
CVPR Nashville
JUNE 11-15, 2025

Clustering in Computer Vision

M. Saquib Sarfraz, Marios Koulakis

What is Clustering

- The purpose of cluster analysis is to group data according to the principle of similarity.

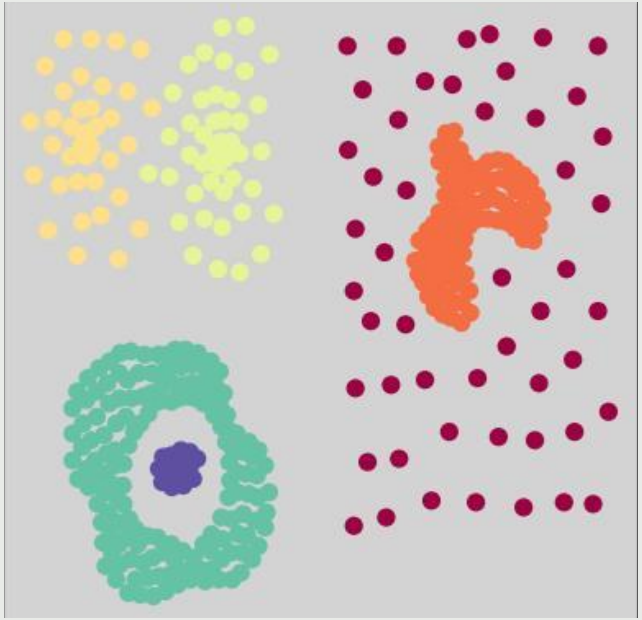
- What is similarity?
    - shape, texture, objects, semantic meaning?
    - grouping of points by similarities is one of the traditional themes extensively investigated by the Gestalt psychologists[1]

[1] Andenberg 1973; Hartigan 1975; Murtagh and Heck 1987; Toussaint 1980; Matula and Sokal 1980)

Gestalt Theory

Perceptual grouping – the law of Prägnanz[2]

- Grouping is key to visual perception
- Elements in a group can have properties that results from relationships
- human perception is biased towards simplicity.



Gestalt Clusters[3]

[2] https://en.wikipedia.org/wiki/Gestalt_psychology
[3] Charles T Zahn. Graph theoretical methods for detecting and describing gestalt clusters. IEEE TOC, 1970.

Gestalt Theory

Perceptual grouping – the law of Prägnanz

- Psychologist identified series of factors that predispose set of elements to be grouped (by human visual system)

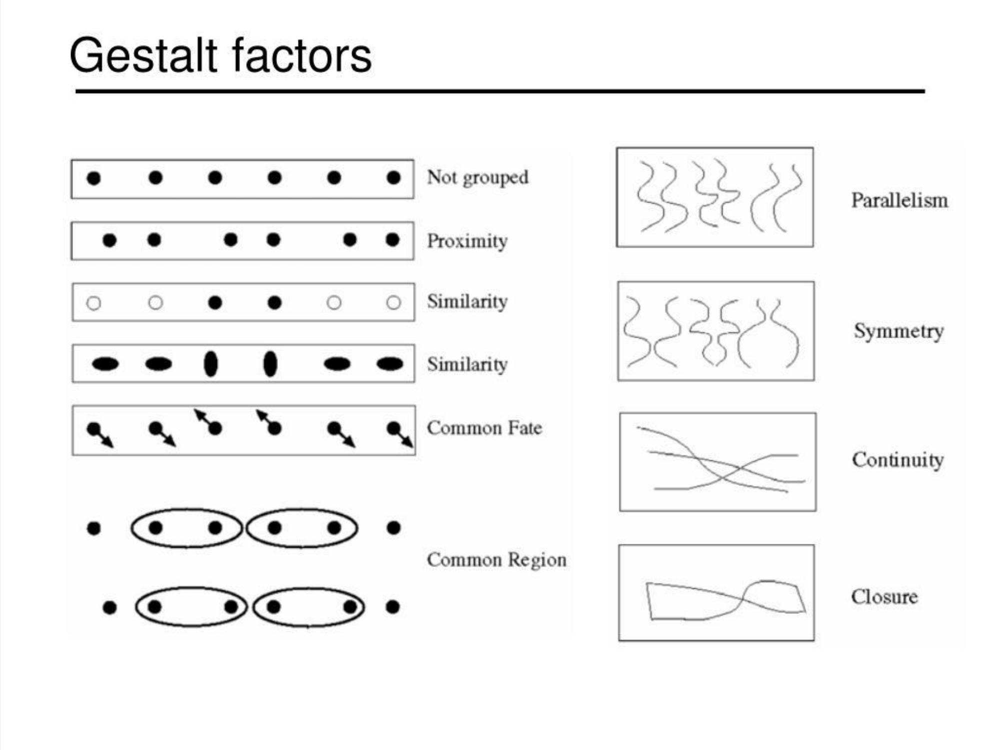


Image Source: Forsyth & Ponce

Gestalt in Computer Vision

Perceptual grouping – the law of Prägnanz

- In computer vision we measure similarity by proximity.

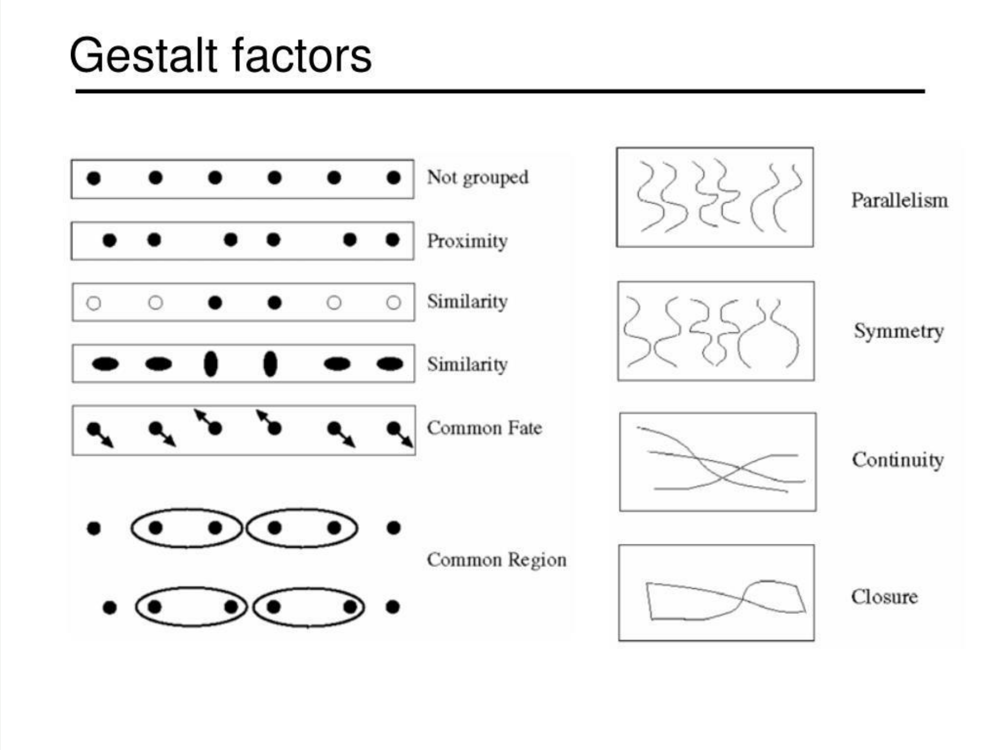


Image Source: Forsyth & Ponce

Gestalt in Computer Vision

Perceptual grouping – the law of Prägnanz

- In computer vision we measure similarity by proximity.
- We encode factors of similarity by representation learning.

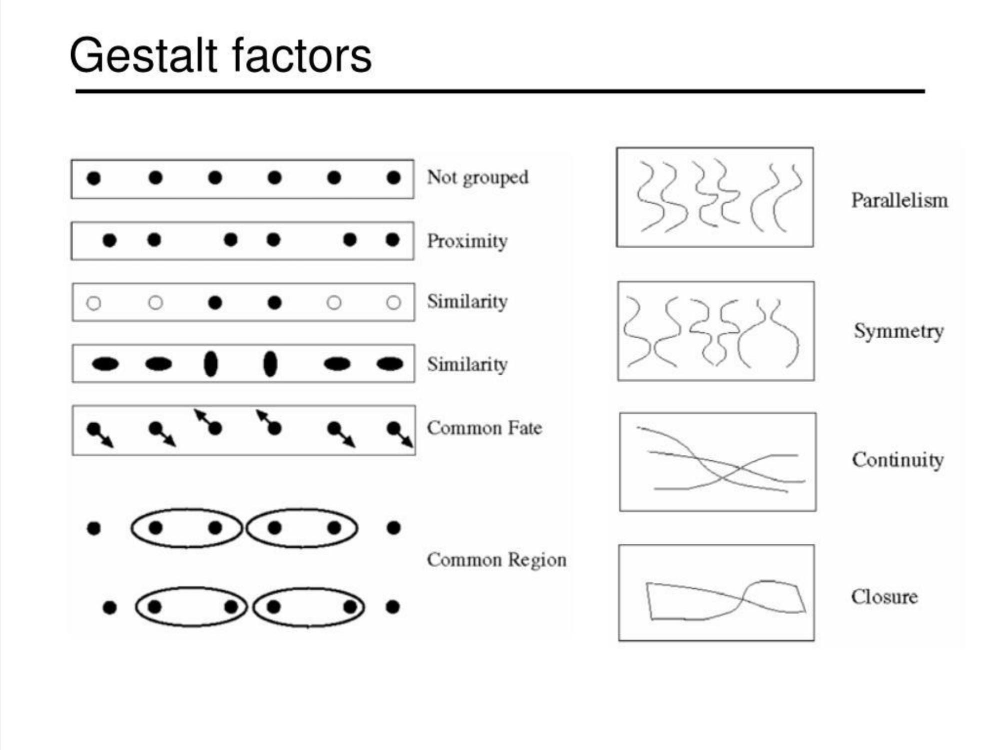
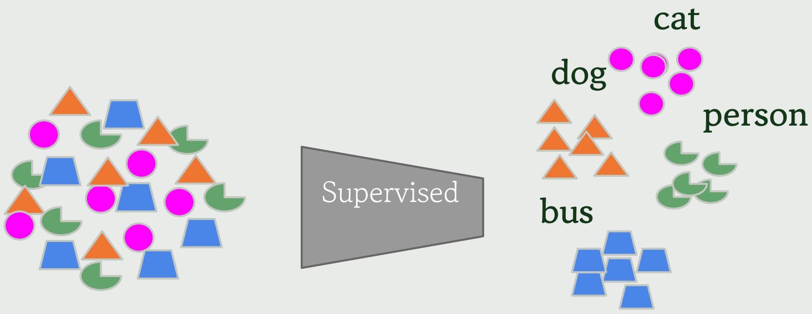


Image Source: Forsyth & Ponce

Clustering or Representation Learning

- Supervised representation learning
  - # of classes (clusters) and their assignments are known

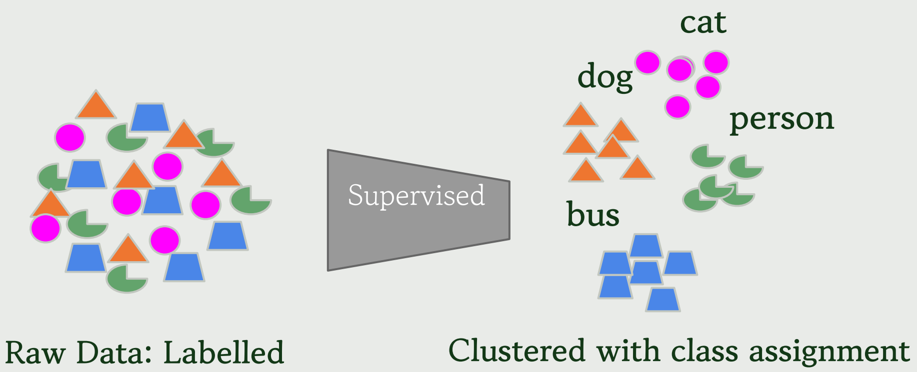


Raw Data: Labelled

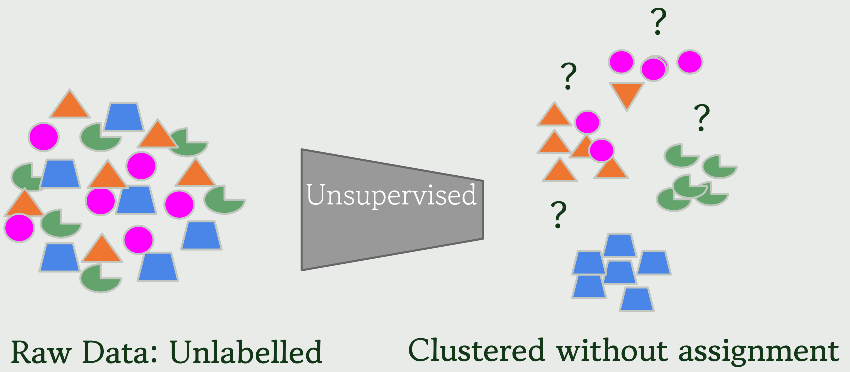
Clustered with class assignment

Clustering or Representation Learning

- Supervised representation learning
  - # of classes (clusters) and their assignments are known

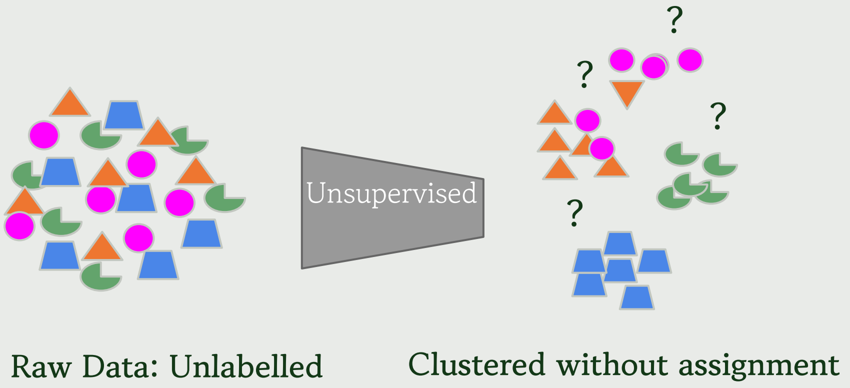


- Unsupervised representation learning
  - # of classes (clusters) and their assignments are NOT known



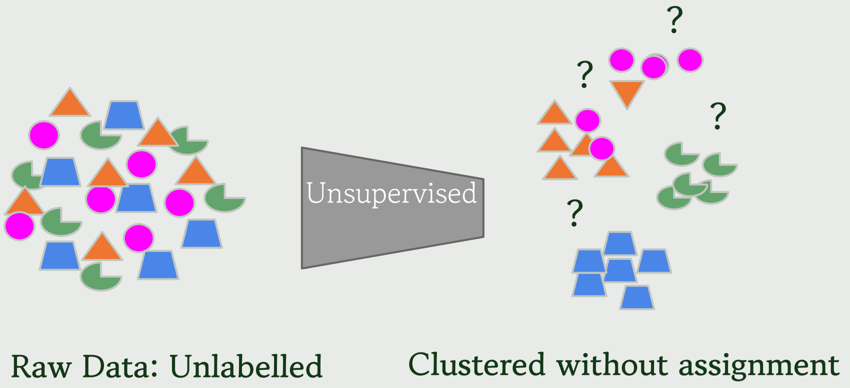
The Clustering Problem

- Unsupervised representation learning
  - # of classes (clusters) and their assignments are NOT known



The Clustering Problem

- Unsupervised representation learning
    - # of classes (clusters) and their assignments are NOT known

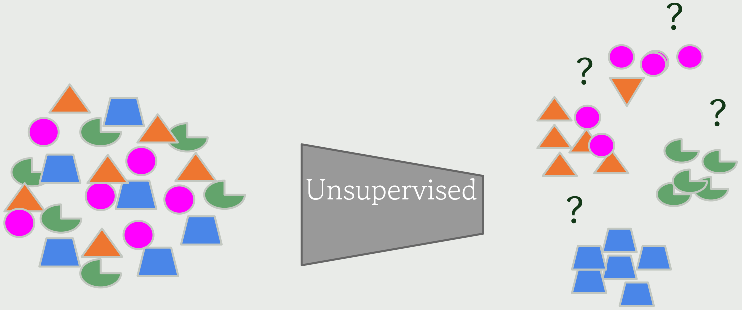


"Distinguish between the disparate clusters when the number of clusters is not known a priori." (Guberman and Wojtkowski 2002)

Guberman and Wojtkowski, "Clustering Analysis As a Gestalt Problem". Gestalt Theory, Vol 24 No.2, 2002

The Clustering Problem

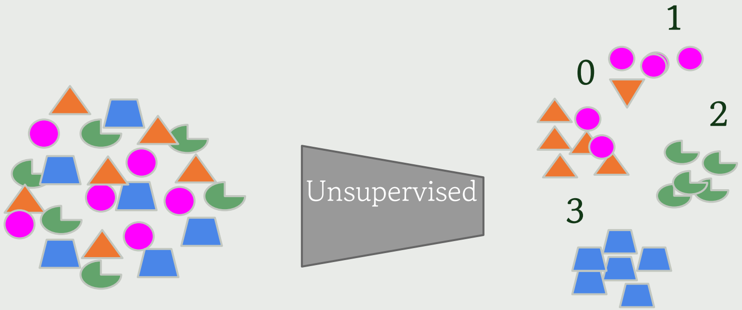
- Unsupervised representation learning
  - # of classes (clusters) and their assignments are NOT known



Raw Data: Unlabelled

Clustered without assignment

- Clustering
  - resolves assignment



Raw Data: Unlabelled

Cluster assignment

The Clustering Problem

- Representation learning clusters data

- Current Self Supervised Learning (SSL) can be thought of as "Deep Clustering" w/o assignment.

- The discovery or assignment of the obtained clusters can be made either directly at the model output or utilizing any clustering mechanism (e.g., K-Means) on top.

Clustering Methods

- Partition Based
- Hierarchy Based
- Density Based
- Hybrid methods

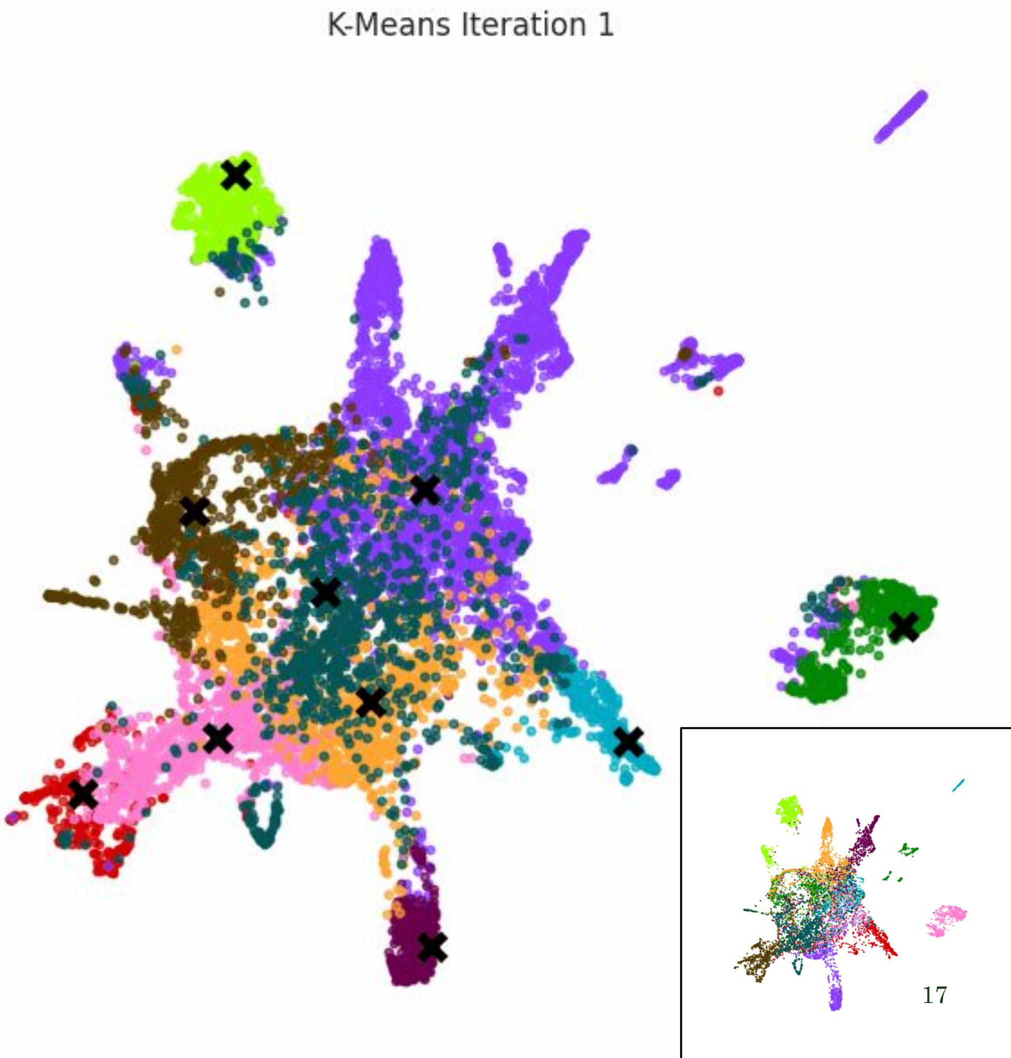
Partition Based Clustering

Partition Based

- Clusters defined as a fixed size partition
- Objectives minimize intra-cluster distances or maximize likelihood
- K-Means, Gaussian Mixture Models (GMMs)
- Example of K-Means++ on **supervised** embedding

Partition Based

- Clusters defined as a fixed size partition
- Objectives minimize intra-cluster distances or maximize likelihood
- K-means, Gaussian mixture models
- Example of k-means++ on **unsupervised** embedding



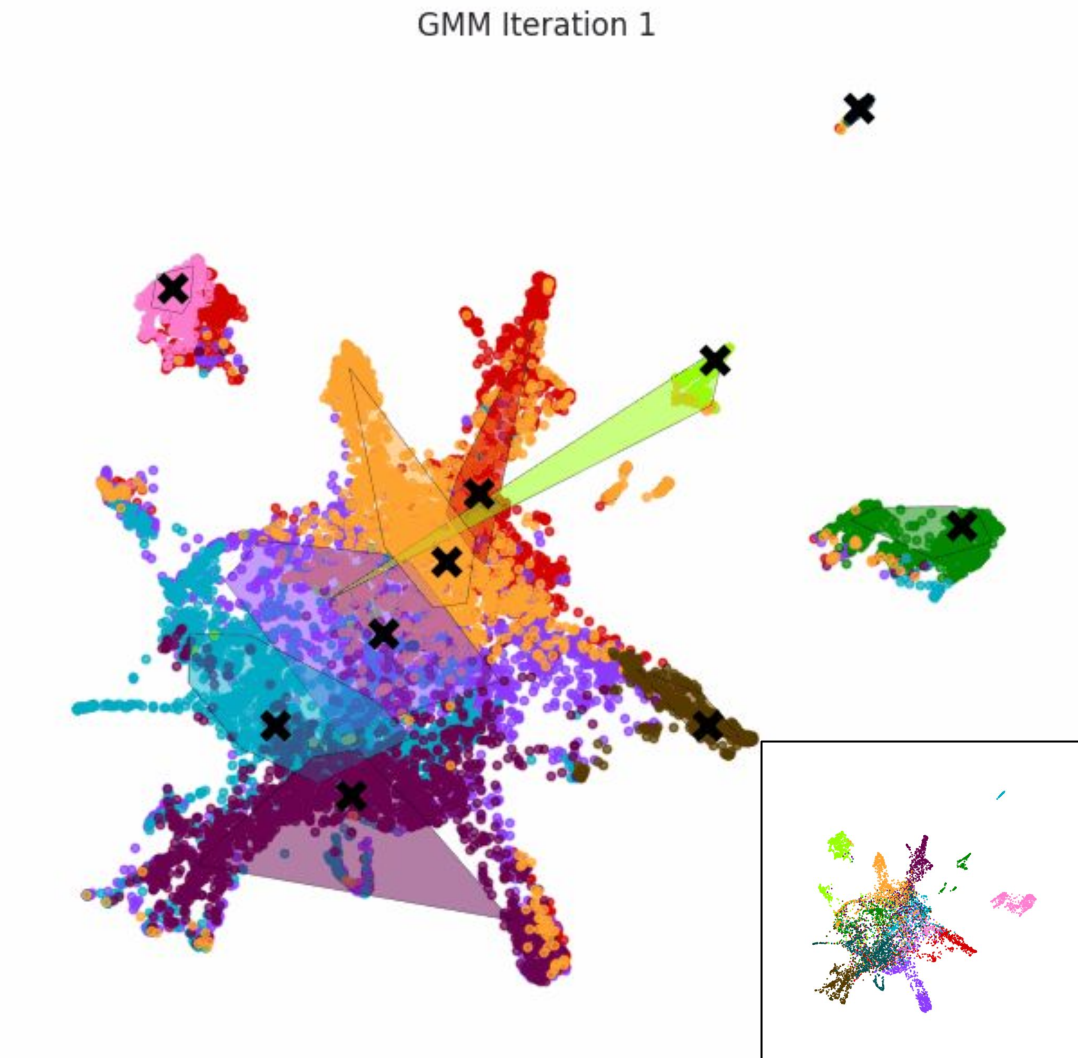
Gaussian Mixture Models (GMMs)

- k Gaussian distributions

$$\mathcal{N}(\mu_j, \Sigma_j)$$

- Sampled probabilities π_j
- Maximize the likelihood of the samples

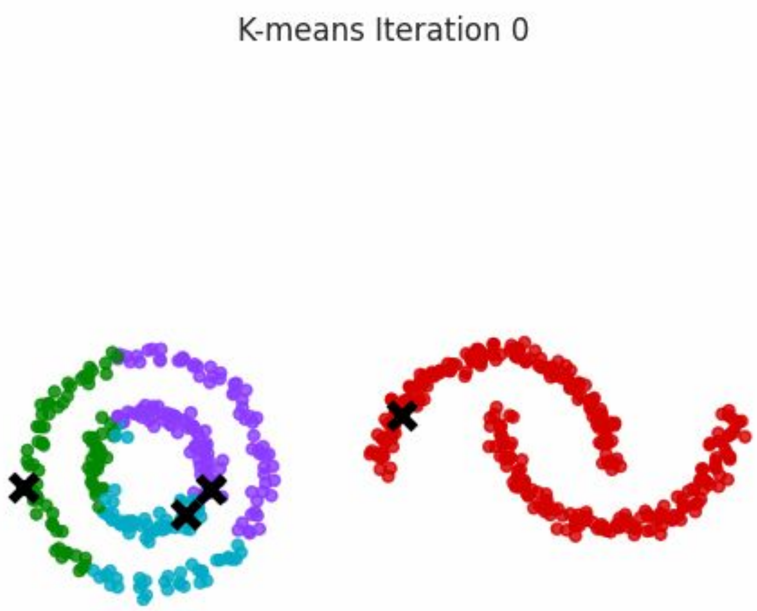
$$\prod_{i=1}^{N} \sum_{j=1}^{k} \pi_j \, p_{\mathcal{N}}(x_i \mid \mu_j, \Sigma_j)$$



What can go wrong?

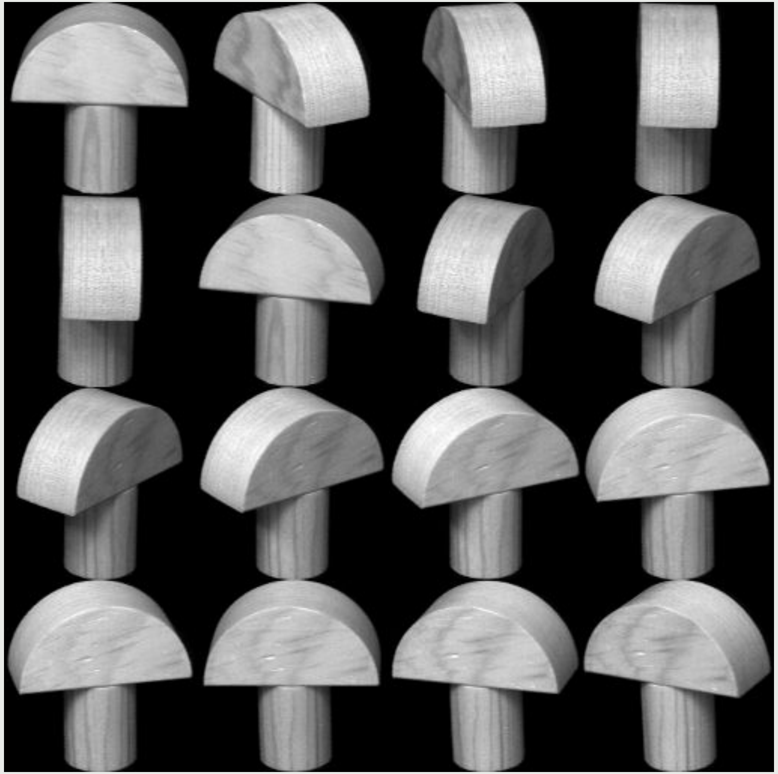
Ignoring Geometry

- Partition models assume specific cluster shapes: spheres ellipsoids
- Topology is ignored
- Foliated clusters are often split incorrectly



Ignoring Geometry

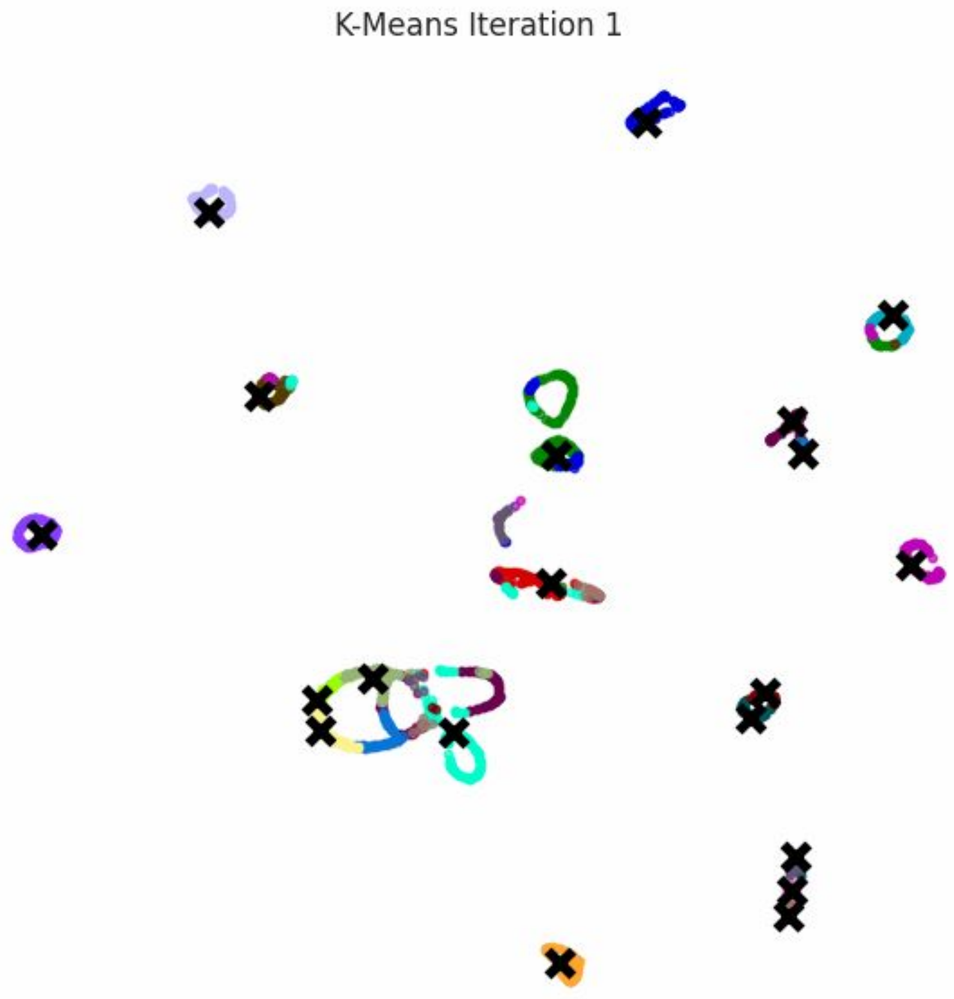
- Often occurring on datasets with transforms
- Novel view synthesis, robotic vision, reinforcement learning, equivariant representation learning, disentanglement
- Dimensionality reduction can simplify shapes



Columbia Object Image Library (COIL-20), S. A. Nene, S. K. Nayar and H. Murase, Technical Report CUCS-005-96, February 1996

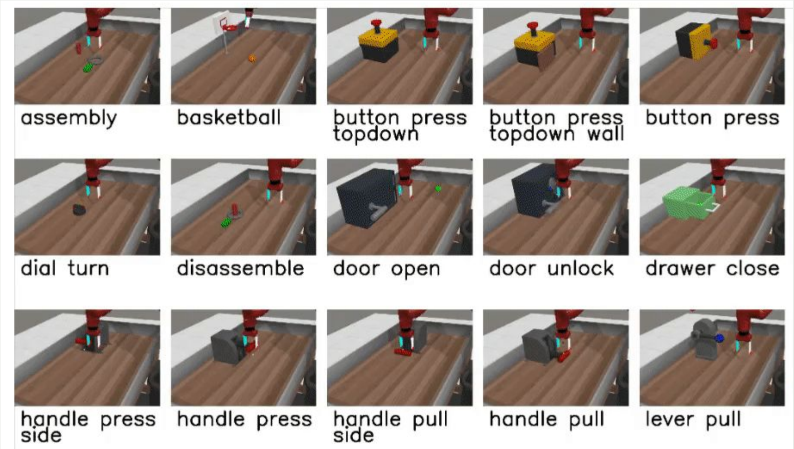
Ignoring Geometry

- Often occurring on datasets with transforms
- Novel view synthesis, robotic vision, reinforcement learning, equivariant representation learning, disentanglement
- Dimensionality reduction can simplify shapes



Ignoring Geometry

https://github.com/Fyusion/LLFF?tab=readme-ov-file



https://meta-world.github.io

https://sunset1995.github.io/HorizonNet

Cluster assignment

- Single clusters split
- Clusters glued together
- Initialization is key: k-means++
- Can use a larger k with a regularized model: Dirichlet Process Gaussian Mixture Model (DPGMM)

Can be slow

- K-means could need multiple iterations to converge
- Using more complex models like GMMs increases the computational cost a lot (inverse of a DxD matrix)
- Initial dimensionality reduction helps

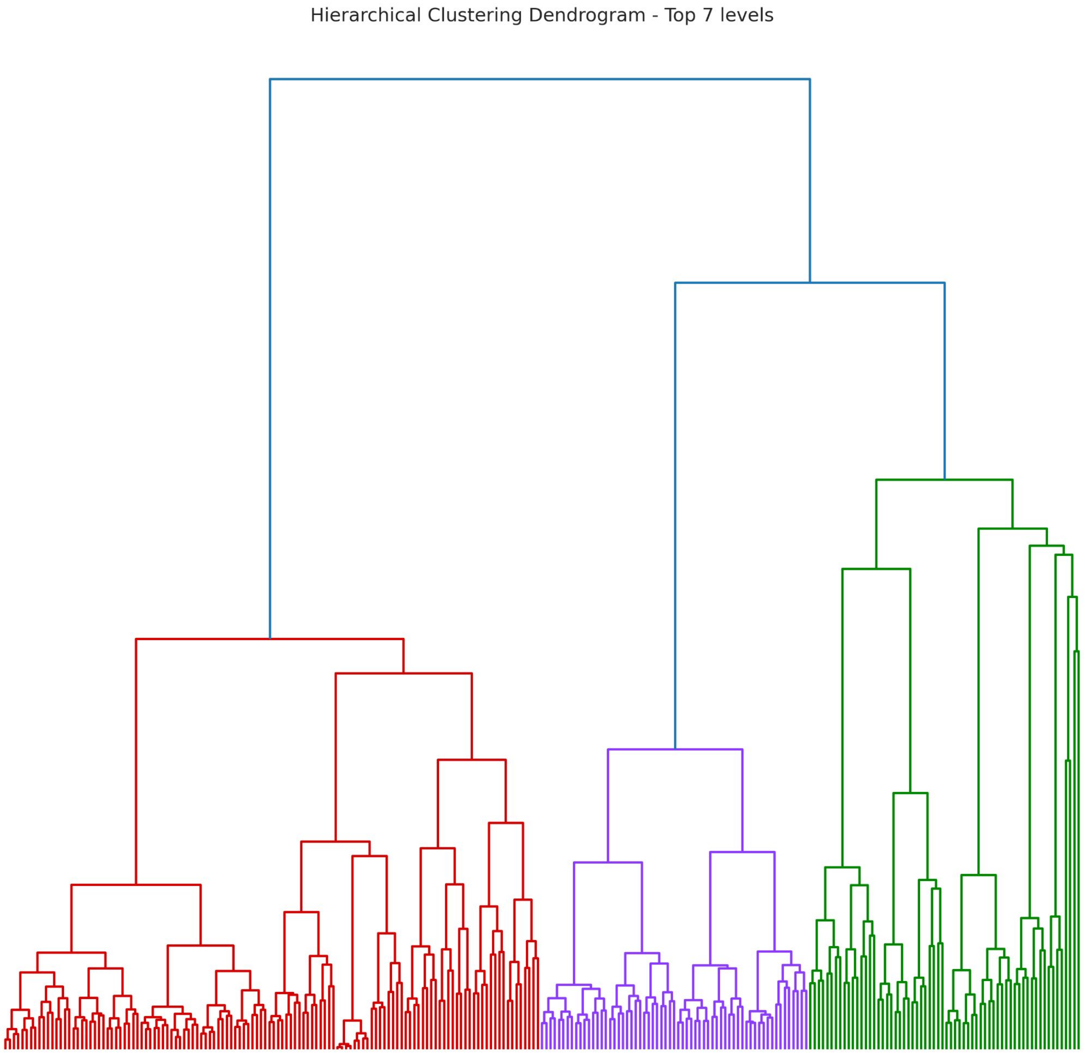
K-Means	0.27s ± 0.037
K-Means++	2.14s ± 0.285
GMM	57.7s ± 6.37
DPGMM	45.7s ± 4.31

Time comparison on DINOv2 embeddings of Imagenette

Hierarchy Based Clustering

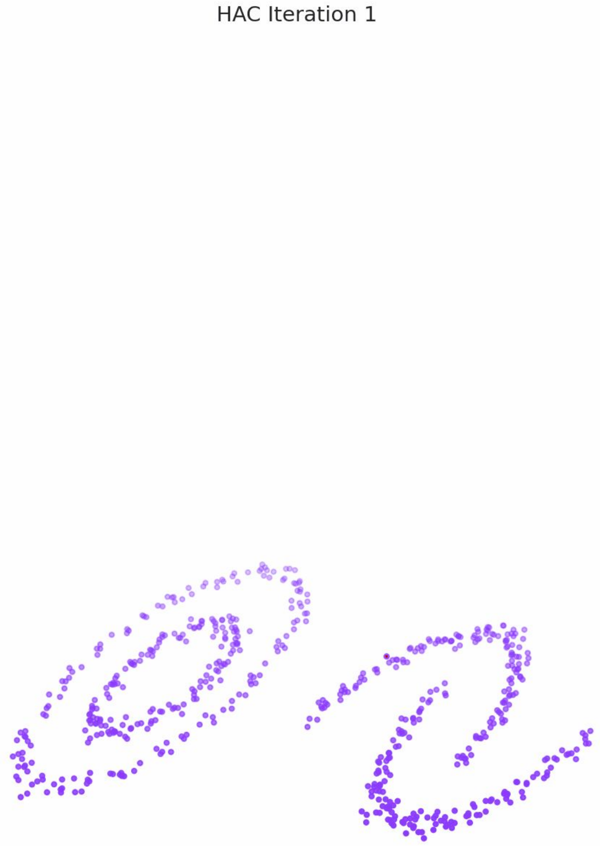
Hierarchy Based

- Top-down: Divisive
- Bottom-up: Agglomerative
- We will focus on the second one



Hierarchical Aggl. Clustering (HAC)

- Start from single points
- On each step merge A, B with a linkage criterion
  - Single $\min_{a \in A, b \in B} d(a, b)$
  - Complete $\max_{a \in A, b \in B} d(a, b)$
  - Average $\mathop{E}_{a \in A, b \in B} (d(a, b))$



Hierarchical Aggl. Clustering (HAC)

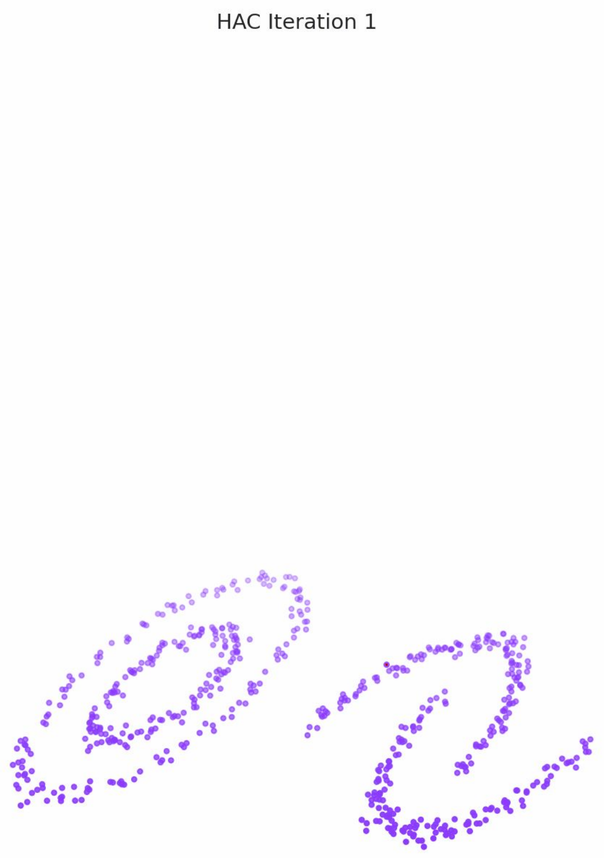
- Start from single points
- On each step merge A, B with a linkage criterion
  - Variance

$$Var(A \cup B) - Var(A) - Var(B)$$

  - Ward

$$\sum_{A \cup B} ||x - E(A \cup B)||^2$$

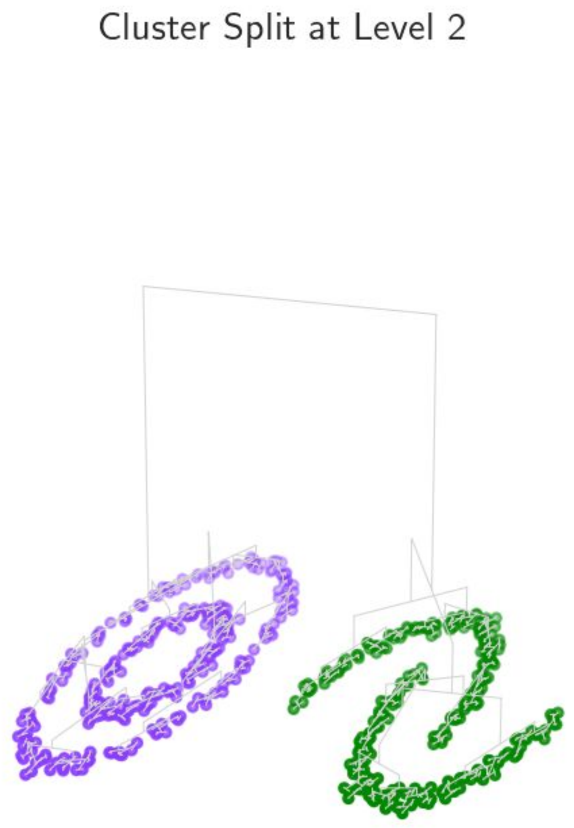
$$- \sum_{A} ||x - E(A)||^2 - \sum_{B} ||x - E(B)||^2$$



What can go wrong?

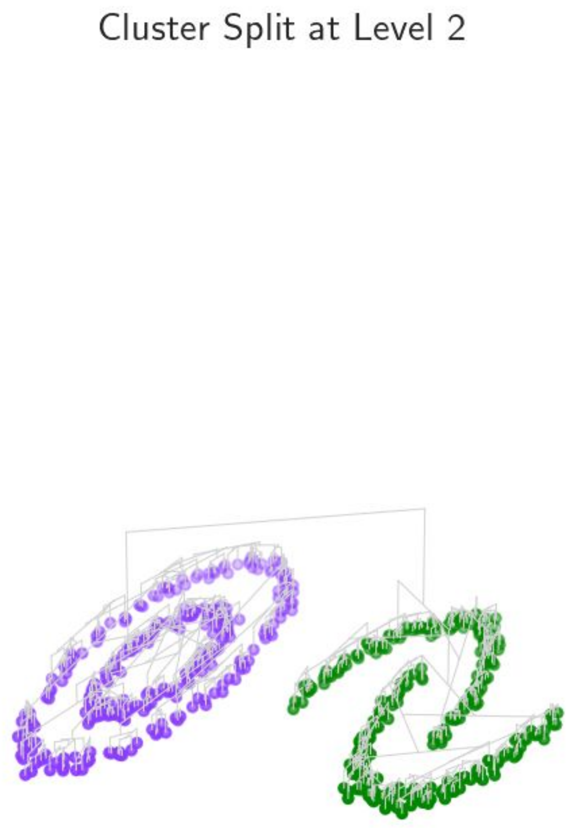
Ignoring Geometry

- Attempt to capture topology with a tree
- Better than partition based
- Mistakes tend to happen on higher levels of the tree
- Method is partly partition based



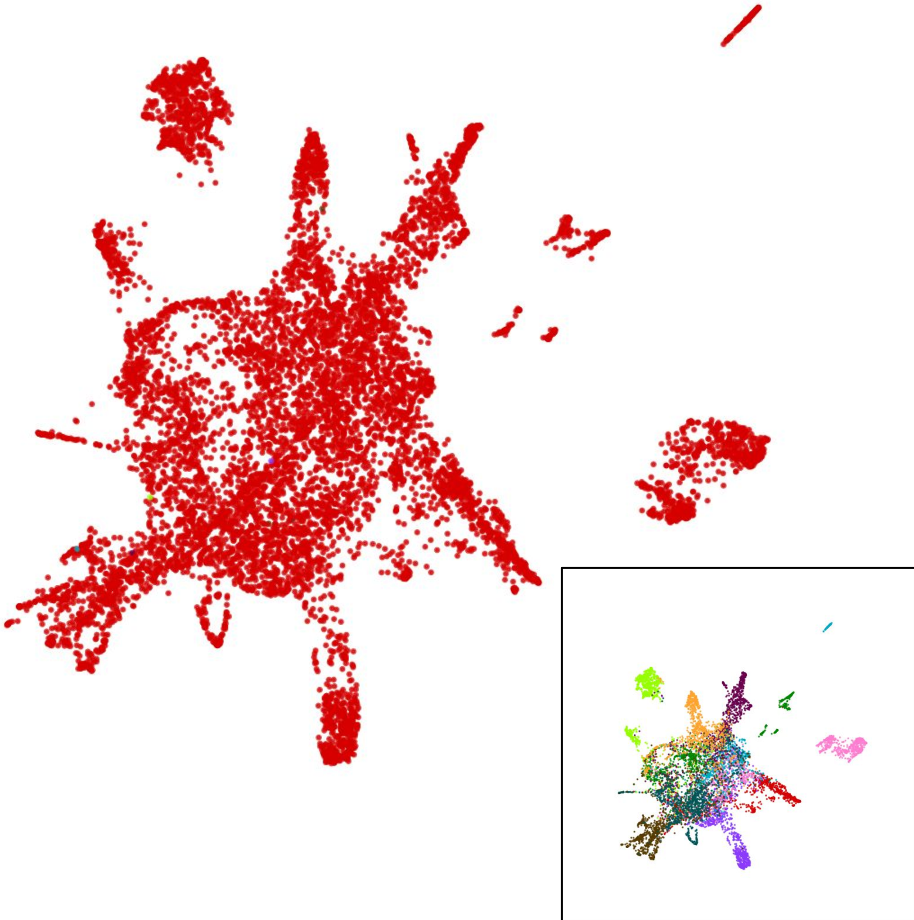
Ignoring Geometry

- Attempt to capture topology with a tree
- Better than partition based
- Mistakes tend to happen on higher levels of the tree
- Using local linkages like single linkage helps



Clusters merged

- Especially on datasets with overlapping clusters
- Here global linkage criteria can help
- Can require a few more clusters than expected and visually separate
- Sensitive to outliers



Density Based Clustering

DBSCAN

- Parameters: ε, minPts
- Core points:

$p \in X, |B(p, \varepsilon) \cap X| \geq minPts$

- Directly reachable:

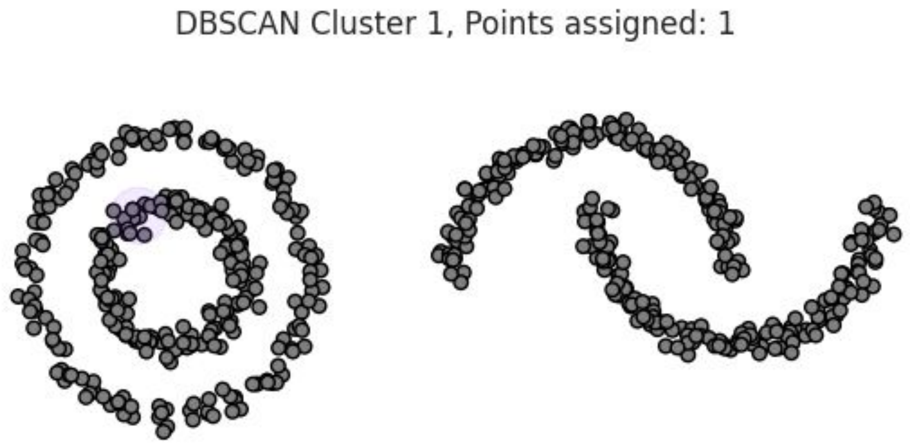
$\exists p,\ p$ is a core point, $d(p, q) < \varepsilon$

- Reachable:

$\exists\ p = p_0, \ldots, p_n = q$

$\forall\ i \leq n - 1\ p_{i+1}$ reachable from p_i

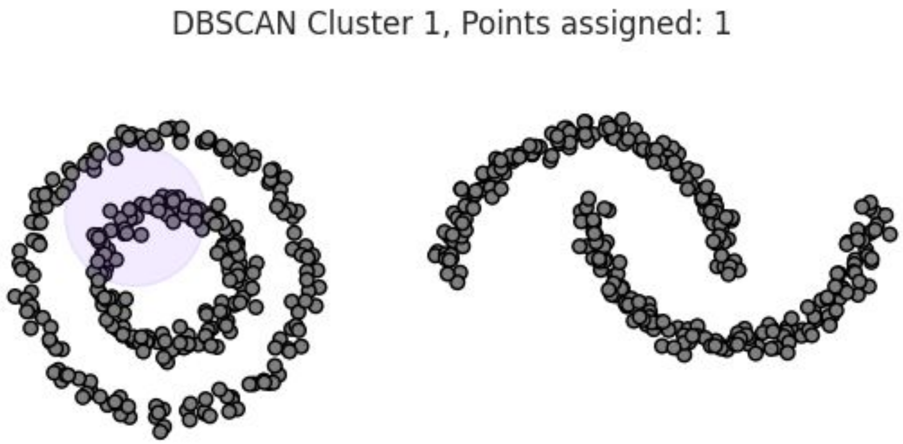
- Start from core points and connect reachable points



What can go wrong?

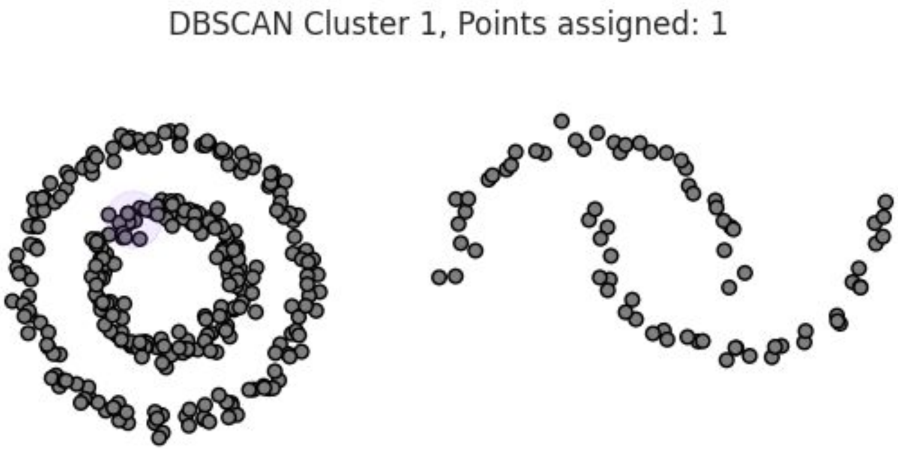
What can go wrong?

- Parameter tuning can be challenging and subjective
- Varying data densities are hard to handle



What can go wrong?

- Parameter tuning can be challenging and subjective
- Varying data densities are hard to handle

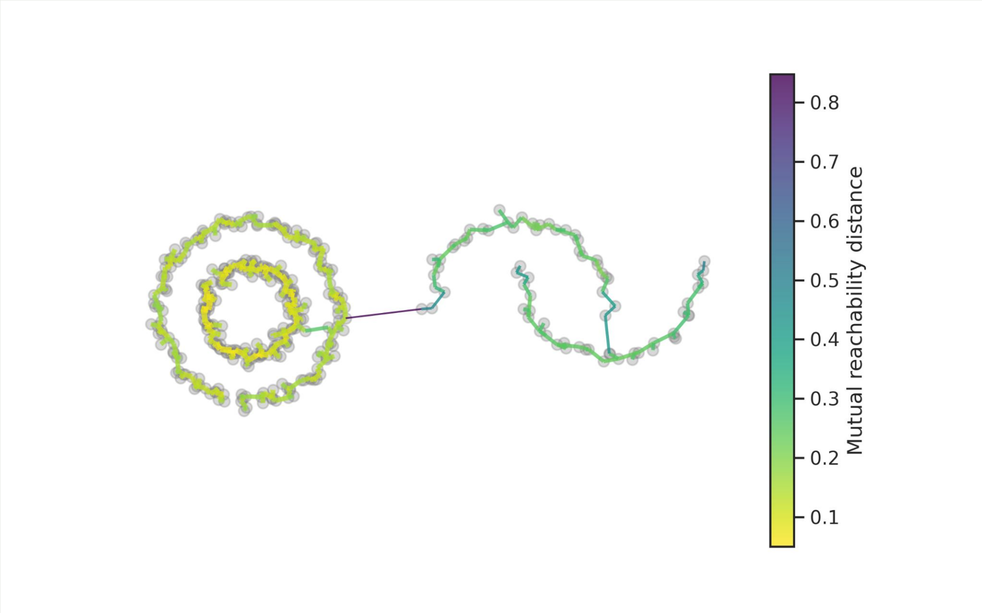


Hybrid Methods

HDBSCAN

- Core distance: $core_k(p)$
- Mutual Reachability Distance

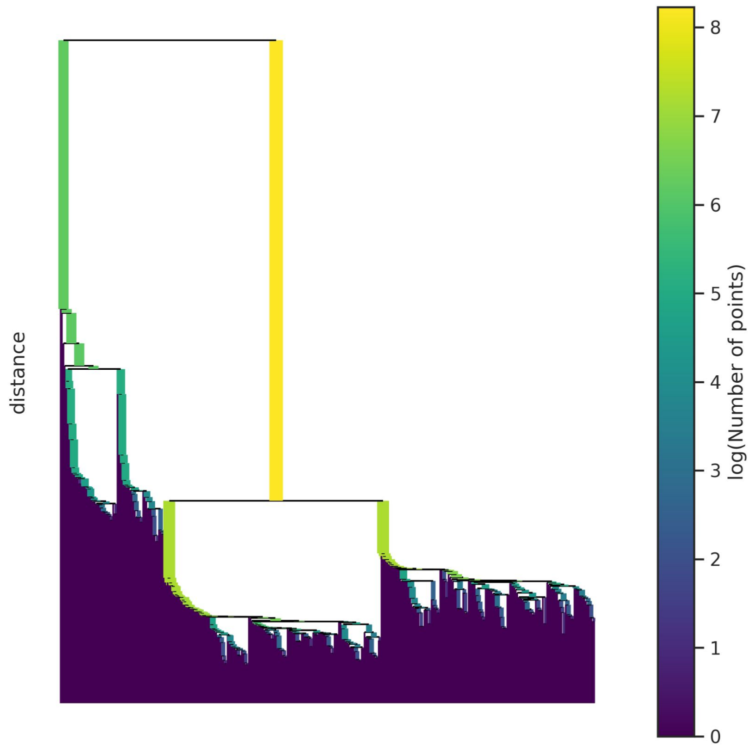
$$d_k(p,q) = \max\{core_k(p), core_k(q), d(p,q)\}$$



HDBSCAN

- Core distance: $core_k(p)$
- Mutual Reachability Distance

$$d_k(p,q) = \max\{core_k(p), core_k(q), d(p,q)\}$$

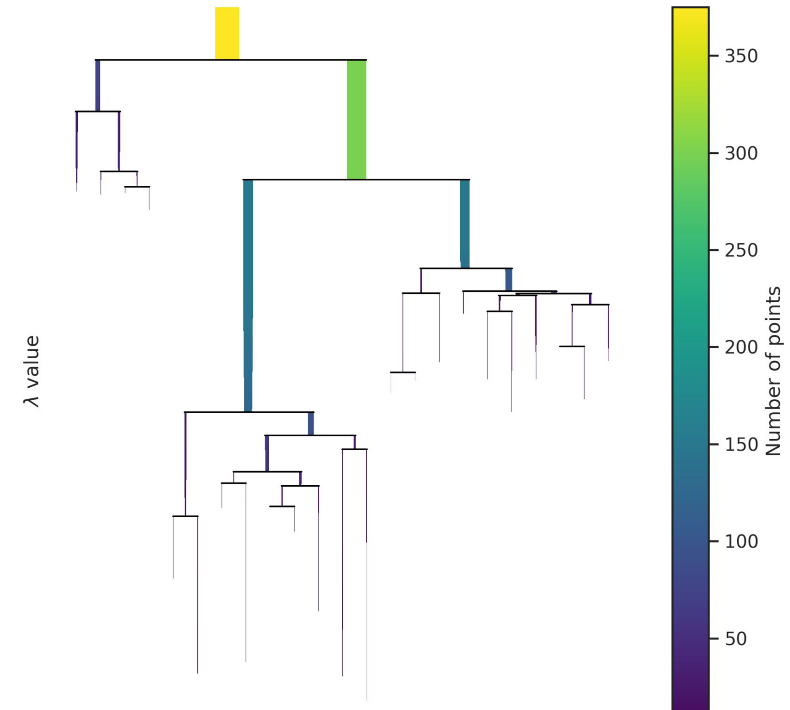


HDBSCAN

- Core distance: $core_k(p)$
- Mutual Reachability Distance

$$d_k(p,q) = \max\{core_k(p), core_k(q), d(p,q)\}$$

- Condensed tree: split only when clusters are formed

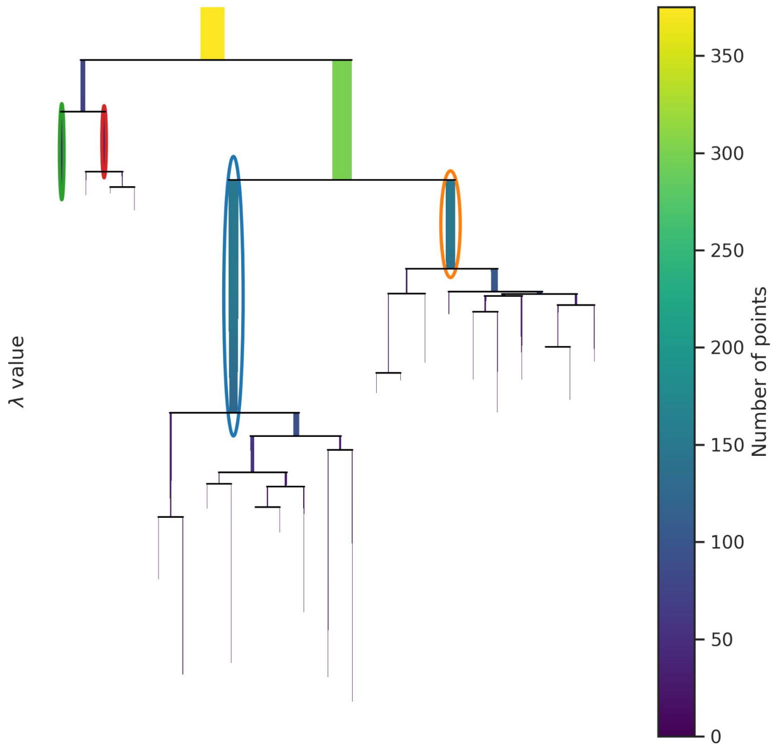


HDBSCAN

- Core distance: $core_k(p)$
- Mutual Reachability Distance

$$d_k(p,q) = \max\{core_k(p), core_k(q), d(p,q)\}$$

- Condensed tree: split only when clusters are formed
- Stability: $\sum_{p \in C} (\lambda_p - \lambda_{birth})$
- Select clusters stabler than their subclusters

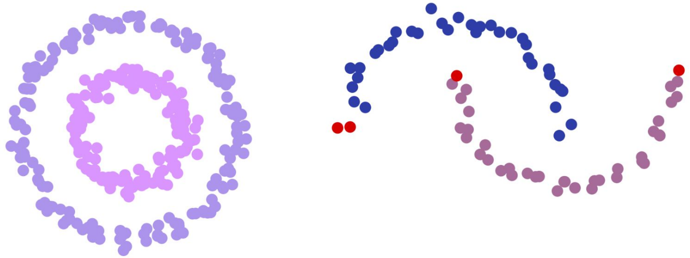


HDBSCAN

- Core distance: $core_k(p)$
- Mutual Reachability Distance

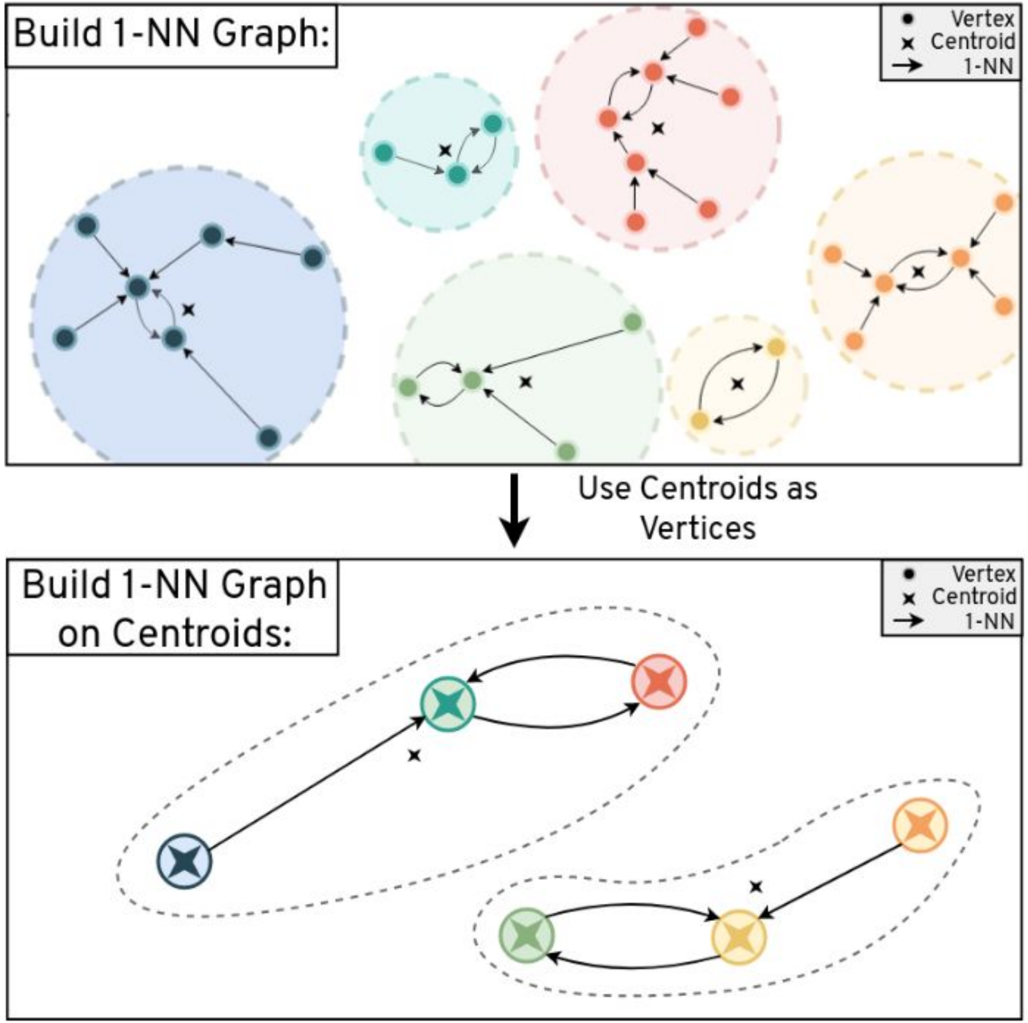
$$d_k(p,q) = \max\{core_k(p), core_k(q), d(p,q)\}$$

- Condensed tree: split only when clusters are formed
- Stability: $\sum_{p \in C} (\lambda_p - \lambda_{birth})$
- Select clusters stabler than their subclusters



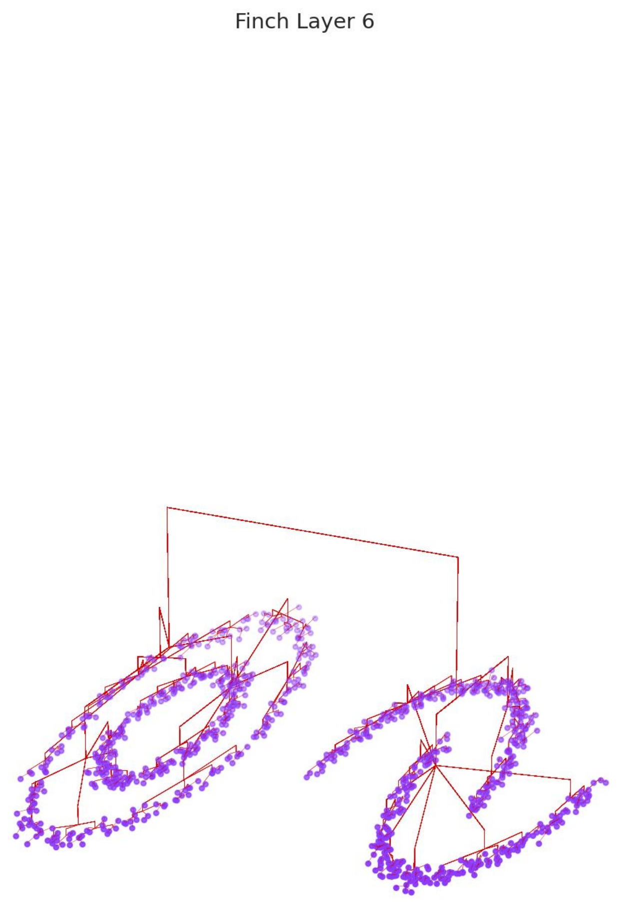
FINCH

- Connect 1-NNs in an agglomerative way
- Reduce the components to their centroids
- Repeat till a hierarchy is built
- Hybrid of hierarchical and partition based clustering



FINCH

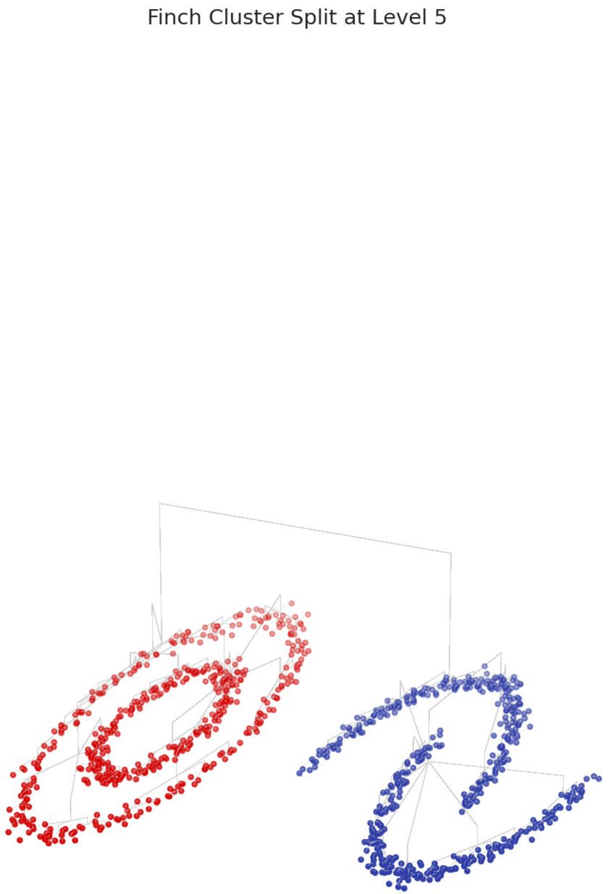
- Connect 1-NNs in an agglomerative way
- Reduce the components to their centroids
- Repeat to get a hierarchy
- Hybrid of hierarchical and partition based clustering
- Designed to be fast
- Based on observations and a theorem of Eppstein et. al, 1-NN graphs are small



What can go wrong?

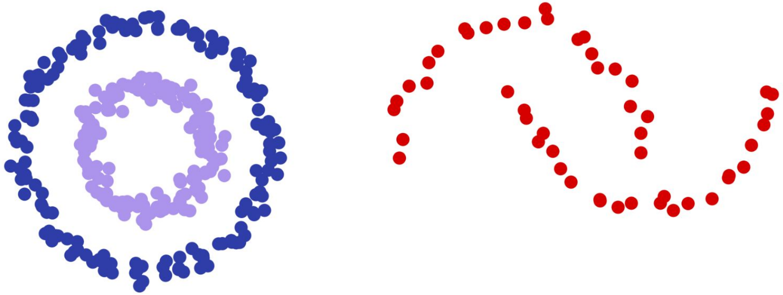
Ignoring Geometry

- Due to the partition-based part of the algorithm
- More intense on the higher levels of the tree
- Can be fixed by applying a more geometry-friendly algorithm after some level



Density Differences

- The algorithm does not detect density on the data manifold, but on the ambient space
- Sparse clusters with point close to each other could be merged
- Reducing dimensionality with a manifold-aware algorithm can help



Evaluation and number of clusters

Clustering Evaluation

- Internal Metrics
  - Measure the quality of the clustering without external information/ground truth (Unsupervised)
  - Examples: Silhouette Score, Davies-Bouldin Index

- External Metrics
  - Compare the clustering against a ground truth (Supervised)
  - Examples: Adjusted Rand Index, Normalized Mutual Information

Number of clusters

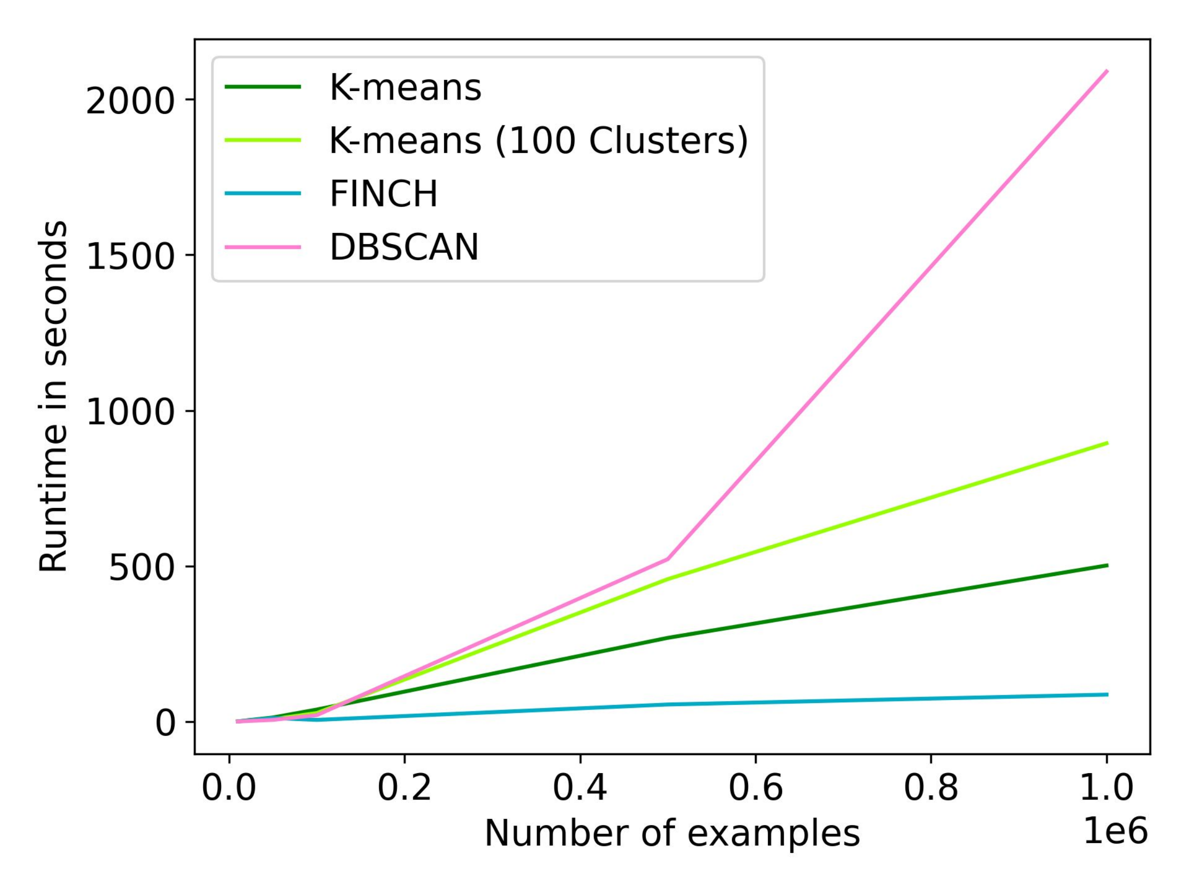
- Subjective, need some prior knowledge to estimate
- Different methods exist like the elbow method, GAP statistic, regularization to discard clusters, selection of hierarchy levels, HDBSCAN
- Do not work that easily on real-world datasets out of the box
- A hierarchy might be enough in some cases - optimal hierarchy

Runtime

K-Means	0.27s ± 0.037
K-Means++	2.14s ± 0.285
GMM	57.7s ± 6.37
DPGMM	45.7s ± 4.31
HAC	8.73s ± 1.014
DBSCAN	0.35s ± 0.099
HDBSCAN	63.82s ± 4.011
FINCH	0.59s ± 0.126

Dataset size: 10,000 points, 10 clusters

Runtime

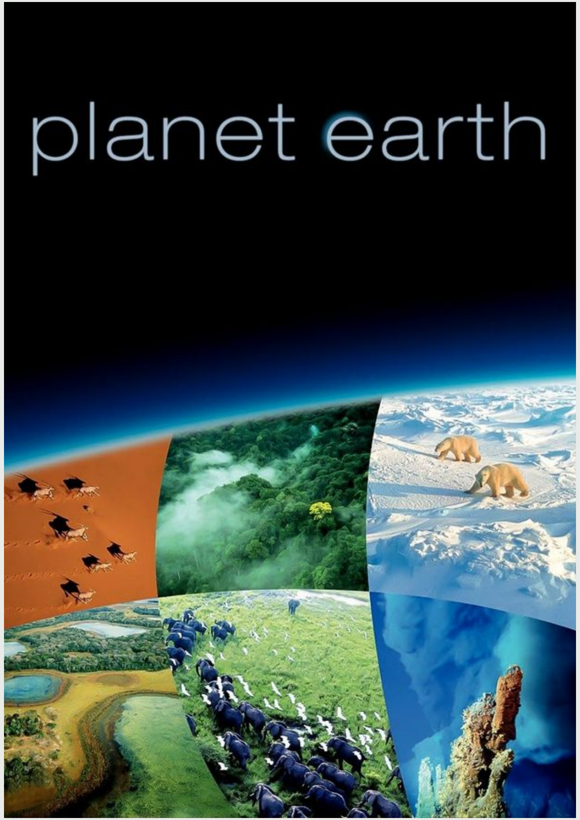


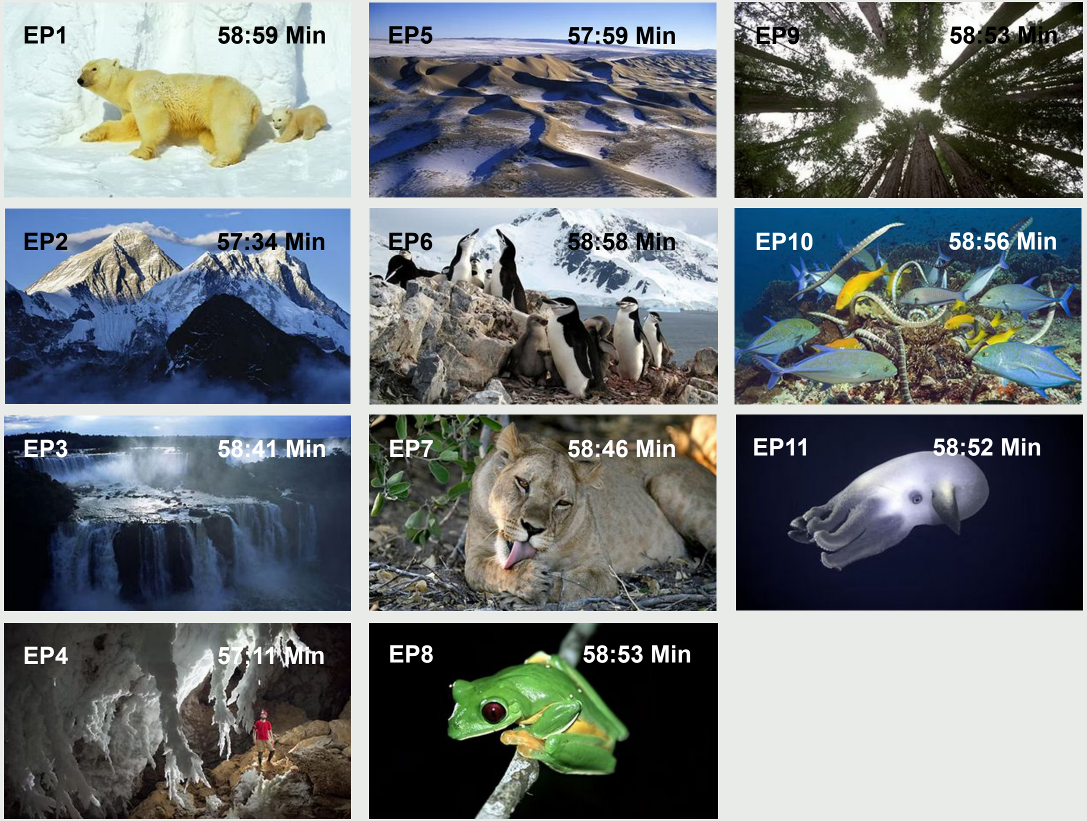
K-means
K-means (100 Clusters)
FINCH
DBSCAN
Runtime in seconds
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Number of examples

Two example Use Cases

Example: Multimodal Retrieval

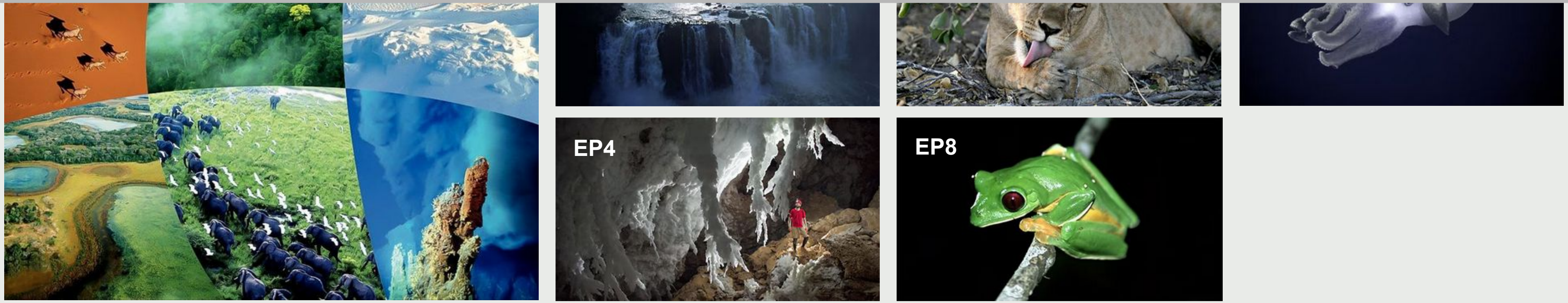
Multimodal Retrieval





Multimodal Retrieval

How can we access all footage of snow leopards? ❄️🐆

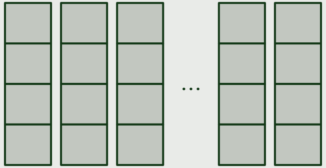


Multimodal Retrieval

Multimodal Retrieval

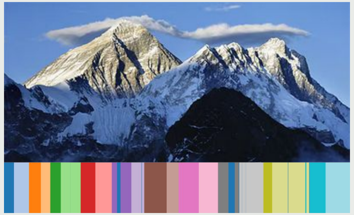
Video

Frames

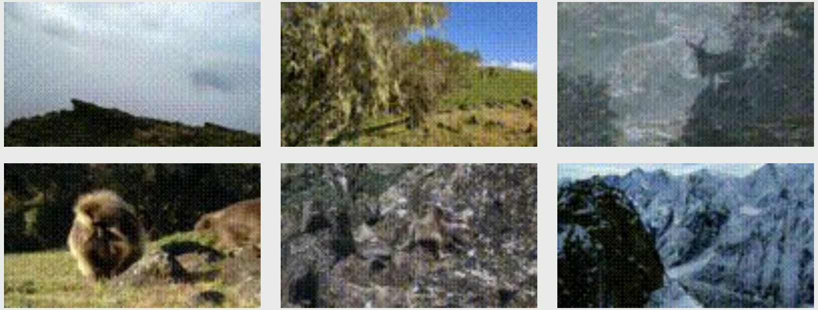


Frame-level Embeddings

Clustering



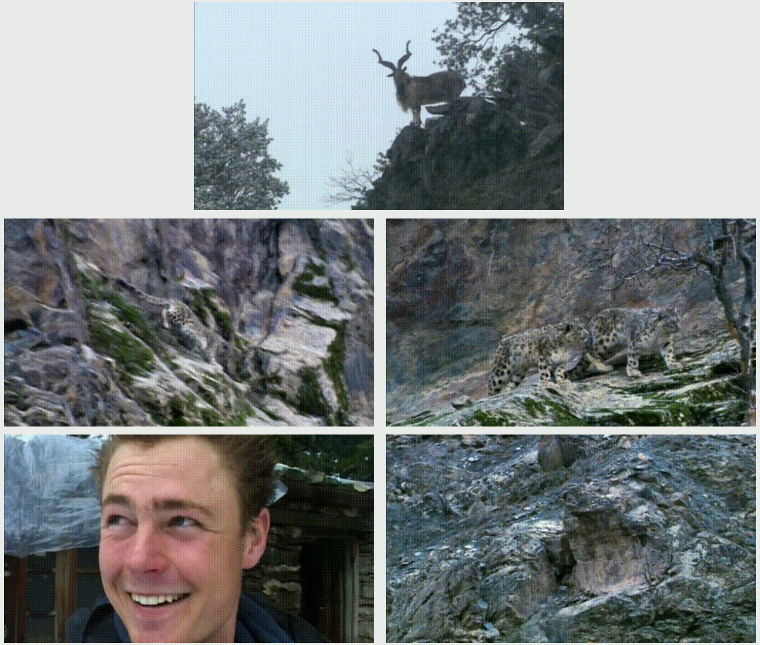
Video Segments



Temporally-Weighted Hierarchical Clustering for Unsupervised ActionSegmentation, Sarfraz et al. CVPR 2021

Multimodal Retrieval

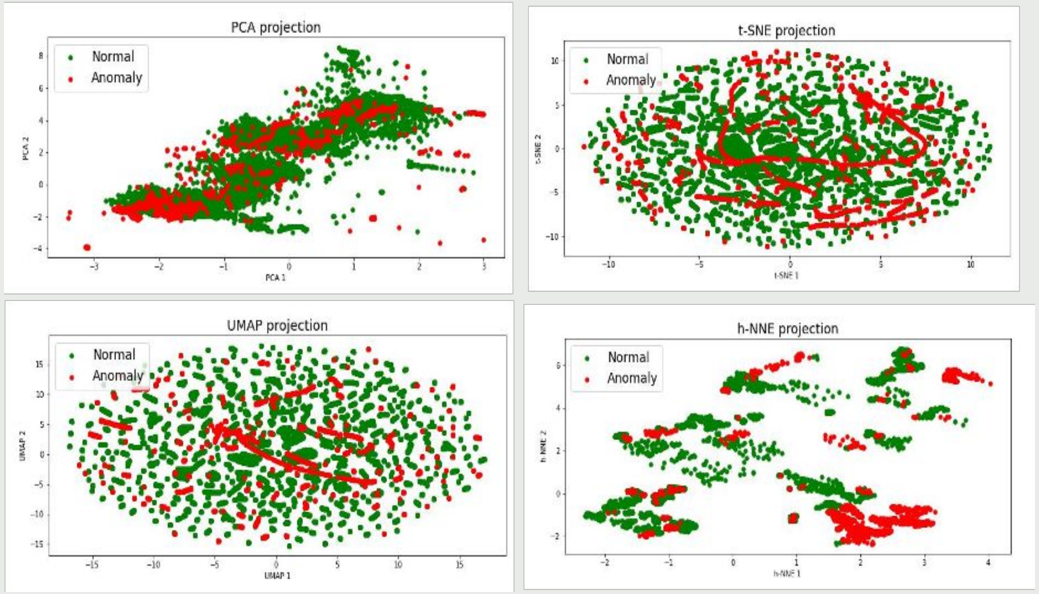
Top-5 results for Query: "footage on snow leopards"



Example: Data Understanding & Annotation

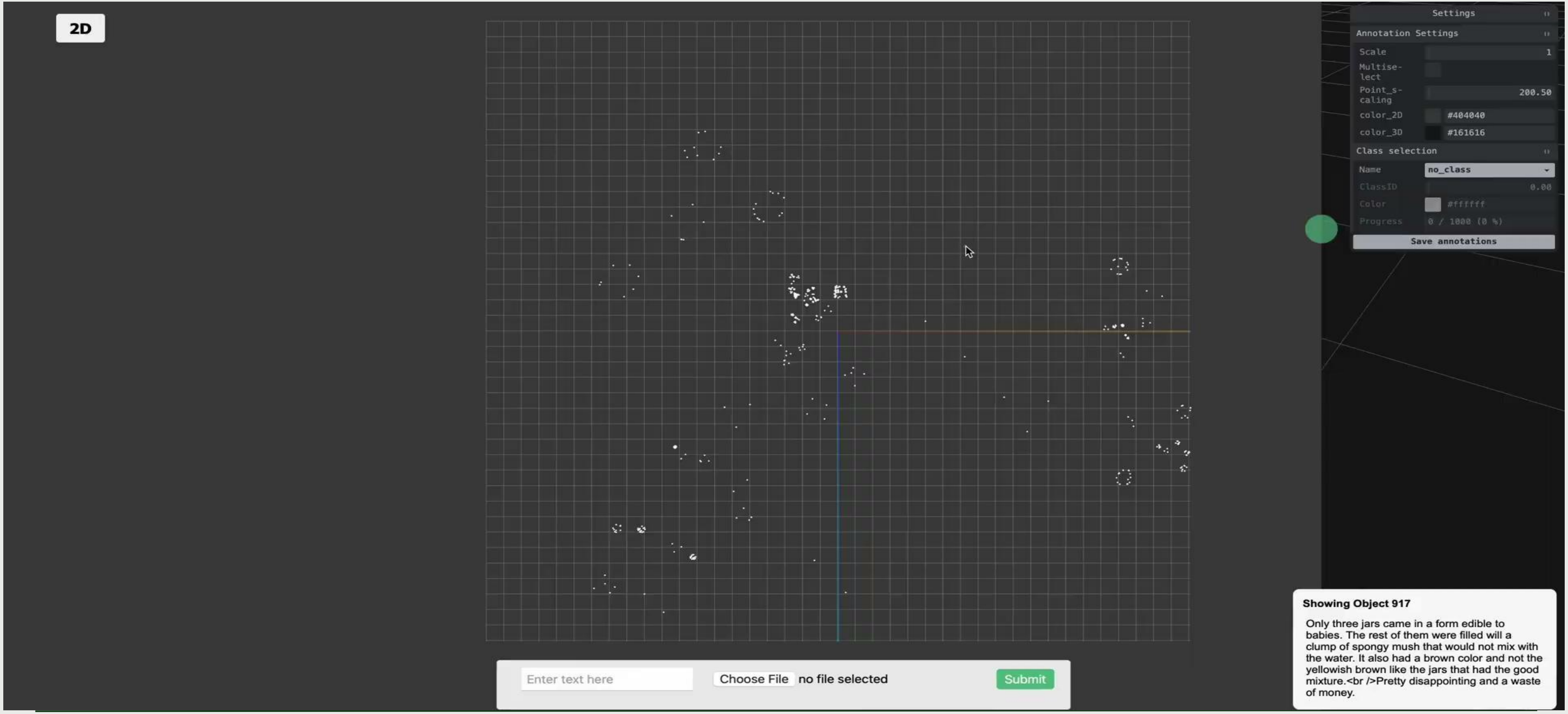
Clustering and Dimensionality Reduction

- PCA: Preserves linear structure
- t-SNE: Preserves local neighbor distributions
- UMAP: Preserves local connectivity
- h-NNE: Preserves hierarchical clustering structures

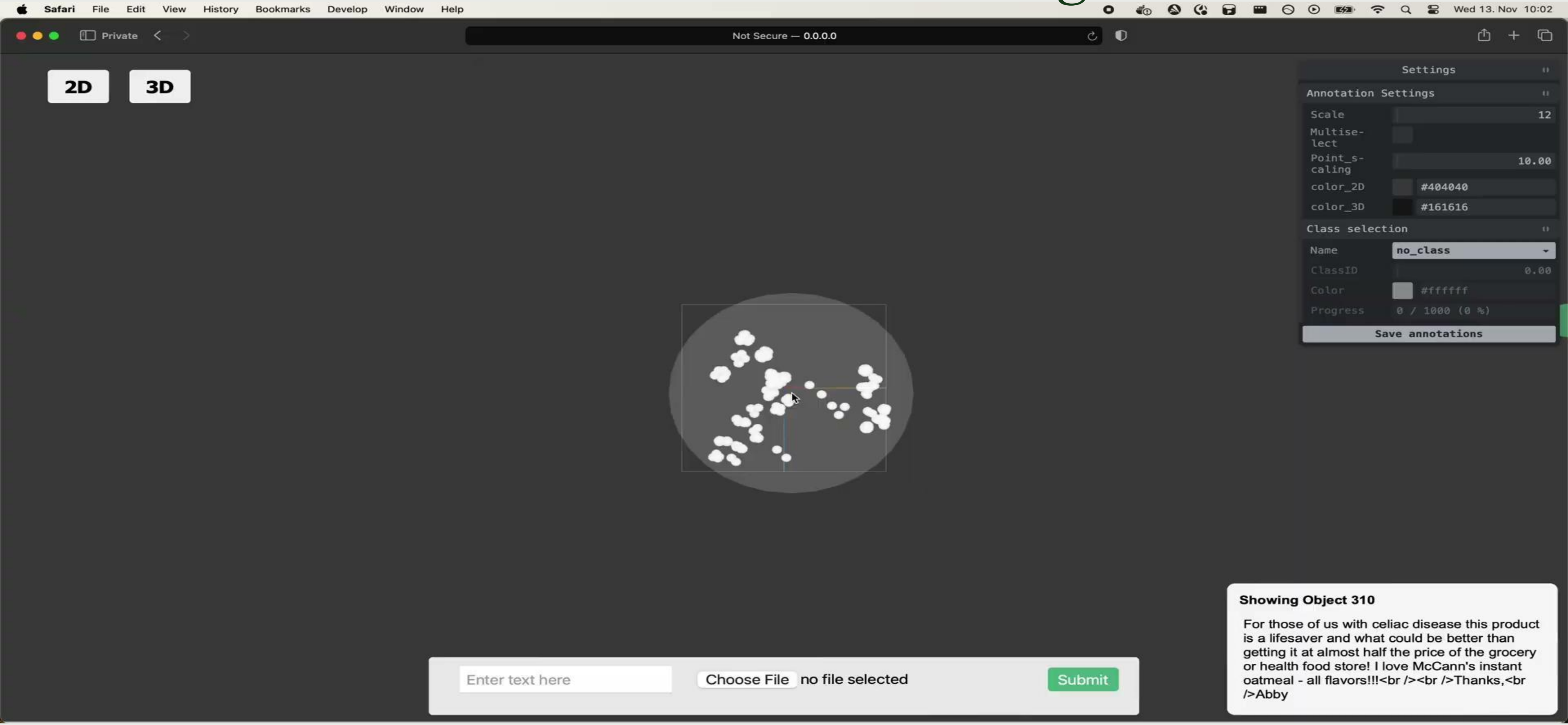


Visualization of an industrial Time Series dataset for Anomaly detection

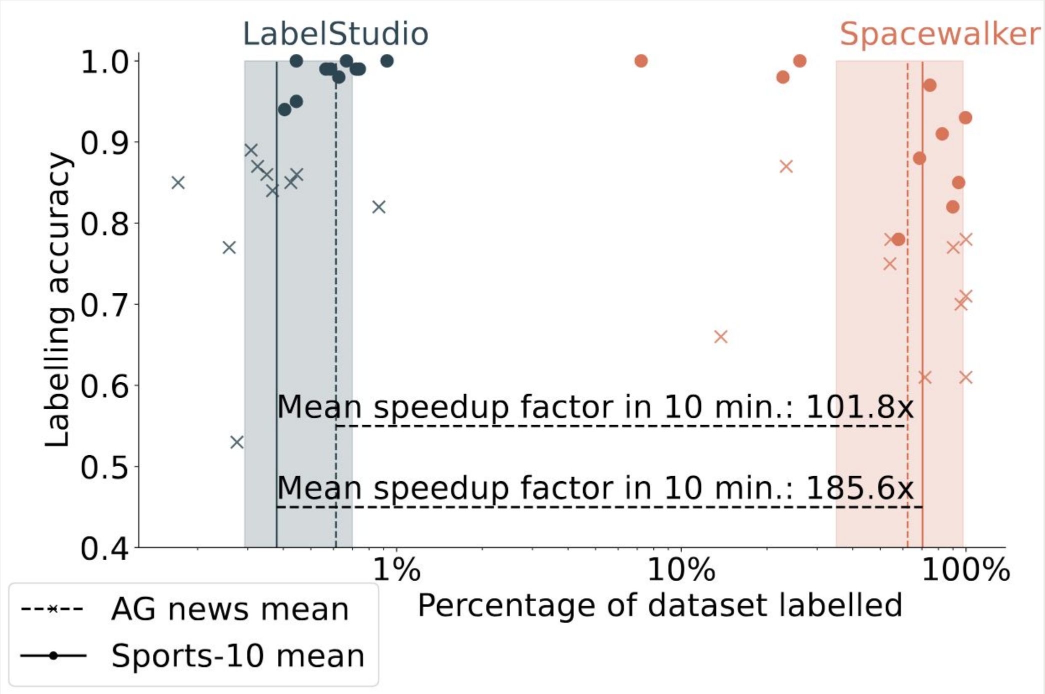
Data Annotation & Understanding



Data Annotation & Understanding



Labeling Speed in Latent Space



SpaceWalker: Traversing Representation Spaces for Fast Interactive Exploration and Annotation of Unstructured Data, Heine et al. MLVis 2025

Takeaways

- A lot of tradeoffs, select a good method for your problem
    - Accuracy vs. Speed
    - Method Complexity

Takeaways

- A lot of tradeoffs, select a good method for your problem
  - Accuracy vs. Speed
  - Model Complexity

- Not all methods or implementations scale well to large data
  - Memory Usage
  - Computational Time

Takeaways

- A lot of tradeoffs, select a good method for your problem
  - Accuracy vs. Speed
  - Model Complexity

- Not all methods or implementations scale well to large data
  - Memory Usage
  - Computational Time

- Importance of Distance Metrics
  - Data Modality Sensitivity
  - Impact on Clustering Results
  - Custom Metrics

Thanks for your attention!

References

References…

Backup

Clustering Evaluation

- Silhouette Score
  - Measures how similar an object is to its own cluster compared to other clusters

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

  - Failure Modes: Can be misleading if clusters have different densities or are not well-separated
  - Considerations: Works best with convex clusters

Clustering Evaluation

- Davies-Bouldin Index
  - Measures the average similarity ratio of each cluster with its most similar cluster

$$DB = \frac{1}{N} \sum_{i=1}^{N} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d_{ij}} \right)$$

  -
  - Failure Modes: Sensitive to the shape and size of clusters
  - Considerations: Lower values indicate better clustering

Clustering Evaluation

- Adjusted Rand Index (ARI)
  - Measures the similarity between two data clusterings, adjusting for chance

$$ARI = \frac{RI - \text{Expected RI}}{\max(RI) - \text{Expected RI}}$$

  - Failure Modes: Can be affected by the number of clusters and the size of the dataset
  - Considerations: Suitable for comparing clustering results with a known ground truth

Clustering Evaluation

- Normalized Mutual Information (NMI)
  - Measures the amount of information shared between the clustering and the ground truth

$$NMI = \frac{2 \times I(C;K)}{H(C) + H(K)}$$

  - Failure Modes: Can be affected by the distribution of cluster sizes
  - Considerations: Higher values indicate better agreement with the ground truth

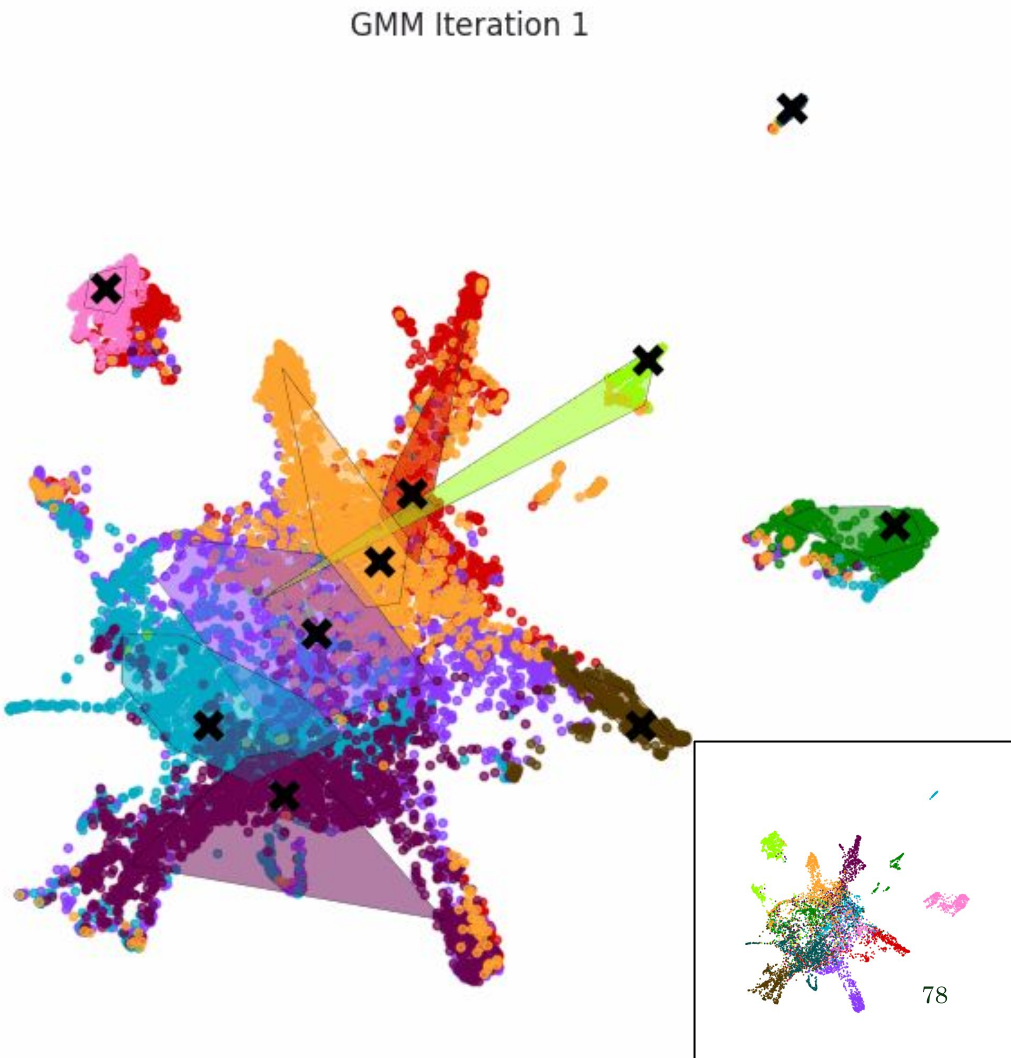
Expectation Maximization

- Expectation Step

$$w_{ij} = \frac{\pi_j \, p_{\mathcal{N}}(x_i|\mu_j,\Sigma_j)}{\sum_{s=1}^{k} \pi_s \, p_{\mathcal{N}}(x_i|\mu_s,\Sigma_s)}$$

- Maximization Step

$$\pi_j^{new} = \frac{1}{N} \sum_{i=1}^{N} w_{ij}$$



Expectation Maximization

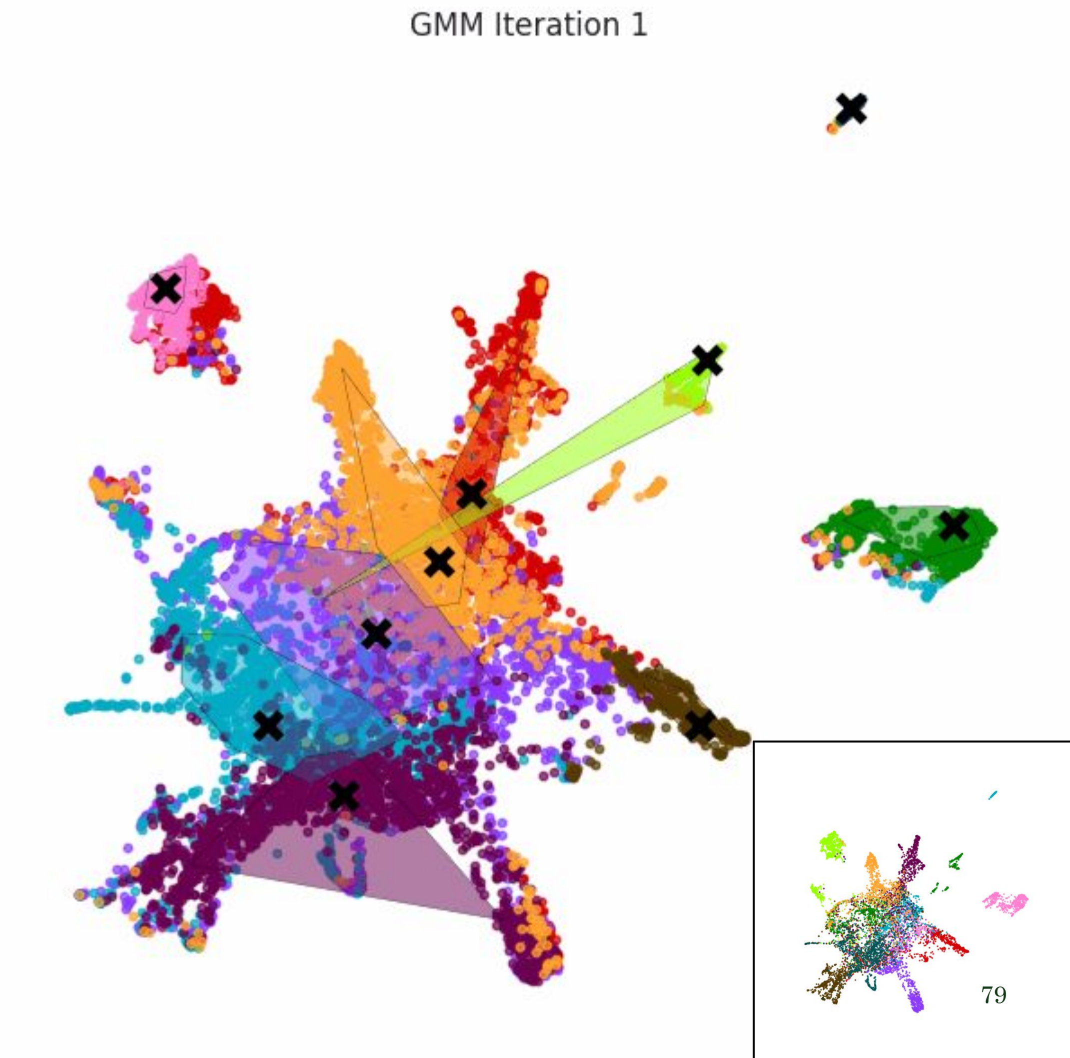
- Expectation Step

$$w_{ij} = \frac{\pi_j \, p_{\mathcal{N}}(x_i | \mu_j, \Sigma_j)}{\sum_{s=1}^{k} \pi_s \, p_{\mathcal{N}}(x_i | \mu_s, \Sigma_s)}$$

- Maximization Step

$$\mu_j^{new} = \frac{\sum_{i=1}^{N} w_{ij} x_i}{\sum_{i=1}^{N} w_{ij}}$$

$$\Sigma_j^{new} = \frac{\sum_{i=1}^{N} w_{ij} (x_i - \mu_j)^\top (x_i - \mu_j)}{\sum_{i=1}^{N} w_{ij}}$$

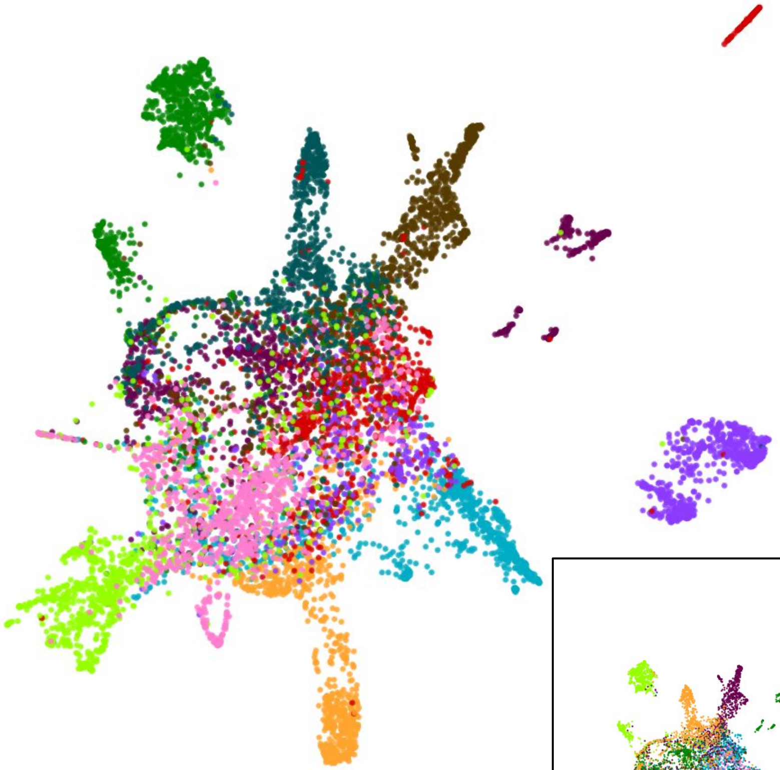


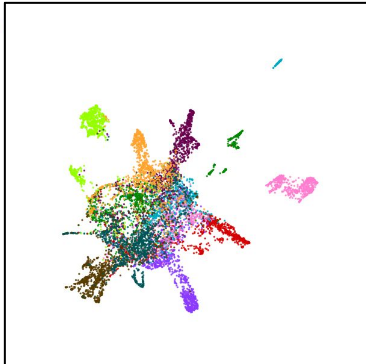
Complex methods tricky to use

- Are slower
- Numerically unstable due to large number of dimensions or non positive-definite covariance matrices
- Can regularize the matrices
- Reducing dimensionality and aligning the dimensions can help a lot, PCA

Clusters merged

- Especially on datasets with overlapping clusters
- Here global linkage criteria can help
- Can require a few more clusters than expected and visually separate
- Sensitive to outliers





DBSCAN

- Parameters: ε, minPts
- Core points:

$p \in X, |B(p, \varepsilon) \cap X| \geq minPts$

- Directly reachable:

$\exists p,\ p$ is a core point, $d(p, q) < \varepsilon$

- Reachable:

$\exists\ p = p_0, \ldots, p_n = q$

$\forall\ i \leq n - 1\ p_{i+1}$ reachable from p_i

- DBSCAN*: Core points