Big Data

VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

AIML427

Manifold Learning/
Nonlinear Dimensionality Reduction (2)

Dr Bach Nguyen

(with thanks to Dr Andrew Lensen)

Bach.Nguyen@vuw.ac.nz

t-distributed Stochastic Neighbor Embedding (t-SNE)

- Last lecture, we focused on preserving distances/rankings
- t-SNE instead uses *probability distributions*
  - How likely would you select a point as a neighbour?

"The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian centred at x_i" [1]

- The "t-" stands for the use of the Student's *t*-distribution

[1] L. Van der Maaten and G. Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9(11) (2008)

Lots of maths, but essentially…

- Calculate neighbour probabilities for each pair of instances
- Symmetrise: $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$ (for N points)

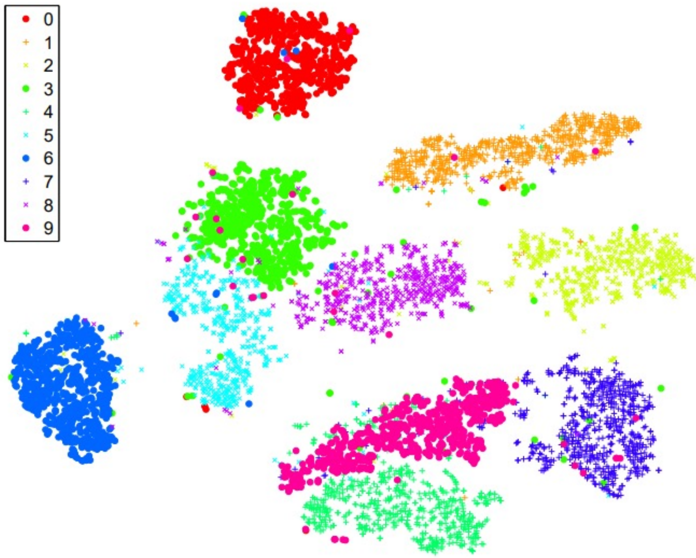
  (NB: $p_{ij} = p_{ji}$ and $p_{ii} = 0$)
- Set of all p_{ij} forms **P**, the probability distribution in high-dimensional space
- Use a similar approach to calculate **Q** (low-dim space)

- Optimise by minimising the (Kullback-Leibler) difference between these two distributions:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- Use gradient descent to optimise low-dim space

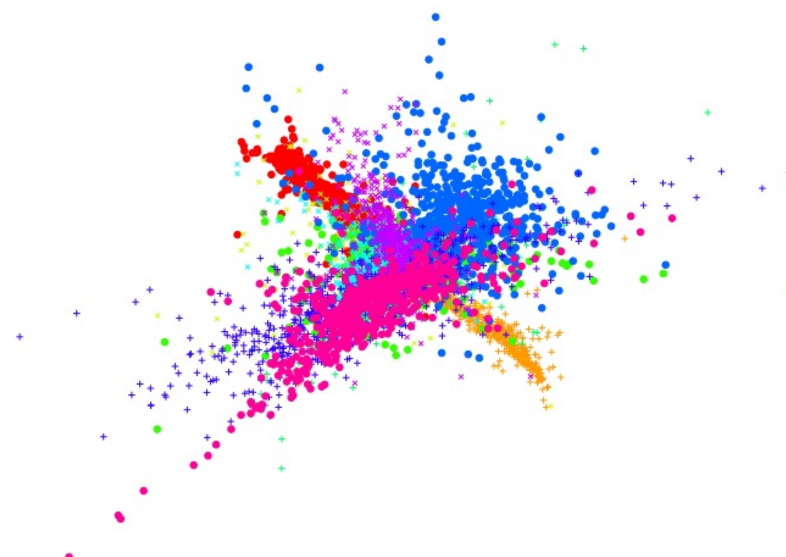
So…is it any good?



(a) Visualization by t-SNE.

(b) Visualization by Sammon mapping.

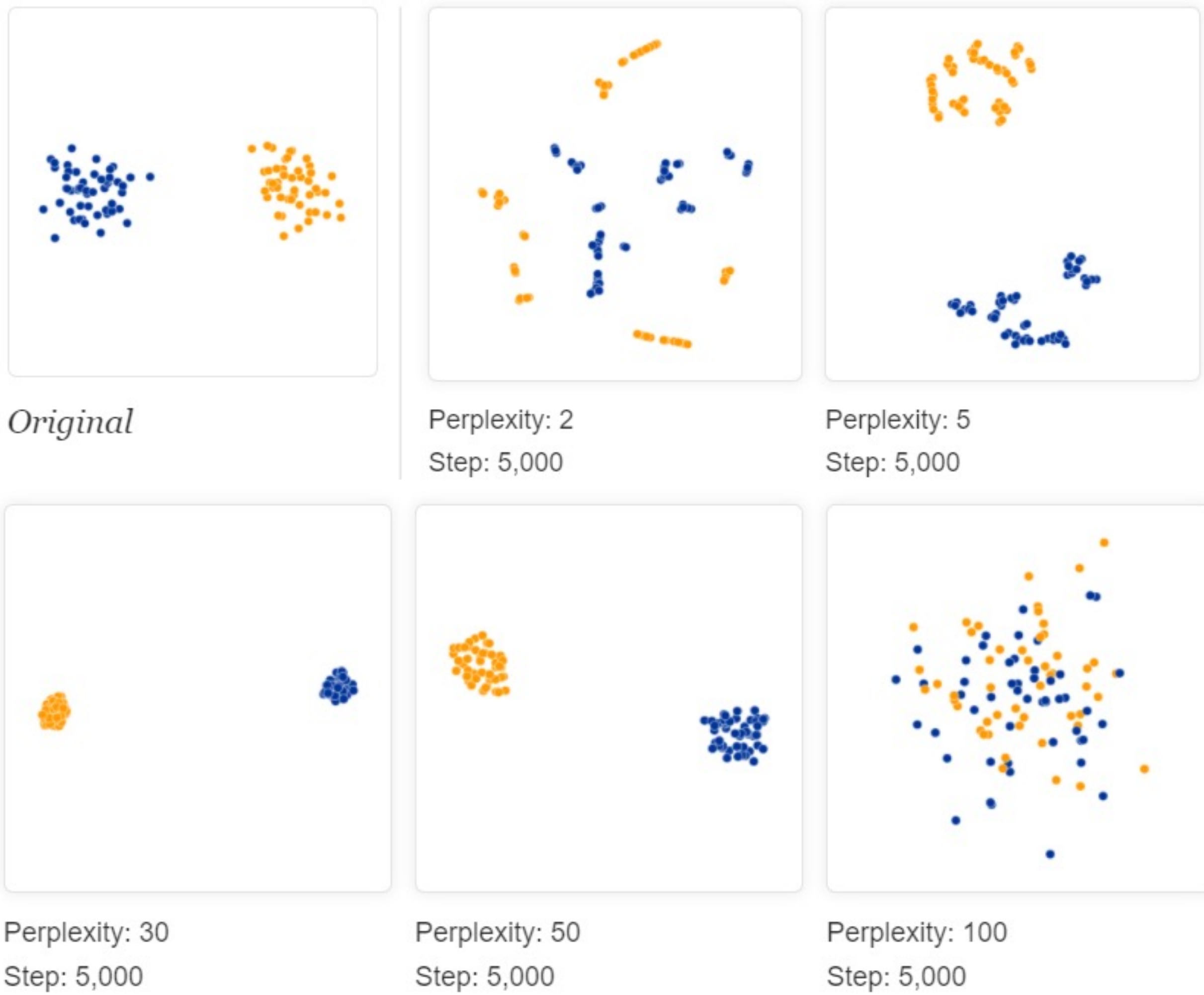
(a) Visualization by Isomap.



(b) Visualization by LLE.

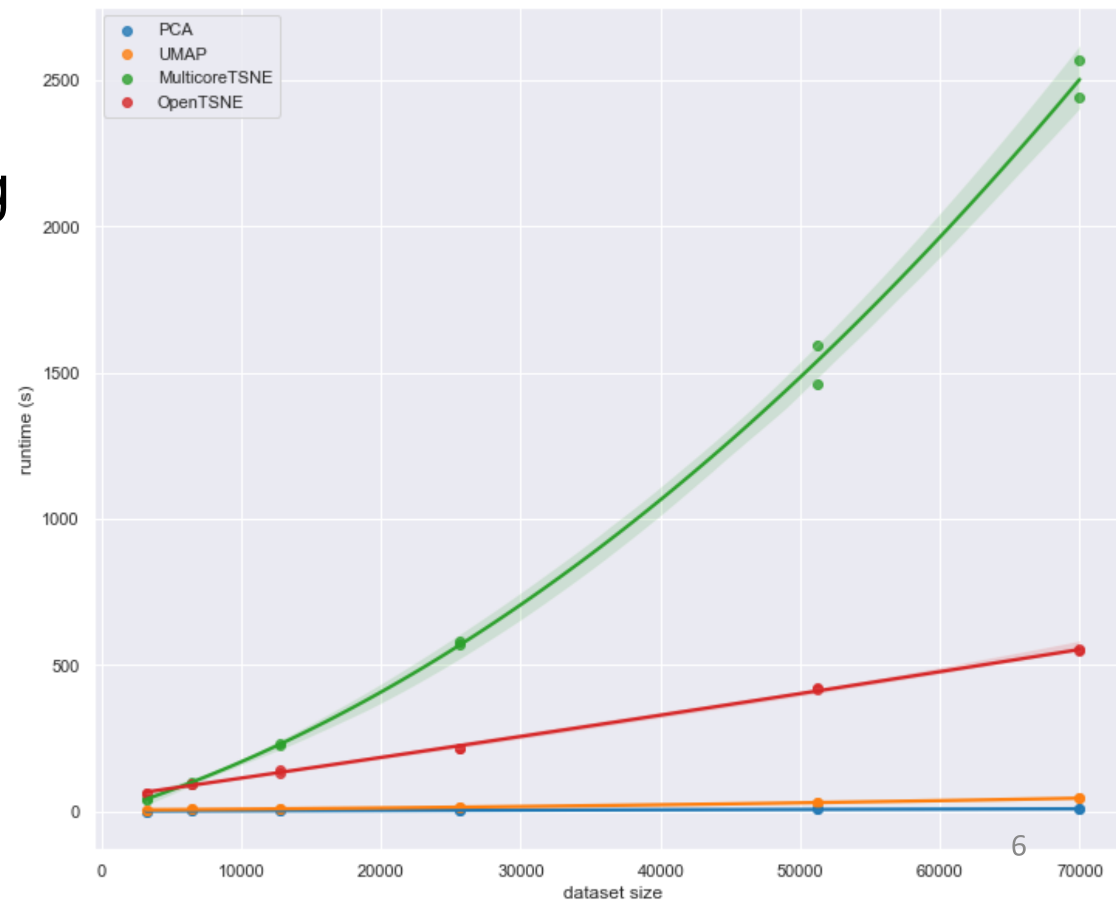
t-SNE: issues

- A stochastic algorithm: different results each run
- Perplexity parameter balances local vs global structure



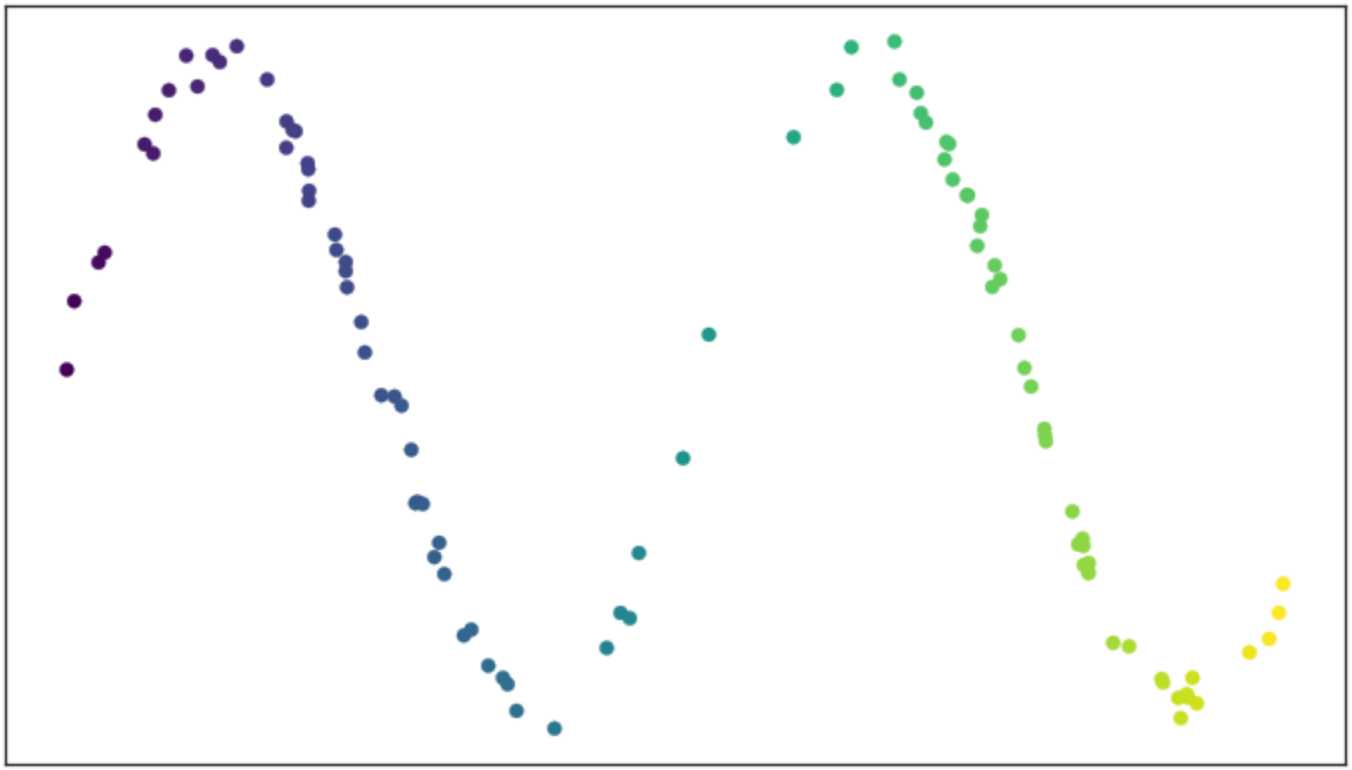
t-SNE: issues

- First and foremost, designed for visualisation
  - Not clear how well it works in d > 3 dimensions
- Hyperparameters can be quite sensitive
- Computationally expensive
- Still no *mapping* from high-dim to embedding
  - Parametric t-SNE exists, but…
- Was state-of-the-art from **2008** until **2018**!
  - *…and then?*



Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)

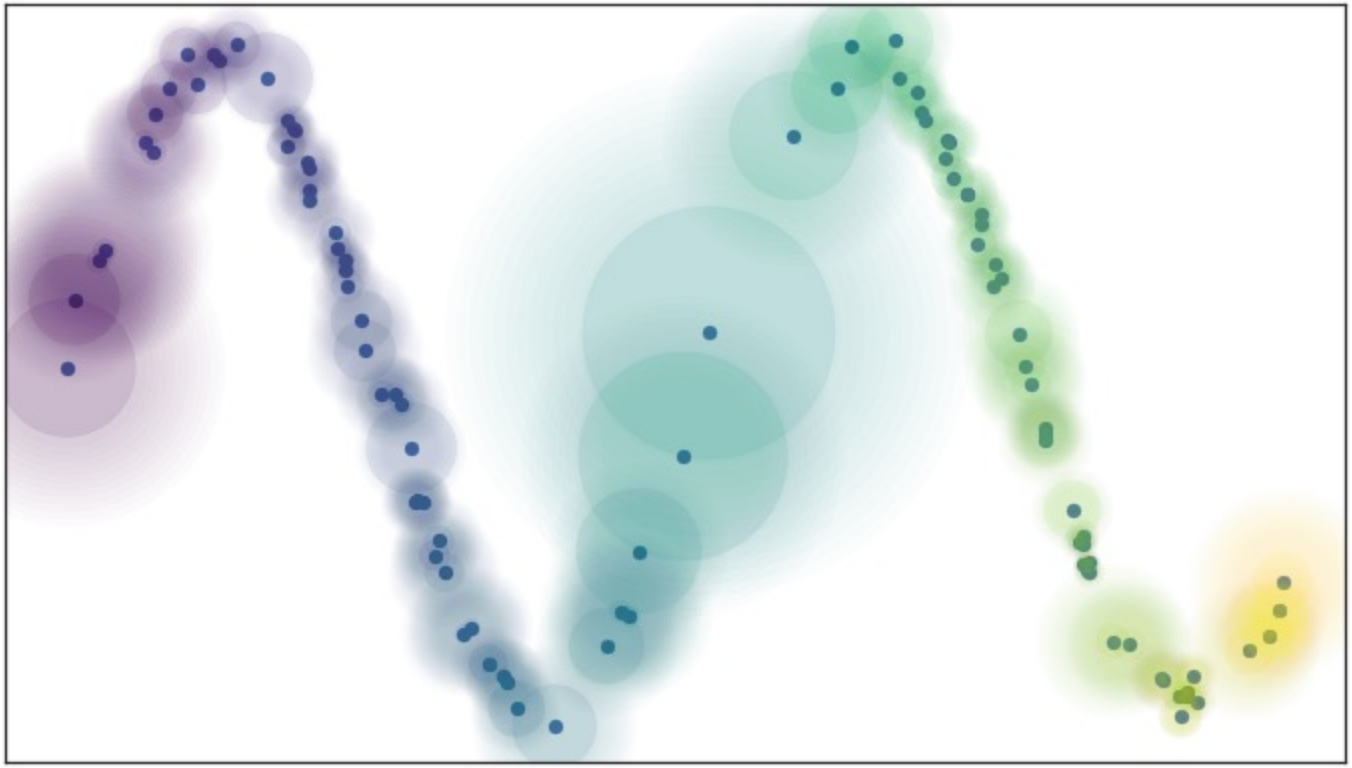
- How does it work? Similar, in ways, to t-SNE
  - Assumes the data is "*uniformly distributed on Riemannian manifold*"



Long version: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP

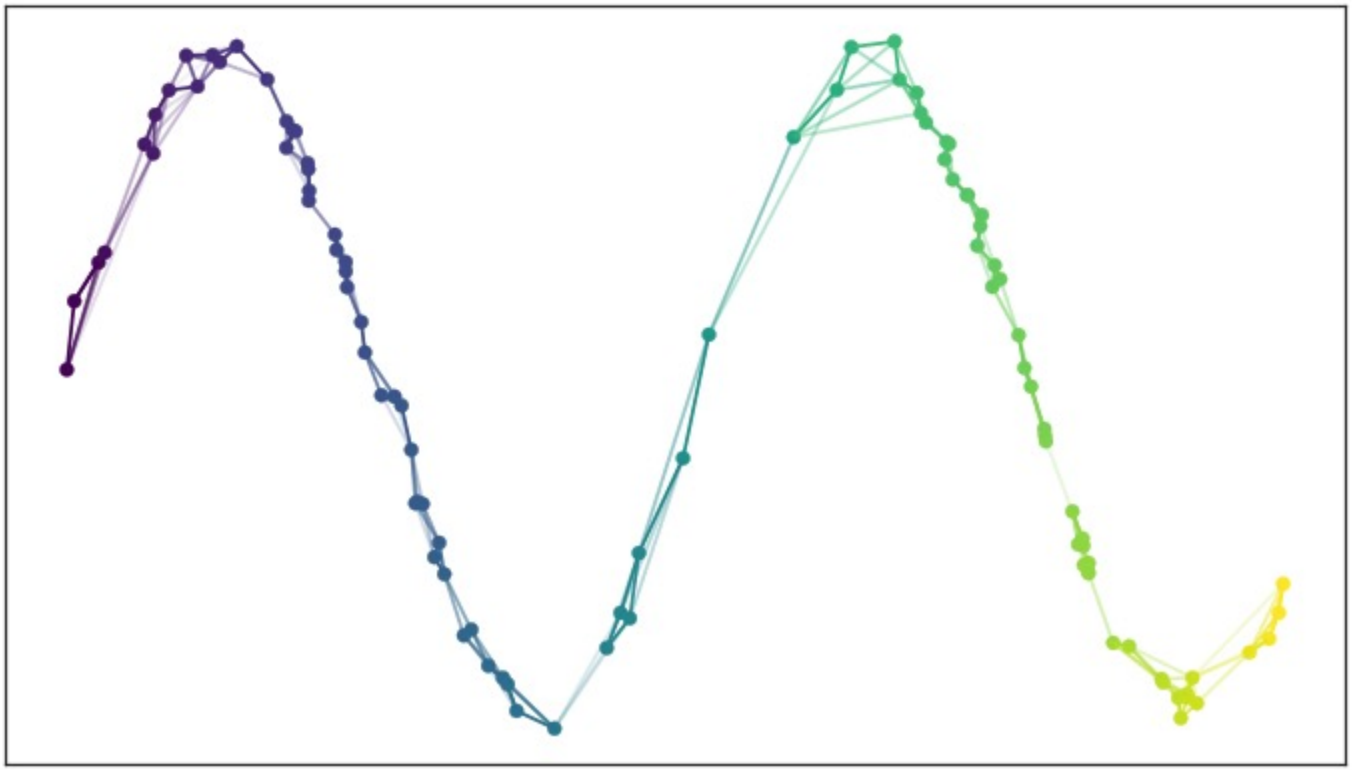
- Each instance is connected to *at least* its nearest-neighbour
- "Fuzzy" connection to neighbours beyond that
- Focuses on differences in distances, not raw: local topology!



Long version: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP

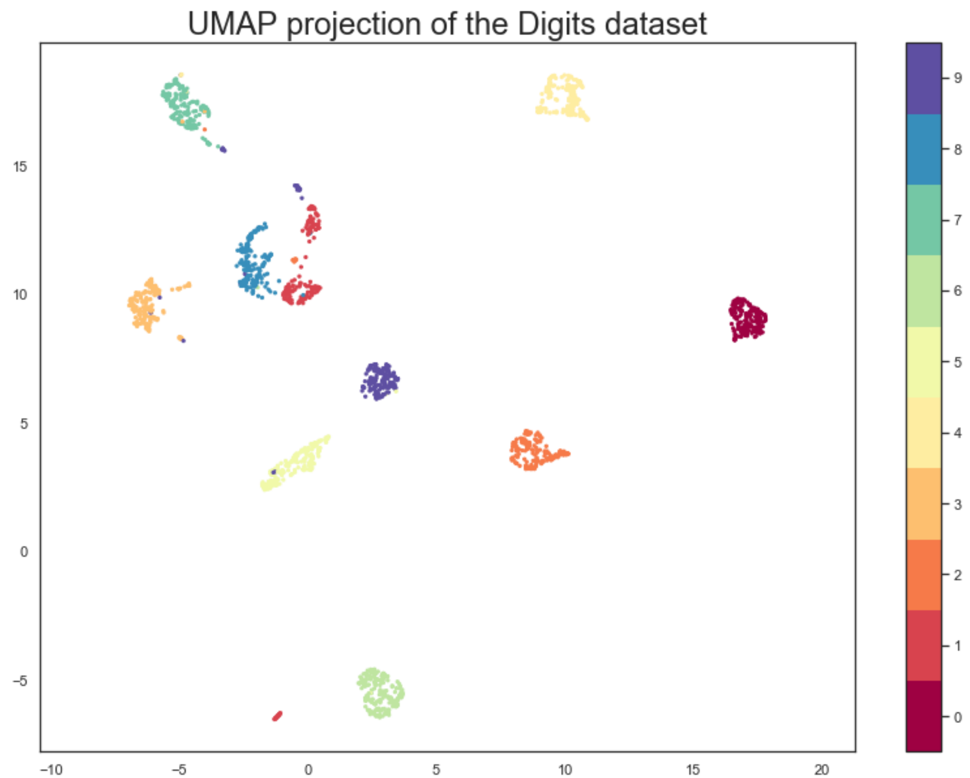
- Process to build a weighted graph
  - Opacity of line represents strength of a relationship
- Nodes “push” and “pull” each other based on size of weights



Long version: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP

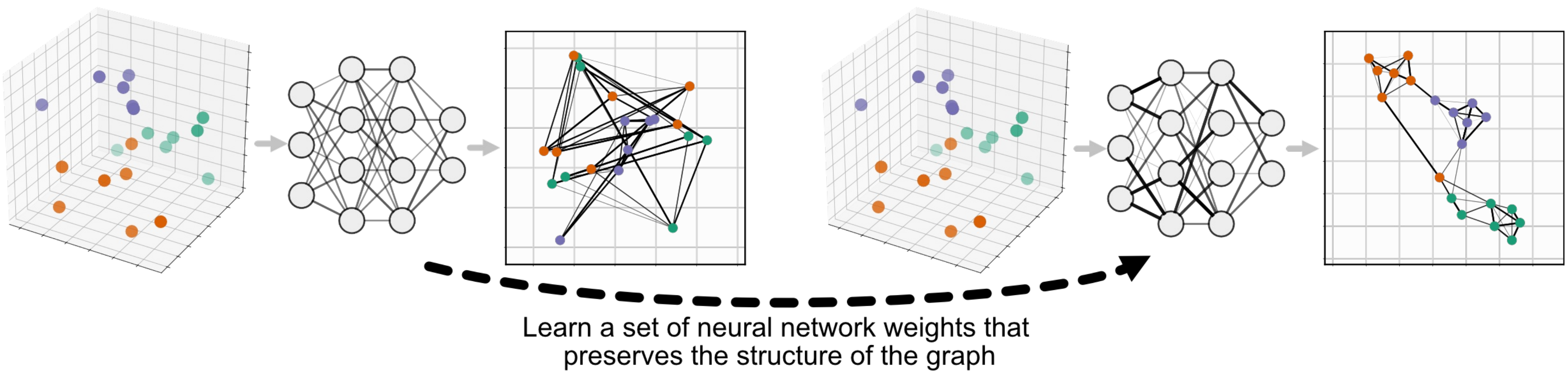
- Uses an approximation of these push/pull forces to give a differentiable objective function: optimise using gradient descent (fast/easy)



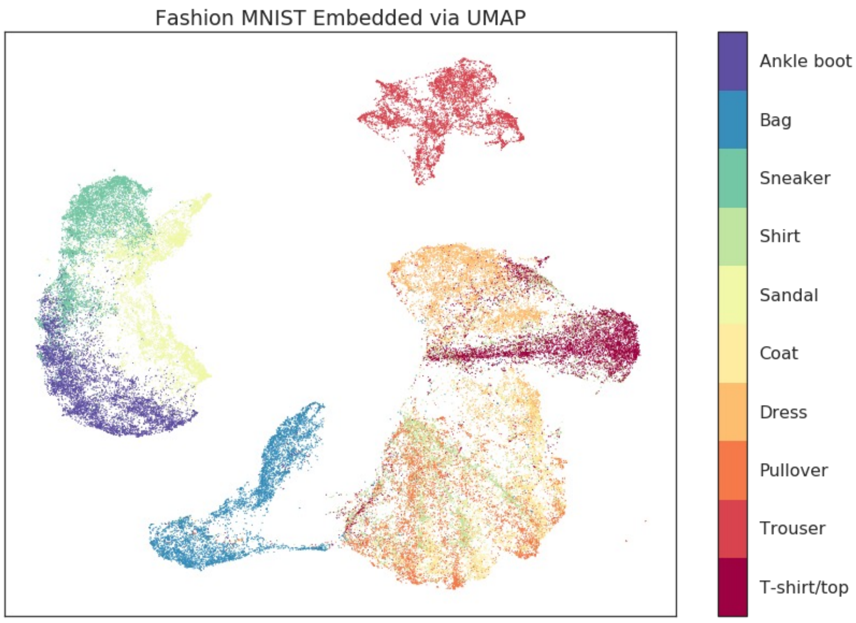
Long version: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

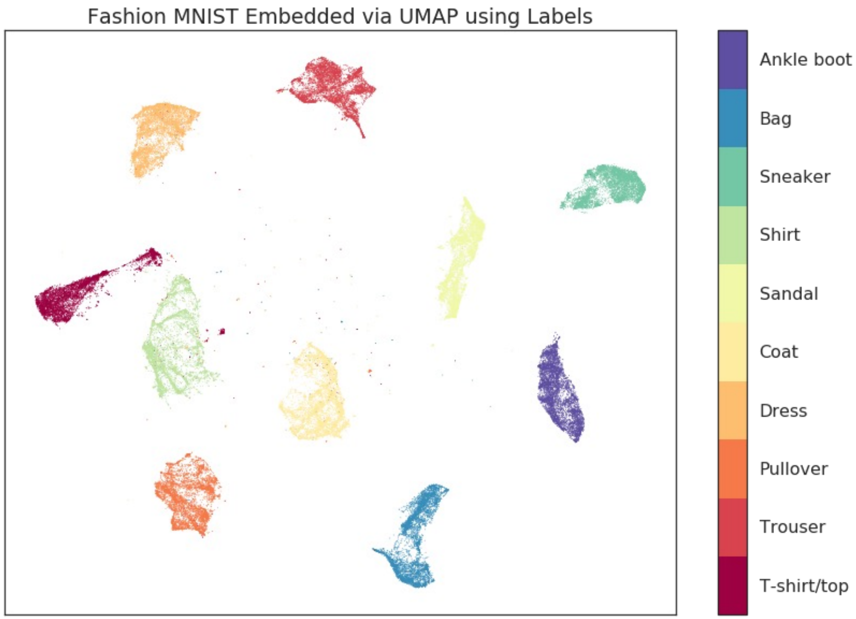
Extensions to UMAP

- Parametric UMAP: trains a NN to create the embedding



- (Semi-)Supervised UMAP: combine the two "spaces"





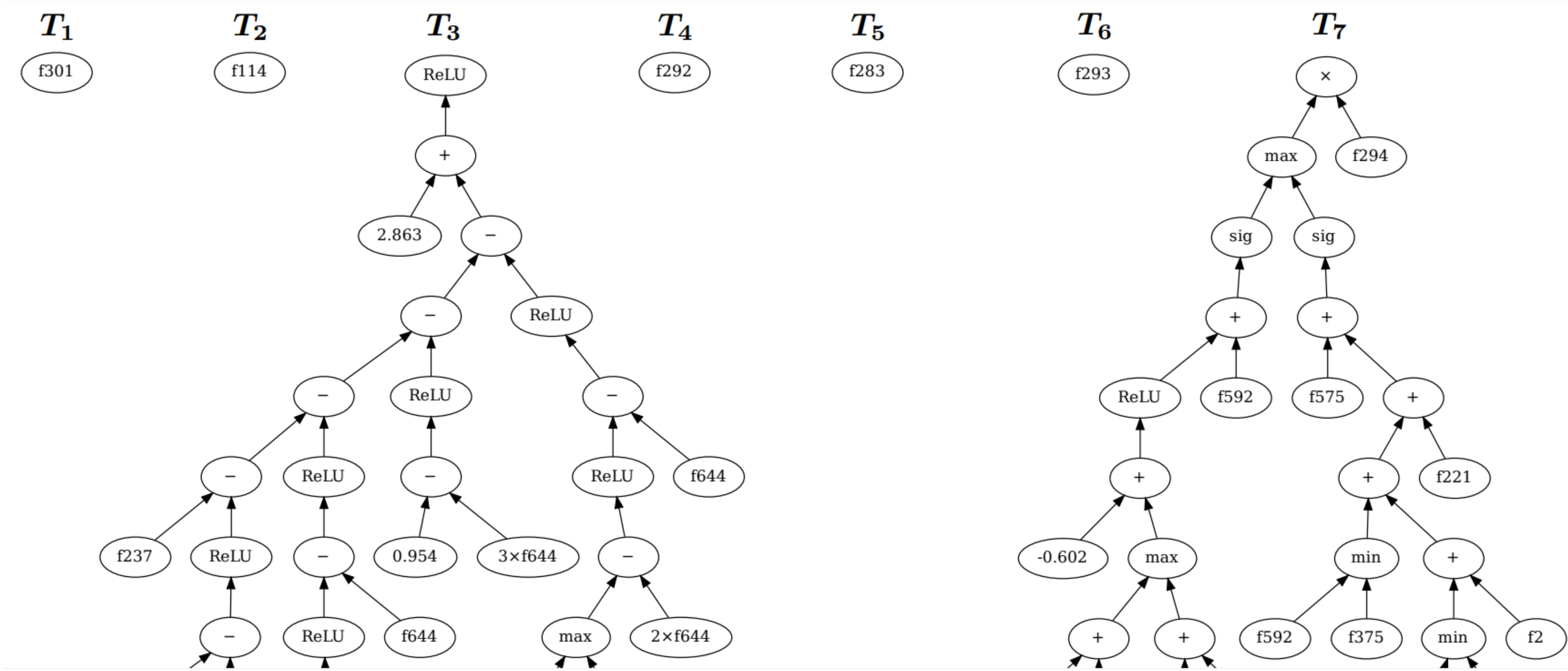
Limitations of UMAP (and t-SNE)?

- Both t-SNE and UMAP *simplified/approximated* in order to make a differentiable objective function
  - What if better non-differentiable objective functions exist?

- Parametric t-SNE is a mapping – i.e. we have a concrete functional model from D to d
  - …but is a 3-layer 100-neuron fully-connected NN at all interpretable?
  - I argue NO!

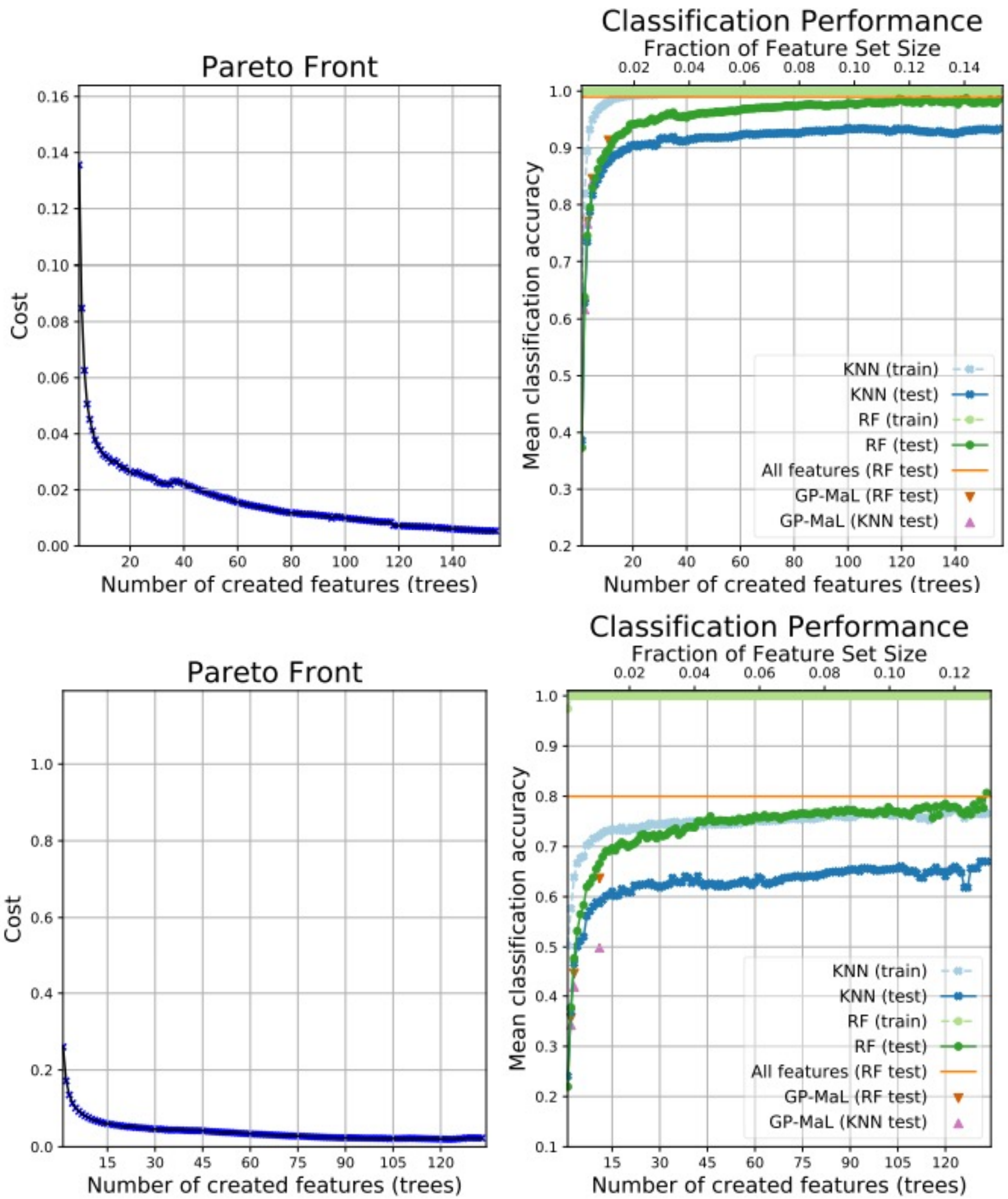
- Using EC/Genetic Programming to find simpler functional models/mappings for Manifold Learning.

Genetic Programming for Manifold Learning



GP-MaL

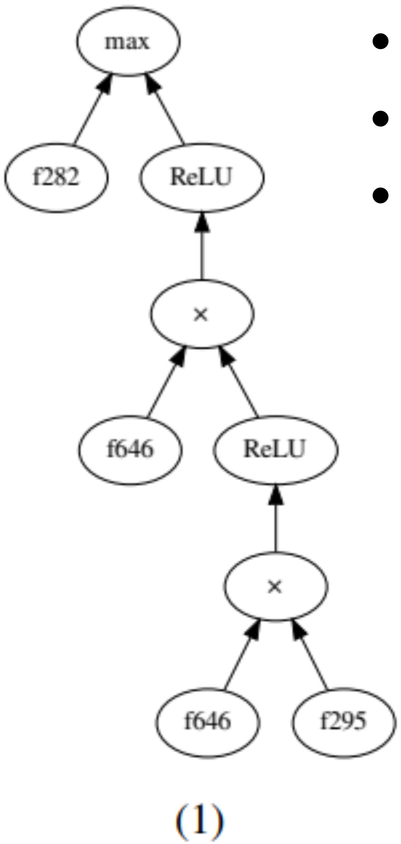
Manifold Learning: Embedding Quality vs Dimensionality



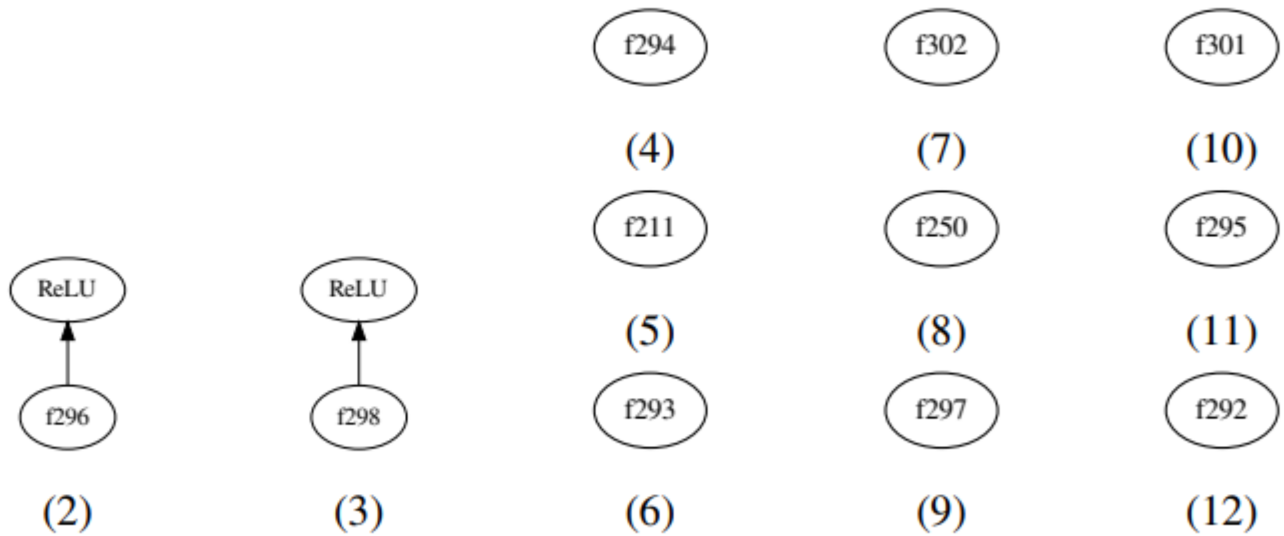
Diminishing returns – most data has low *intrinsic dimensionality.*

A. Lensen, M. Zhang, and B. Xue. "[Multi-Objective Genetic Programming for Manifold Learning: Balancing Quality and Dimensionality]" in *Genet Program Evolvable Mach* **21,** 399–431 (2020). https://doi.org/10.1007/s10710-020-09375-4

Manifold Learning: Embedding Quality vs Dimensionality



- *649*-dimensional MFEAT dataset;
- *12* evolved trees;
- 95% test accuracy (post-hoc).

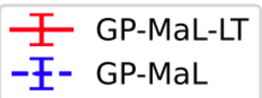


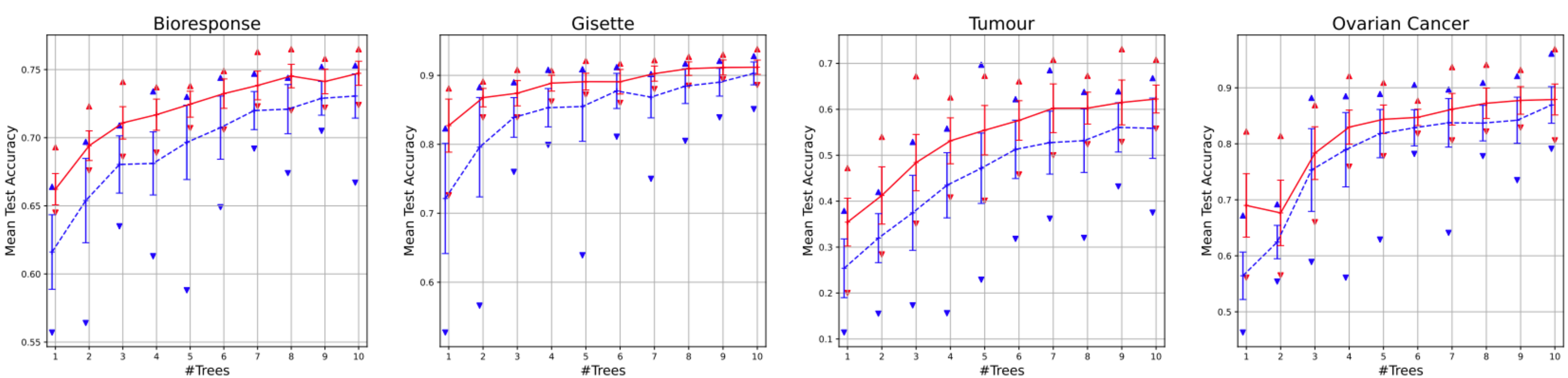
A. Lensen, M. Zhang, and B. Xue. "**Multi-Objective Genetic Programming for Manifold Learning: Balancing Quality and Dimensionality**" in *Genet Program Evolvable Mach* **21,** 399–431 (2020). https://doi.org/10.1007/s10710-020-09375-4

Manifold Learning: Preserving Local Topology

A. Lensen, M. Zhang, and B. Xue. "**Genetic Programming for Manifold Learning: Preserving Local Topology**" in *IEEE Trans. Evolutionary Computation* (Early Access)
DOI: 10.1109/TEVC.2021.3106672

- There is an inherent trade-off between preserving *global* and *local* *topology*
- In many tasks, local topology preservations is more important
  - E.g. image segmentation, semi-supervised learning, …
- Prioritising local topology preservation can retain more crucial structure



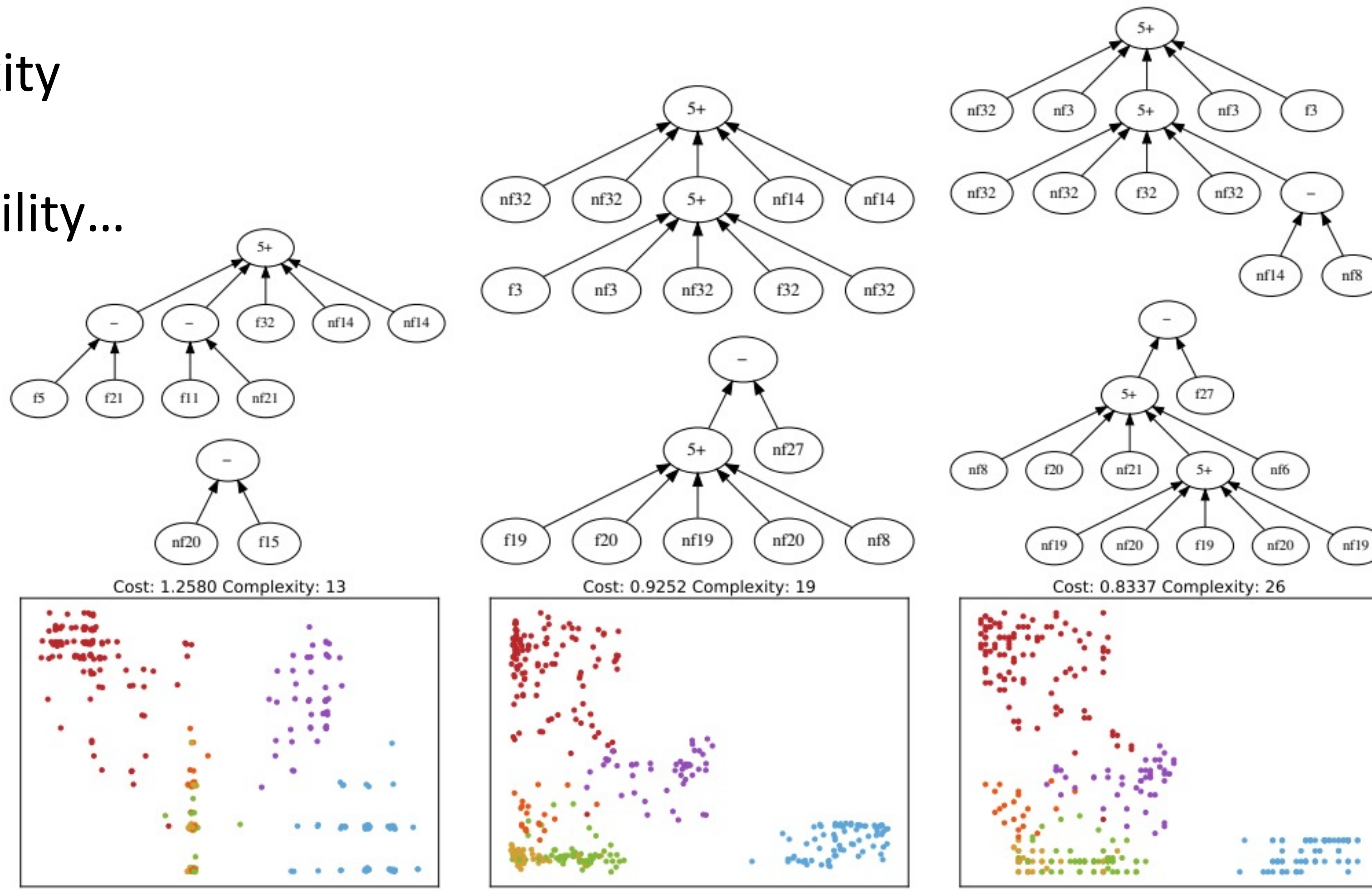


Explainable Unsupervised Learning

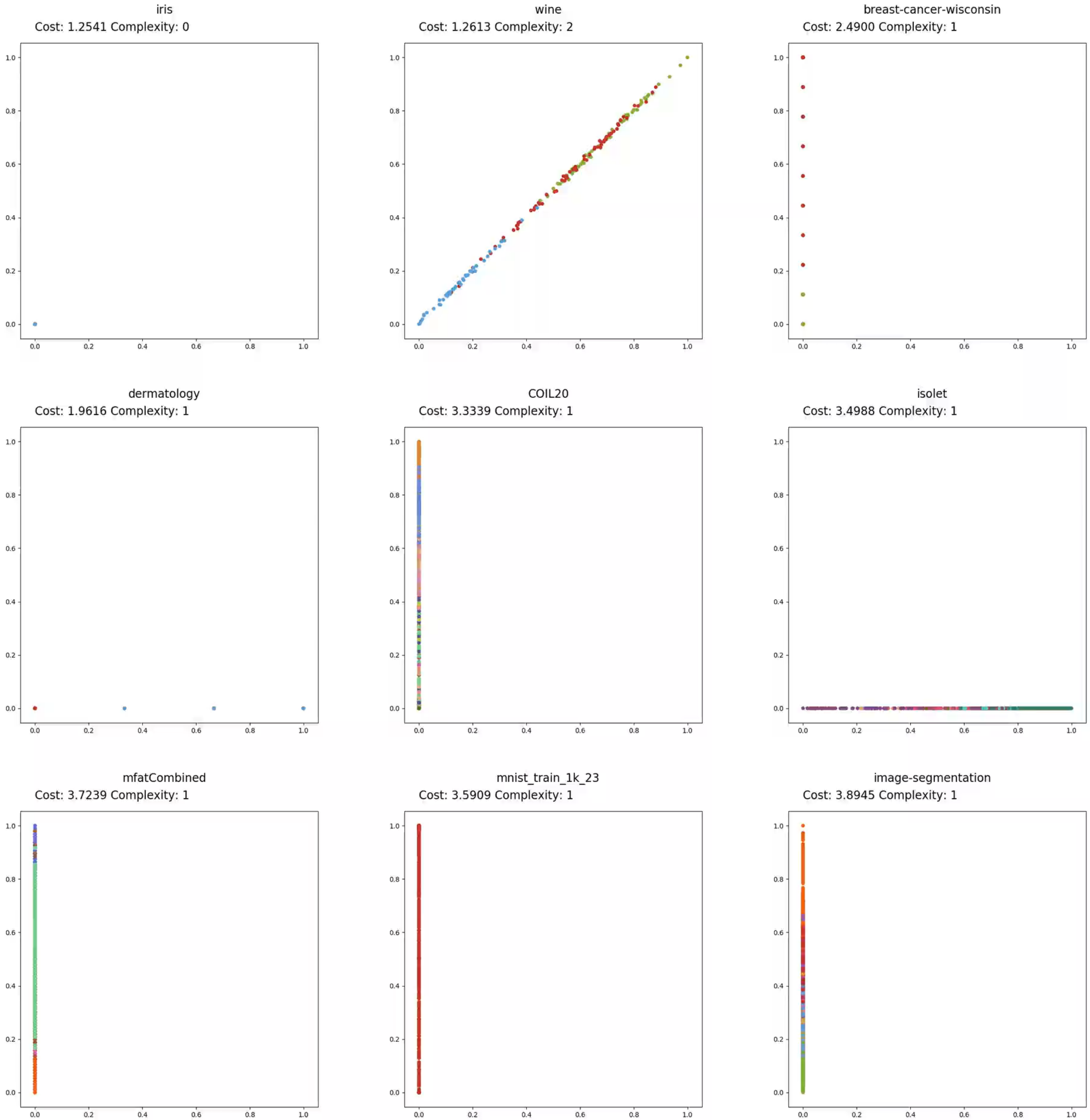
What do these visualisations actually ***mean***?

2D Manifold Learning ⇔ Visualisation?

Complexity
vs
Interpretability...



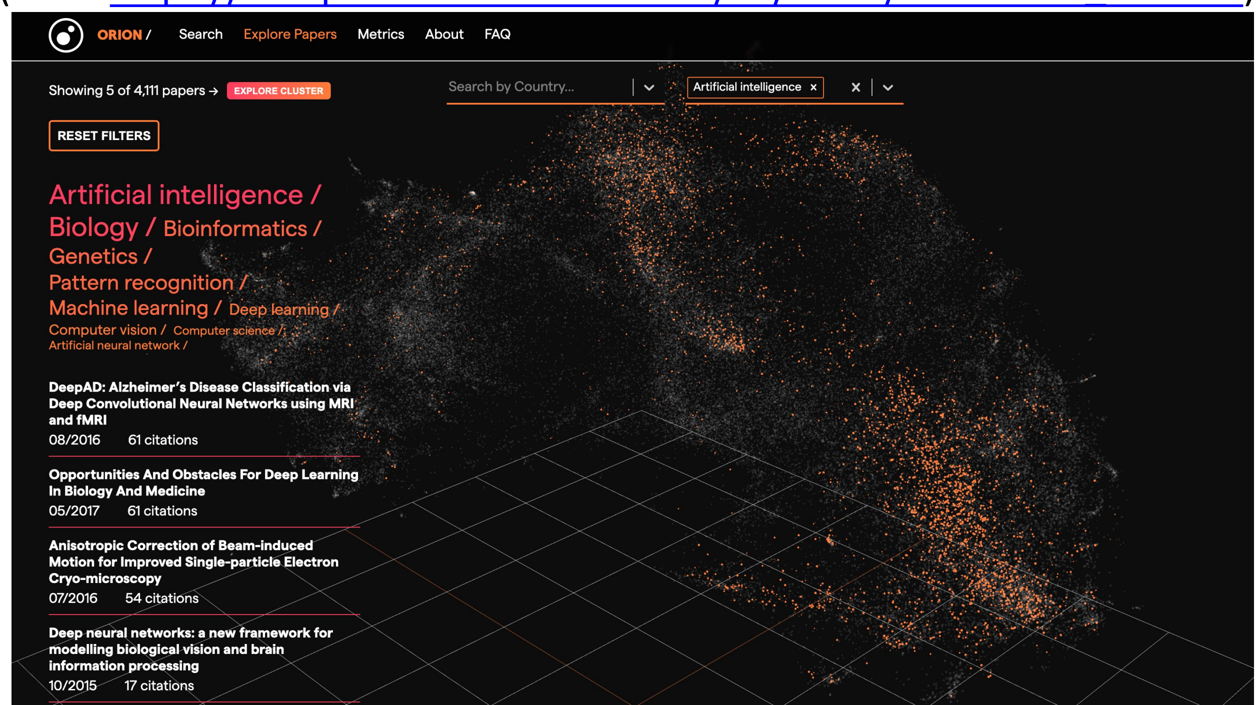
A. Lensen, B. Xue and M. Zhang, "**Genetic Programming for Evolving a Front of Interpretable Models for Data Visualisation**" in *IEEE Transactions on Cybernetics*, doi: 10.1109/TCYB.2020.2970198.


iris
Cost: 1.2541 Complexity: 0
wine
Cost: 1.2613 Complexity: 2
breast-cancer-wisconsin
Cost: 2.4900 Complexity: 1
dermatology
Cost: 1.9616 Complexity: 1
COIL20
Cost: 3.3339 Complexity: 1
isolet
Cost: 3.4988 Complexity: 1
mfatCombined
Cost: 3.7239 Complexity: 1
mnist_train_1k_23
Cost: 3.5909 Complexity: 1
image-segmentation
Cost: 3.8945 Complexity: 1

Cool uses of UMAP

- Modelling 3D animals (wireframes) in 2D
- Compare t-SNE vs UMAP vs PCA on big datasets
- PixPlot: Embeds >27,000 historical photographs (2,048px) in 2D
- Orion Search: Embedding of academic paper abstracts

(From https://umap-learn.readthedocs.io/en/latest/interactive_viz.html)



Demo Time

https://colab.research.google.com/drive/1rF_gFIU7s5DGT3rHsAhP61ElDYYJMAys?usp=sharing