

Unsupervised Learning in Seismology

René Steinmann

GFZ Discovery Fellow

Leipzig, 8 May 2025

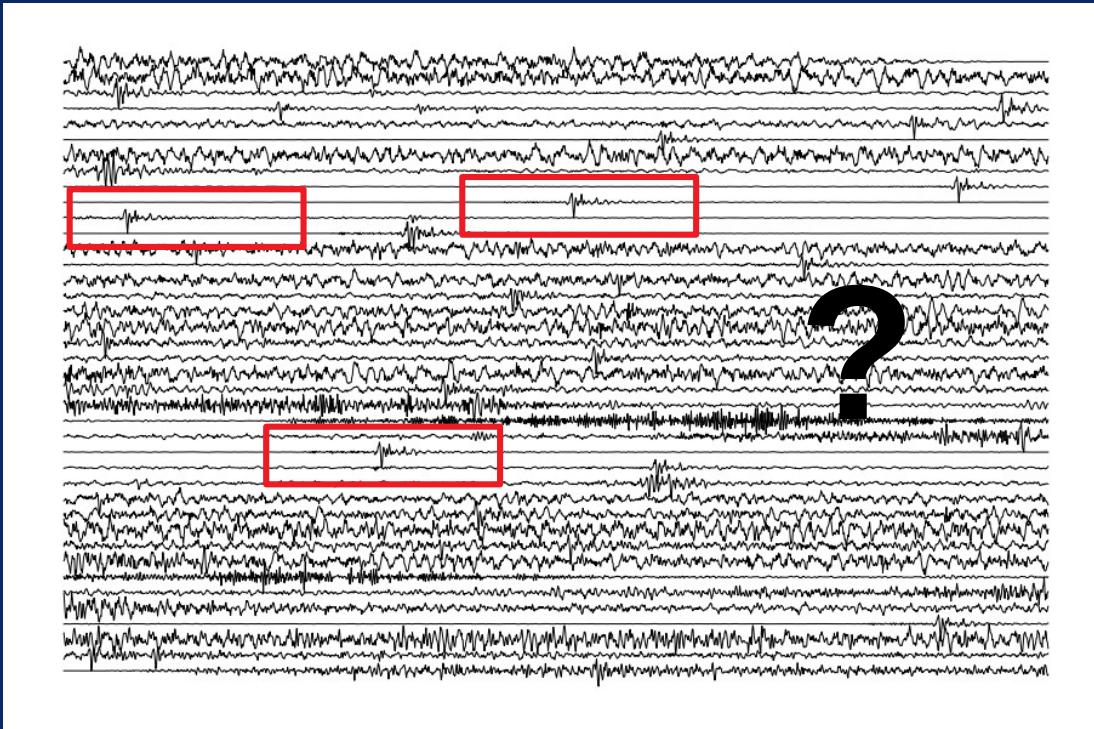
Outline

1. What is unsupervised learning and why is it interesting?
2. How do we do unsupervised learning?
3. Applications
4. Challenges & Solutions

Outline

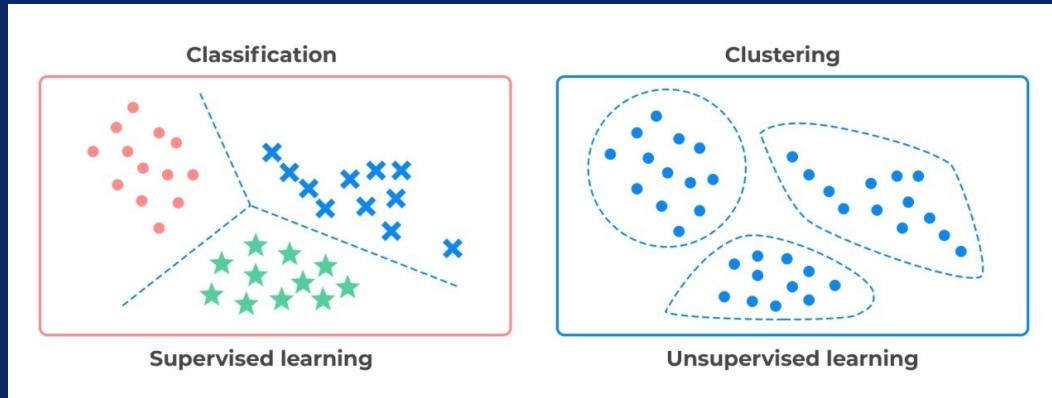
- 1. What is unsupervised learning and why is it interesting?**
2. How do we do unsupervised learning?
3. Applications
4. Challenges & Solutions

Why unsupervised learning?



20 min
What if we
have 10
years of
recordings?

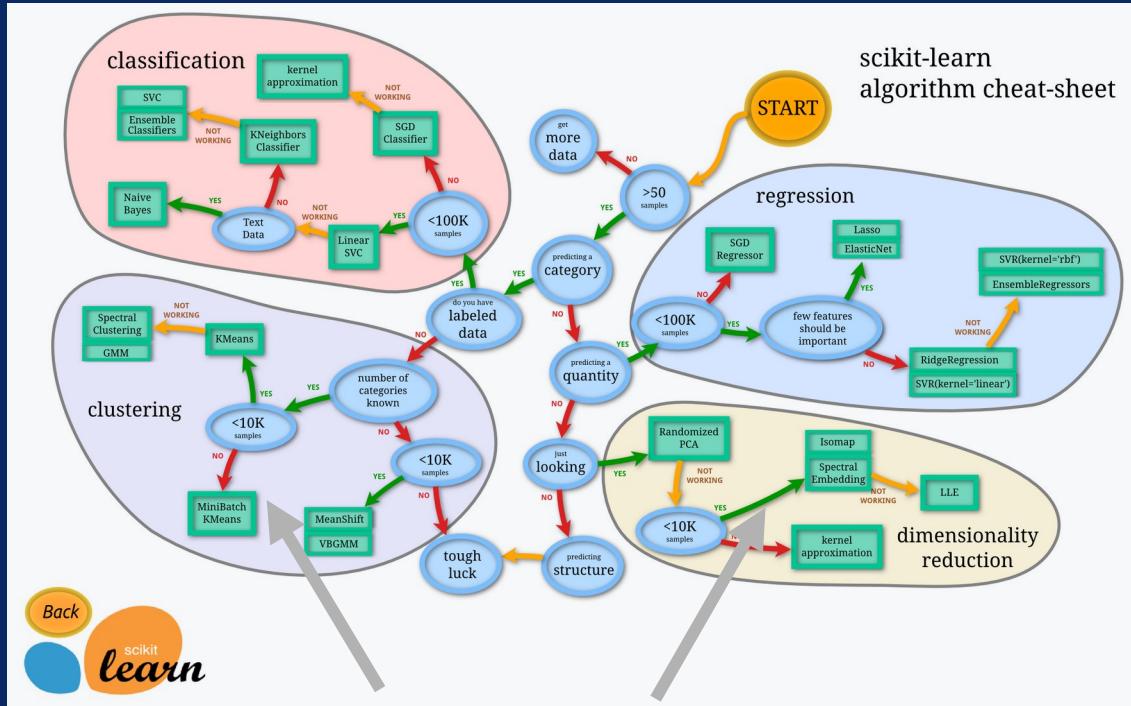
Unsupervised learning is about finding hidden patterns, structures, or groupings in data without using labels



<https://vocal.media/01/supervised-and-unsupervised-learning>

- We have labels
- Task: predict a category
- Earthquake vs. Noise
- No labels needed
- Task: find signal families
- Are there different type of noise sources?

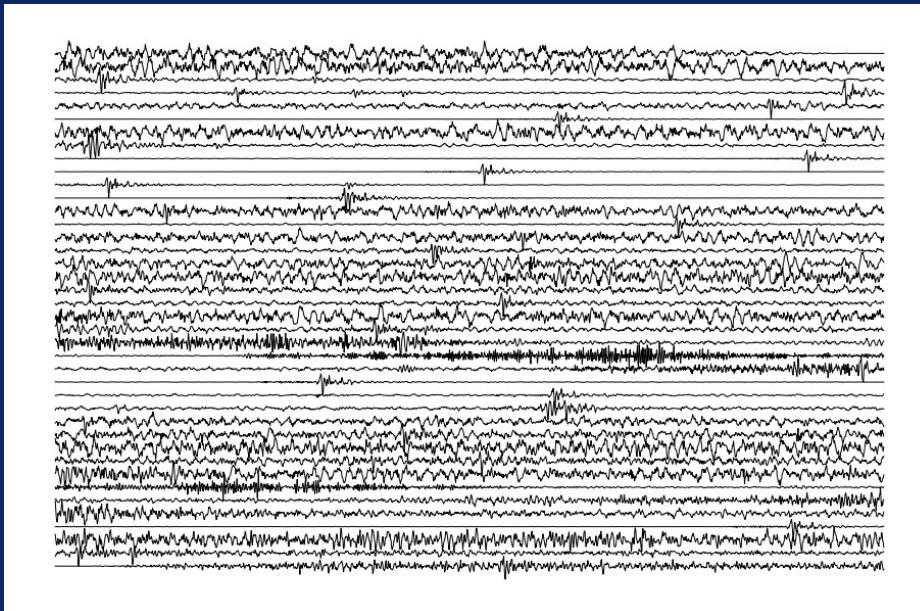
What methods of unsupervised learning exist?



Other methods:

- Anomaly detection
(e.g. Isolation Forest)

How could unsupervised learning be useful here?

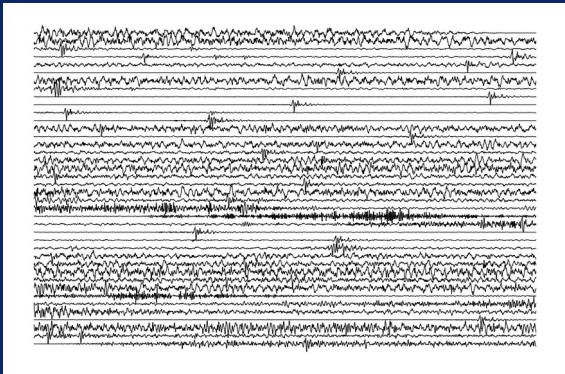


- **Clustering:** Finding groups of similar signals
- **Dimensionality reduction:** Visualize large time series
- **Anomaly detection:** any weird/rare signals in my data?

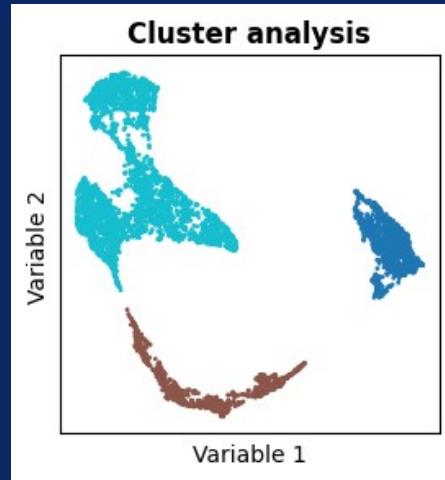
Outline

1. What is unsupervised learning and why is it interesting?
- 2. How do we do unsupervised learning?**
3. Applications
4. Challenges & Solutions

How do we do unsupervised learning (clustering)?

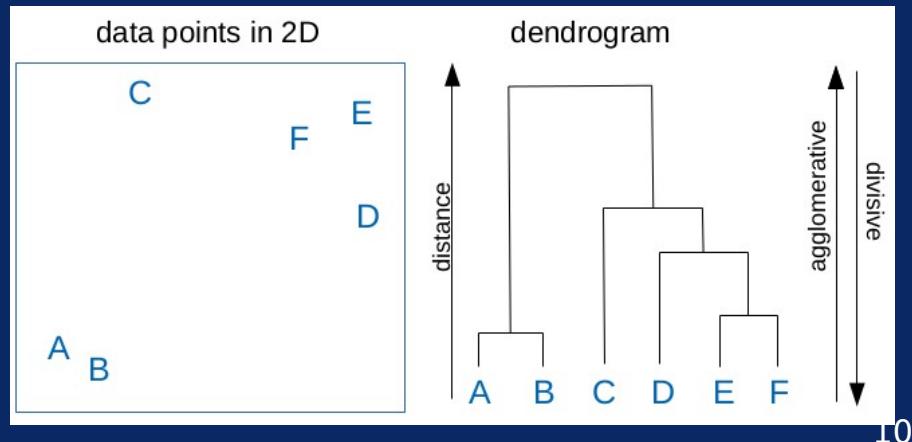
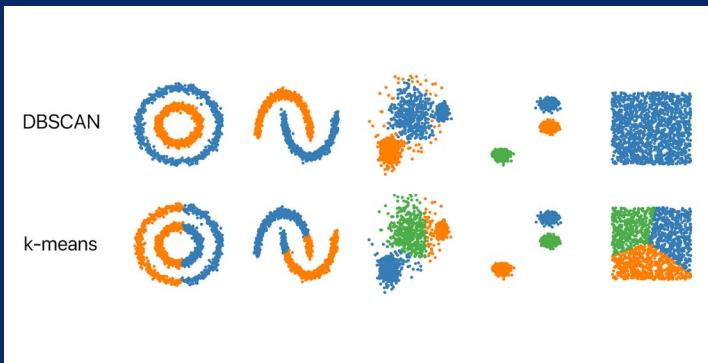
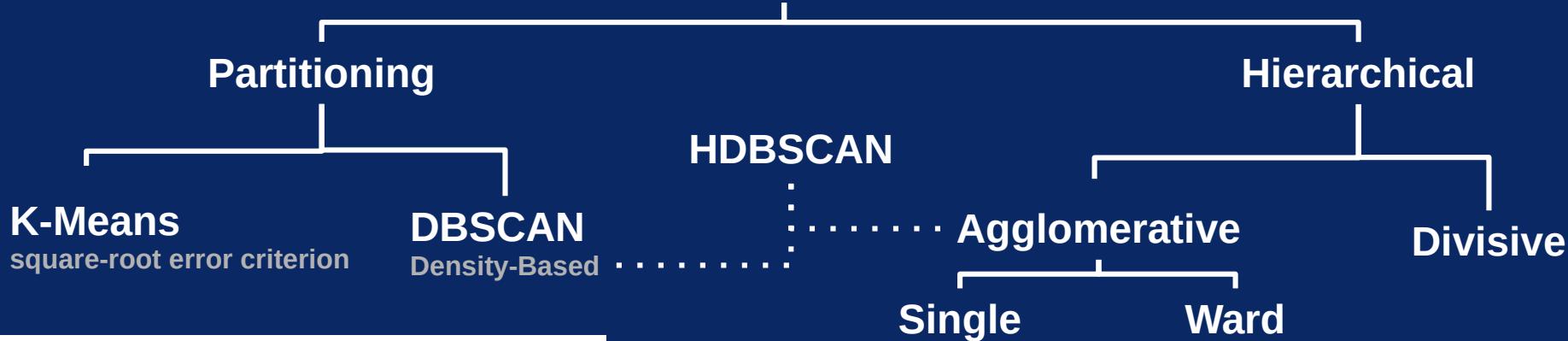


How do we get there?

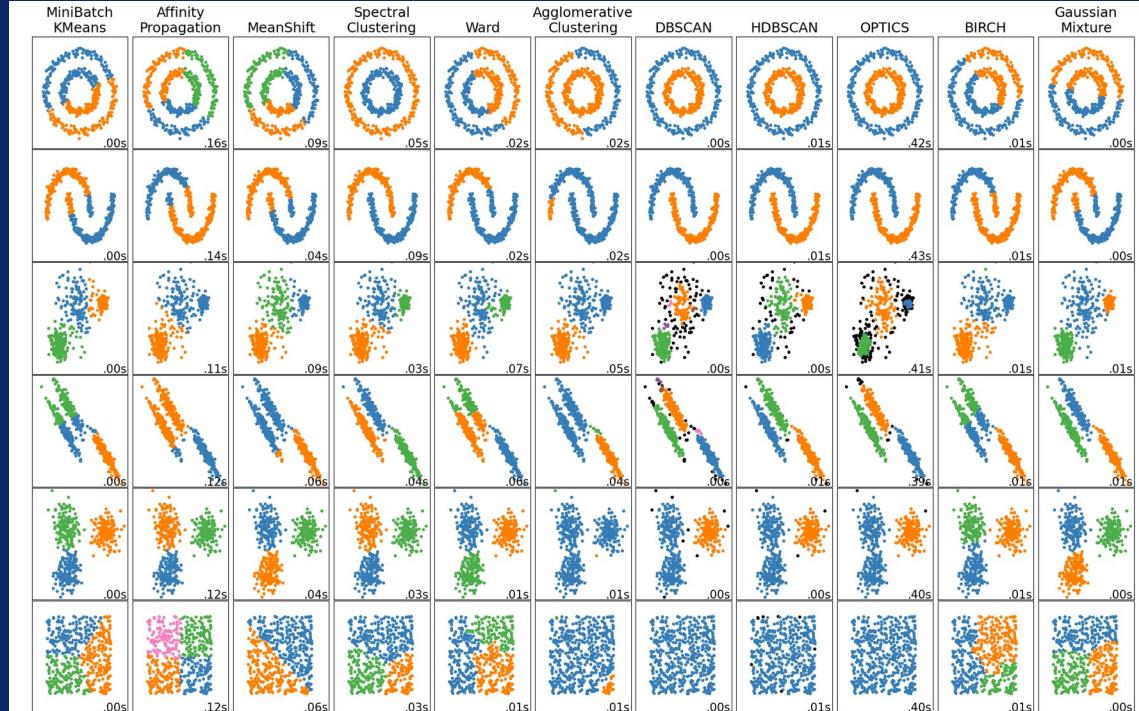


How does clustering work?

A rough overview of clustering methods



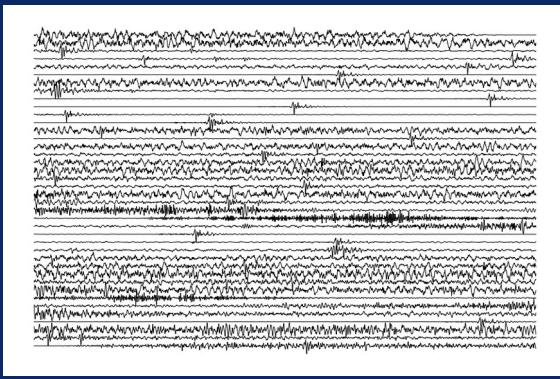
Clustering Overview in Scikit-Learn



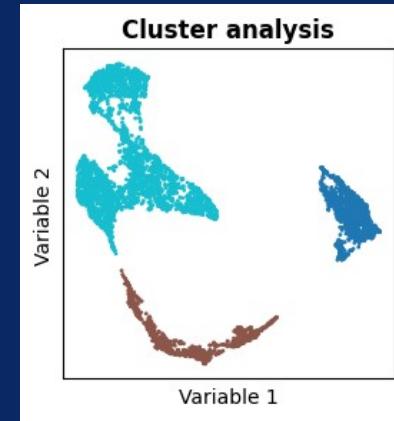
Understanding K-Means

Understanding DBSCAN

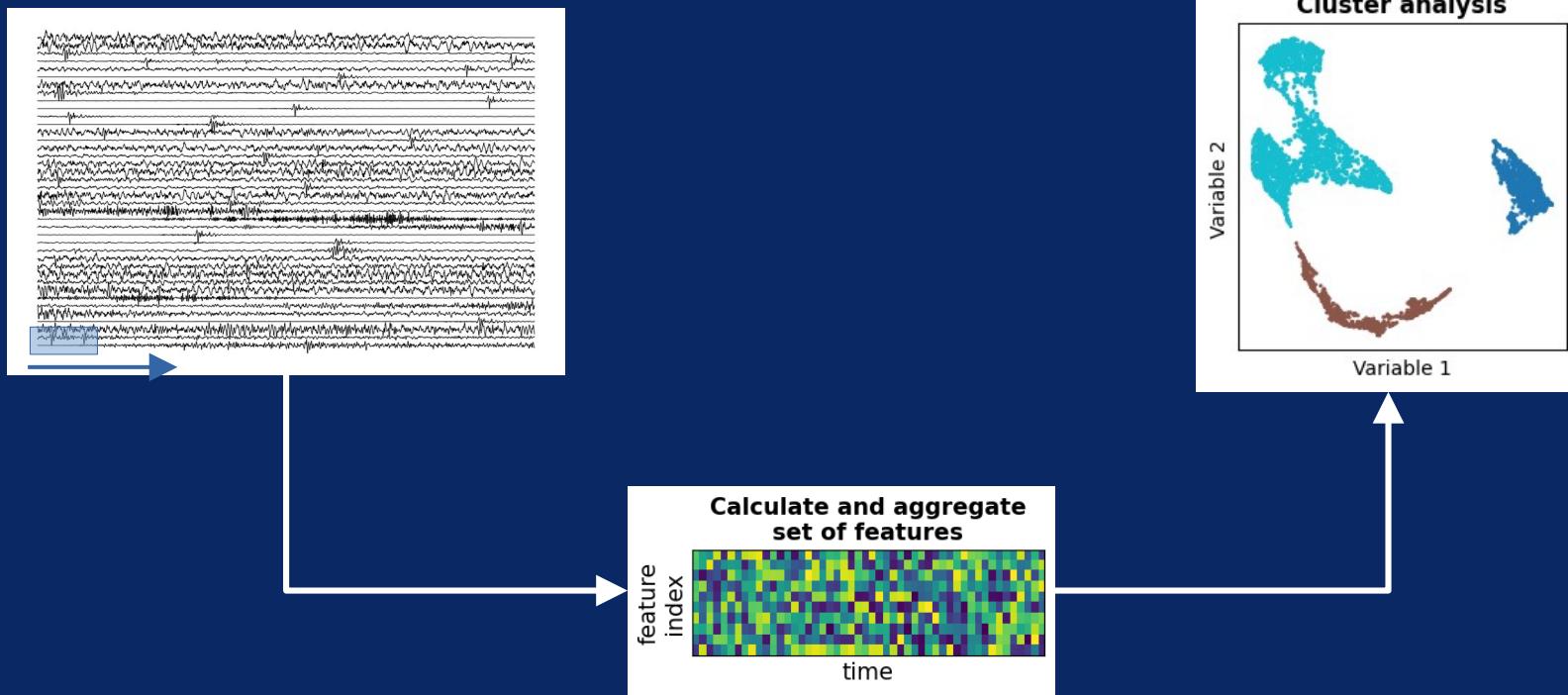
How do we do unsupervised learning (clustering)?



How do we get there?



How do we do unsupervised learning (clustering)?



Why not working directly with the time series data?



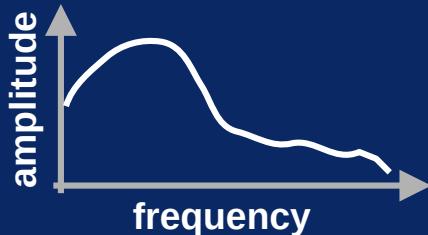
- ✗ **Time-invariant**
- ✗ **High-dimensional (curse of dimensionality)**
- ✗ **Sensitive to small changes**

What options of data representation do we have?

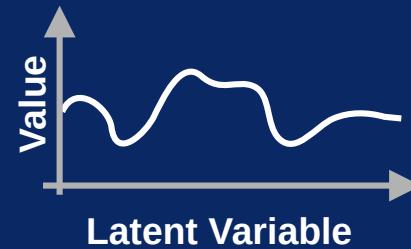
Hand-designed Features
 $f_{\max} = 1 \text{ Hz}$
Peak Amplitude = 10^{-7} m
Energy = 10^{-4} m^2
...



Frequency representation
(spectral or wavelet coefficients,
Mel Frequency cepstral coefficients)



Learned representation
(auto-encoders, contrastive learning)



easy to interpret
low performance

hard to interpret
high performance

How does an auto-encoder look like?

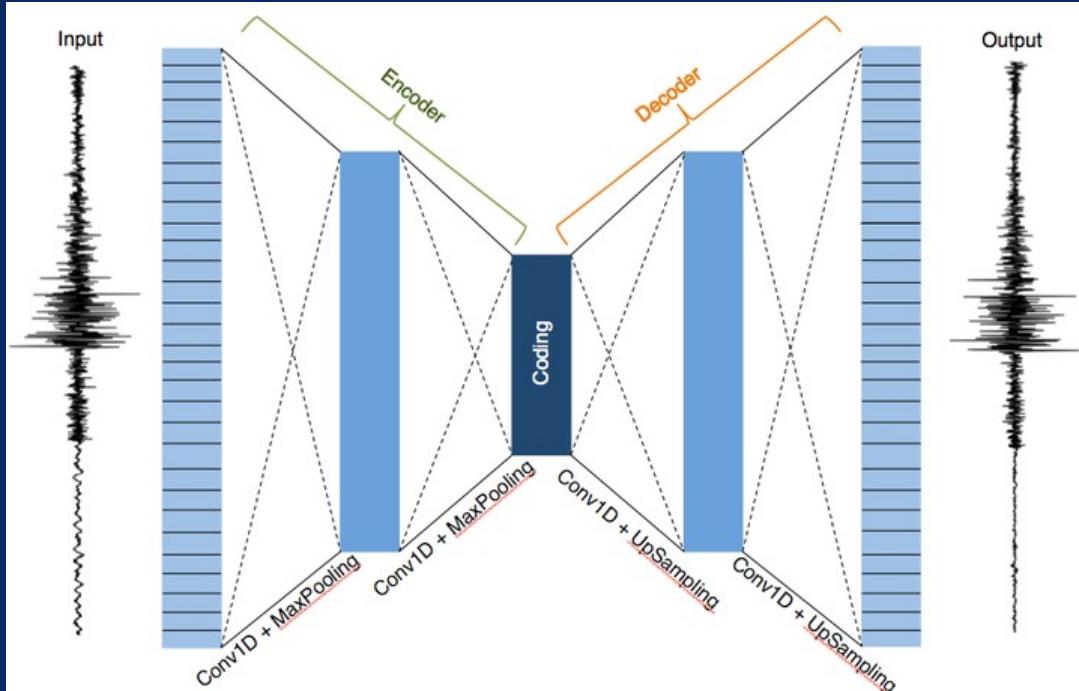
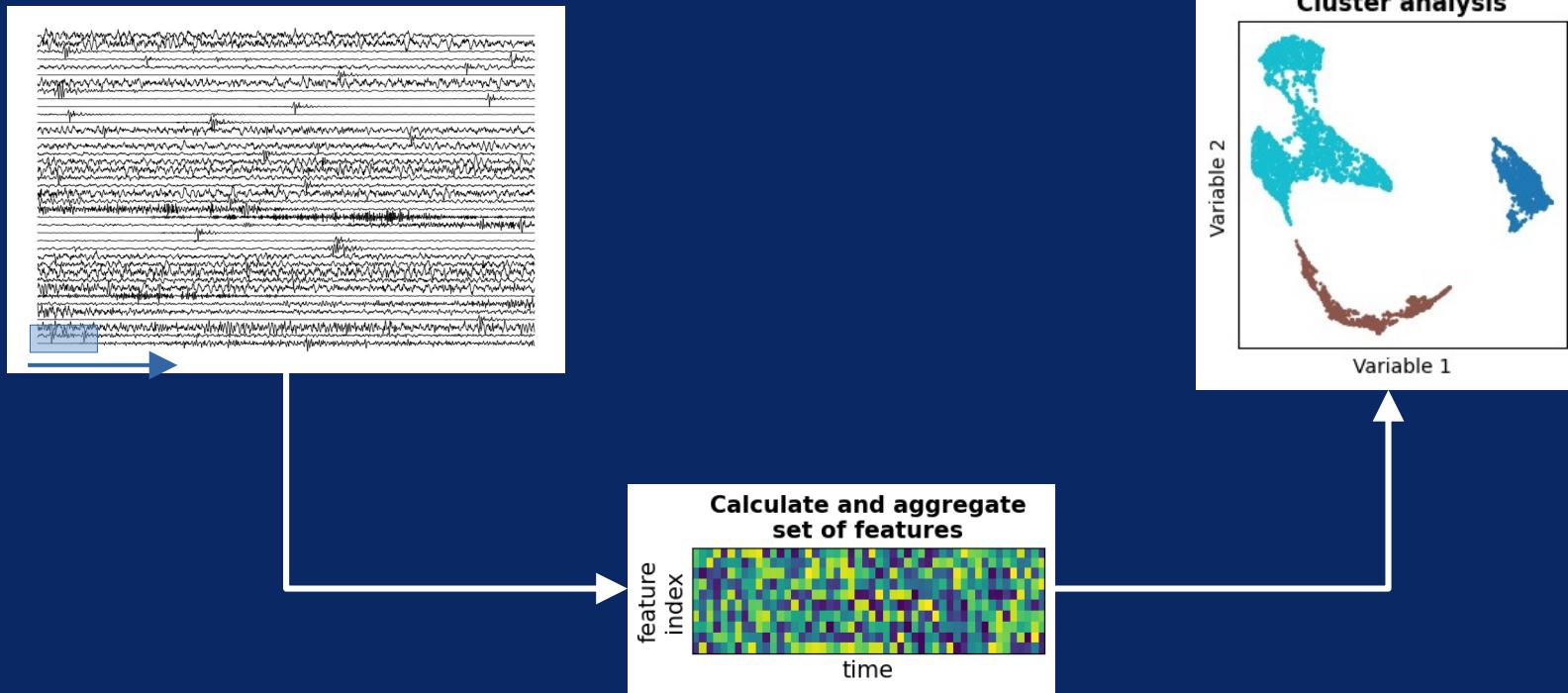


Image: <https://ngrayluna.github.io/project/denoiser/>

- Coding = Latent Variables
- Learning Objective = Reconstruction of input (no labels)
- Classic use-case: denoising

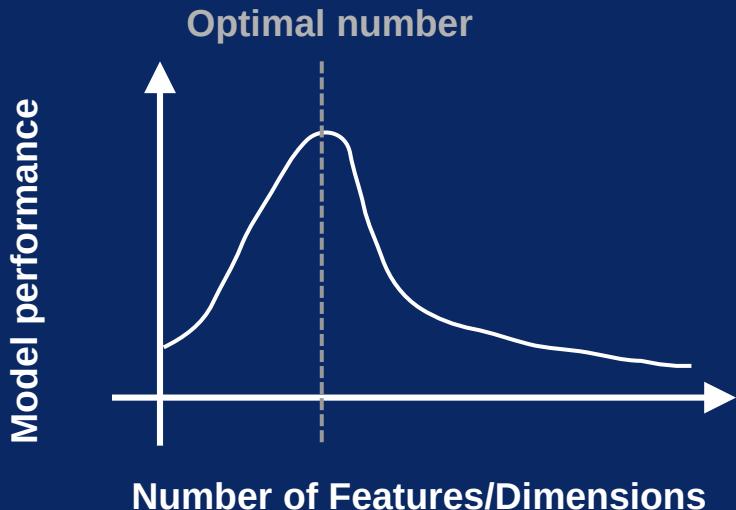
How do we do unsupervised learning (clustering)?



Often high-dimensional

Why do we need to reduce dimension?

The curse of dimensionality



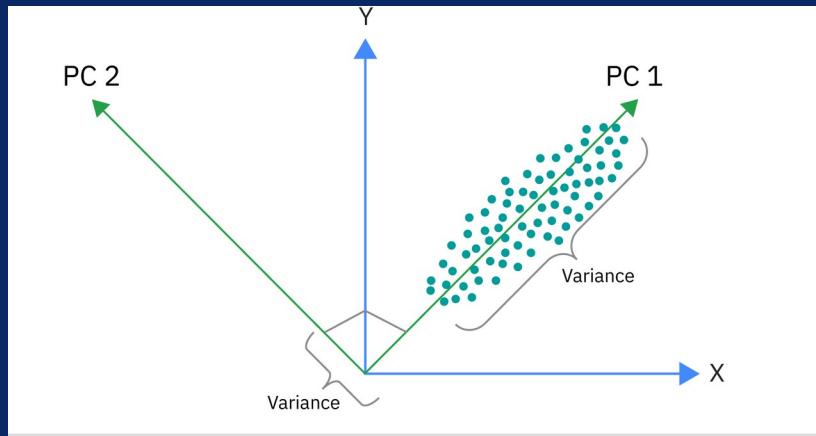
Why?

- distances become meaningless
- data points become sparse

How to reduce the dimensions?

- Principal Component Analysis (PCA)
- Manifold learning (T-SNE, UMAP)
- ...

Principal Component Analysis (PCA)



<https://www.ibm.com/think/topics/principal-component-analysis>

- PCA finds the directions in which the data varies the most
- Provides the eigenvectors (direction) and the eigenvalues (variance)
- Principal components are ranked according to variance
- Principal components are orthogonal to each other
- Considered a linear method

Principal Component Analysis (PCA)

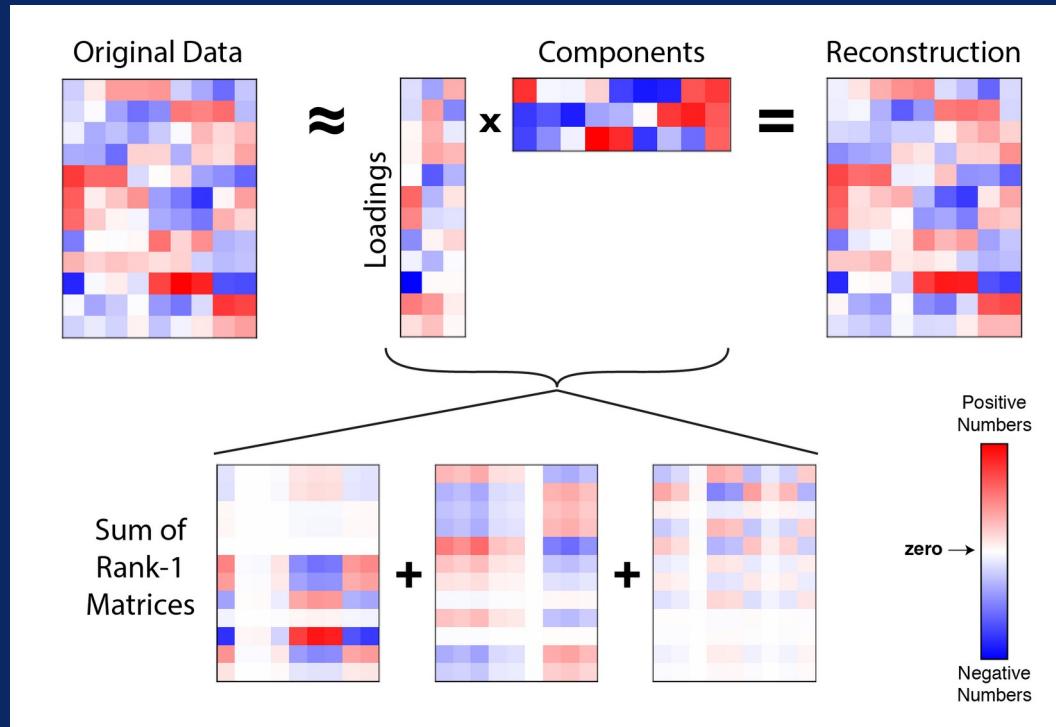
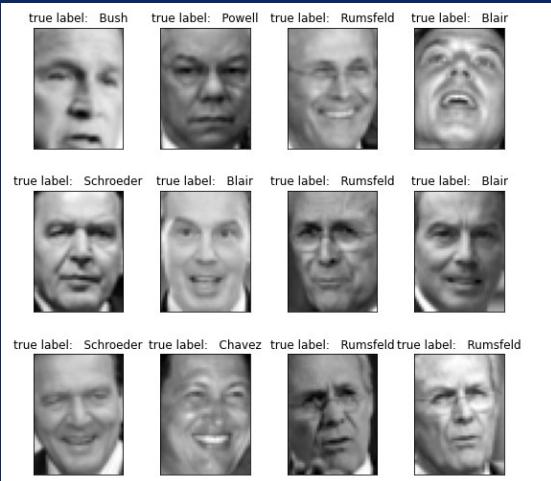


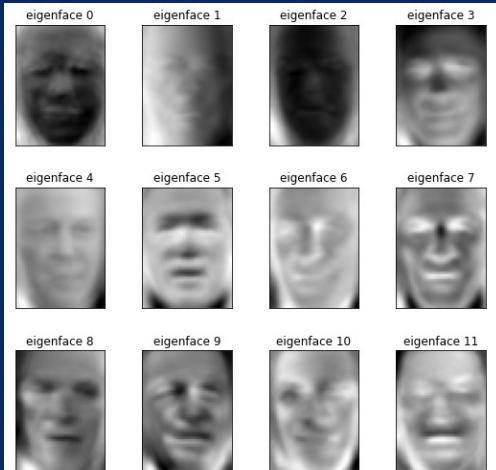
Image: <https://alexhwilliams.info/itsneuronalblog/2016/03/27/pca/>

PCA in Facial Recognition

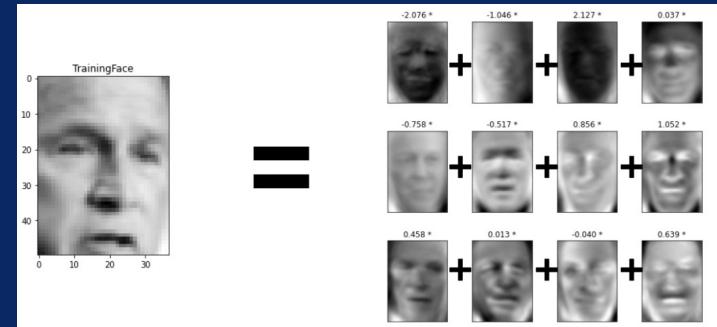
Dataset of Faces



The Eigenfaces (Principal Components)

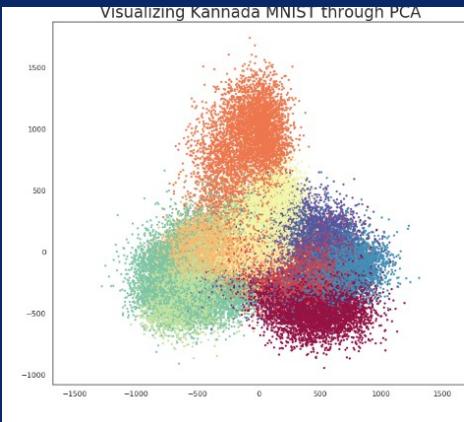


Reconstruction of input data

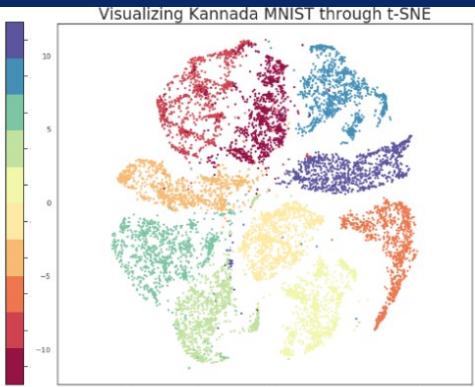


Sometimes PCA fails . . .

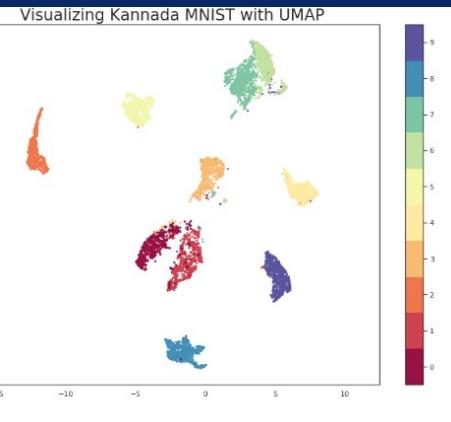
PCA



Visualizing Kannada MNIST through t-SNE



UMAP (Uniform Manifold Approximation and Projection)

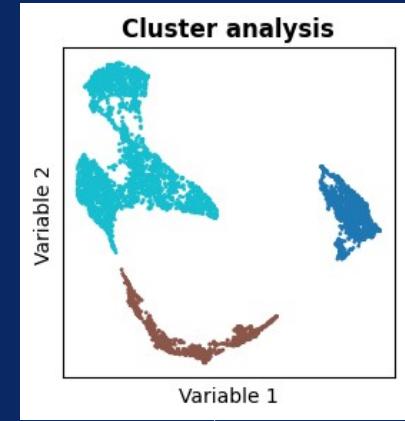
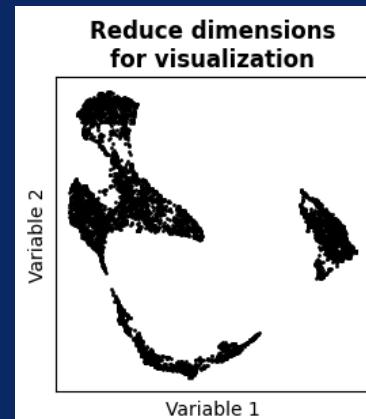
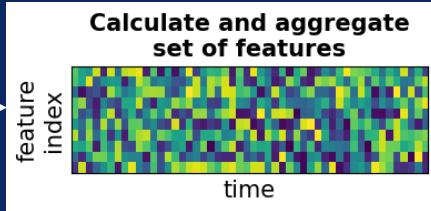
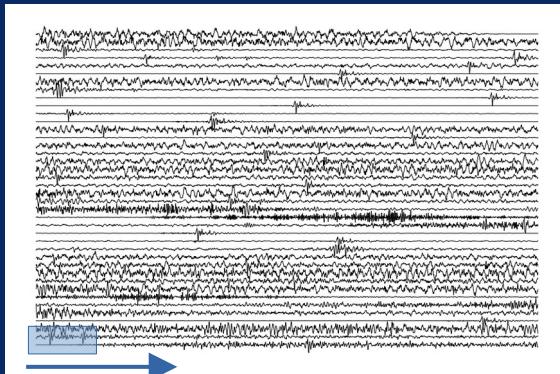


T-SNE (t-distributed Stochastic Neighbor Embedding)

	PCA	t-SNE	UMAP
Type	Linear	Nonlinear, probabilistic	Nonlinear, topological
Preserves	Global structure	Local structure	Local and Global
Dimensionality	Any	Mostly 2 or 3	Any (often 2 or 3)
Axes	Meaningful, ranked	arbitrary	arbitrary
Clustering?	Sure	Difficult	Less difficult

Regular framework for unsupervised learning in seismology

Time series data



Outline

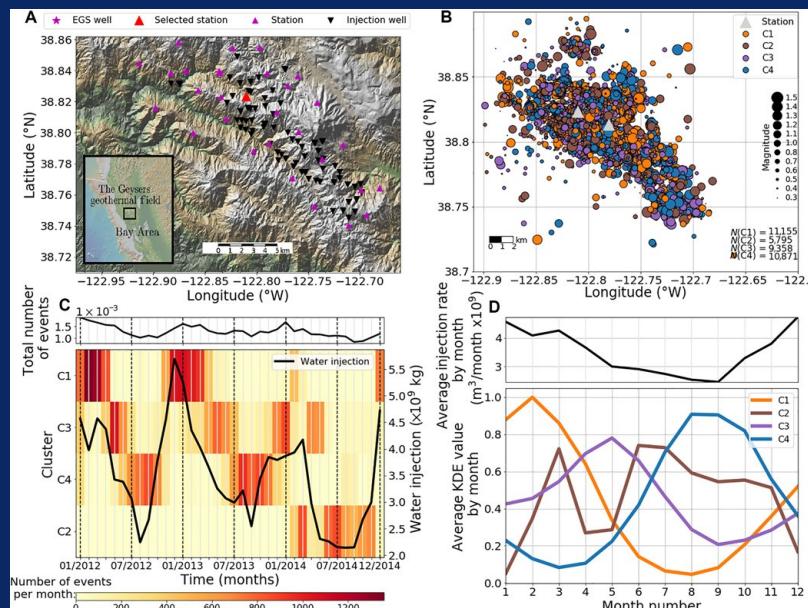
1. What is unsupervised learning and why is it interesting?
 2. How do we do unsupervised learning?
- 3. Applications**
4. Challenges & Solutions

Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field

BENJAMIN K. HOLTZMAN , ARTHUR PATÉ , JOHN PAISLEY, FELIX WALDHAUSER , AND DOUGLAS REPETTO Authors Info & Affiliation

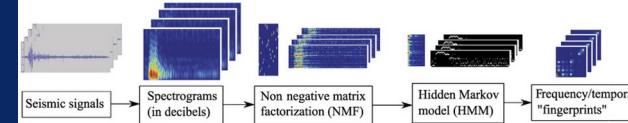
SCIENCE ADVANCES • 23 May 2018 • Vol 4, Issue 5 • DOI: 10.1126/sciadv.aao2929

Analysis/Interpretation of clustering output



Input: seismic waveforms from catalog

A ML method



B NMF

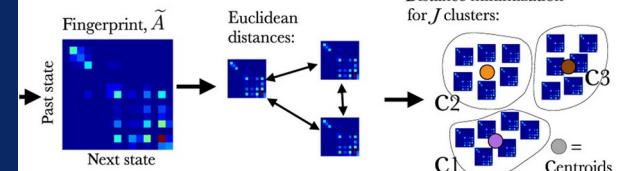
The diagram illustrates the computation of a spectrogram F (size $D \times M$) as a product of three components:

- Frequency**: The vertical axis of the spectrogram.
- Time**: The horizontal axis of the spectrogram.
- Dictionary of spectral patterns** (size $D \times K$): A sparse matrix where non-zero entries represent spectral patterns.
- Activation coefficient matrix** $\text{diag}(\mathbf{a})V$ (size $K \times M$): A matrix where each column is a scaled version of a vector \mathbf{a} .

The diagram shows the spectrogram F being approximated by the product of the dictionary and the activation matrix, with a residual term \mathbf{e} representing the difference between the original spectrogram and the reconstructed one.

• HMM:

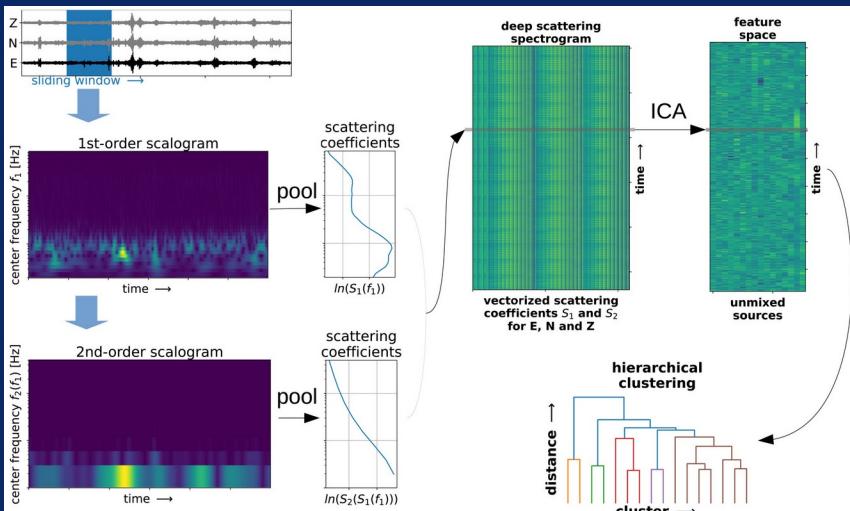
④ K-means clustering



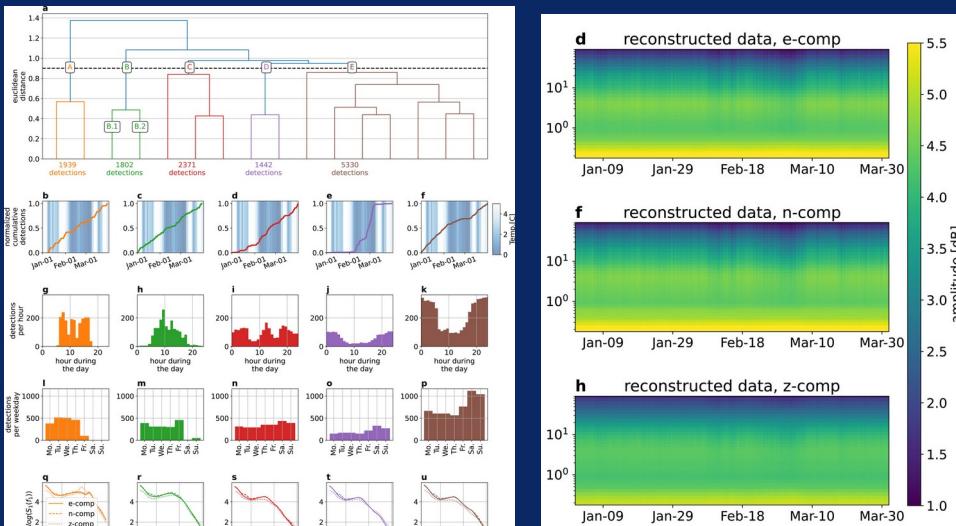
AI-Based Unmixing of Medium and Source Signatures From Seismograms: Ground Freezing Patterns

 René Steinmann , Léonard Seydoux, Michel Campillo

Input: continuous 3-comp seismograms

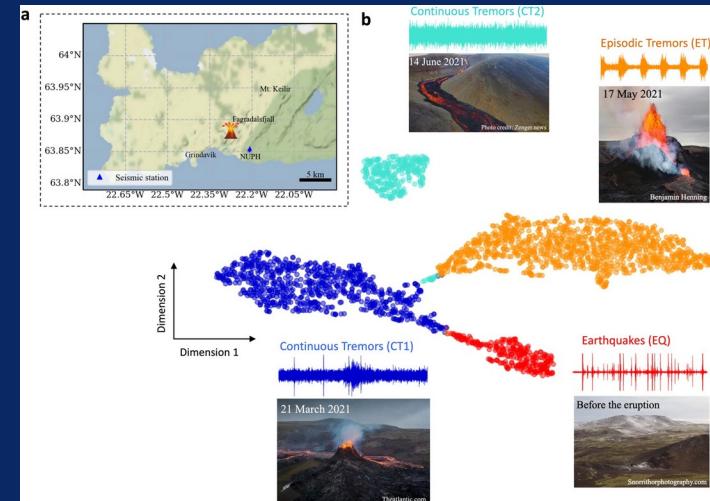
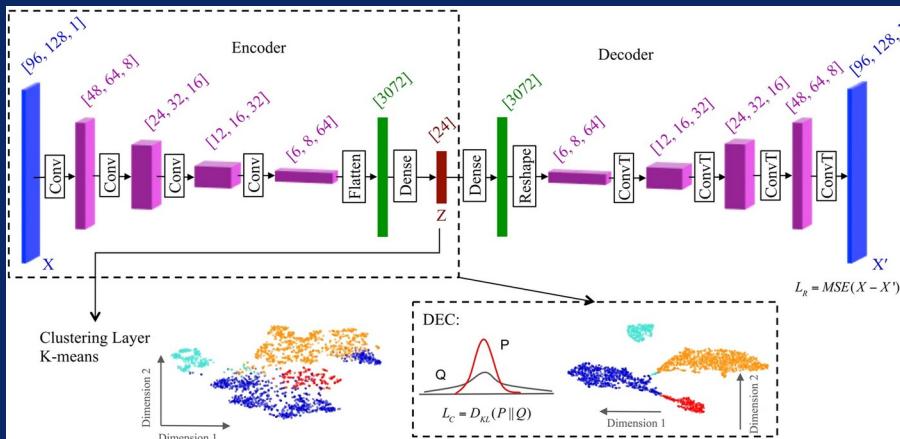


Interpretation of clusters together with „Eigenfaces“



Tremor clustering reveals pre-eruptive signals and evolution of the 2021 Geldingadalir eruption of the Fagradalsfjall Fires, Iceland

Zahra Zali , S. Mostafa Mousavi, Matthias Ohrnberger, Eva P. S. Eibl & Fabrice Cotton

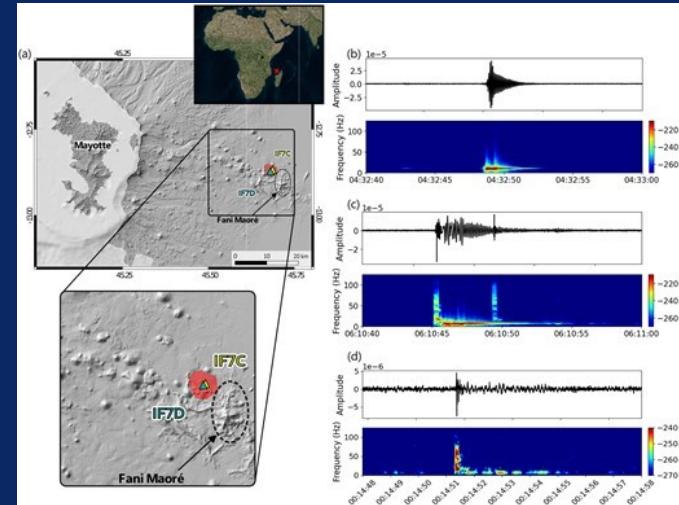


Self-supervised learning of seismological data reveals new eruptive sequences at the Mayotte submarine volcano ⚡

Joachim Rimpot ✉, Clément Hibert, Lise Retailleau, Jean-Marie Saurel, Jean-Philippe Malet, Germain Forestier, Jonathan Weber, Tord S Stangeland, Antoine Turquet, Pascal Pelleau

Geophysical Journal International, Volume 240, Issue 1, January 2025, Pages 1–12,

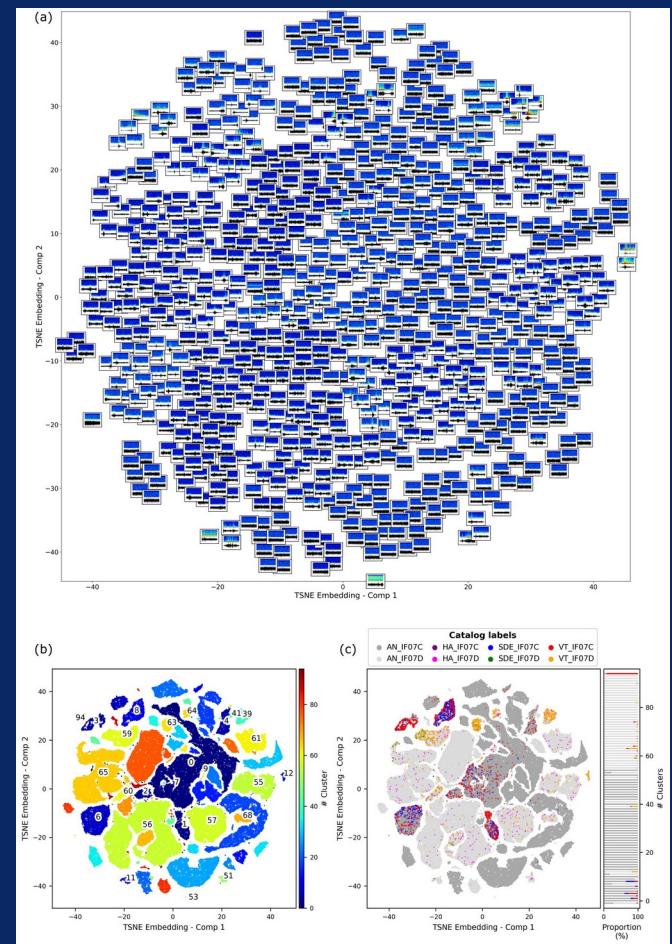
Input: continuous seismograms



Feature Extraction with Contrastive Learning

Dimension Reduction with T-SNE

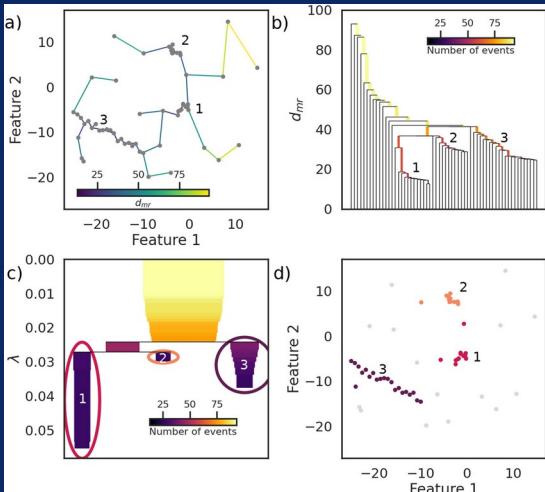
Clustering with DBSCAN



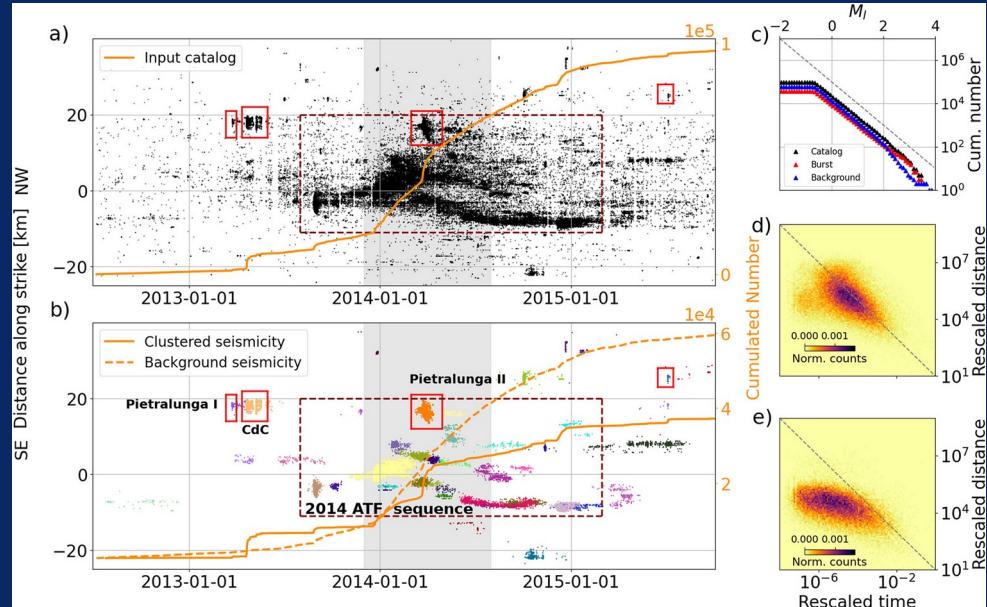
Unraveling Earthquake Clusters Composing the 2014 Alto Tiberina Earthquake Swarm via Unsupervised Learning

David Essing, Piero Poli

Hierachical-DBSCAN



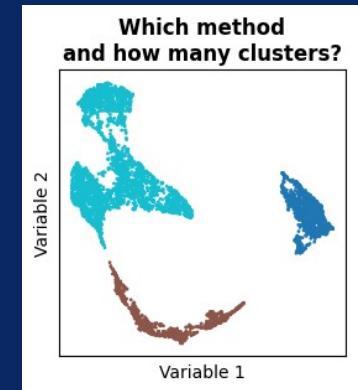
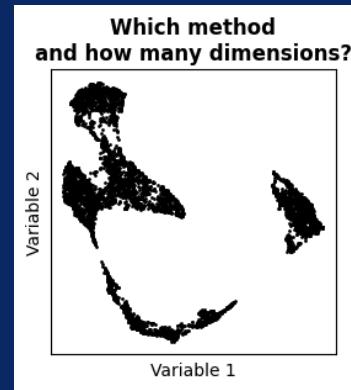
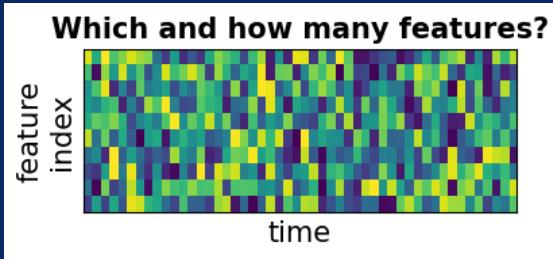
Input: spatial and temporal features of earthquake catalog



Outline

1. What is unsupervised learning and why is it interesting?
2. How do we do unsupervised learning?
3. Applications
- 4. Challenges & Solutions**

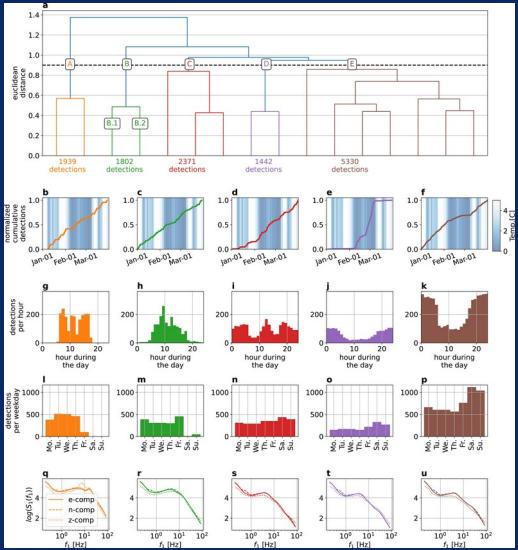
Making a choice



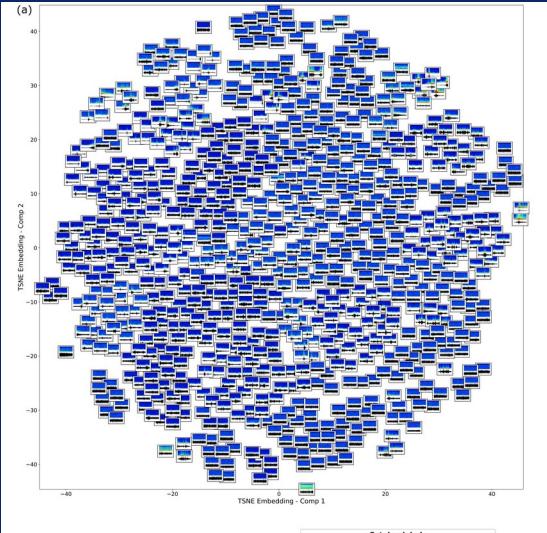
Solution:

- adapt method to goal and data
- If possible, use diagnostic methods like Silhouette Analysis or Explain Variance Ratio
- If possible ,visualize along the way

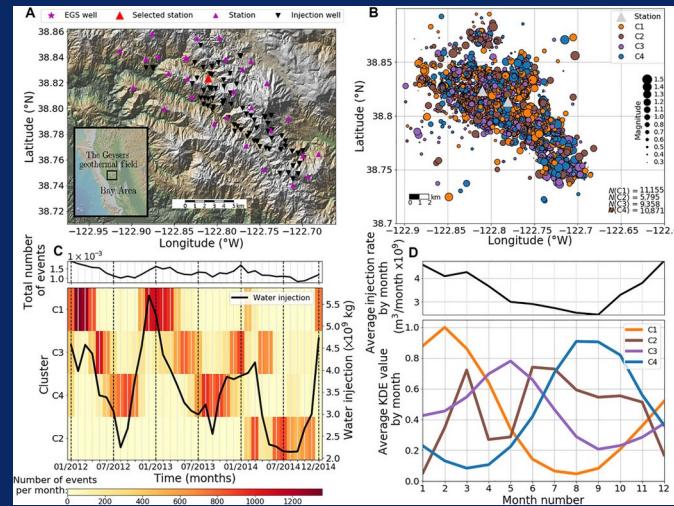
Link the results to the real world



Look at data characteristic
of clusters!



Try to visualize your data in
different ways.



Any other data available? Water
injection rate, temperature,
GNSS time series?

Recap

- **When to use unsupervised learning?**
 - To find patterns in data **without labels** — great for **exploratory analysis, anomaly detection, and new discoveries.**
- **What sort of data?**
 - **Anything**, including continuous seismograms, seismogram snippets, catalogs
- **Most common pipeline**
 - Raw Data → Features → Dimensionality Reduction → Clustering
- **Challenges**
 - Choosing the right approach, setting parameters, and interpretation

Contact

jane.public@gfz.de

www.gfz.de

+49 (0)3316264-0

Key Messages

- Point 1
- Point 2
- ... to be shown at the end for the discussion

Last slide

→ Example 2 for content

Contact

jane.public@gfz.de
www.gfz.de
+49 (0)3316264-0

Follow us on



<https://bsky.app/profile/gfz.bsky.social>



https://www.instagram.com/gfz_potsdam/



<https://www.youtube.com/user/GFZvideos>



<https://www.linkedin.com/company/gfz-helmholtz-zentrum-fuer-geoforschung>