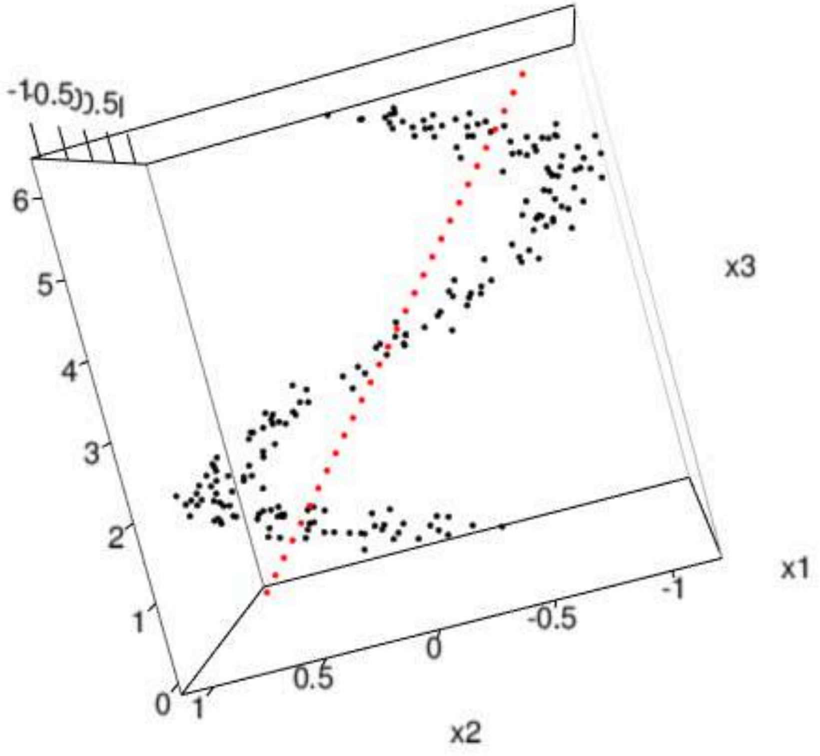
Original Data



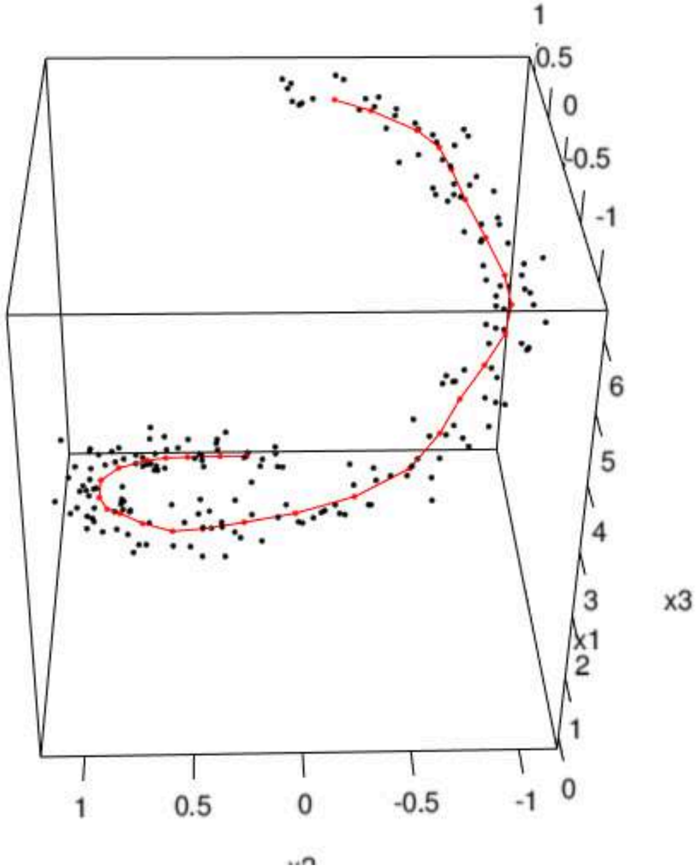
Polynomial Kernel



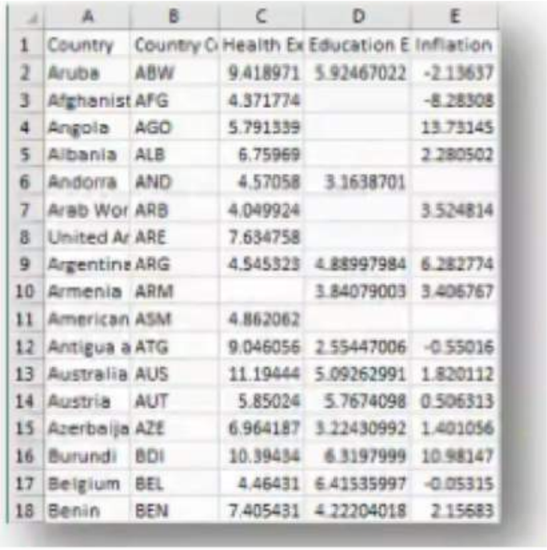
Self Organizing Maps (SOM)



Prototypes initialized along 1st principal component axis



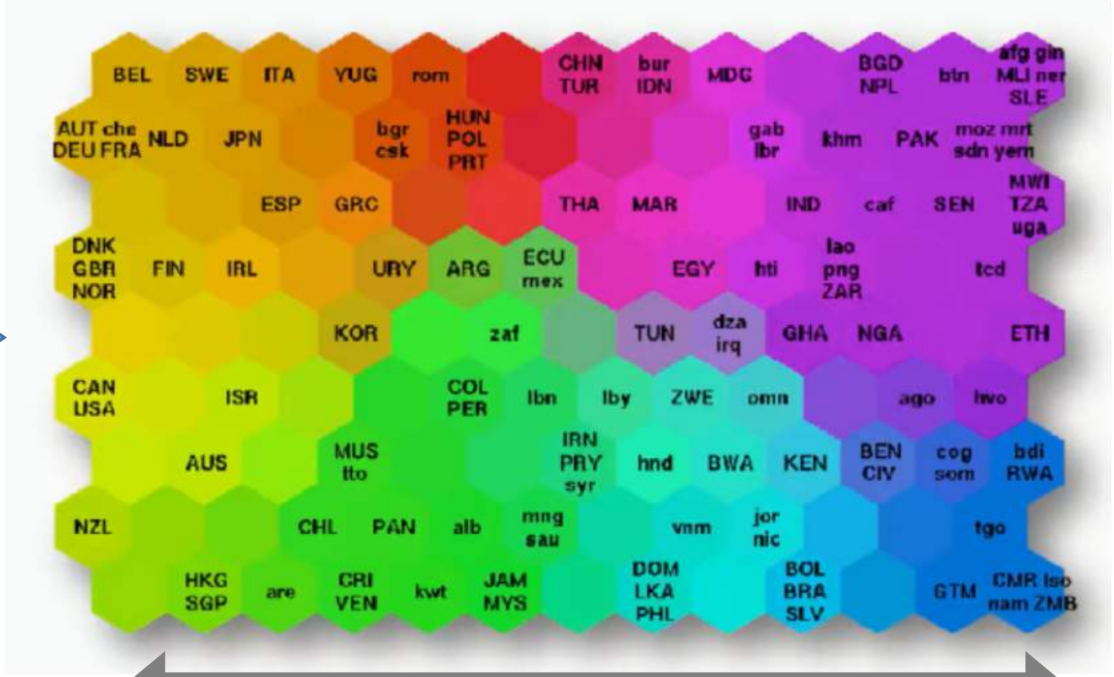
Final result of SOM iteration



	A	B	C	D	E
1	Country	Country C	Health Ex	Education E	Inflation
2	Aruba	ABW	9.418971	5.92467022	-2.13637
3	Afghanist	AFG	4.371774		-8.28308
4	Angola	AGO	5.791339		13.73145
5	Albania	ALB	6.75969		2.280502
6	Andorra	AND	4.57058	3.1638701	
7	Arab Wor	ARB	4.049924		3.524814
8	United Ar	ARE	7.634758		
9	Argentina	ARG	4.545323	4.88997984	6.282774
10	Armenia	ARM		3.84079003	3.406767
11	American	ASM	4.862062		
12	Antigua a	ATG	9.046056	2.55447006	-0.55016
13	Australia	AUS	11.19444	5.09262991	1.820112
14	Austria	AUT	5.85024	5.7674098	0.506313
15	Azerbaija	AZE	6.964187	3.22430992	1.401056
16	Burundi	BDI	10.39434	6.3197999	10.98147
17	Belgium	BEL	4.46431	6.41535997	-0.05315
18	Benin	BEN	7.405431	4.22204018	2.15683

39 features

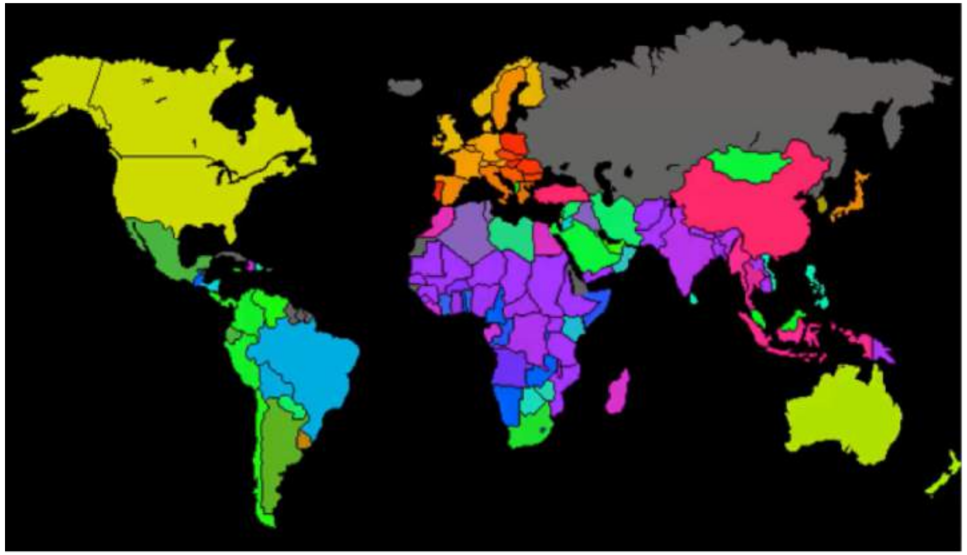
(development indicators)



2D map

Higher income

Lower income

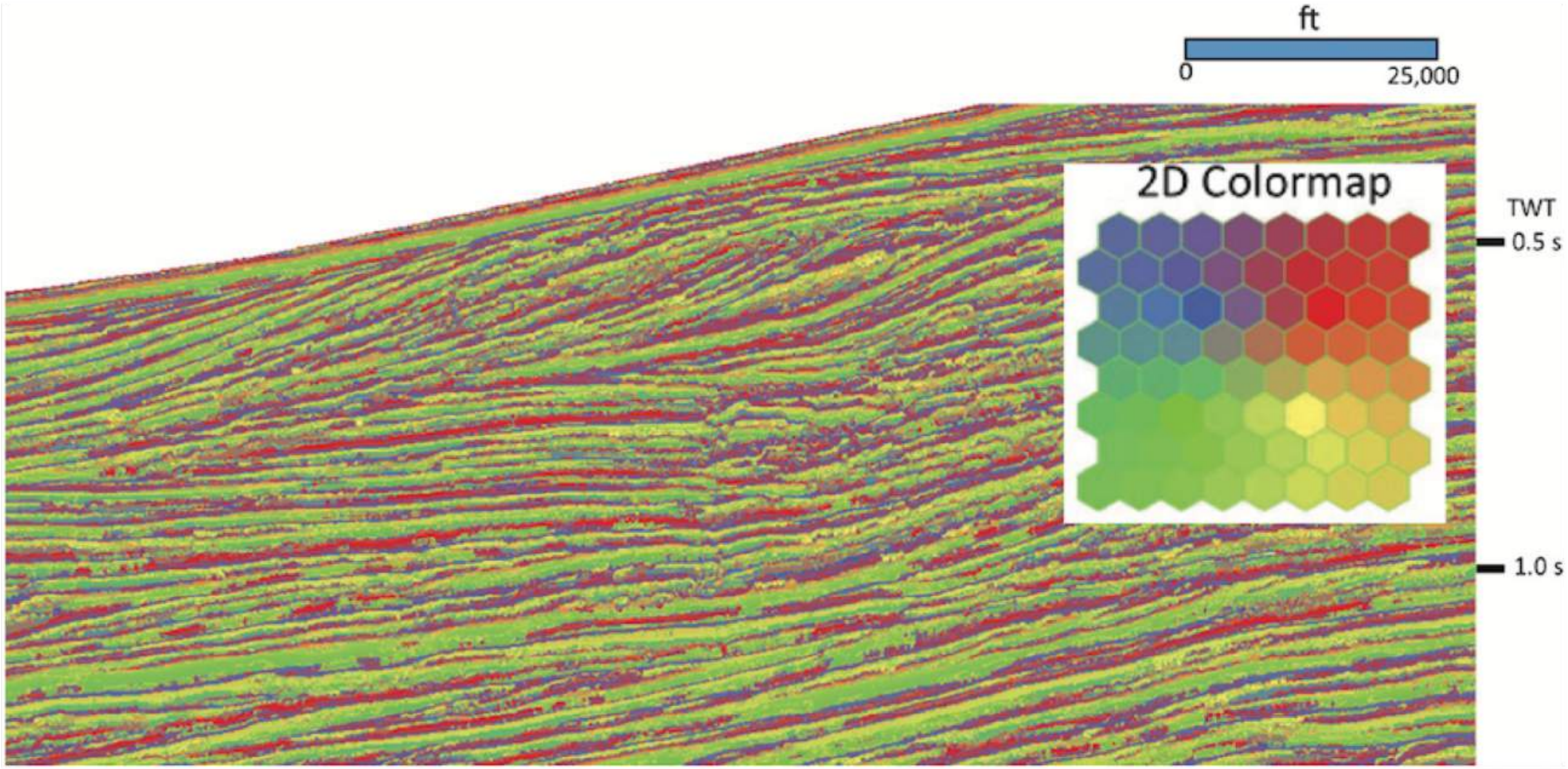


Self Organizing Map in higher dimensions

Geoscience Example 2:

PCA & SOM for interpretation of seismic reflection data

Roden et al. (2015)

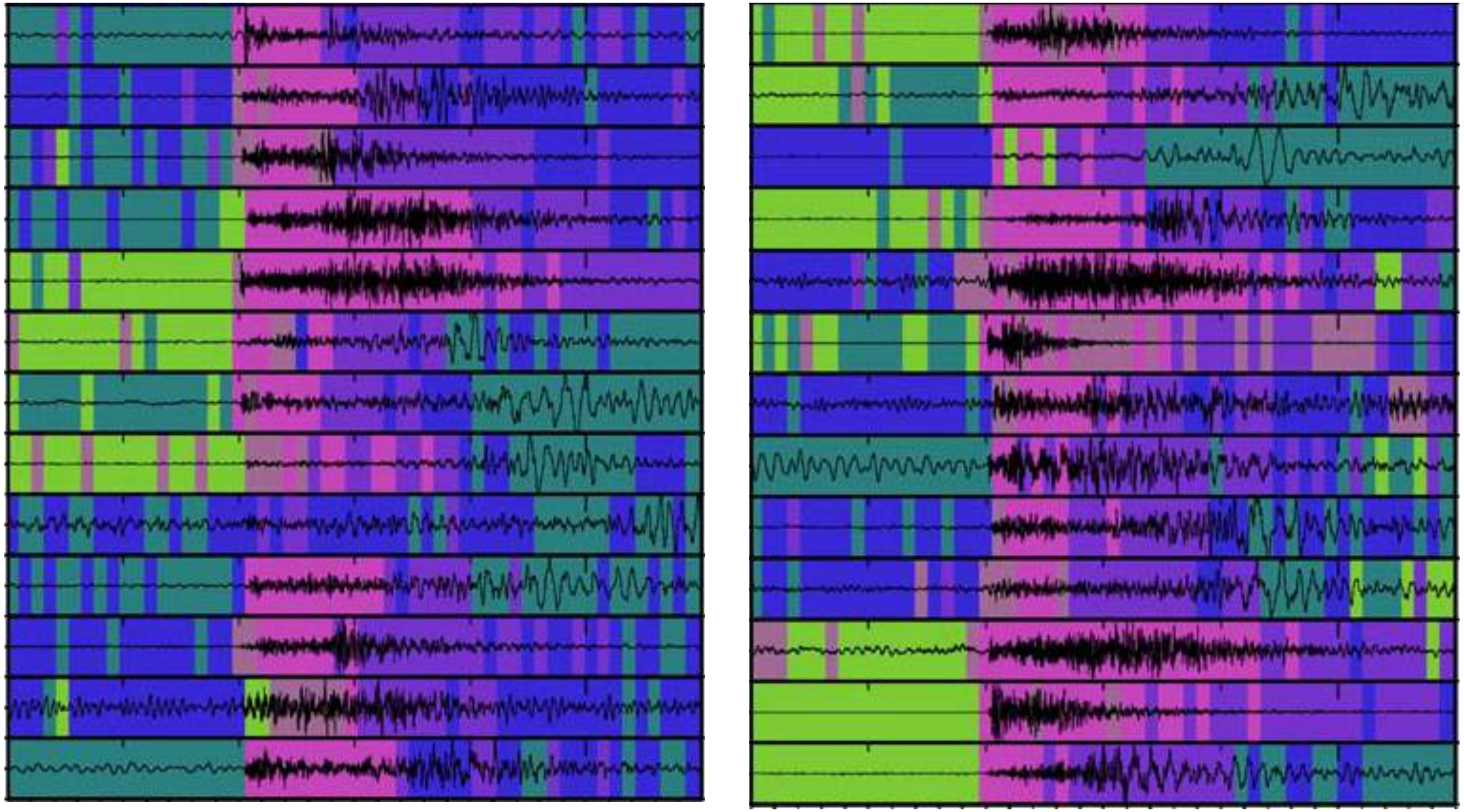


1) PCA used to select subset of seismic attributes
2) SOM (64 prototypes in 8x8 grid) identifies geologic features

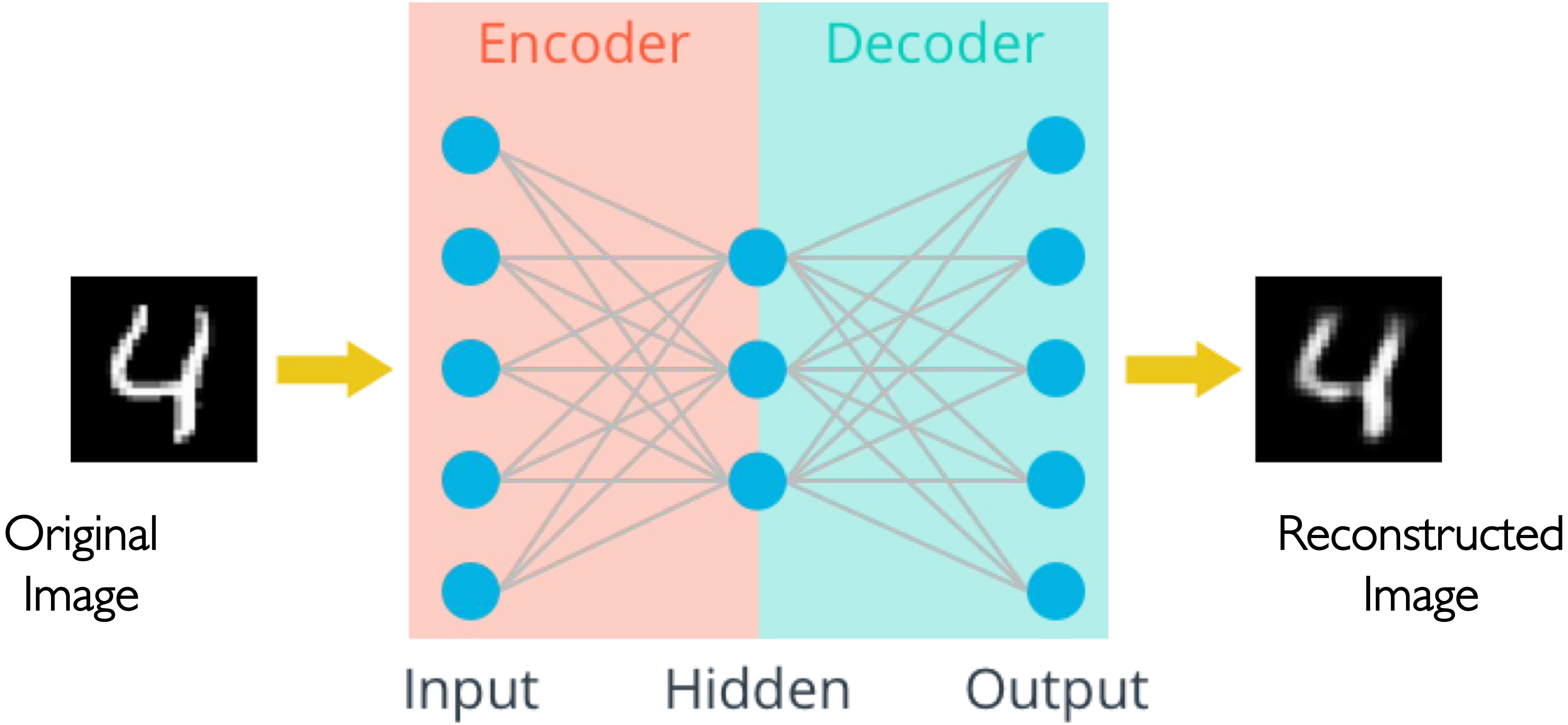
Geoscience Example 3:
SOM clustering to visualize and discriminate wave phases

Köhler et al. (2009)

SOM +
Hierarchical clustering

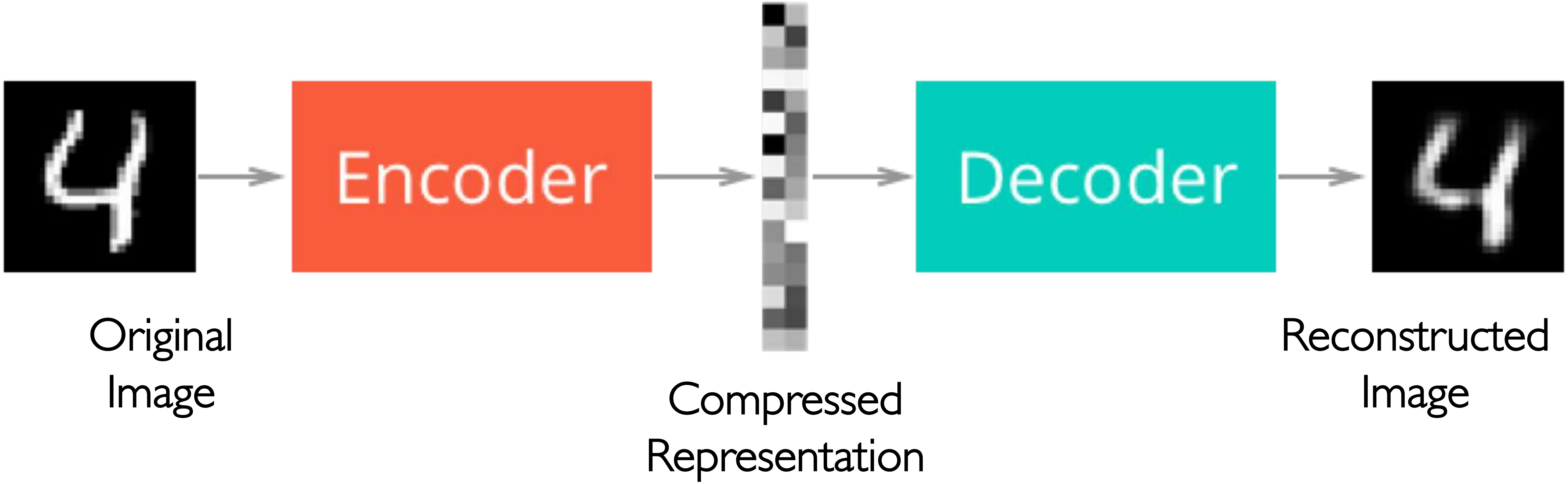


Neural Network: Autoencoder



Autoencoder learns an approximate identity operator, composed of an encoder (reduces dimensionality) and a decoder

Neural Network: Autoencoder

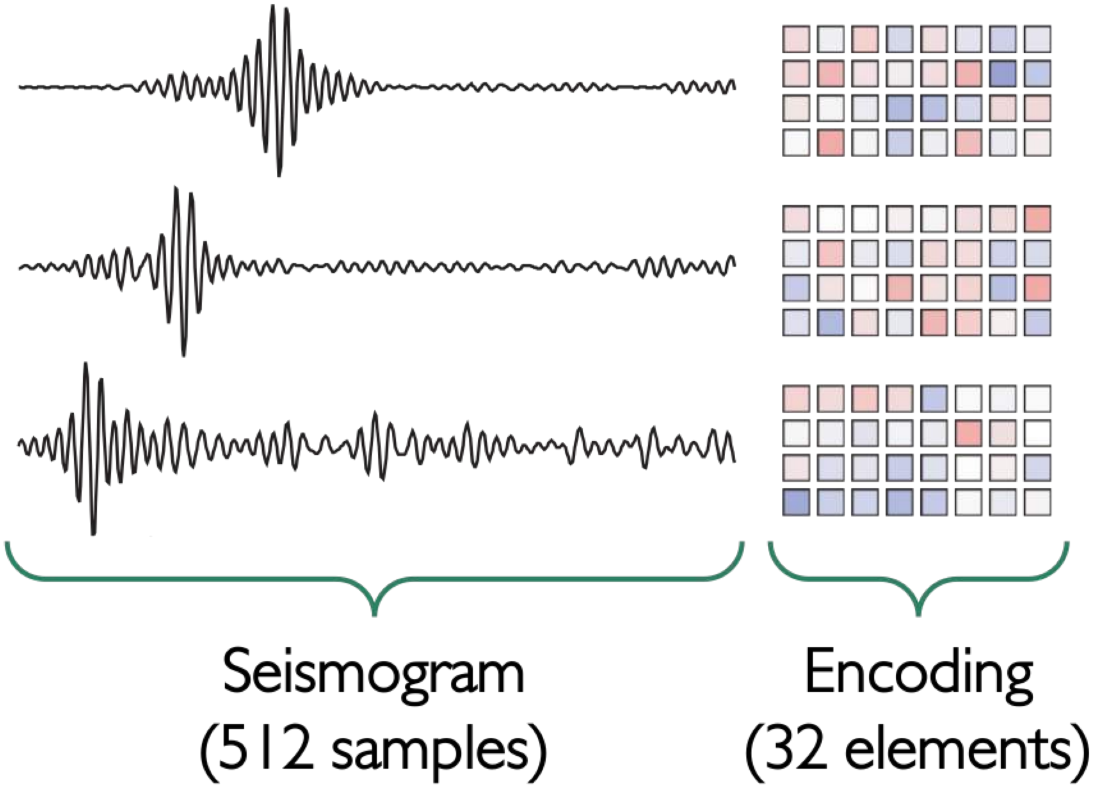


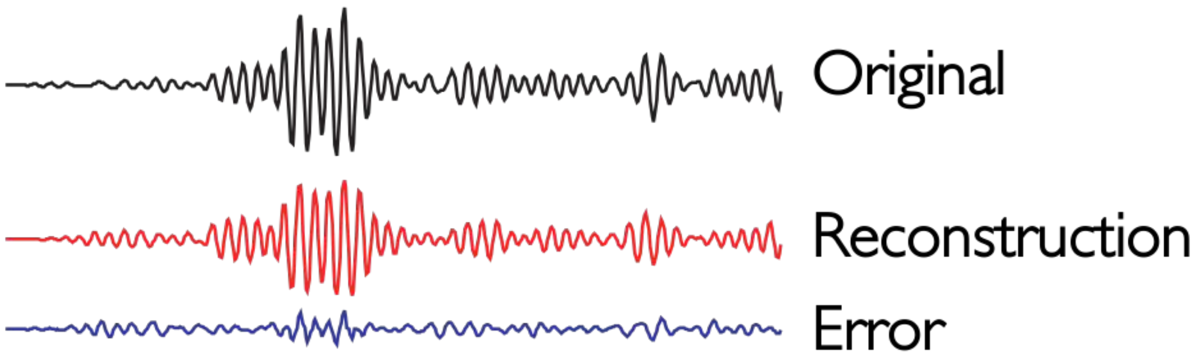
Autoencoder learns an approximate identity operator, composed of an encoder (reduces dimensionality) and a decoder

Geoscience Example 4:
Autoencoder for waveform data

Valentine & Trampert (2012)

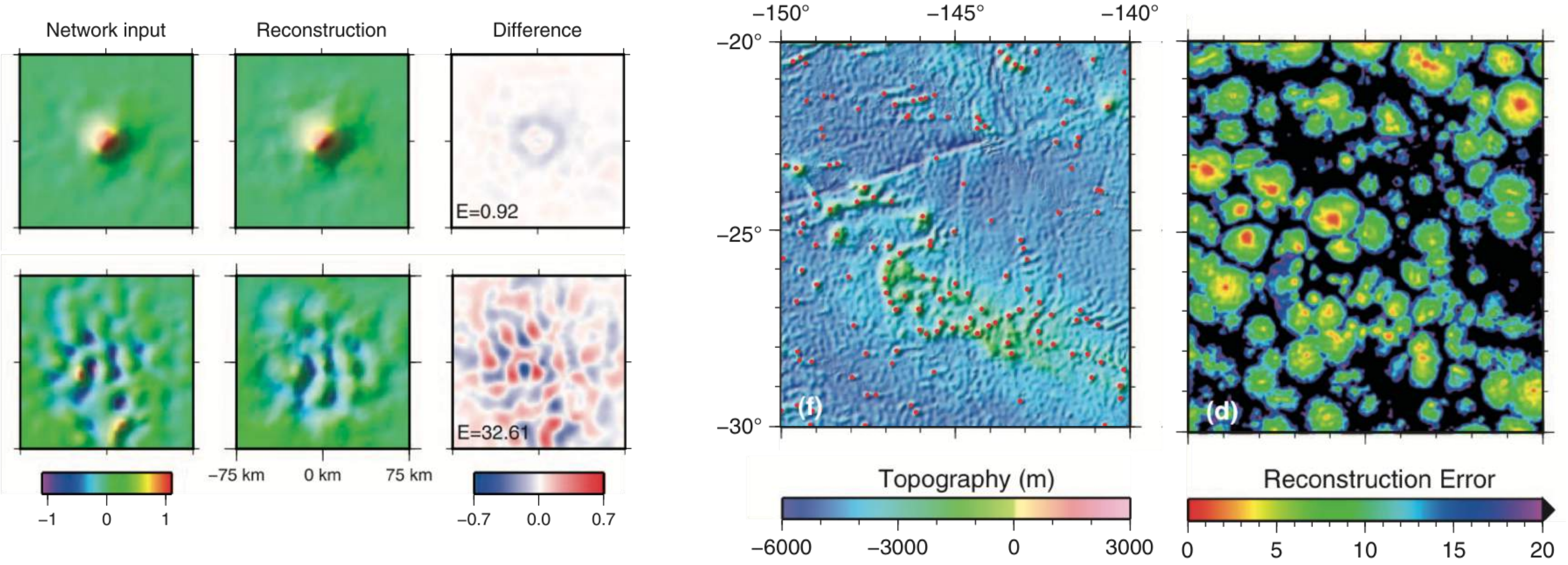
Autoencoder learns to reconstruct earthquake waveforms from low-dimensional representation (encoding)





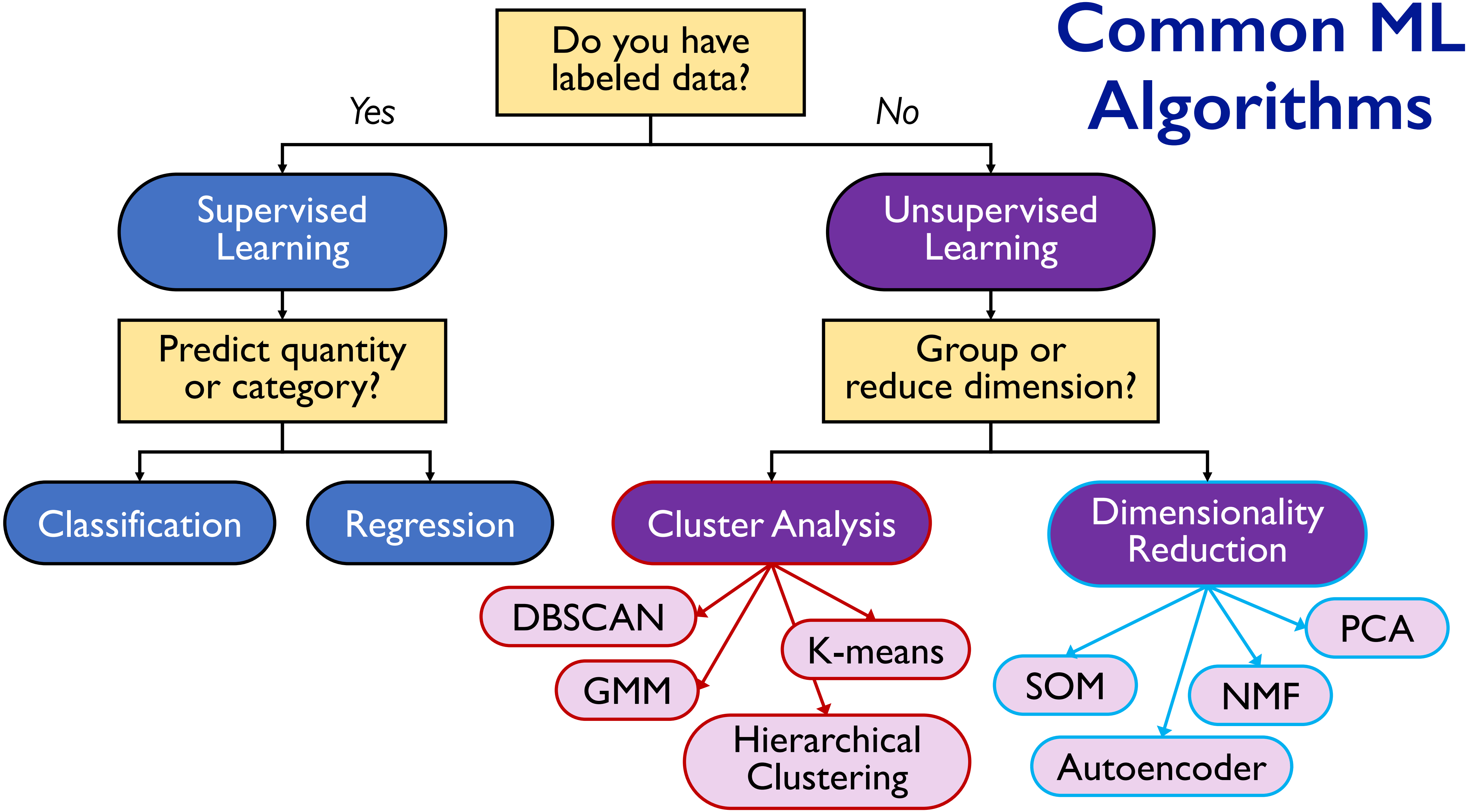
Geoscience Example 5:
Autoencoder for finding seamounts in bathymetric data

Valentine et al. (2013)



1) Autoencoder learns features to reconstruct seamount bathymetry
2) Seamount discovery → reconstruction quality as classification metric

Common ML Algorithms



Questions?

karianne_bergen@fas.harvard.edu