Data Mining on the Cloud 101: Unsupervised learning on the cloud

SSA Anchorage 2024

Theresa Sawi, Nate Groebner, Felix Waldhauser, Kaiwen Wang, Eric Beaucé
Last edited 2024/04/28

Outline

- **Introduction to unsupervised machine learning (UML)**
  - feature extraction & dimensionality reduction
  - clustering

- **SpecUFEx tutorial: Amatrice 2016**

Outline

- **Introduction to unsupervised machine learning (UML)**
  - feature extraction & dimensionality reduction
  - clustering

- **SpecUFEx tutorial: Amatrice 2016**

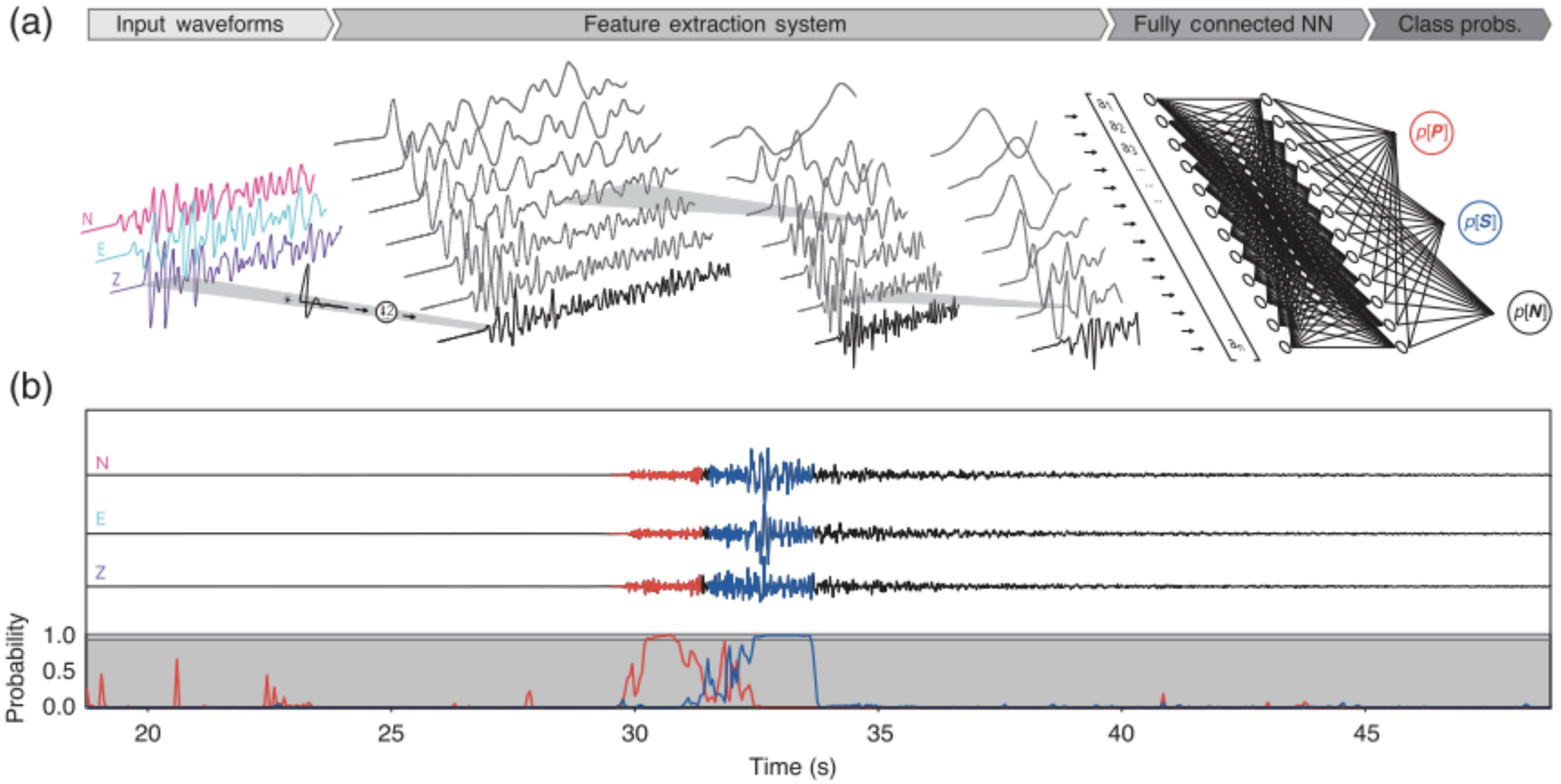
Machine Learning

Supervised	Unsupervised
Train a function that best approximates mapping between input data (x) and labels (y) **f(x) = y**	

Supervised Machine Learning

f(x) trained on millions of P- and S-wave arrivals

x **f(x)** **y**



Ross et al., 2018, BSSA

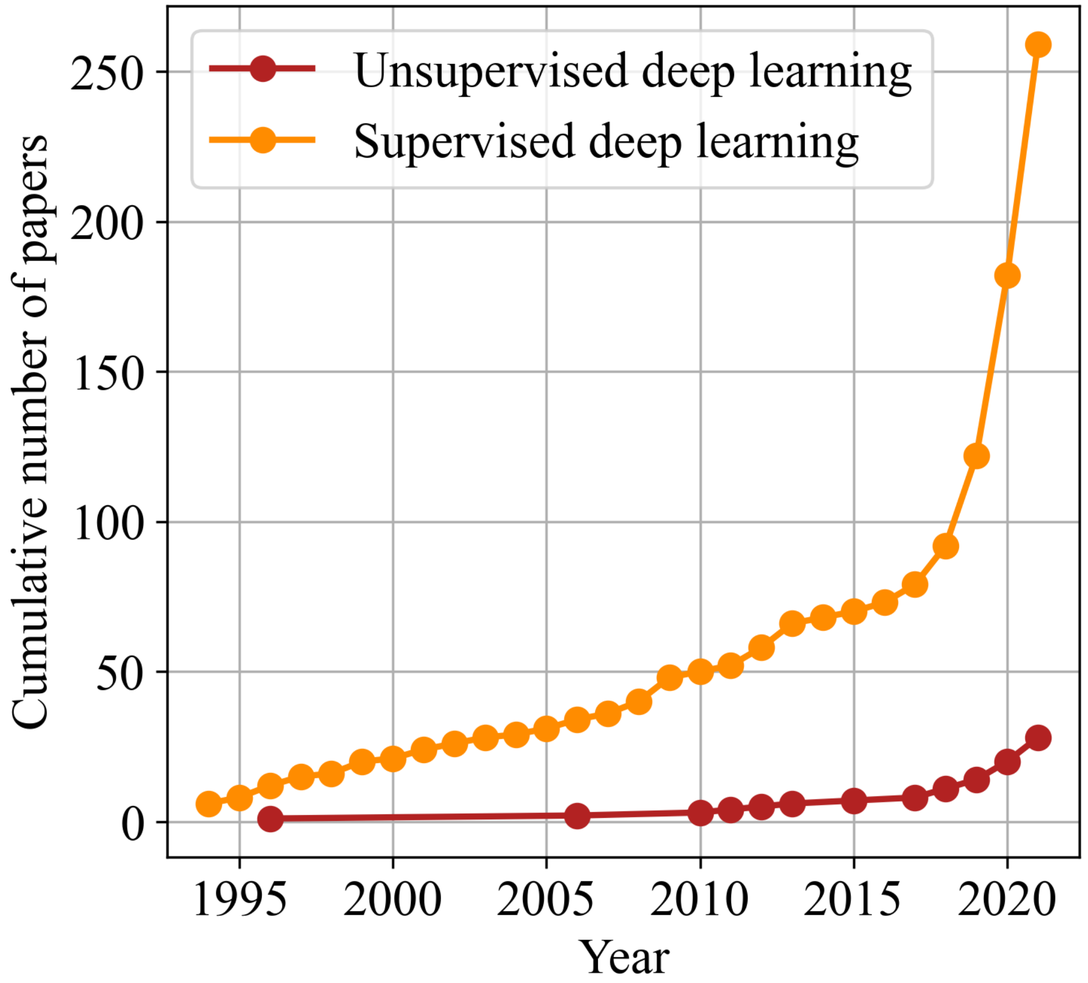
Machine Learning

Supervised	Unsupervised
Train a function that best approximates mapping between input data (x) and labels (y)	Learn a function that *infers* natural structure within the data
f(x) = y	**f(x) = y**

Machine Learning

Supervised	Unsupervised
Train a function that best approximates mapping between input data (x) and labels (y)	Learn a function that *infers* natural structure within the data
f(x) = y	**f(x) = y**
• Needs labeled data • Only finds patterns that it has learned • Routine work	• Unlabeled data • Infers patterns in data • Exploratory work

Unsupervised machine learning (UML) in seismology



Sawi Dissertation, 2023. Adapted from Mousavi & Beroza, 2022; Science

ML Algorithm Examples

Unsupervised ML	Supervised ML
Dimensionality reduction: Principal component analysis (PCA) Singular value decomposition (SVD) Nonnegative matrix factorization (NMF)*	**Regression:** Linear Polynomial
Clustering: Kmeans* Hierarchical DBSCAN	**Classification:** Decision trees Random forest K-nearest neighbors Support vector machines (SVM)

Outline

- Introduction to unsupervised machine learning (UML)
  - **feature extraction & dimensionality reduction**
  - clustering

- **SpecUFEx tutorial: Amatrice 2016**

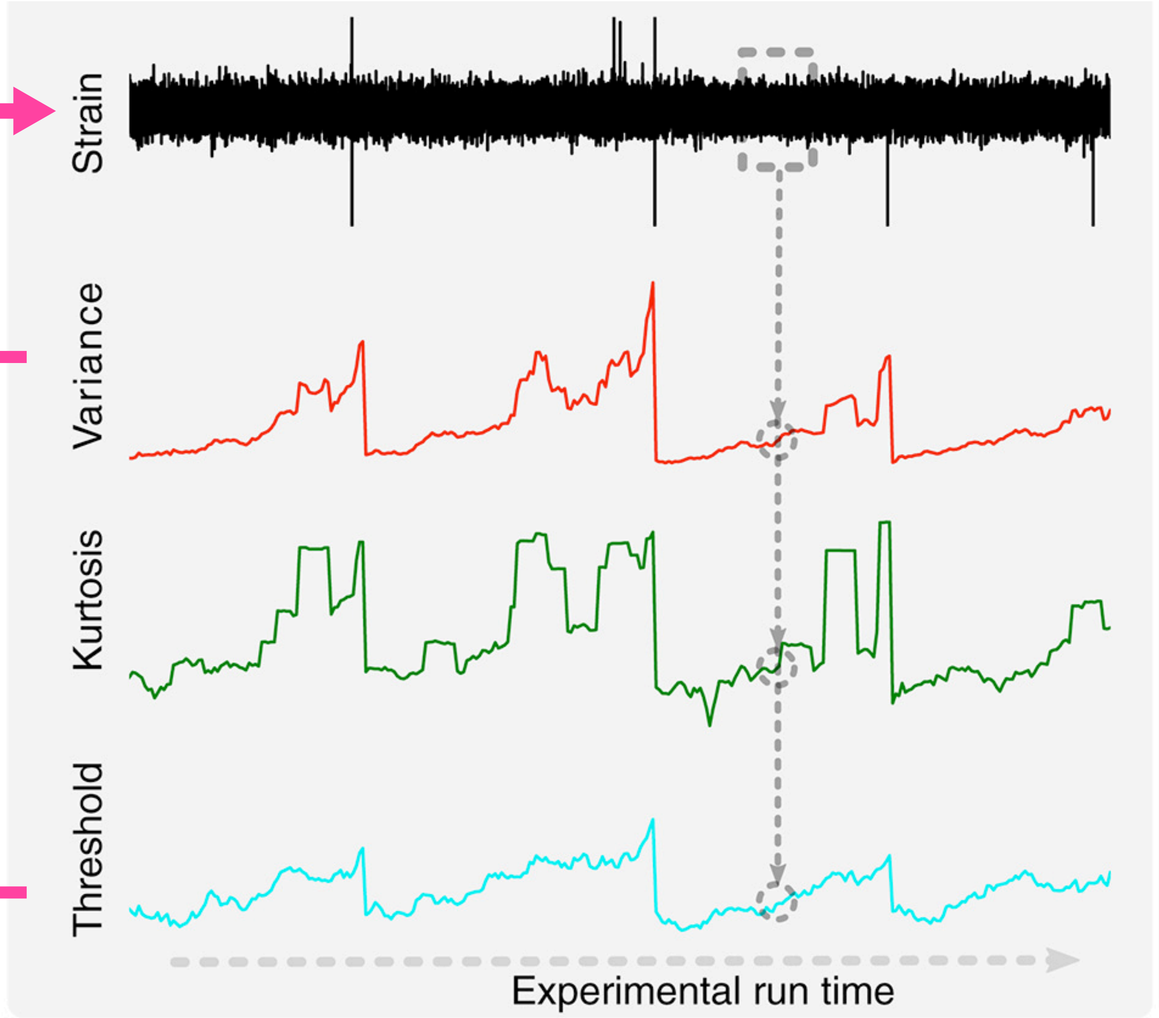
Feature Extraction & Dimensionality Reduction

"Garbage in, garbage out"

Extracting information from data

- Reduce noise and overfitting
- Aid interpretation
- Reduces dimension
- Faster computations

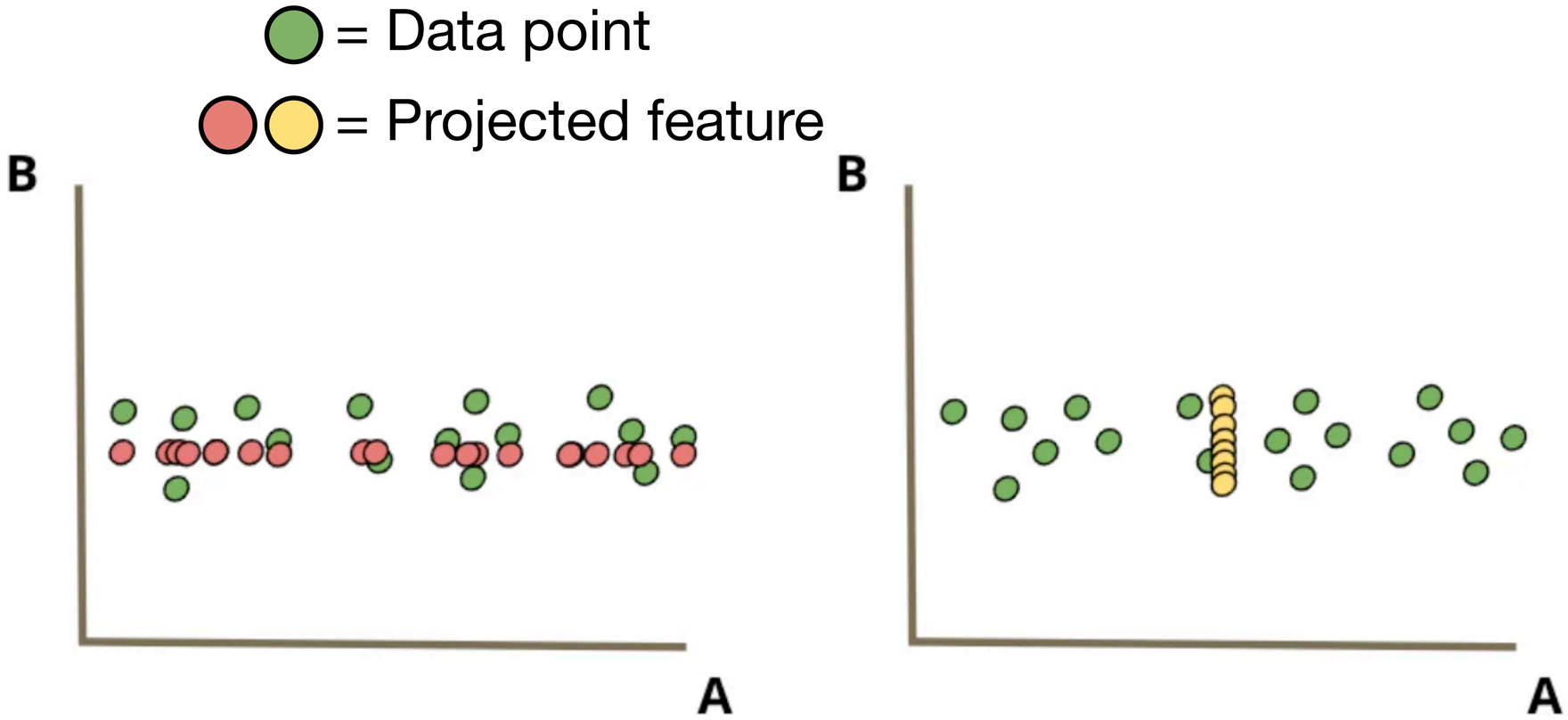
Techniques depend on data type and domain knowledge



Rouet-Leduc et al., (2017) GRL

Feature Extraction & Dimensionality Reduction

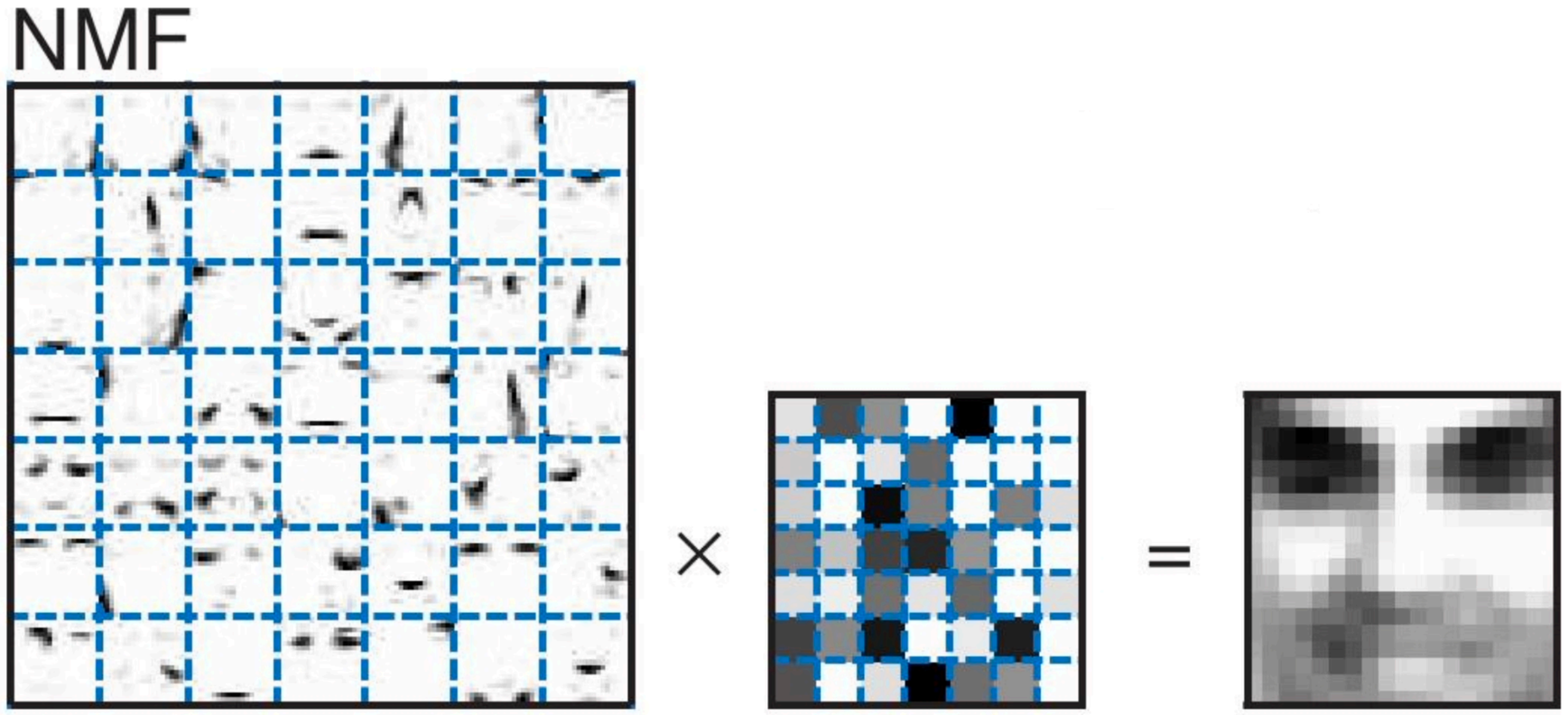
Example: find features with the highest variance / spread



https://medium.com/@gallettilance/singular-value-decomposition-787c08b8931b

Feature Extraction & Dimensionality Reduction

Example: Non-negative matrix factorization (NMF)



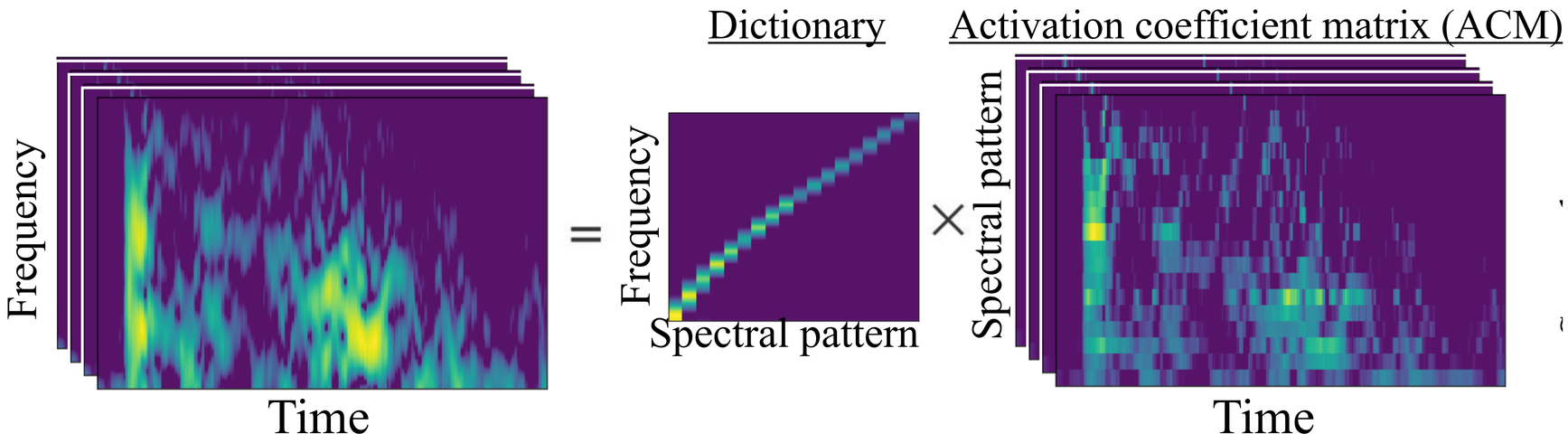
NMF learns a "parts-based" representation. Each column captures something interpretable. This is a result of the nonnegativity constraint.

Lee and H.S. Seung (2001)

Spectral Unsupervised Feature Extraction (SpecUFEx)

Holtzman et al., 2018; Sci Adv

Step 1. Non-Negative Matrix Factorization



Extract spectral "patterns" (frequency distributions) through time

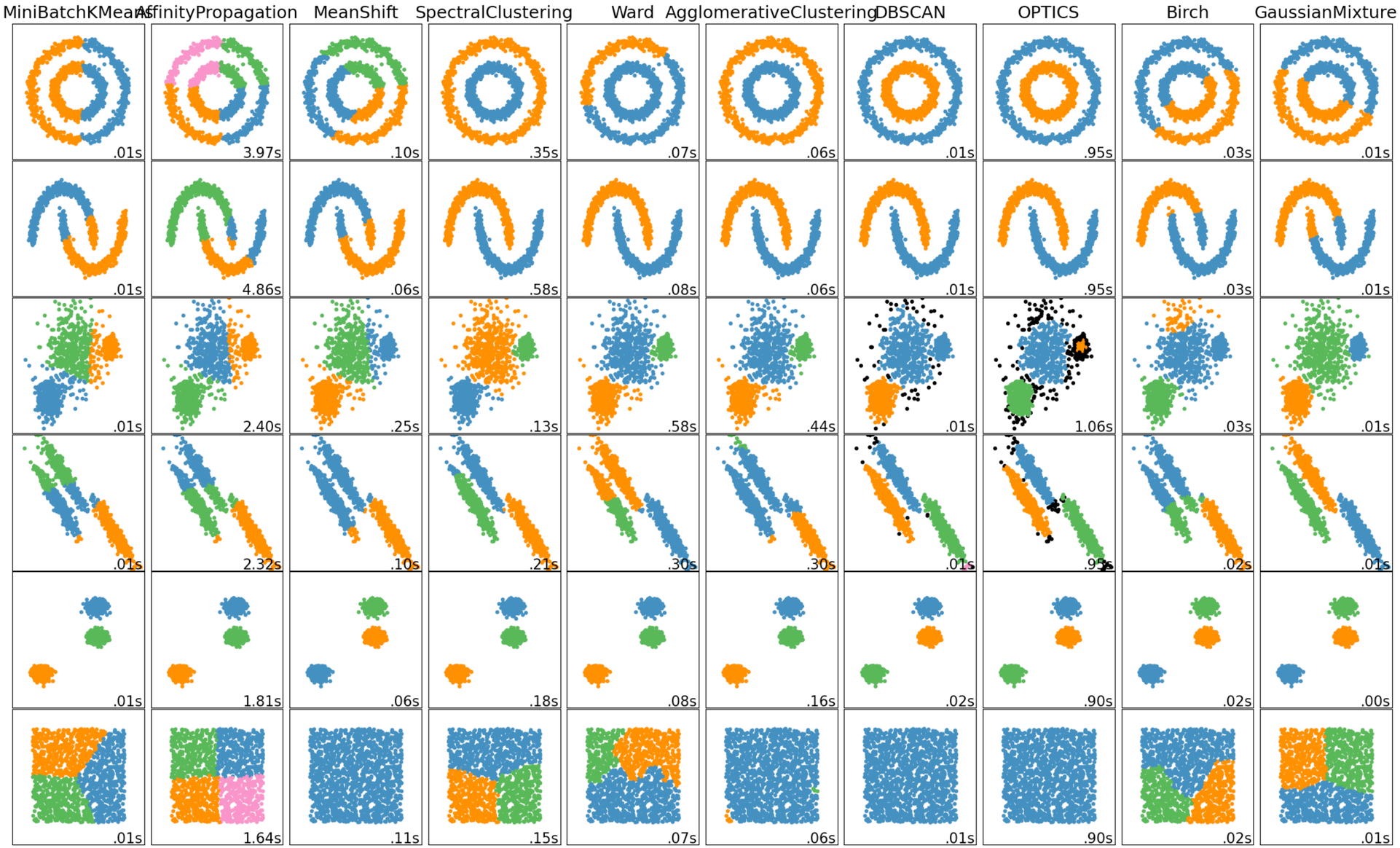
Spectrograms (units in dB) → Non-negative Matrix Factorization (NMF)

Outline

- Introduction to unsupervised machine learning (UML)
  - feature extraction & dimensionality reduction
  - clustering

- SpecUFEx tutorial: Amatrice 2016

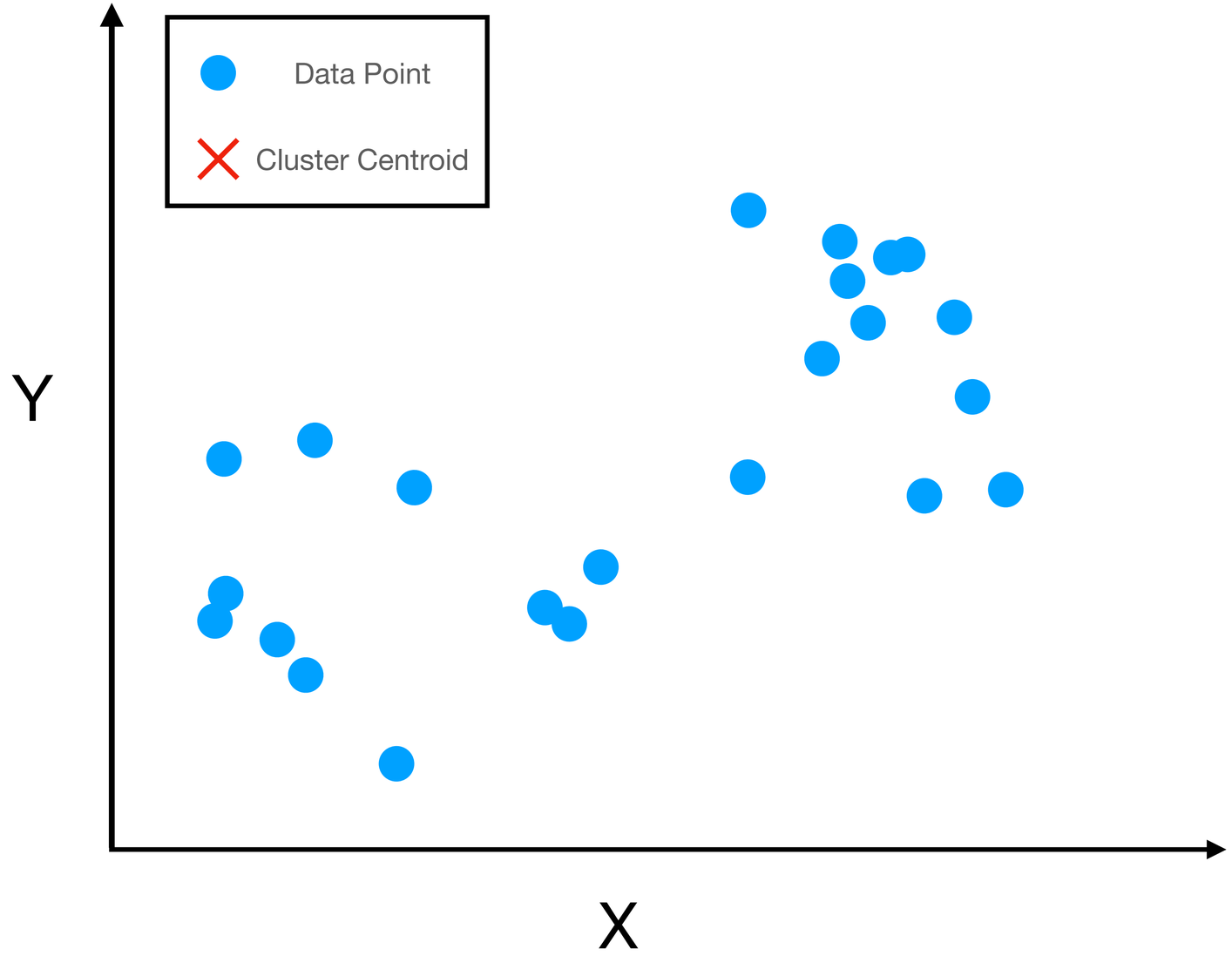
Clustering



https://scikit-learn.org/0.23/auto_examples/cluster/plot_cluster_comparison.html

Clustering: K-means

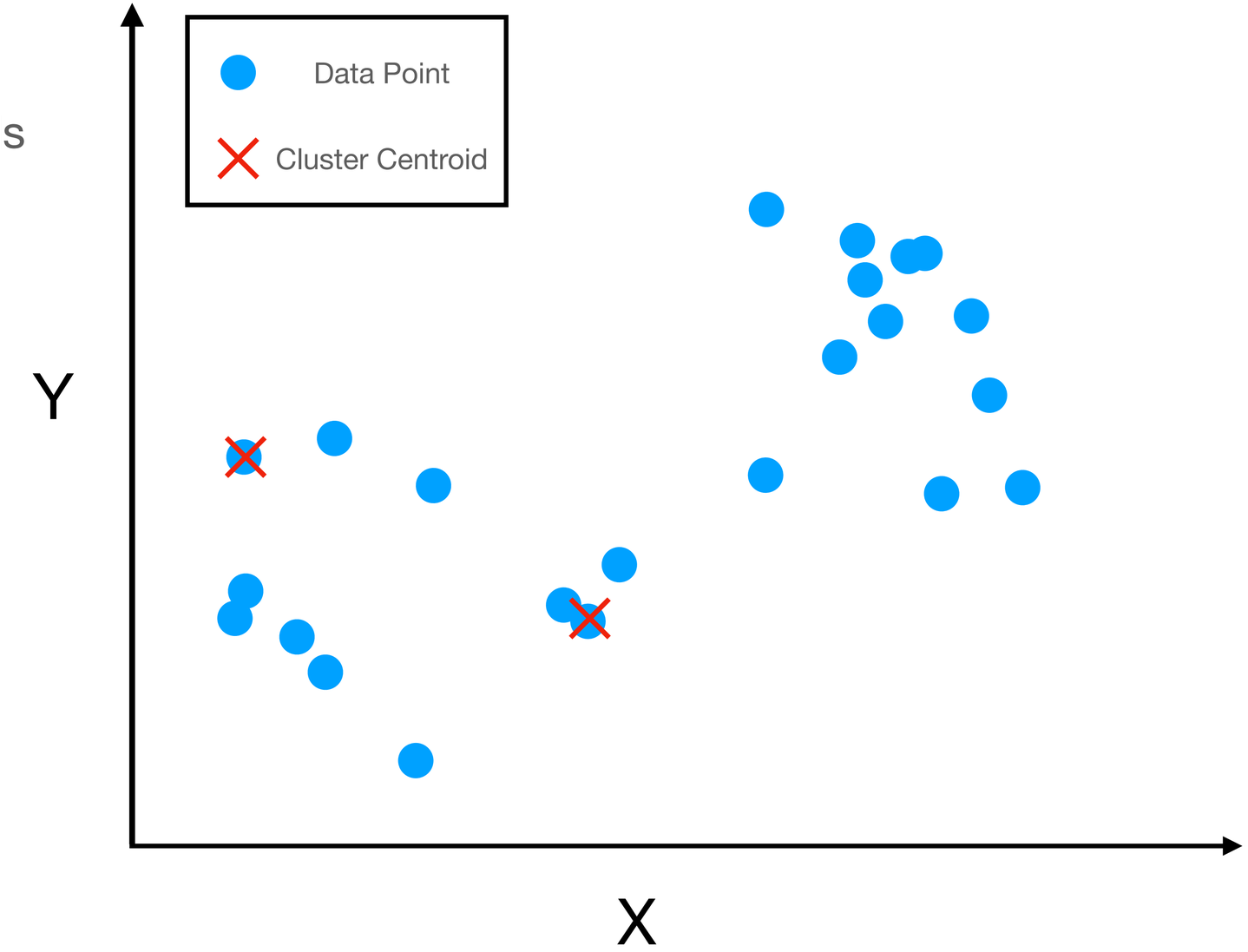
Lloyd,1982, IEEE



Clustering: K-means

Lloyd,1982, IEEE

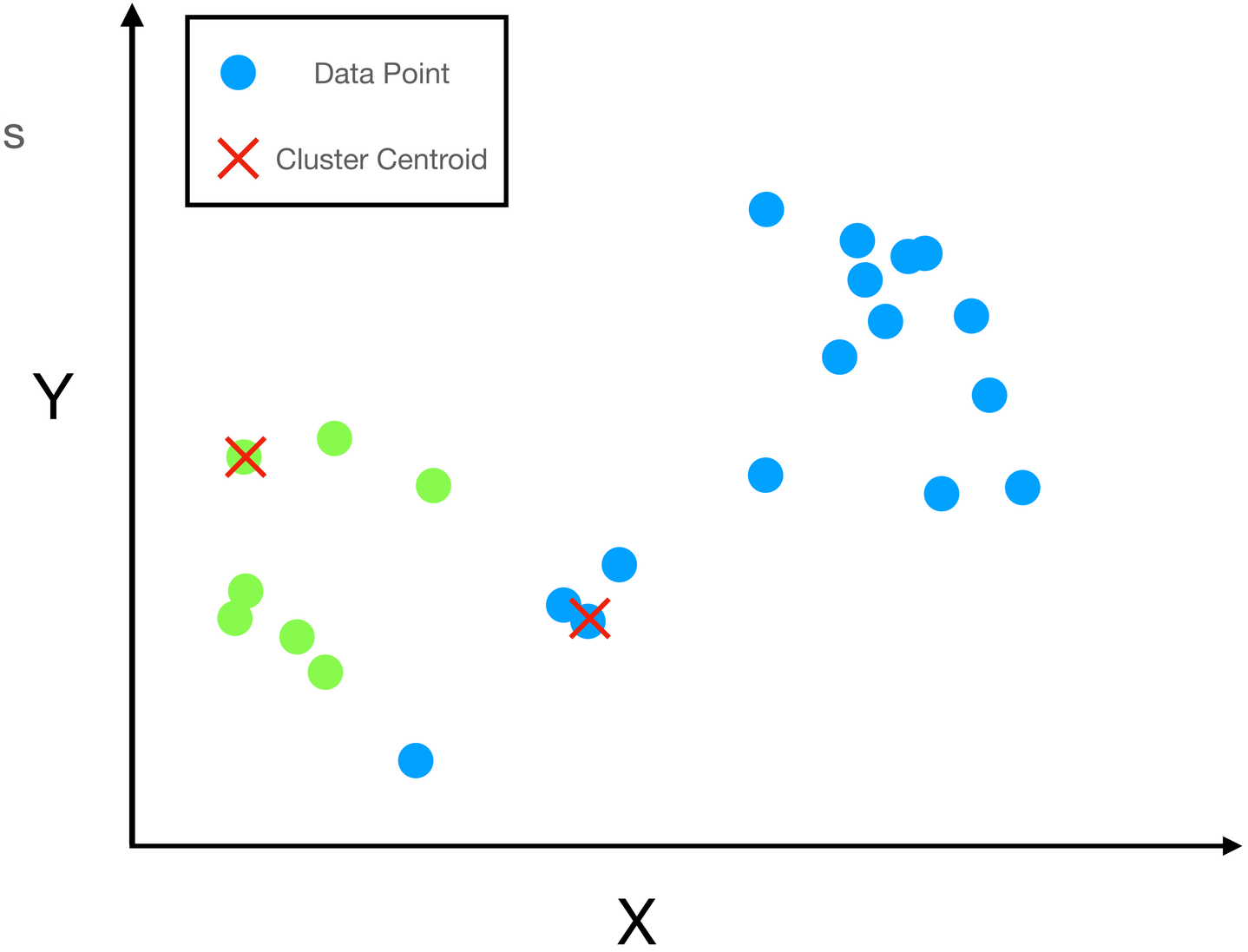
Step 1: Define centroids



Clustering: K-means

Lloyd,1982, IEEE

Step 1: Define centroids
Step 2: Cluster data

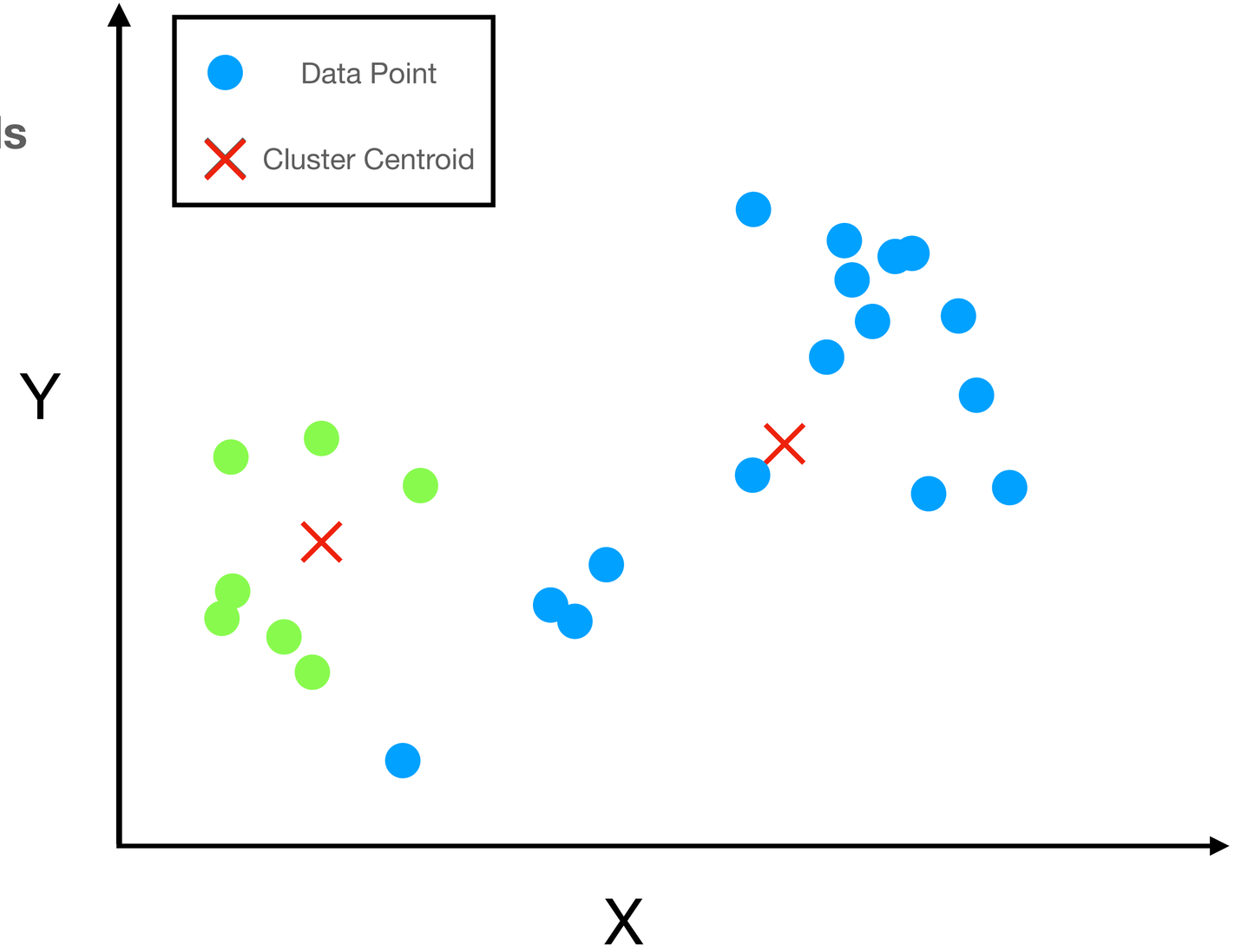


Clustering: K-means

Lloyd,1982, IEEE

Step 1: Define centroids

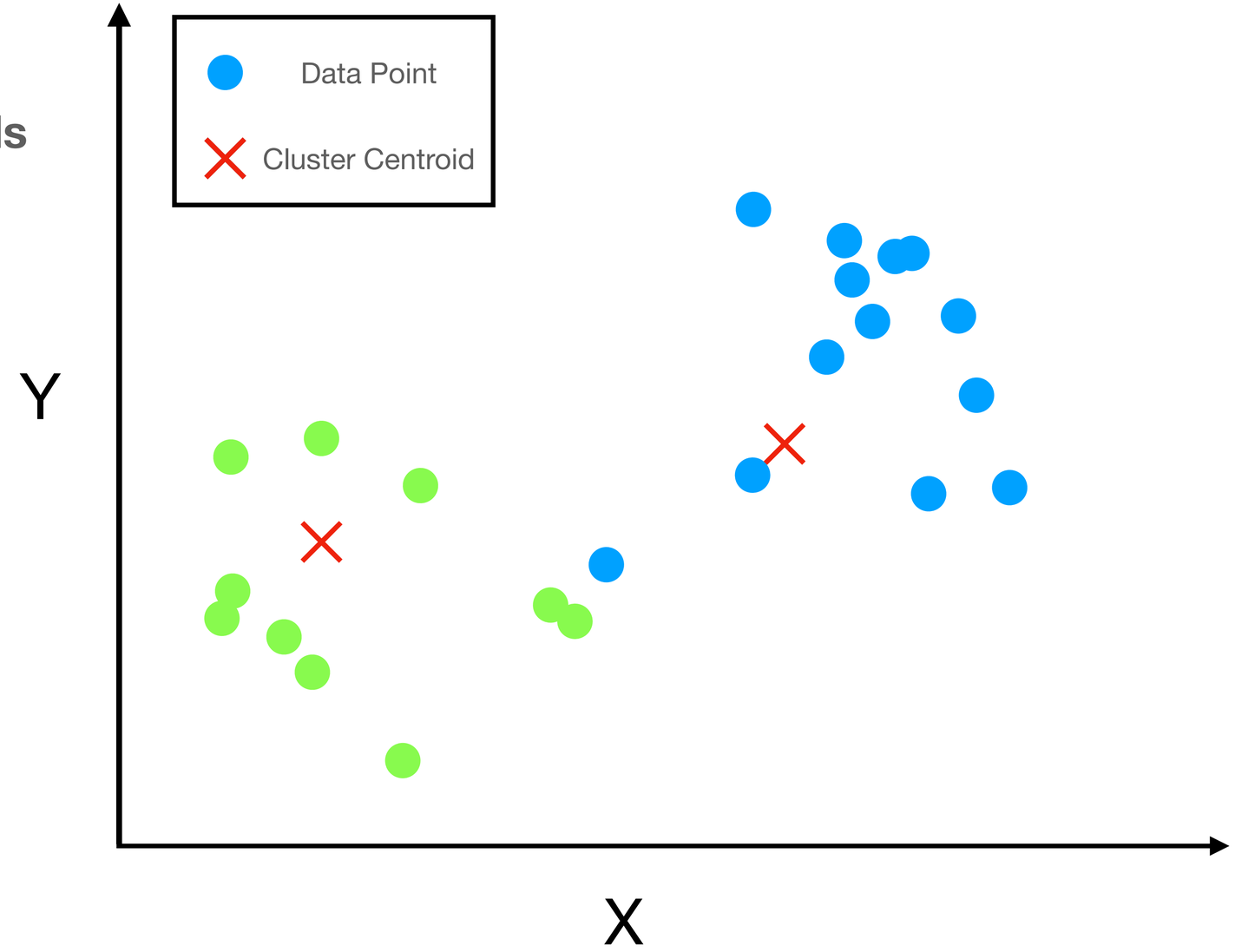
Step 2: Cluster data



Clustering: K-means

Lloyd,1982, IEEE

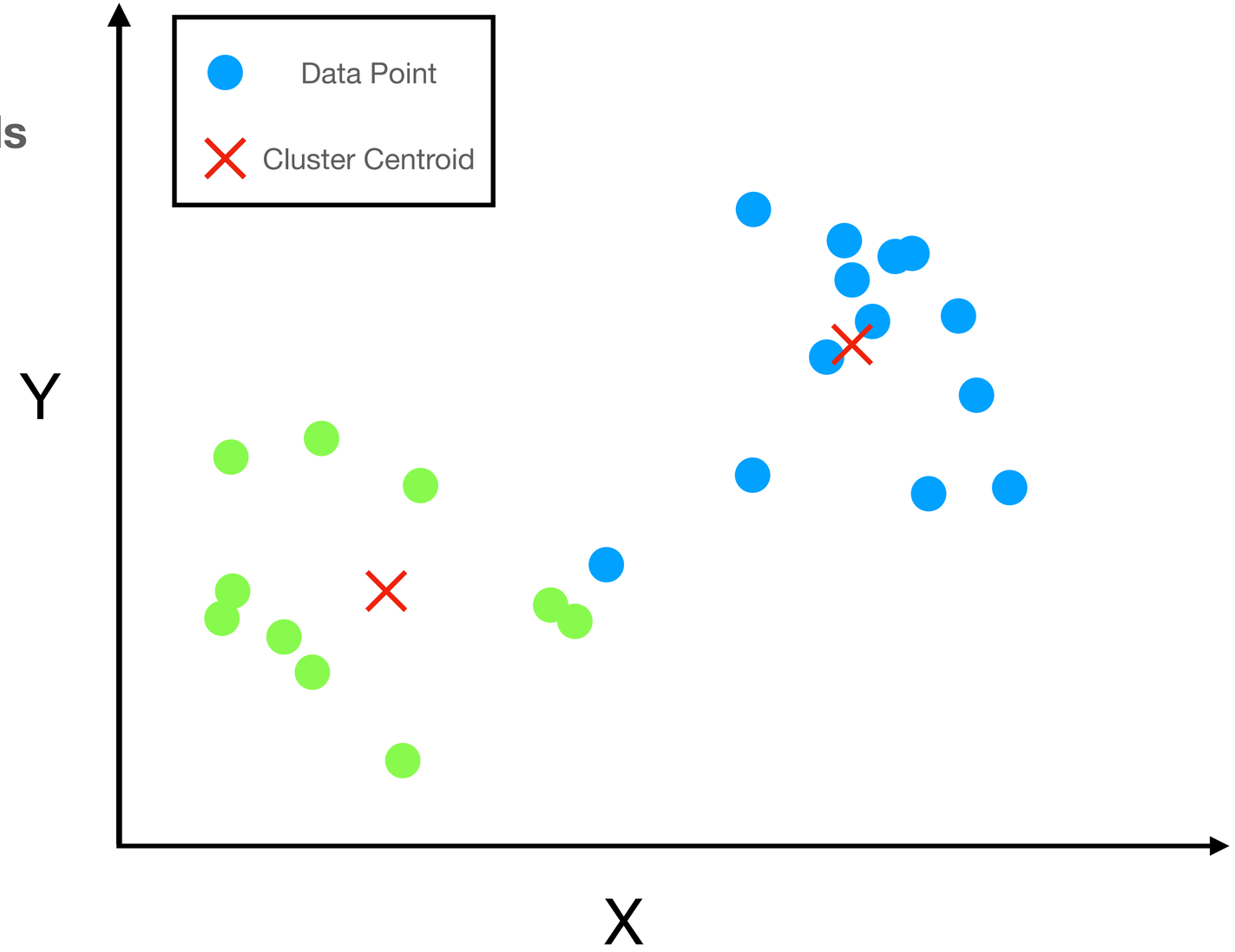
Step 1: Define centroids
Step 2: Cluster data



Clustering: K-means

Lloyd,1982, IEEE

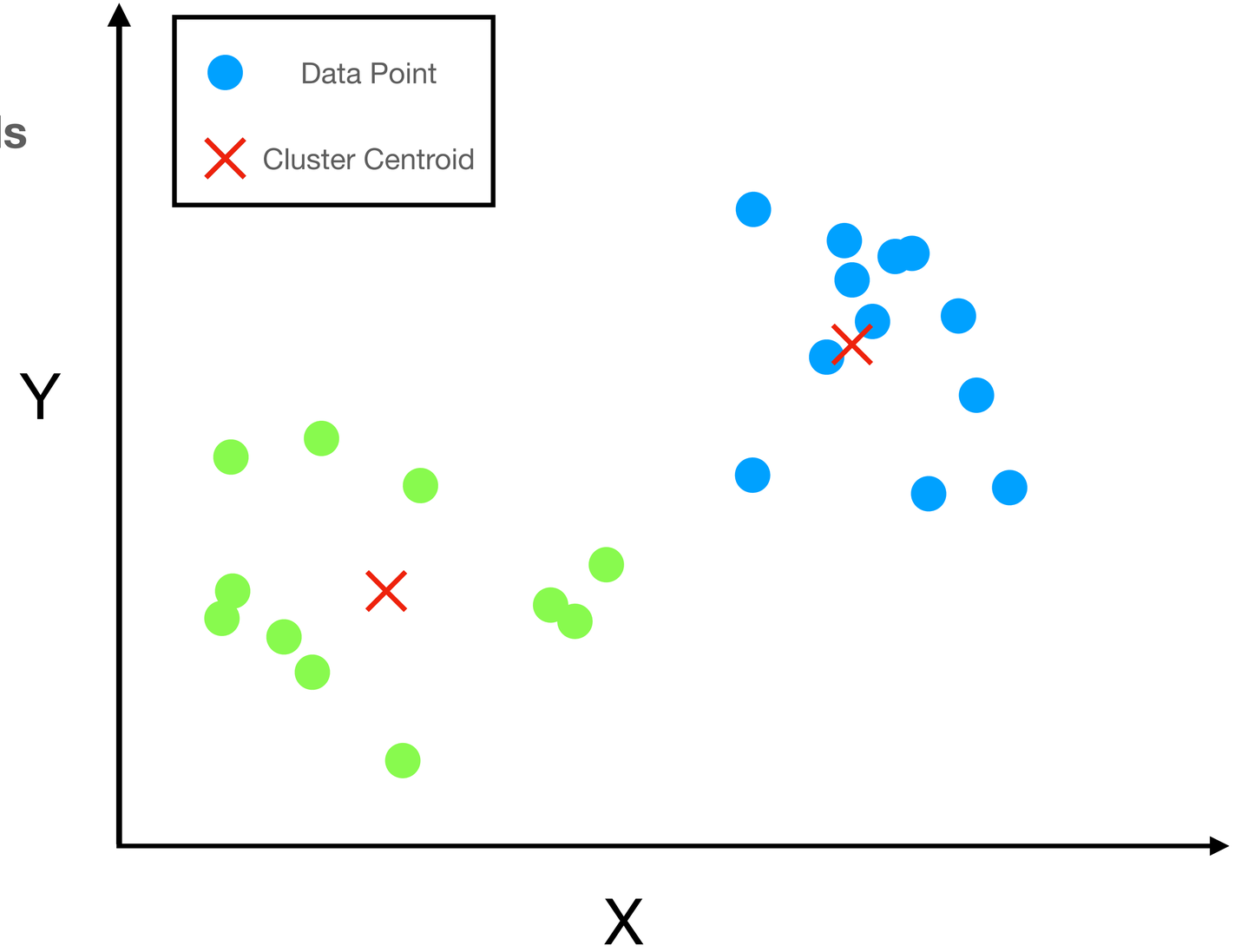
Step 1: Define centroids
Step 2: Cluster data



Clustering: K-means

Lloyd,1982, IEEE

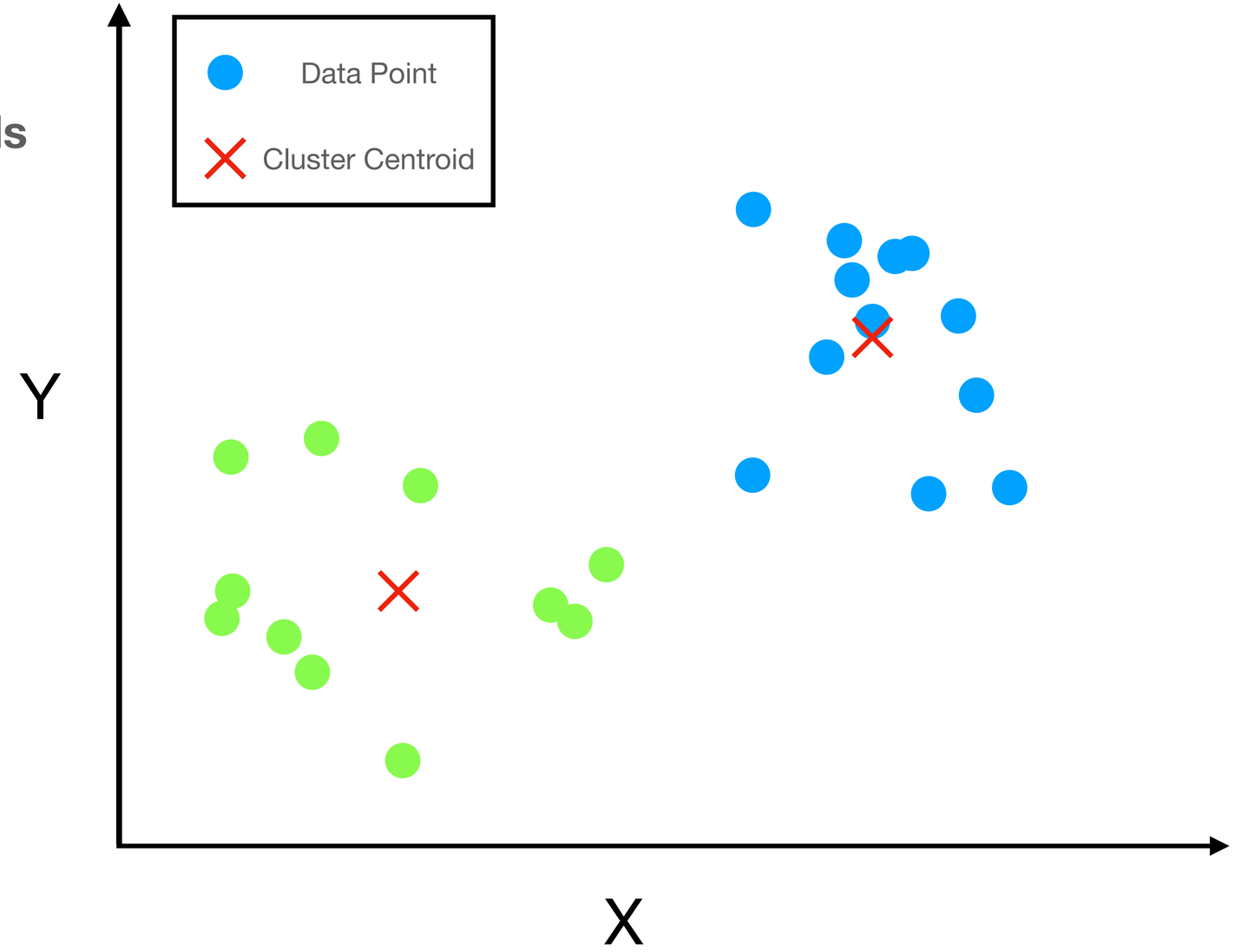
Step 1: Define centroids
Step 2: Cluster data



Clustering: K-means

Lloyd,1982, IEEE

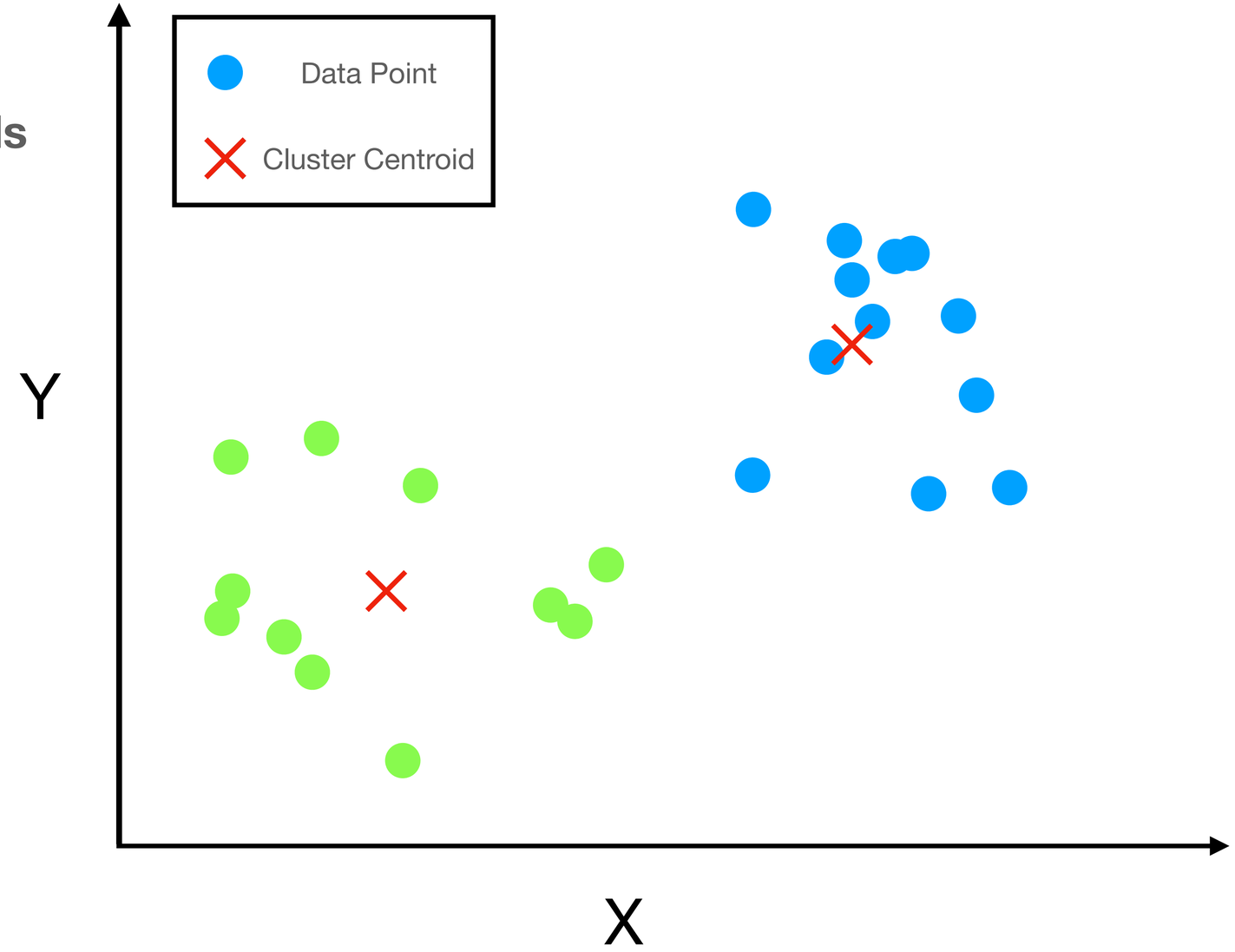
Step 1: Define centroids
Step 2: Cluster data



Clustering: K-means

Lloyd,1982, IEEE

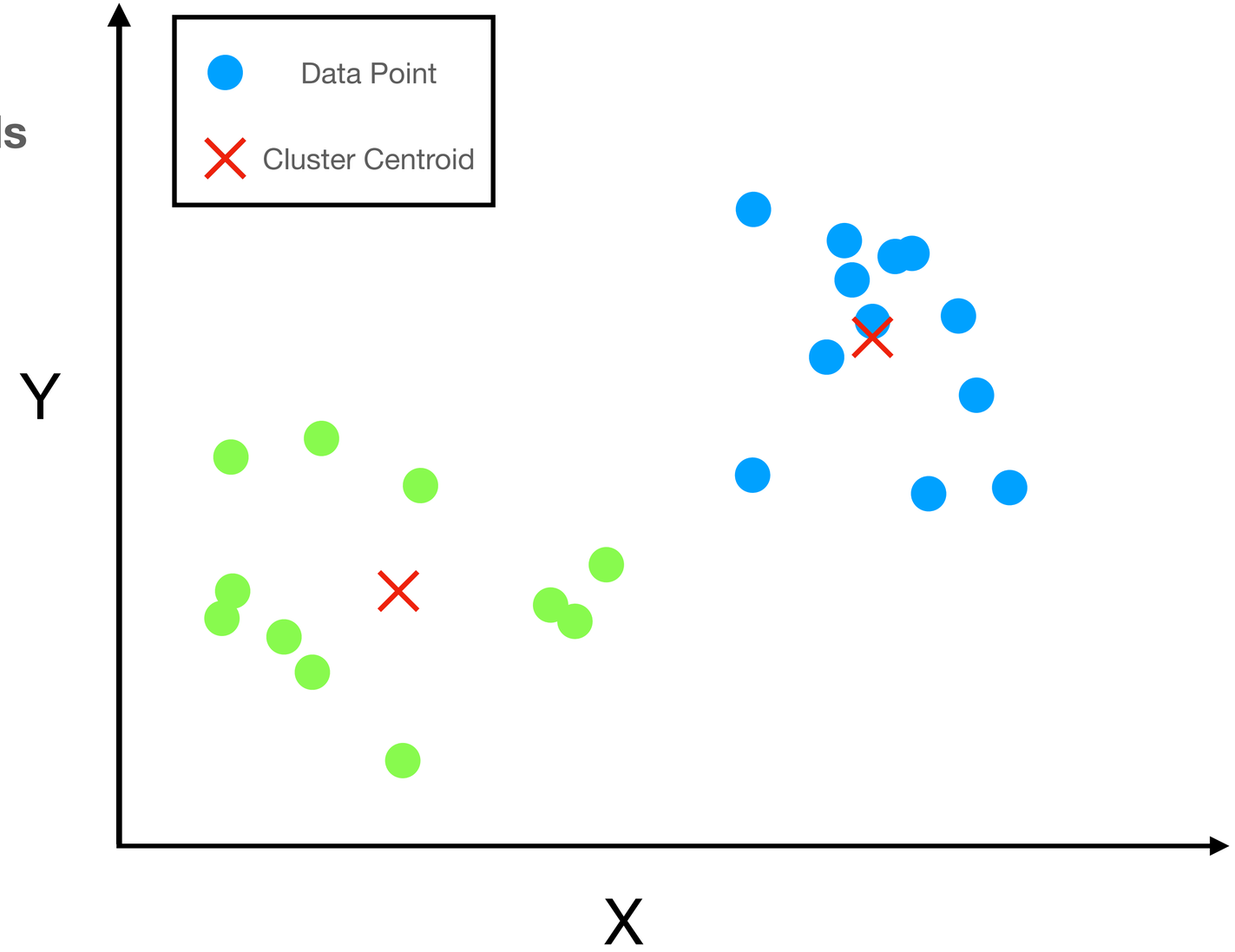
Step 1: Define centroids
Step 2: Cluster data



Clustering: K-means

Lloyd,1982, IEEE

Step 1: Define centroids
Step 2: Cluster data



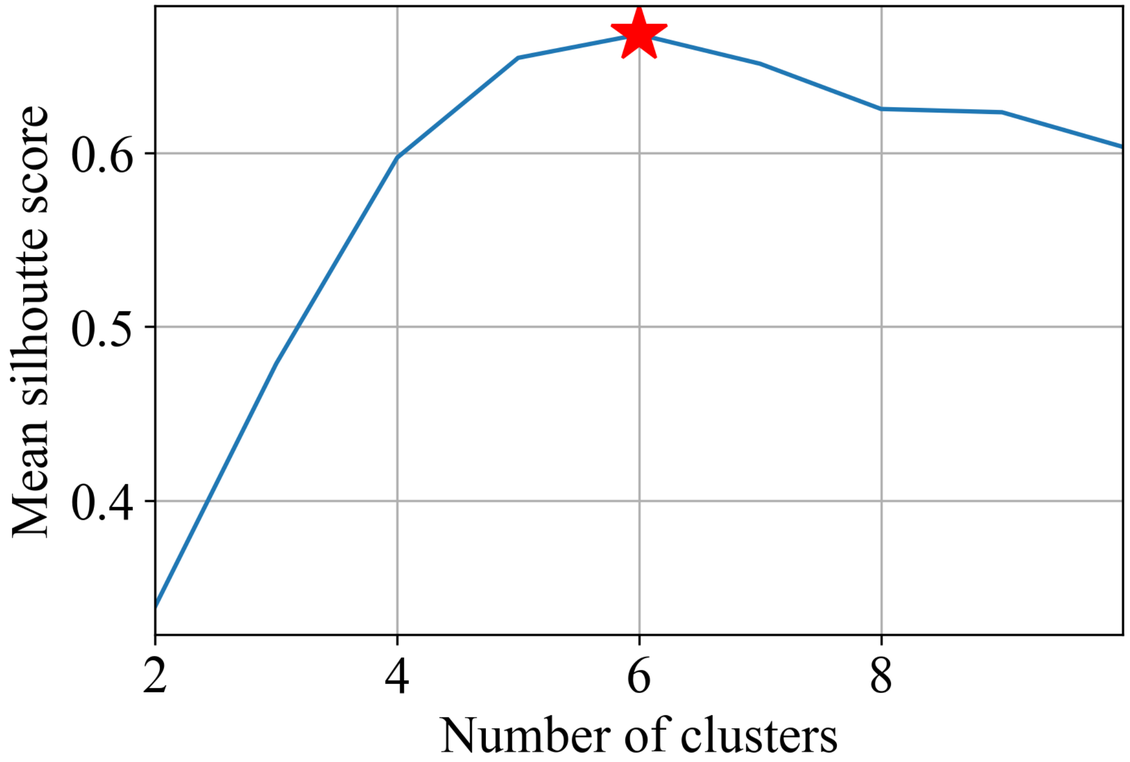
How many clusters: Silhouette Scores (SiS)

a : average Euclidean distance of x to every other data point in its own cluster

b : average Euclidean distance of x to every data point in the closest neighboring cluster

$$SiS(x) = \frac{b - a}{max(a, b)}$$

Rousseeuw 1987



Summary

- Unsupervised ML infers patterns in large data sets without the need for prior training labels
- Feature extraction and dimensionality reduction can aid clustering and interpretation
  - *Examples: Nonnegative matrix factorization (NMF), K-means clustering*

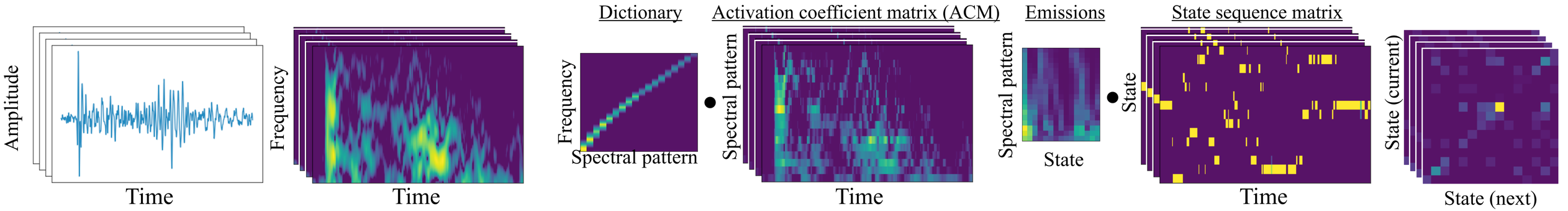
Outline

- **Introduction to unsupervised machine learning (UML)**
  - feature extraction & dimensionality reduction
  - clustering

- **SpecUFEx tutorial: Amatrice 2016**

Spectral Unsupervised Feature Extraction (SpecUFEx)

Holtzman et al., 2018; Sci Adv



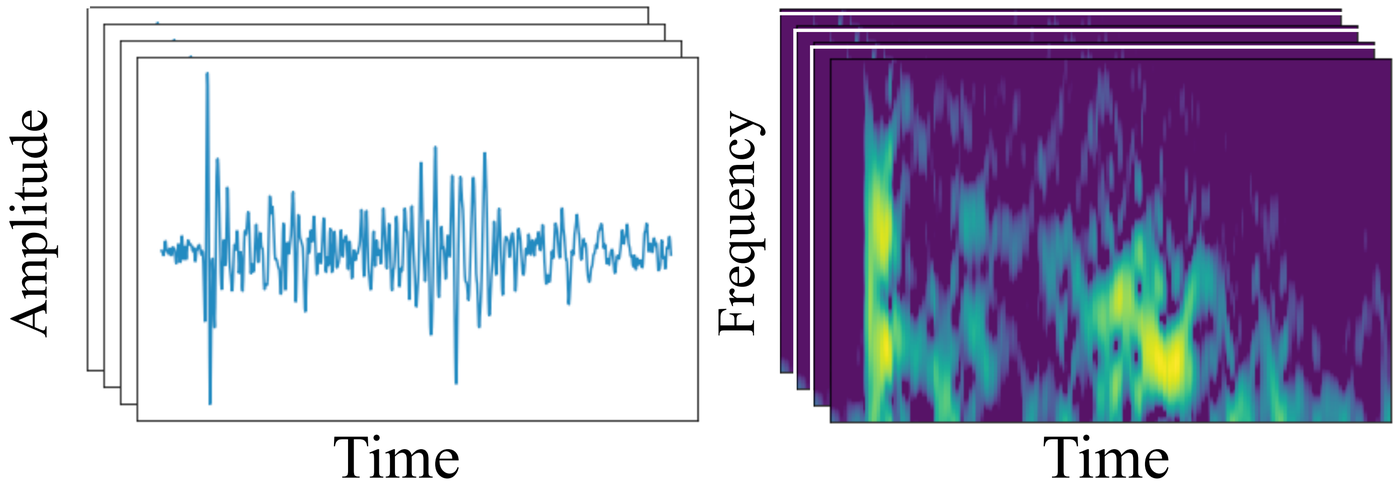
- J. Paisley, D. Blei and M.I. Jordan. Bayesian nonnegative matrix factorization with stochastic variational inference. In E.M. Airoldi, D. Blei, E.A. Erosheva & S.E. Fienberg (Eds.), Handbook of Mixed Membership Models and Their Applications. Chapman and Hall/CRC Handbooks of Modern Statistical Methods, 2014

- M. Hoffman, D. Blei, C. Wang and J. Paisley. Stochastic variational inference, Journal of Machine Learning Research, vol. 14, pp. 1303-1347, 2013.

Spectral Unsupervised Feature Extraction (SpecUFEx)

Holtzman et al., 2018; Sci Adv

Generate spectrogram

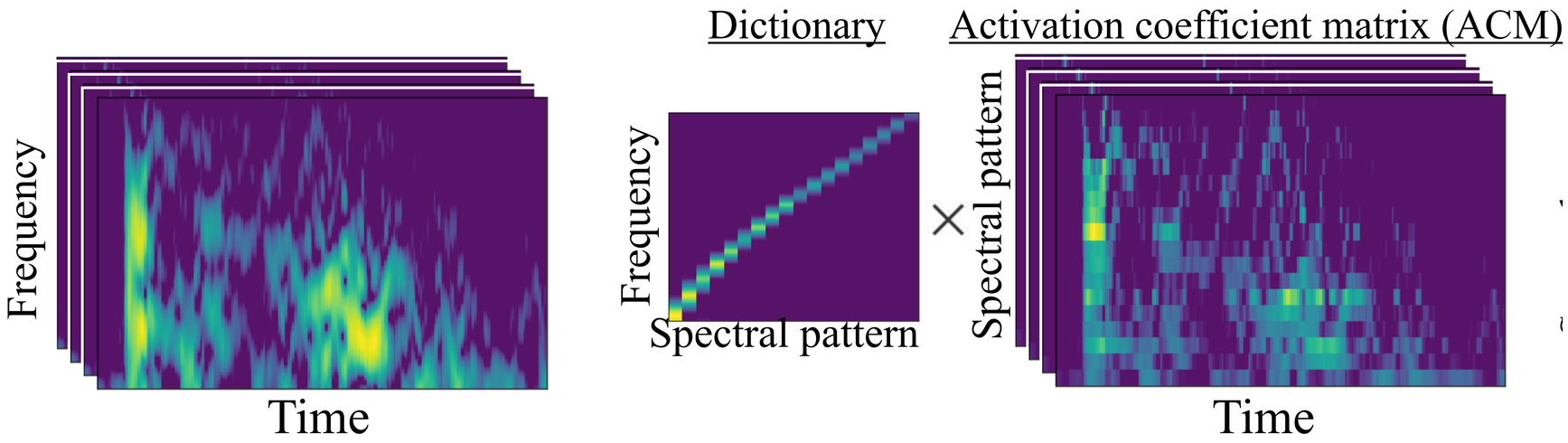


Ground amplitude through time ->
Frequency content through time

Spectral Unsupervised Feature Extraction (SpecUFEx)

Holtzman et al., 2018; Sci Adv

Non-negative Matrix Factorization



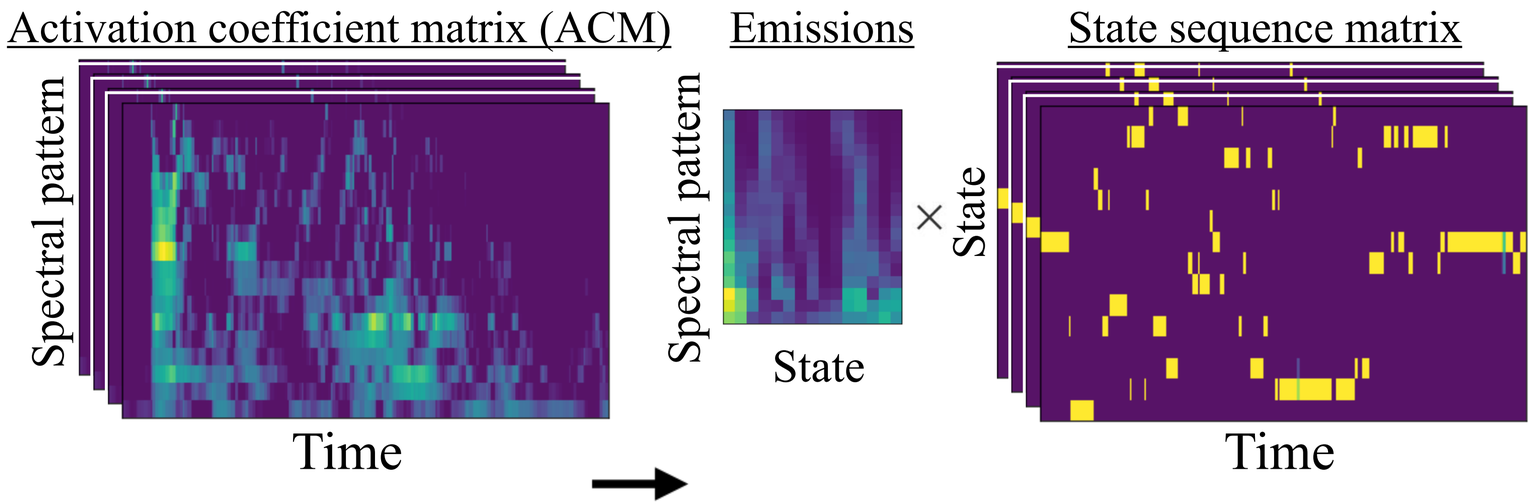
Extract spectral "patterns" (frequency distributions) through time

Spectrograms (units in dB) → Non-negative Matrix Factorization (NMF)

Spectral Unsupervised Feature Extraction (SpecUFEx)

Holtzman et al., 2018; Sci Adv

Hidden Markov Model

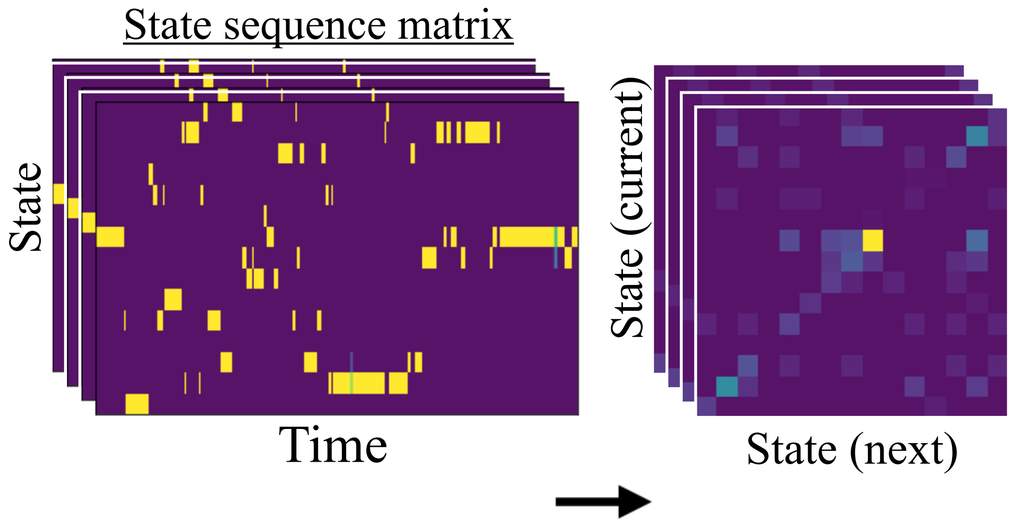


Hidden states control concurrent spectral patterns

Spectral Unsupervised Feature Extraction (SpecUFEx)

Holtzman et al., 2018; Sci Adv

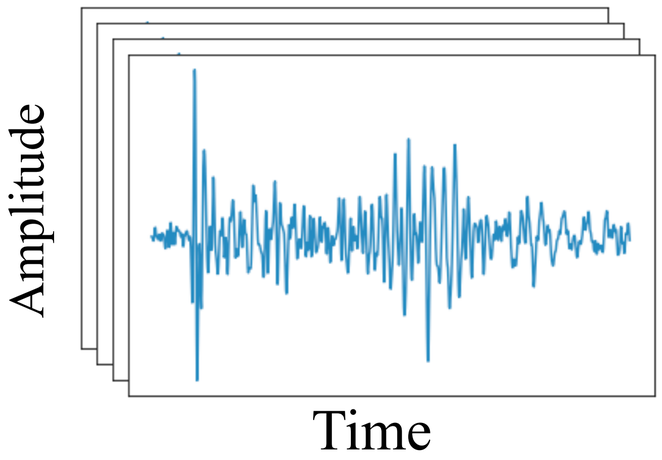
Fingerprints



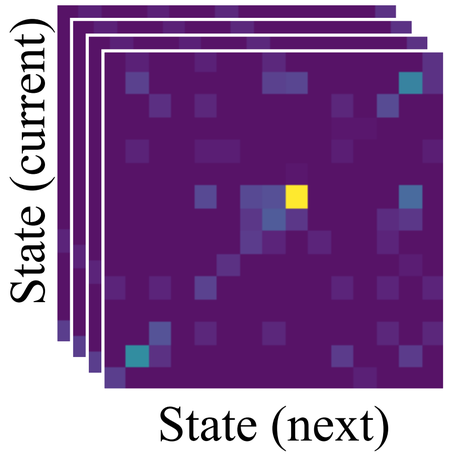
Count state transitions

SpecUFEx Workflow

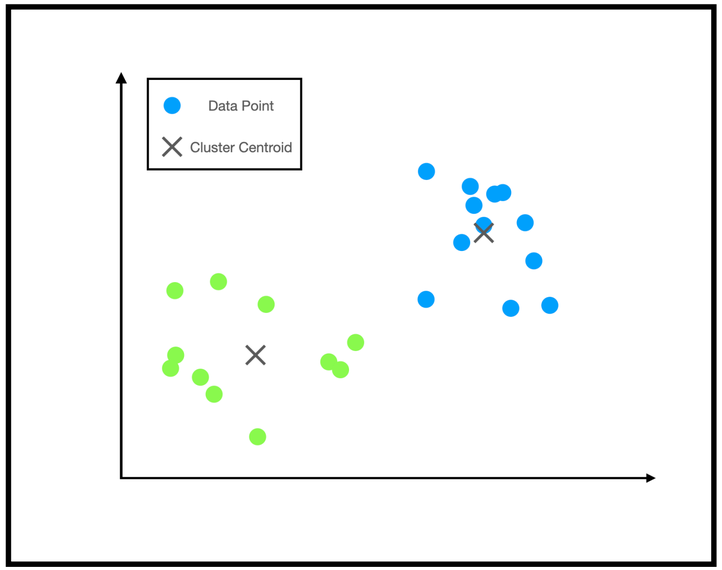
Data:
Earthquake waveforms



Feature Extraction:
NMF and hidden Markov model

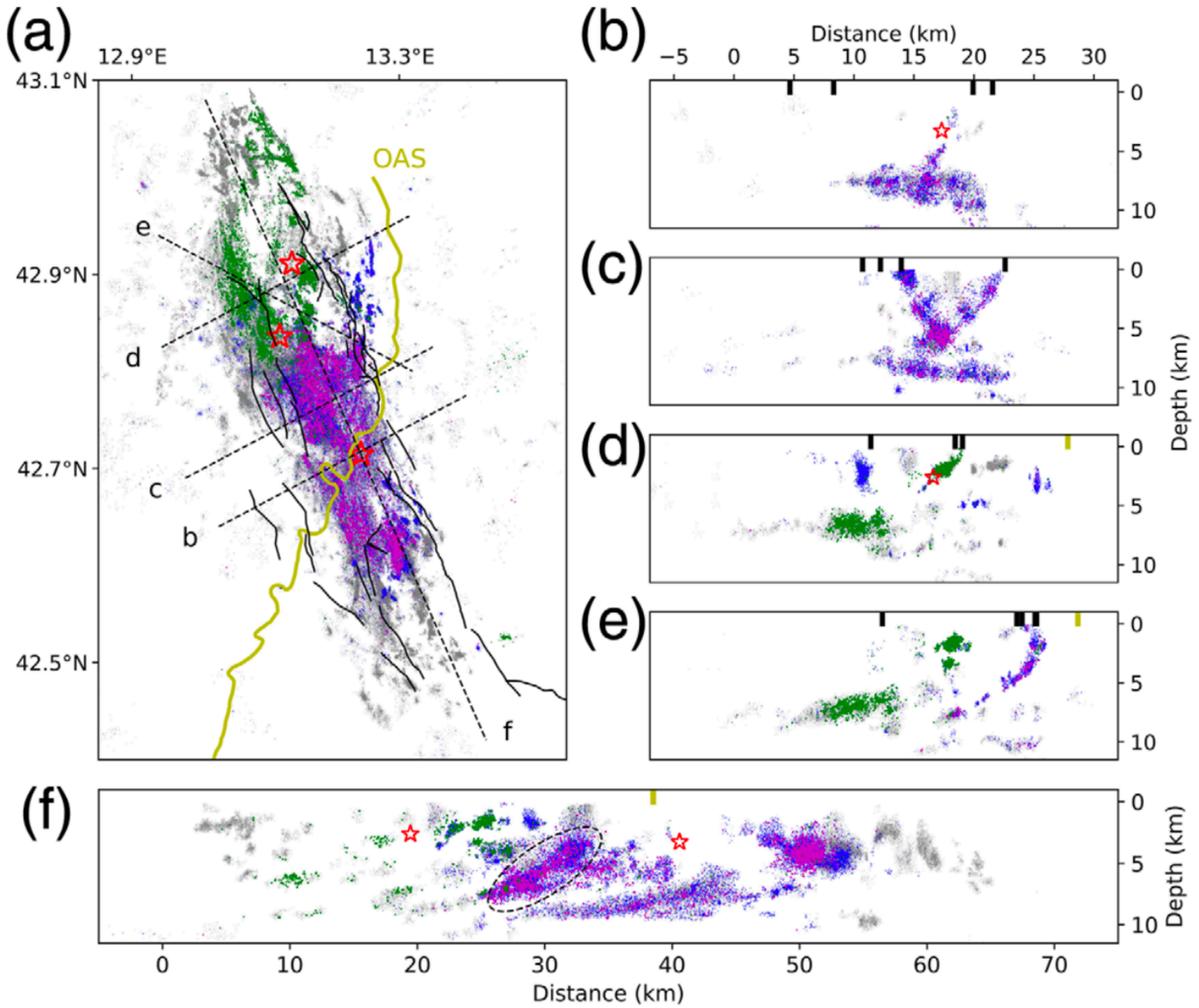


Clustering:
K-means with silhouette scores



Amatrice-SpecUFEx Tutorial

(subset = 6 days, ~1000 earthquakes)



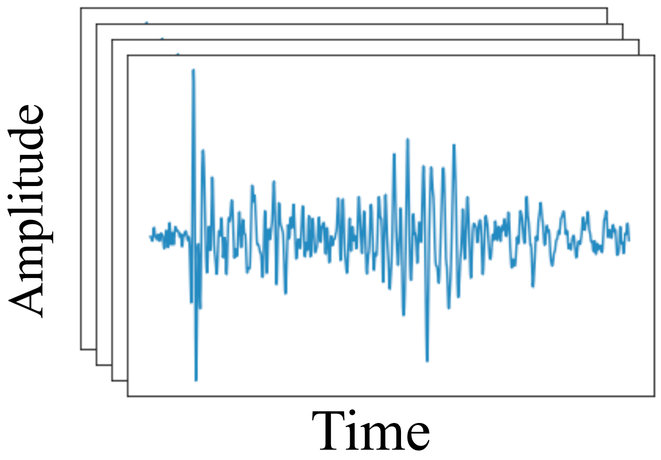
Tan et al., 2021; The Seismic Record

Amatrice-SpecUFEx Tutorial

Workflow

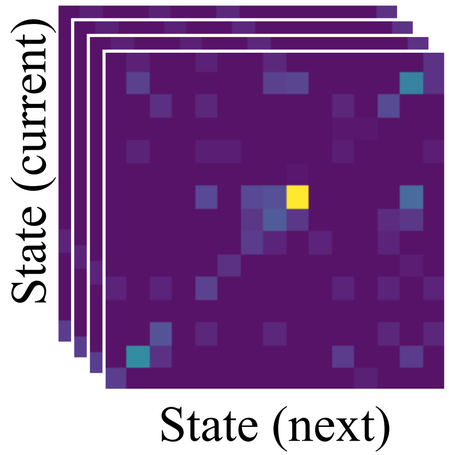
Data:

Earthquake waveforms



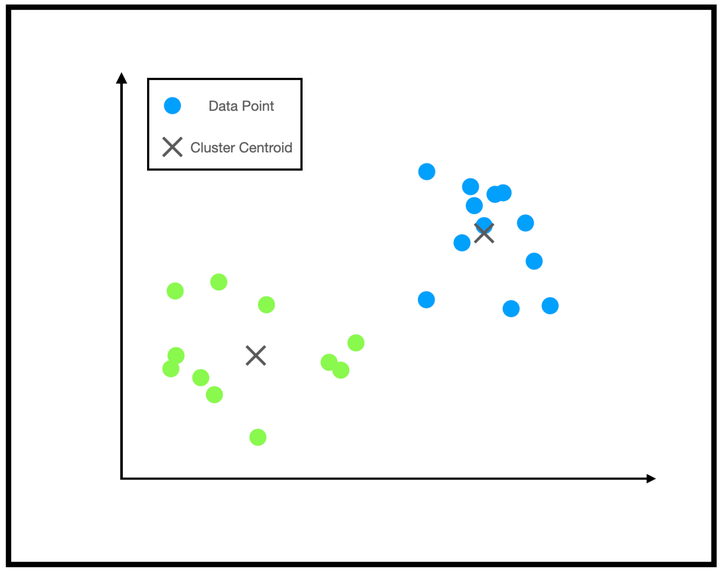
Feature Extraction:

NMF and hidden Markov model



Clustering:

K-means with silhouette scores



Tutorial Notebook (static)

sawilabs.com/cloud.html

These Slides

sawilabs.com/slides.pdf