Machine Learning

Clustering and Manifold Learning

Dr Chris Willcocks
Department of Computer Science

Lecture Overview

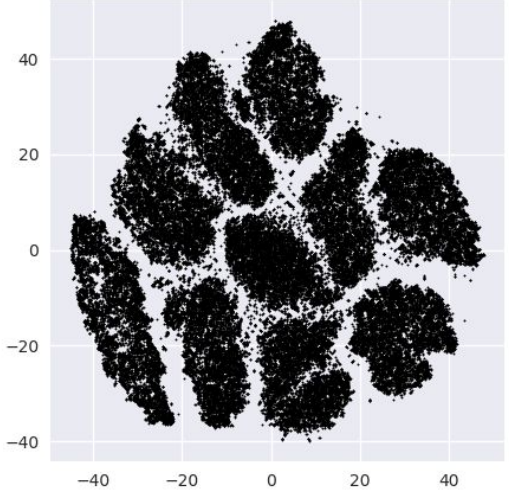
Durham University

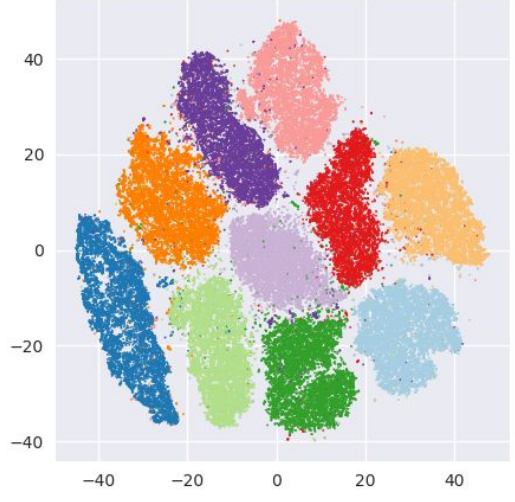
Recap

- Reinforcement Learning, RNNs
- Density Estimation

Today's lecture

- Clustering
  - Theory
  - k-Means
  - DBSCAN
  - ...
- Manifold Learning
  - T-SNE, UMAP
  - AE Embeddings, ...





Relation to Density Estimation

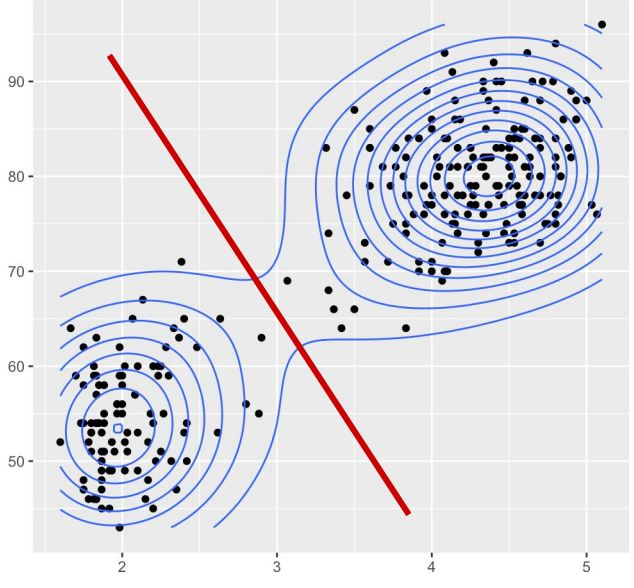
Durham University

Manifold Learning

- Given some unorganised (high-dimensional) data can we learn some (low-dimensional) structure?

Clustering

- Given some (structured) data, can we learn the decision boundaries?

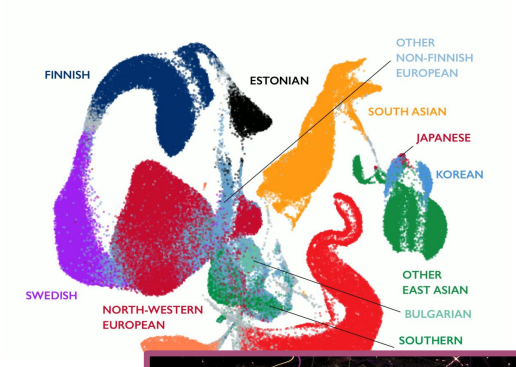


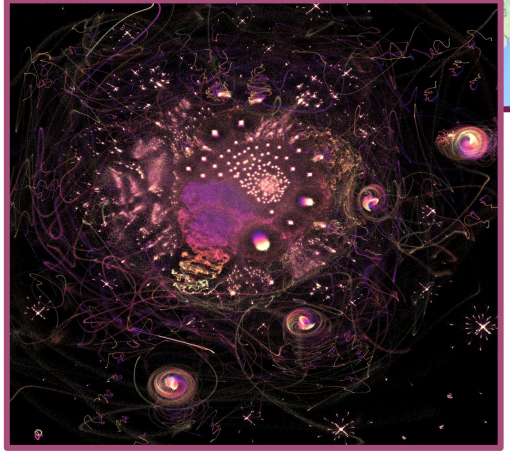
Manifold Learning & Clustering Applications

Durham University

- Plant and animal ecology
- Transcriptomics
- Business and marketing
- Credit risk & insurance
- Medicine
- Crime analysis
- Education
- Politics and demographics
- Chemistry
- Climatology
- *Pointless* but beautiful math
- ...

https://macarthurlab.org/2018/10/17/gnomad-v2-1/





https://johnhw.github.io/umap_primes/index.md.html

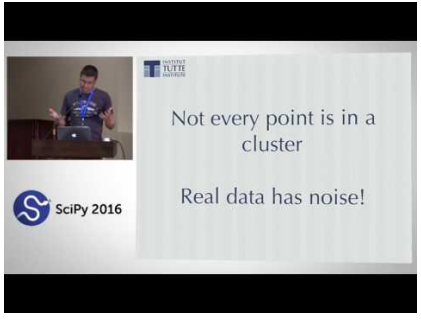
Clustering

The **task**

- Assign observations of a PDF to coherent clusters or label them as noise

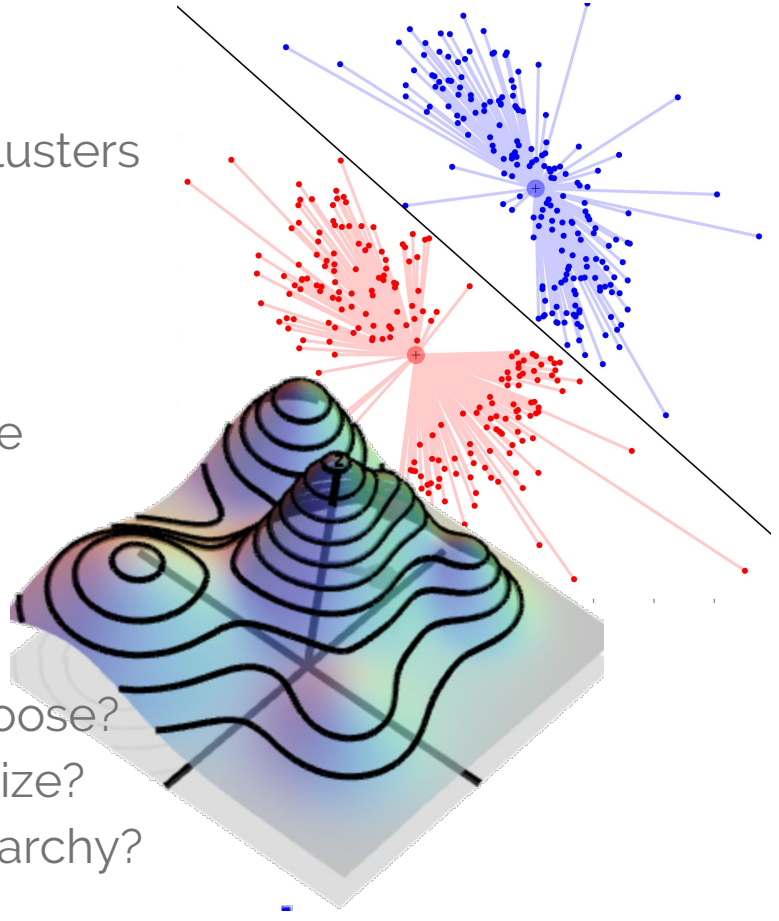
A **cluster** is a...

- **Connected-component** of a **level set** of the **unknown PDF** over our data **observations**



Some of the **challenges**

- What algorithm to choose?
- What metric to minimize?
- What scale, what hierarchy?



Types of Clustering Algorithm

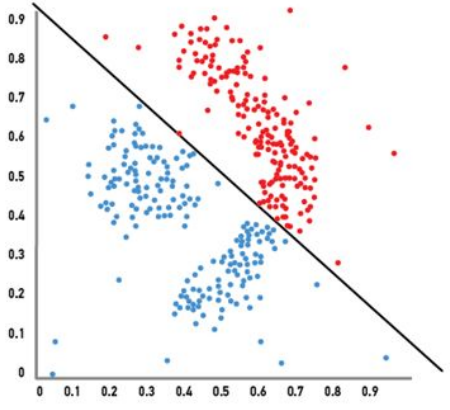
Durham University

Assignment:

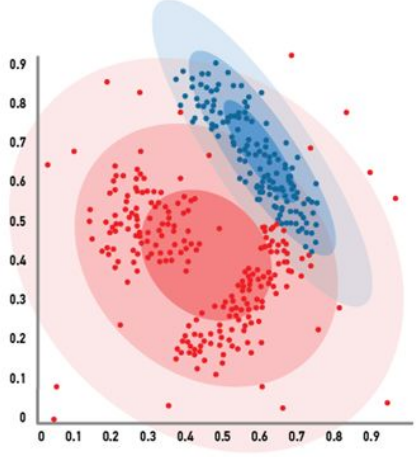
1. **Hard** clustering
2. **Soft** (fuzzy) clustering

Categories

- Centroid-based
- Density-based
- Distribution-based
- Connectivity-based
  - Bottom-up
  - Top-down



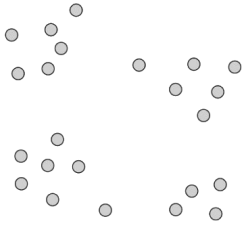
Hard clustering



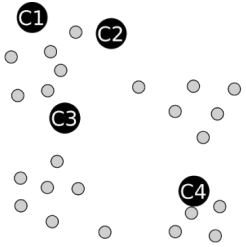
Fuzzy clustering

k-Means Clustering Algorithm

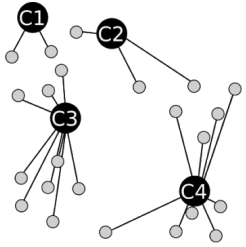
Durham University



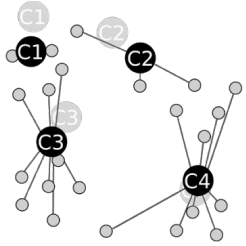
Input data



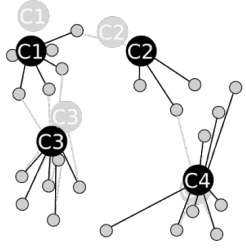
(1) Pick *k* random centroids



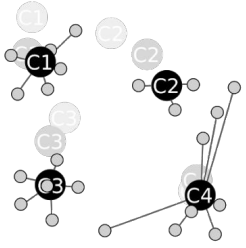
(2) Assign each point to nearest centroid



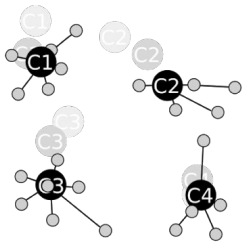
(3) Update centroids to mean of assigned



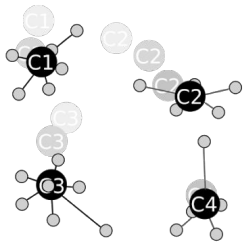
Repeat (2)



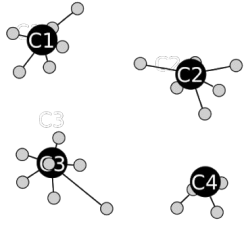
Repeat (3)



Repeat (2)

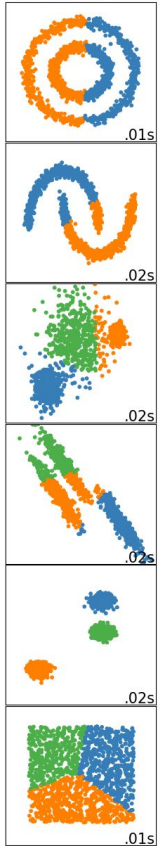


Repeat (3)

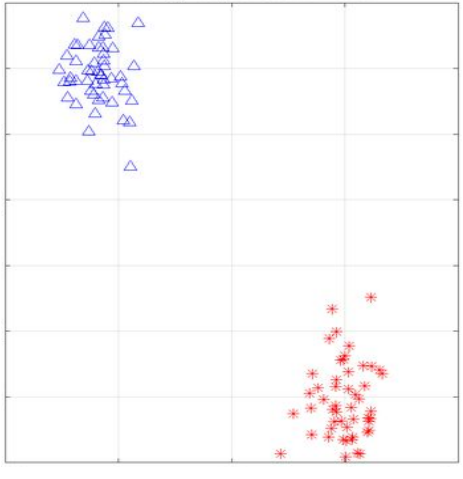


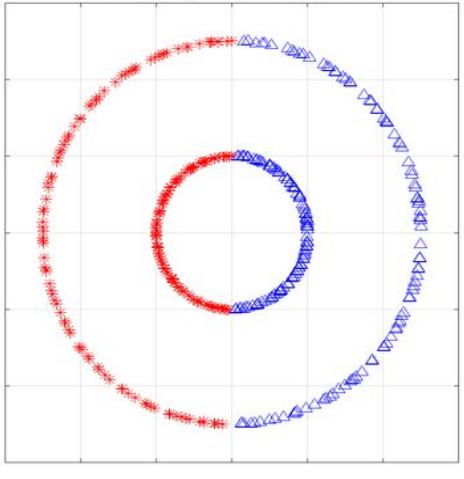
No movement = convergence

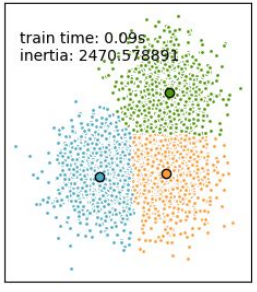
Properties (limitations) of *k*-Means



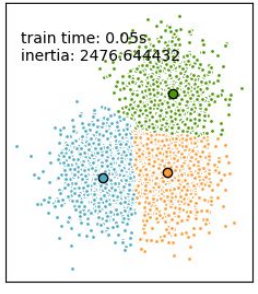
- Requires *k* to be known in advance (a prior)
- Assigns noise to clusters
- Can exhibit slow convergence
  - Mini-batch *k*-means, spatial index data structure
- Requires data to be <u>linearly</u> separable (convex clusters):



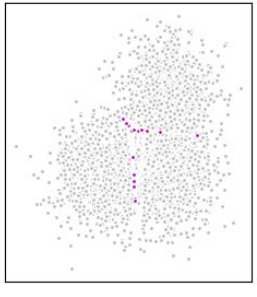




k-Means



Mini-batch k-Means



Difference

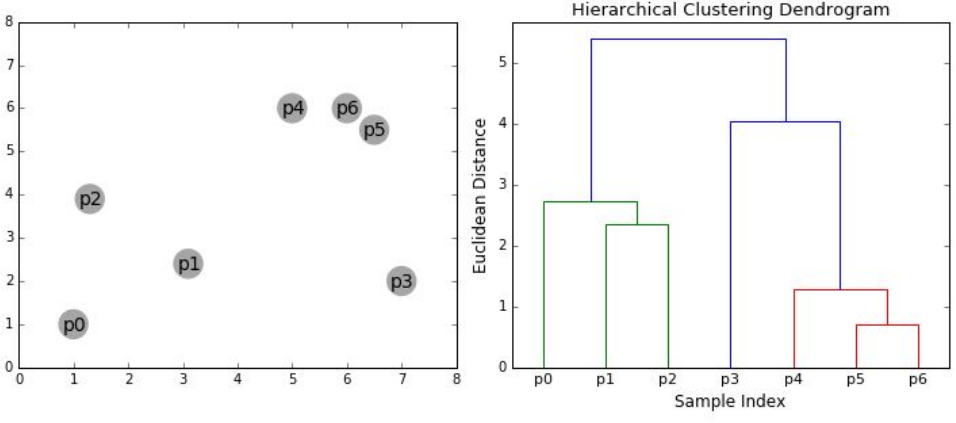
Hierarchical Clustering

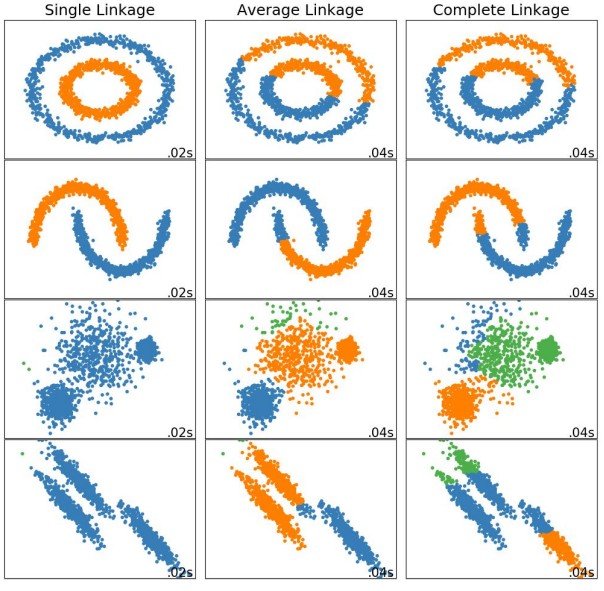
Durham University

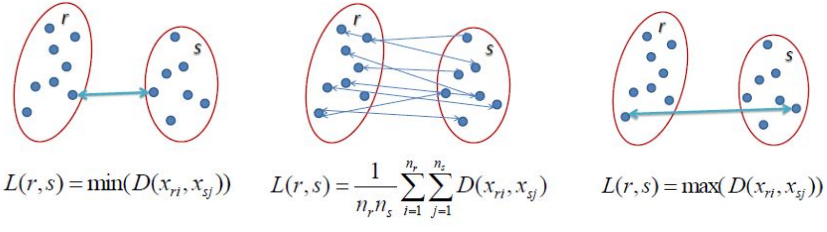
Hierarchical clustering methods are:

1. **Top-down** (divisive)
2. **Bottom-up** (agglomerative)

Various splitting/merging strategies….







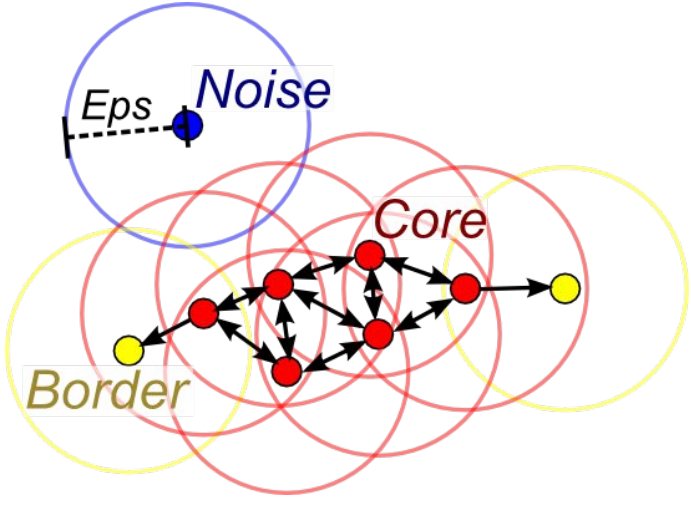
Density-based (DB)
Clustering

A Density-Based Algorithm for Discovering Clusters in Large Spatia
https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf
by M Ester - Cited by 13952 - Related articles
In this **paper**, we present the new clustering algorithm **DBSCAN** relying on density-based notion of clusters which is designed to dis- cover clusters of

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

Two parameters:

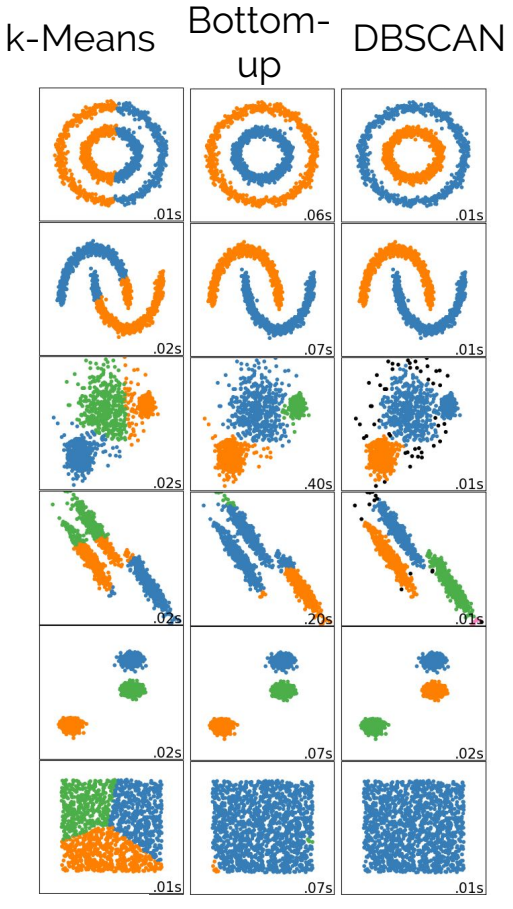
- Maximum radius $\varepsilon = 0.4$
- Minimum points $\delta = 5$



Algorithm:

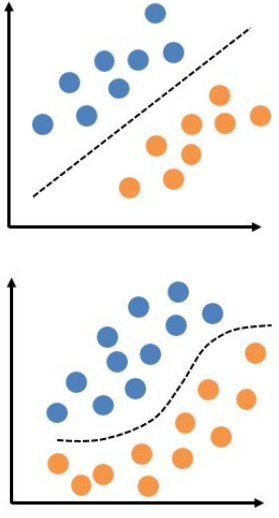
1. For all points, identify core points
2. Find core points connected components, these are the clusters
3. Assign density-reachable borders also to clusters, otherwise its noise

Properties of DBSCAN



Main **strengths**

- Creates arbitrarily shaped clusters
- Classifies noise
- Average runtime $n \log n$ with spatial indexing structure



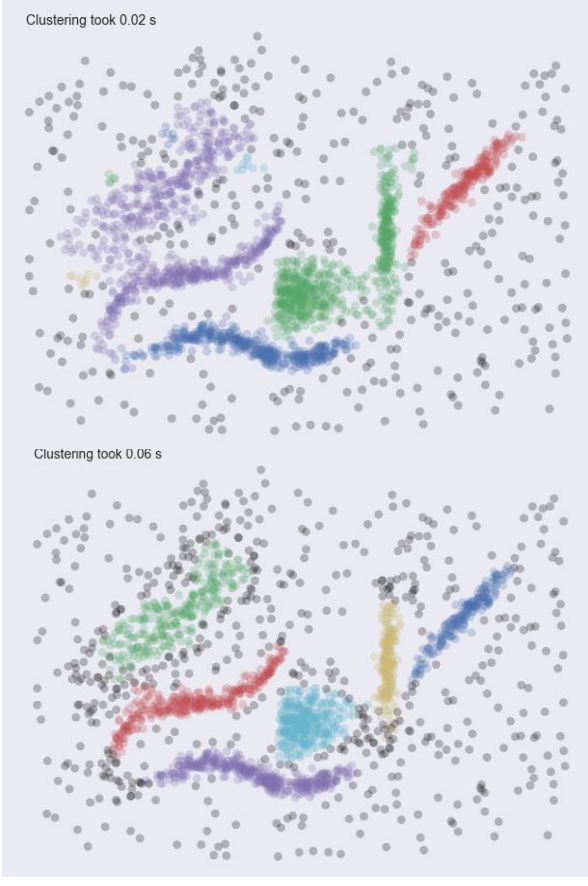
Main **limitations**

- Quite difficult to tune ε and δ
- Non-deterministic
  - Although modification exists
- Relies on euclidean distance measure

HDBSCAN (Hierarchical DBSCAN)

Durham University

- More recent algorithm by some of same authors of DBSCAN
- Combines Hierarchical clustering with DBSCAN
- Just one intuitive parameter:
  - **minClusterSize**
- (Also has a *minSamples* parameter but is insensitive to it and can be set with sane defaults)
- Supports clusters of varying density
- Can be implemented reasonably efficiently

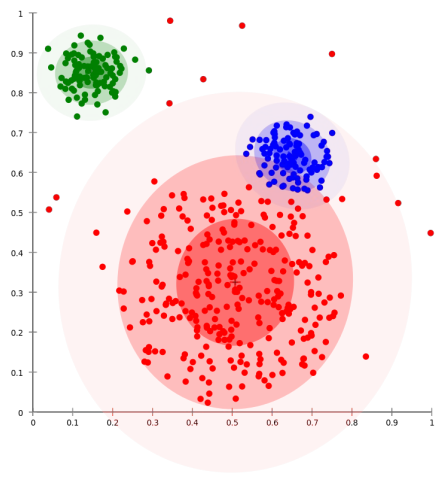


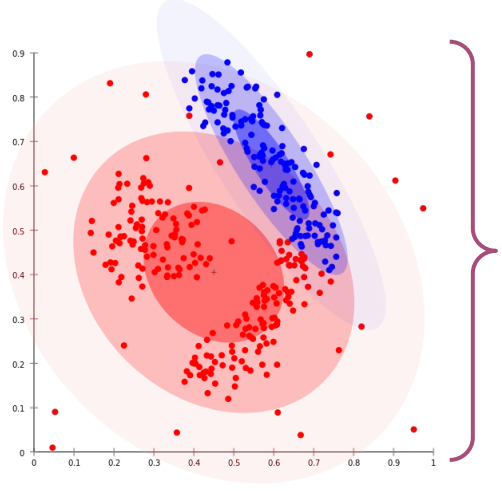
Distribution-based Clustering

Supposing our data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters, we can...

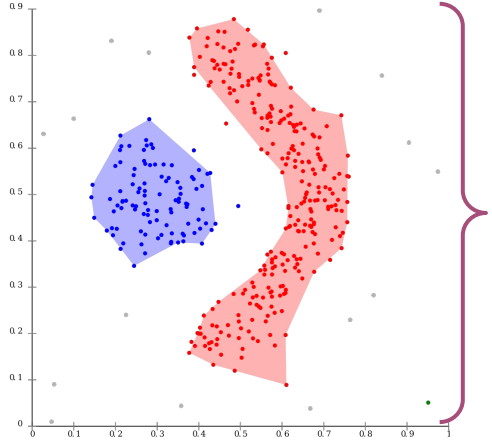
- Assign to distribution in a k-Means like fashion
- Optimise the parameters of the prior distributions such as to maximize the likelihood of the data given those assignments. Then repeat.

GMMa (good case)





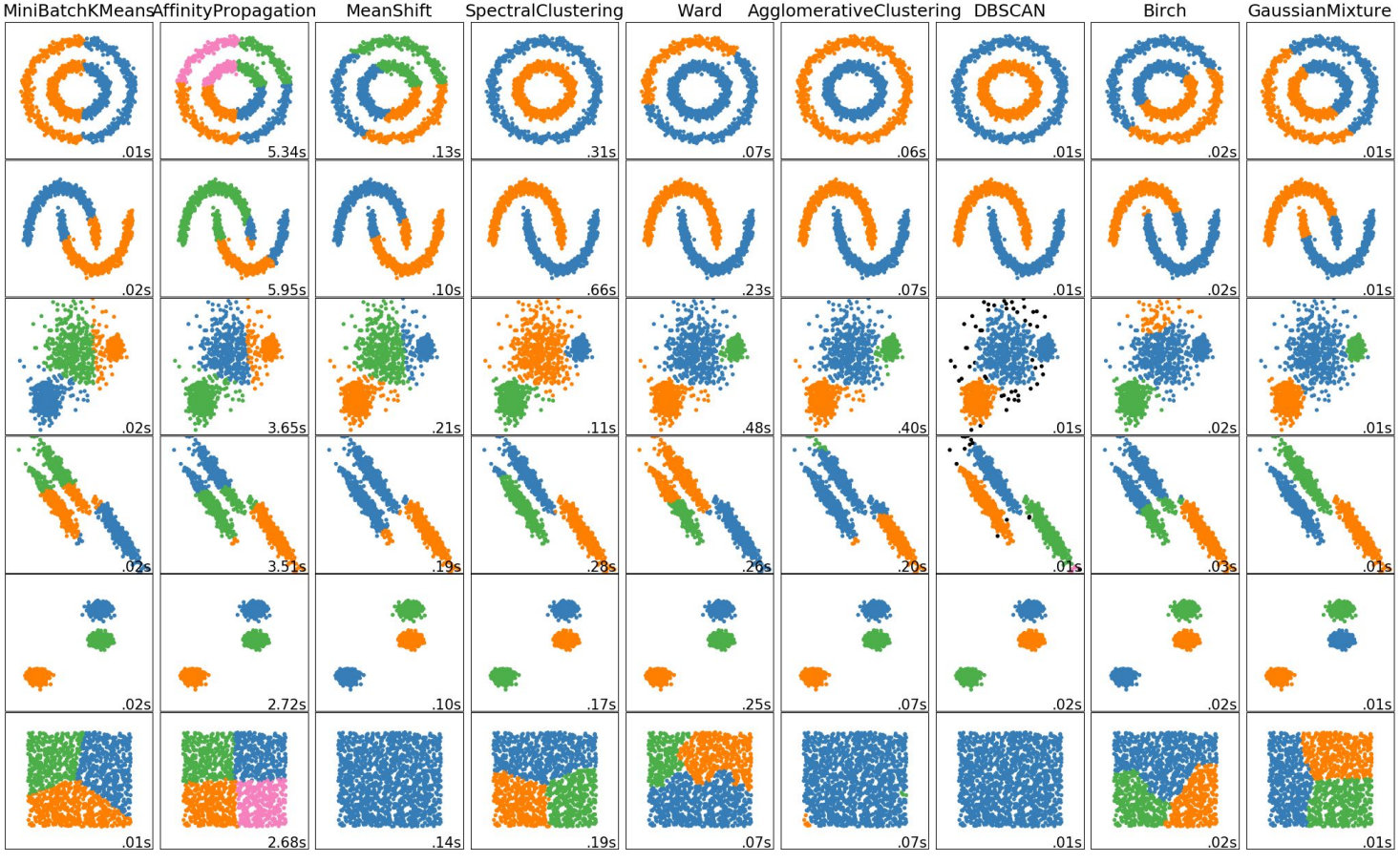
GMM (bad case)



DBSCAN

Clustering Algorithms...

Durham University



Python implementations in **scikit-learn**

https://scikit-learn.org/stable/modules/clustering.html

HDBSCAN separate: https://hdbscan.readthedocs.io/

Manifold Learning (or NLDR)

Durham University

The **task**

- Given some unorganised (high-dimensional) data can we learn some organised (low-dimensional) manifold that captures the data distribution?

The **challenges**

- What should the manifold look like, and why?
- Curse of dimensionality
  - Distances in high-dimensional spaces are all similar
- What neighborhood to define?
- What metric to minimize?

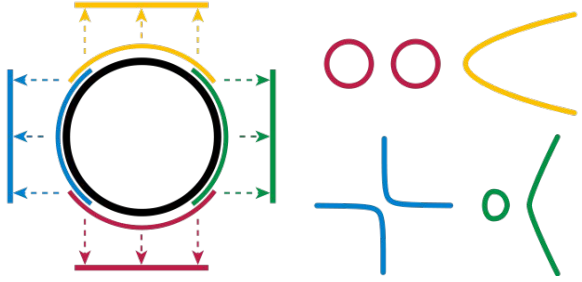
The **application:**

- Save space, understand data patterns, ...

Terminology

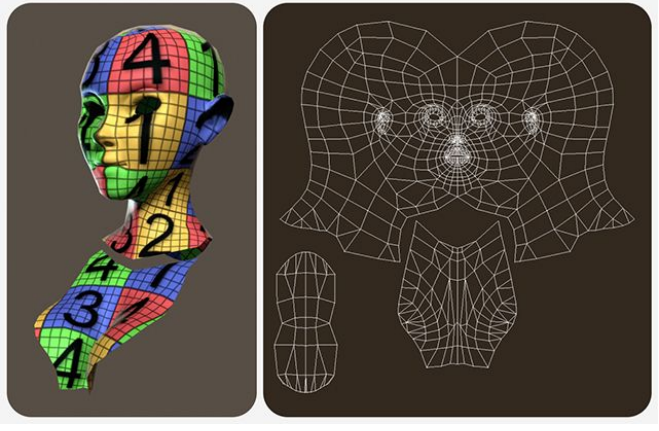
Manifold

- A space that locally resembles Euclidean space.
- Hierarchy of types of manifold:
  - Topological
  - Differentiable
  - Riemannian



Embedding

- An injective map $\phi : \mathcal{M} \to \mathcal{N}$ that preserves the underlying structure.

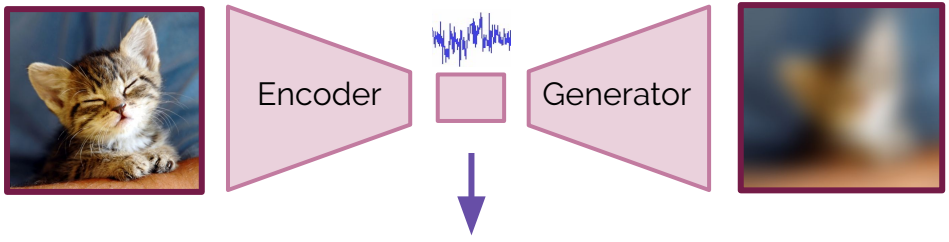


"Recovering the riemannian metric"

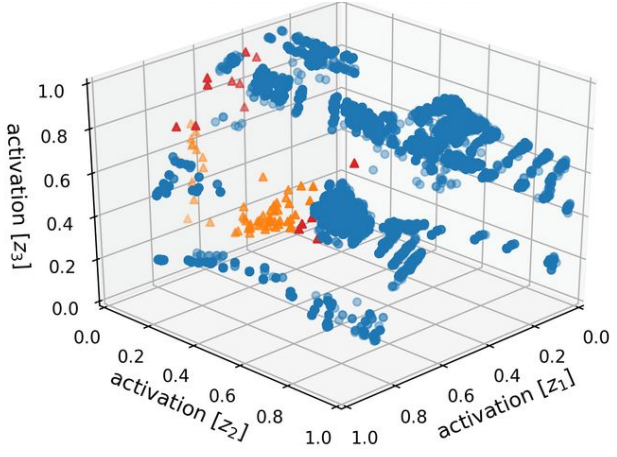
Autoencoders

Durham University

Learn a non-linear mapping from high-dimensional to an embedded space



- Local variations with $L2$
- **Can fracture** a manifold into many different domains (giving different regions for nearby data points)
- If we have **prior knowledge** about data, we can split the manifold into lovely categories easily
  - Global coherency

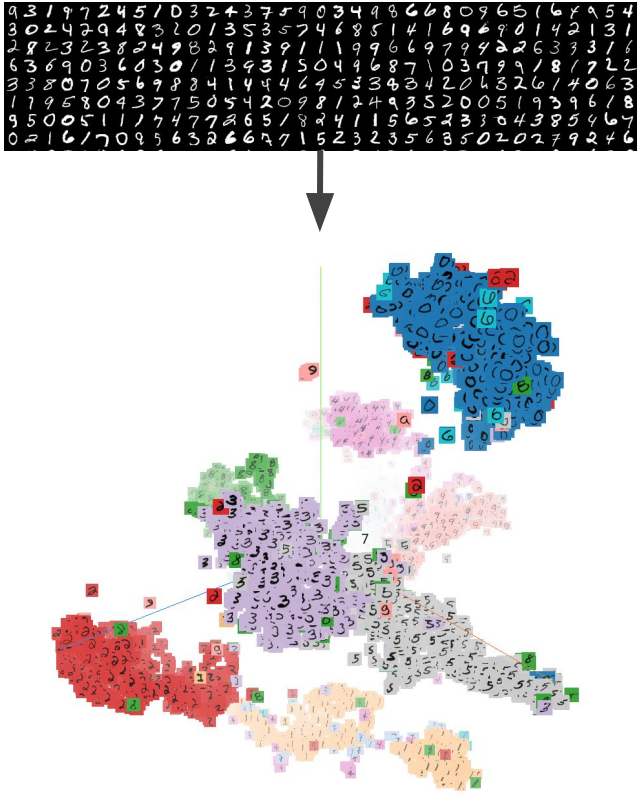


T-**D**istributed **S**tochastic **N**eighbor **E**mbedding

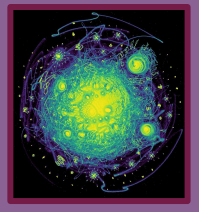
Visualizing Data using t-SNE - Journal of Machine Learning Research
www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf ▾
by L Maaten - 2008 - Cited by 7224 - Related articles
We present a new technique called "**t-SNE**" that visualizes high-dimensional da
.... present in the **paper** are sufficient to demonstrate the superiority of **t-SNE**.

The t-SNE algorithm:

- Uses local relationships between points to create a low-dimensional mapping.
  - Definies this relation by assuming an **adaptive Gaussian prior** based on a "perplexity" hyperparameter
- Defines a distribution over points in low-dimensional space, and minimizes KL-divergence between the two distributions
  - Ensures dissimilar points are far apart



UMAP



UMAP: Uniform Manifold Approximation and Projection for Dimension ...
https://arxiv.org › stat
by L McInnes - 2018 - Cited by 78 - Related articles
9 Feb 2018 - The **UMAP** algorithm is competitive with t-SNE for visualization quality, and ... Furthermore, **UMAP** has no computational restrictions on

Durham University

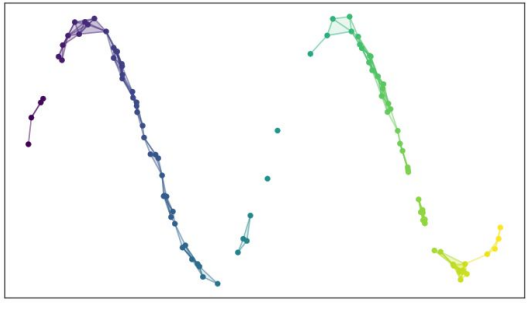
The UMAP algorithm:

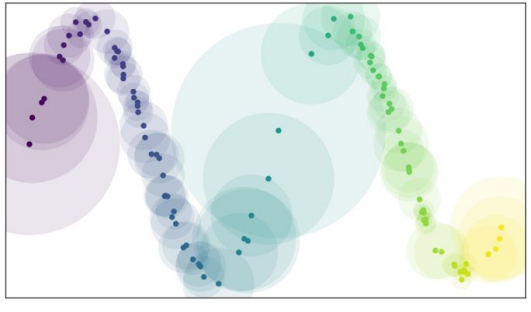
- Finds k-nearest neighbors in dataset
  - Creates edge weights based on diameter of neighborhood graph and distance to nearest neighbor
- Minimize:

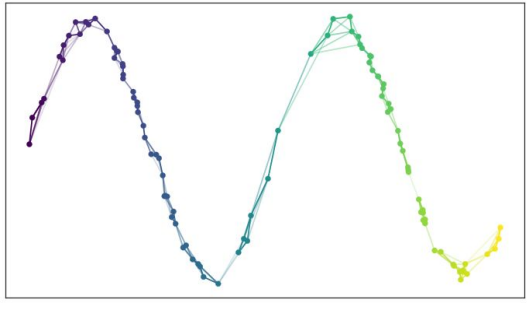
$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

Get the clumps right

Get the gaps right

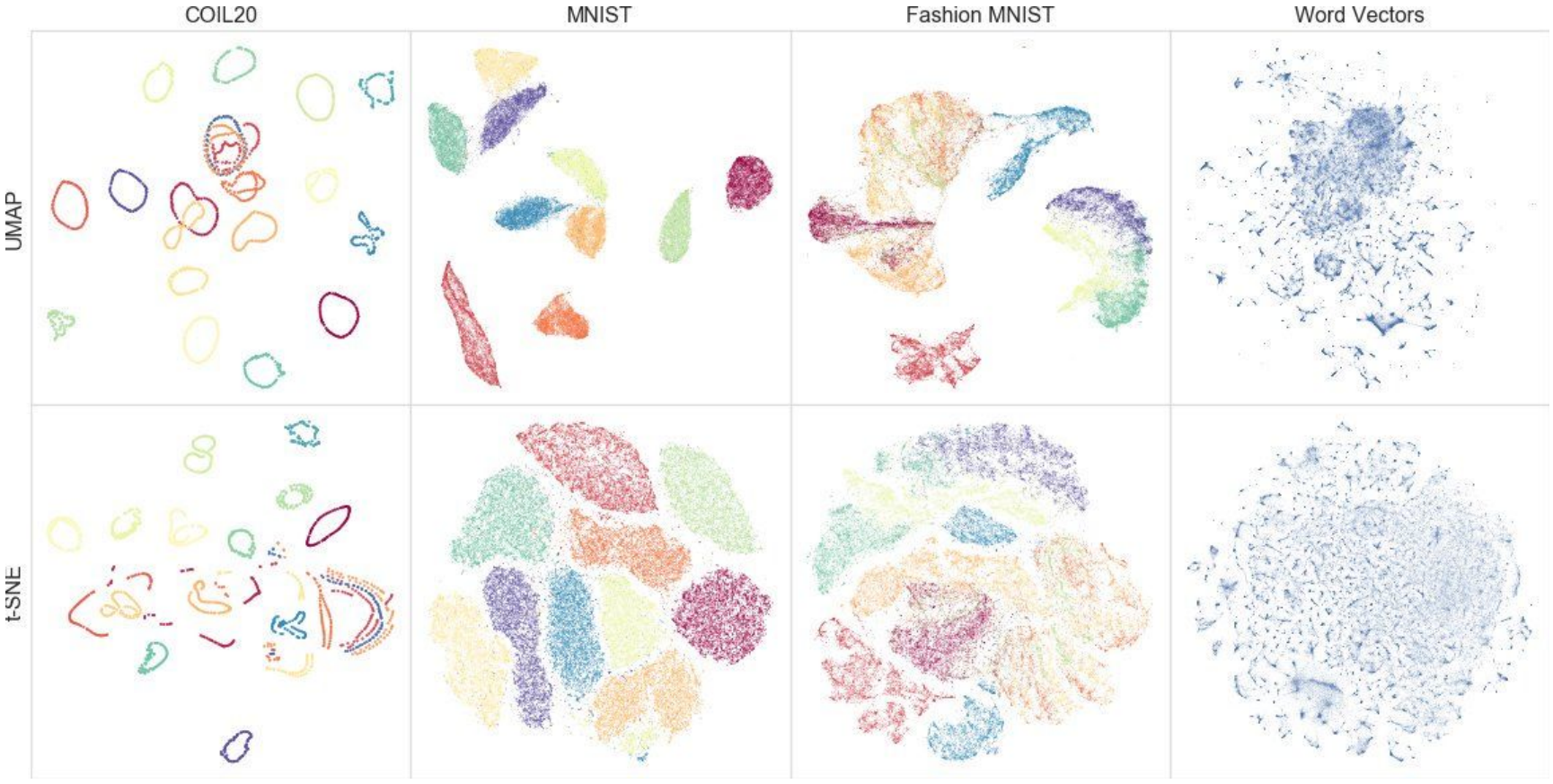






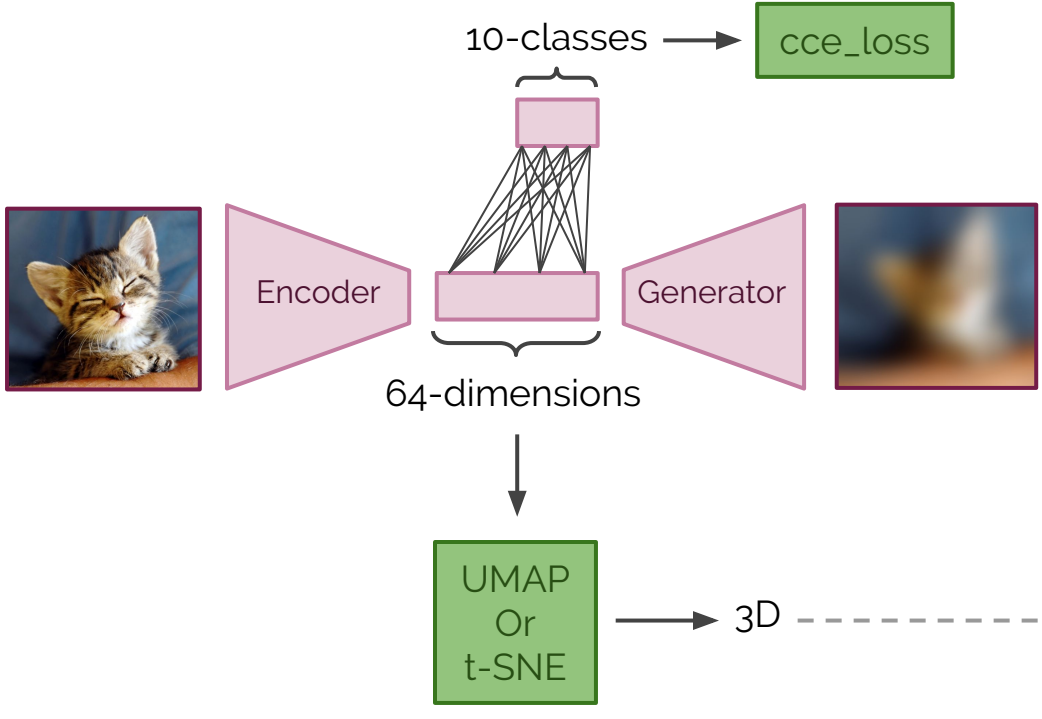
UMAP and t-SNE Examples

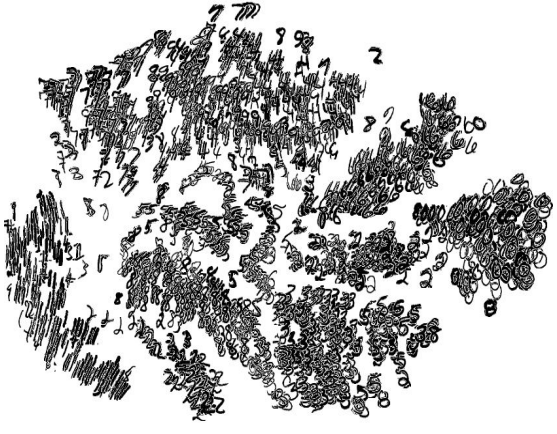
Durham University



Learning Tailored Manifolds

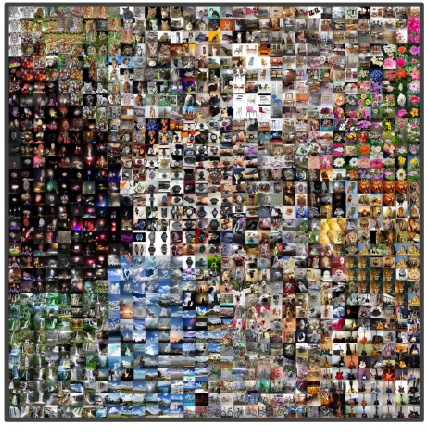
Durham University

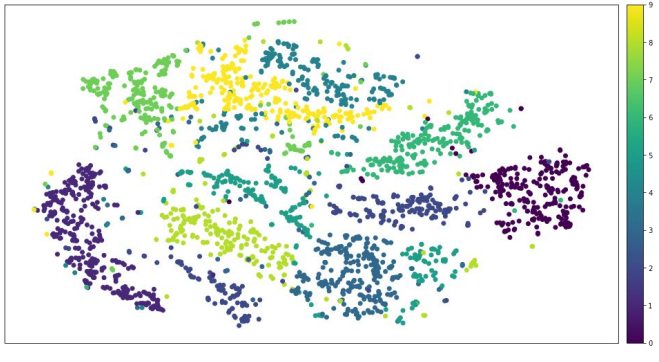


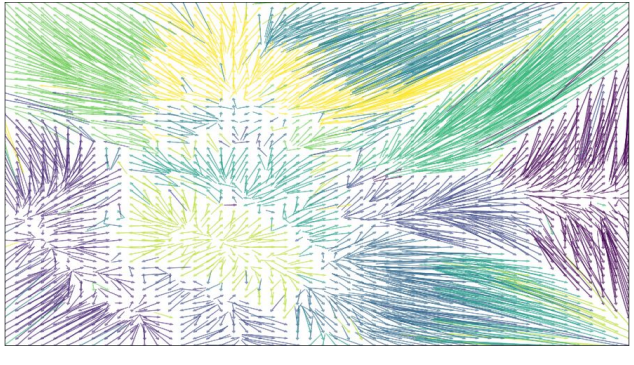


Take Away Points

- There is no one clustering algorithm that outperforms all…
  - But density-based stuff is generally pretty good in low-dimensions
- There's an infinite number of great manifolds out there
  - **UMAP** is currently a pretty good NLDR algorithm
- Good density estimators underpin a lot of this field
  - Especially defining local neighborhoods is important
- In practice, UMAP works pretty well with **HDBSCAN**.
  - See: https://umap-learn.readthedocs.io/en/latest/clustering.html







Jonker-Volgenant algorithm