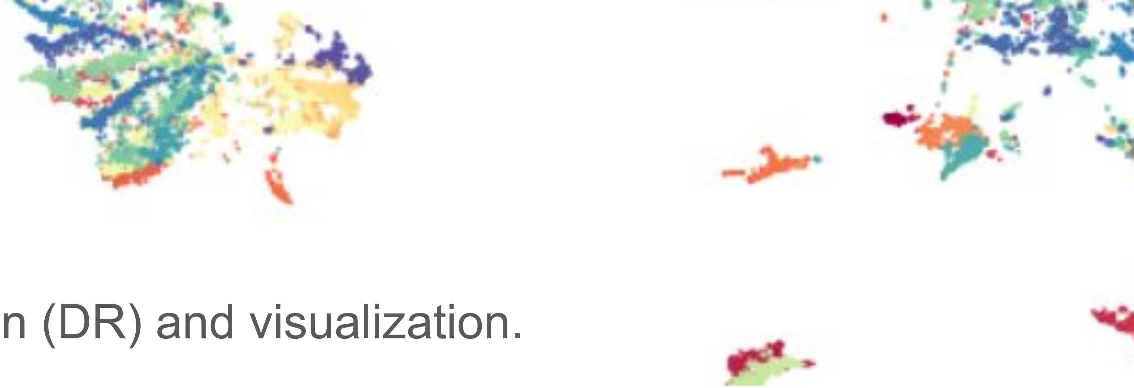
UMAP

Uniform Manifold Approximation and Projection

UMAP Quick Facts



What: Dimensionality reduction (DR) and visualization.

Speed: Very fast. Slower than PCA and faster than tSNE

Good For: DR of any finite metric space. Preserving local and global structure of data throughout DR.

Poor For: Datasets with isolated points.

```
conda install -c conda-forge umap-learn
```

OR

```
pip install umap-learn
```

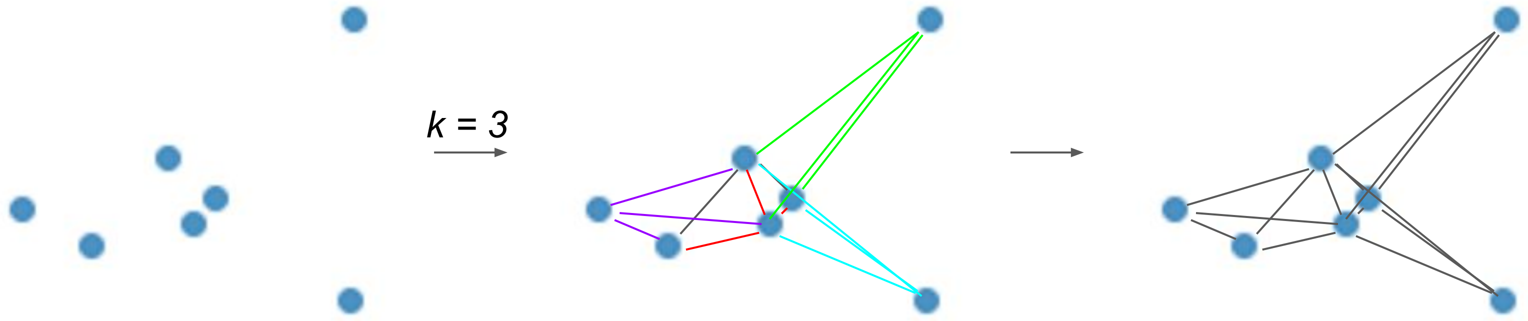
UMAP Algorithm

1) **Compute Neighborhood Graph**
   a) **KNN Searching,**
   b) **Determine local metrics.**
   c) **Construct non-symmetric weighted adjacency matrix from local metrics.**
   d) **Combine local metrics to construct symmetric weighted adjacency matrix.**
2) **Optimize Graph Layout in Reduced Space**
   a) **Uses a force directed graph layout algorithm.**

```
conda install -c conda-forge umap-learn
```

OR

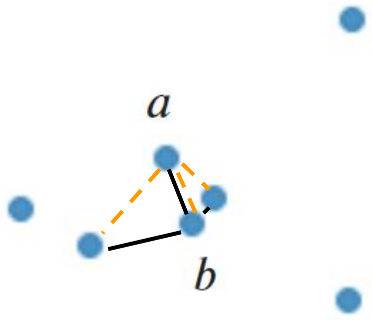
```
pip install umap-learn
```



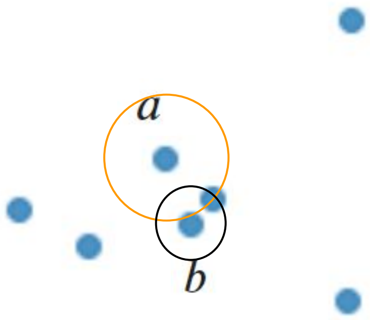
Space’s metric is applied to find all edge lengths between k-NN.

Distances are symmetric, so a symmetric adjacency matrix results.

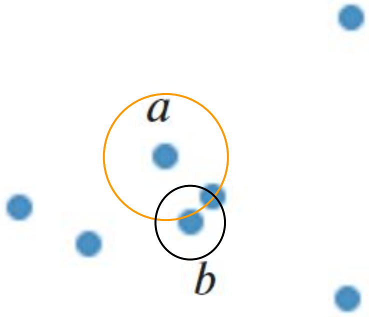
Local Metrics and KNN



For each point find KNN and…



… declare the ball which contains one of them to be weight one.



$$d_a(a, b) \neq d_b(a, b)$$

Local Metric Spaces For Everyone!

A collection of local metric spaces $\{(X_i, d_i)\}_{i \in I}$.

Immediate consequences:

- Dense regions of data have ‘short’ rulers.
- Sparse regions of data have ‘long’ rulers.
- Instead of having one distance between two points we have two.
- In each cluster of k-points, data are approximately uniformly distributed.

Asymmetric Adjacency Graph

At last we construct an adjacency graph: $\overline{G} = (V, E, w)$

V is our entire set of data points. $E = \{(x_i, x_{i_j}) | 1 \leq j \leq k, 1 \leq i \leq N\}$

Local distance from x_i*'s* perspective to j^{th}-NN.

Distance to closest NN.

$$w(x_i, x_{i_j}) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

Normalizing constant ensuring the sum of weights to all NN equals $\log_2 k$.

Symmetric Adjacency Matrix

$$B = \overline{G} + \overline{G}^T - \overline{G} \circ \overline{G}^T$$

Things to consider:

- In $\overline{G} = (V, E, w)$ weights were probabilities and could be interpreted as, 'the probability of an edge being included'.
- By the construction of *B*, entries are:

$$w(x_i, x_{i_j}) + w(x_{i_j}, x_i) - w(x_i, x_{i_j}) \cdot w(x_{i_j}, x_i) = \mathbb{P}(\text{ include edge } i \text{ or } i_j)$$

- *B* is symmetric.

Where: Kaggle

What: Grayscale images of Hindi characters.

Image Size: 32x32 (1024 dimensions)

Image Count: ~304,000 in total (792 images of each of the 384 character combinations).

Data pared to 158 images each of: अ अं अः आ ओ औ

$$\mathbb{R}^{948\times1024}$$

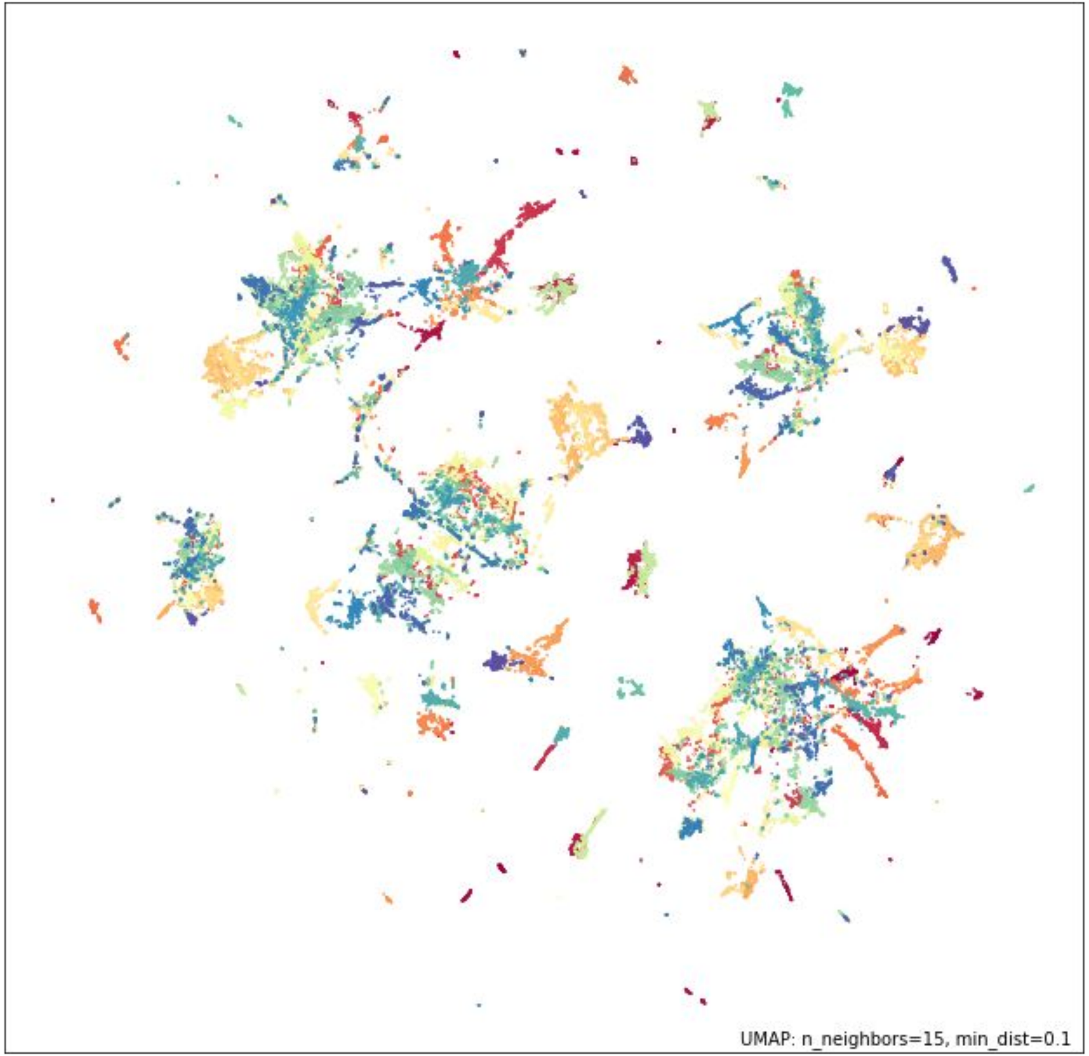
SHABD Results

Data Size: $\mathbb{R}^{60672 \times 1024}$

All 384 characters (color).

158 images of each.

Time: 26.73s



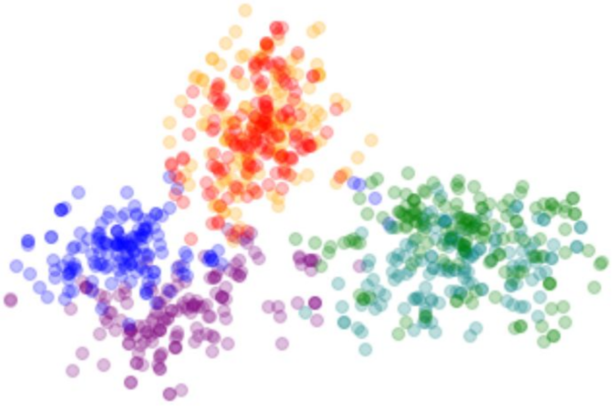
SHABD Results

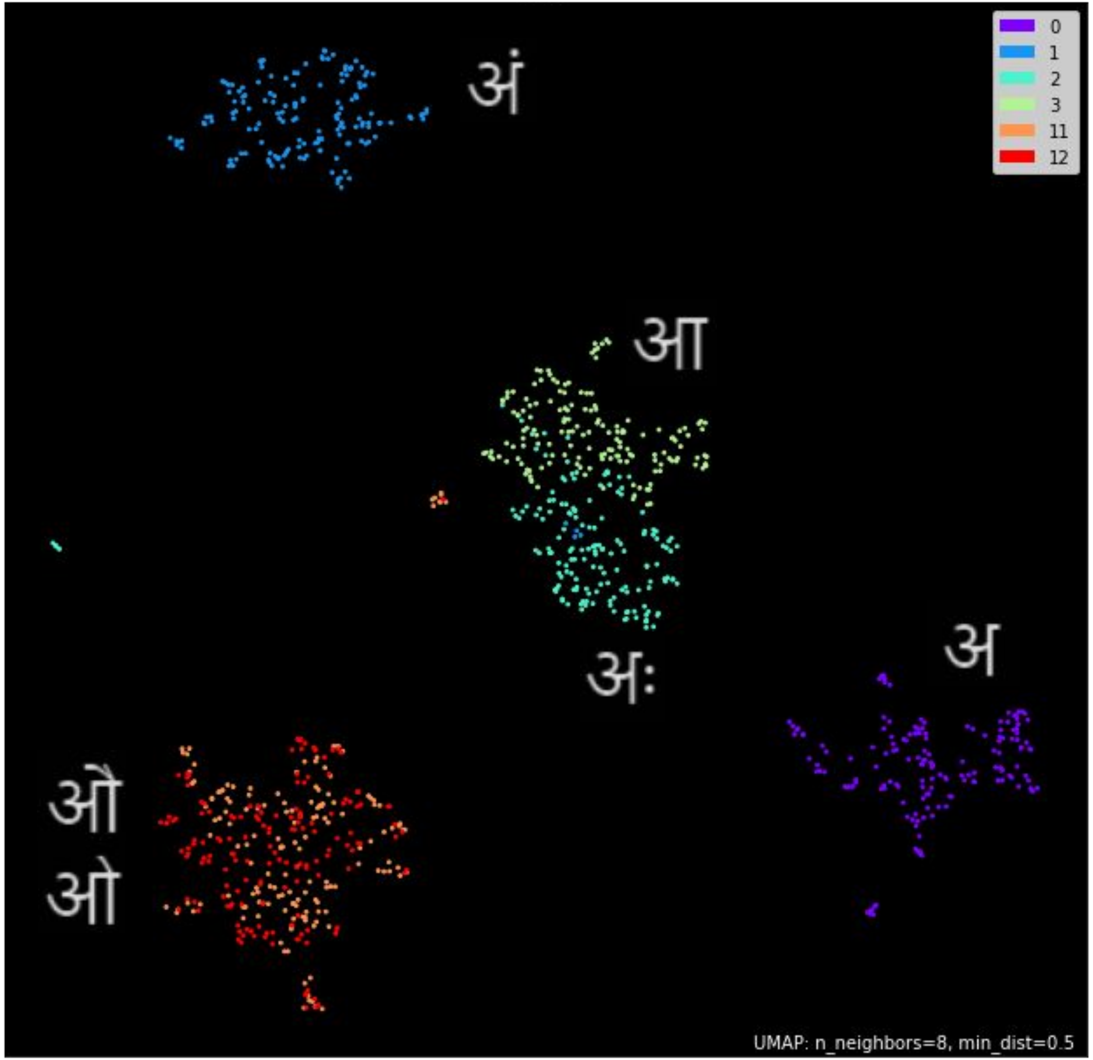
Data Size: $\mathbb{R}^{948\times1024}$

Time: 2.70s

Right:UMAP

Below: PCA





Aerial Visible/Infrared Imaging Spectrometer (AVIRIS)

Where: Salinas Valley, California

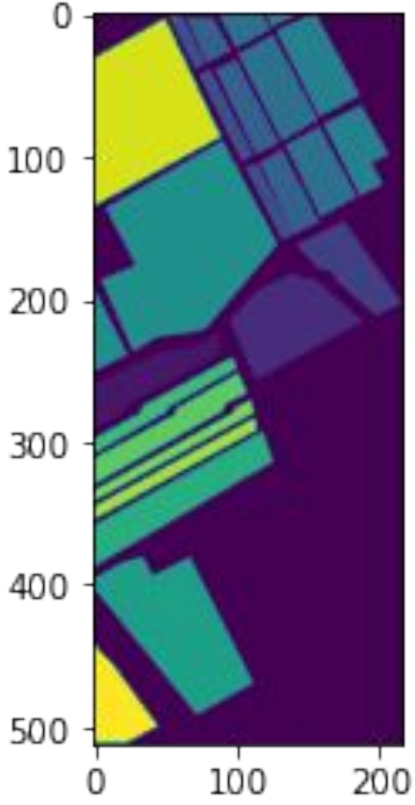
What: Image data with 224 dimensions per pixel, each pixel represents a 3.7m by 3.7m patch of Earth.

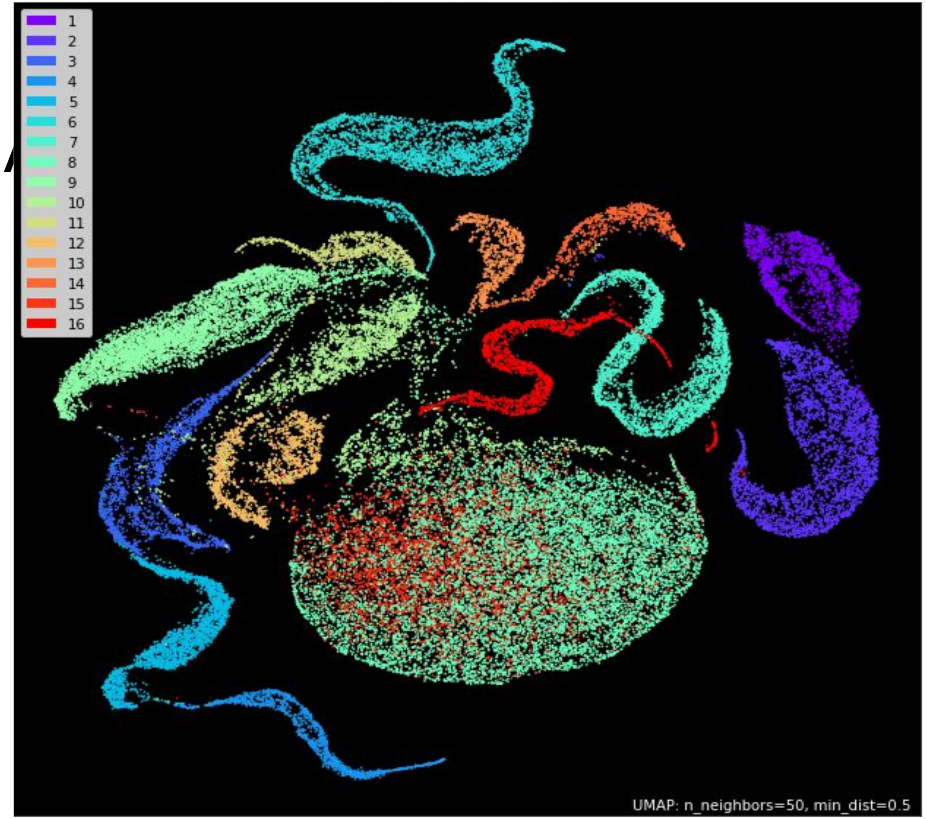
Pixel Count: ~111,000 with ~54,000 labeled by crop.

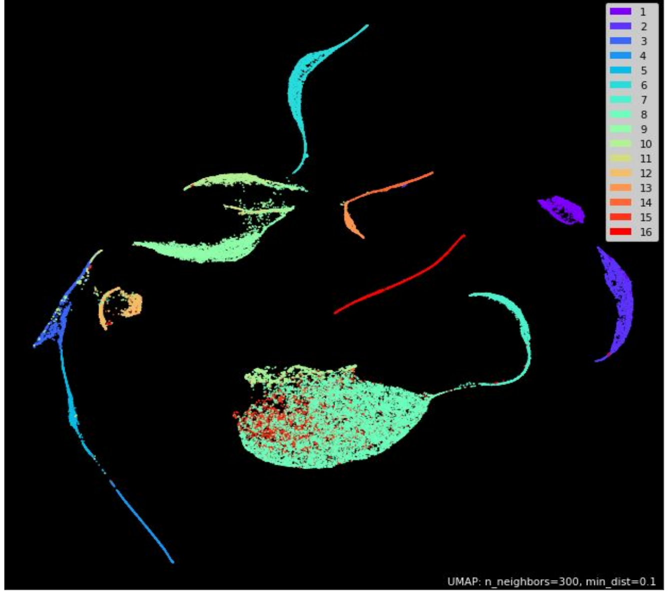
How: Big plane, big camera.

$\mathbb{R}^{54129\times224}$

Right: A section of Salinas Valley labeled by color with intended crop in a given pixel.







UMAP **Left:** 47.56s, **Top Right:** 118.83s

$\mathbb{R}^{54129 \times 224}$ **Bot Right:** PCA 0.44s

References

McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018

Leland McInnes. *UMAP-Learn Documentation*. 2018. Revision 300cbba8 https://umap-learn.readthedocs.io/en/latest/index.html

John Williamson. *What do numbers look like? https://johnhw.github.io/umap_primes/index.md.html*

GitHub Repo

This presentation and ipynb files: https://github.com/AntoineLove?tab=repositories

Data Sources

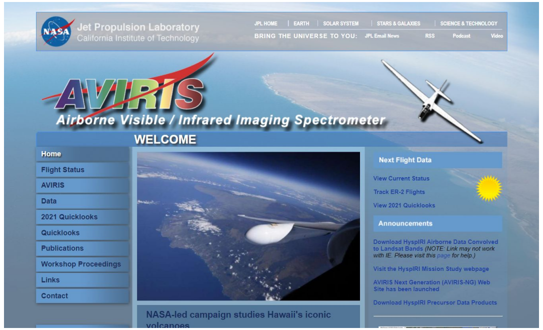
AVIRIS Website: https://aviris.jpl.nasa.gov/

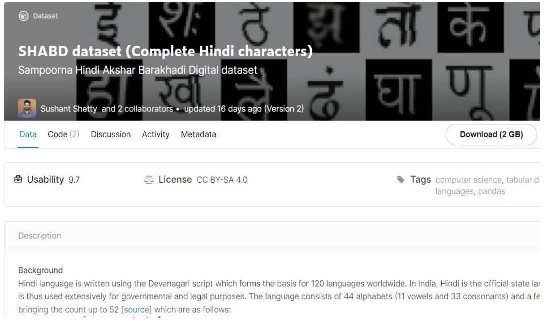
Salinas Dataset:

AVIRIS Data of Salinas Valley, California

Kaggle Dataset

Sampoorna Hindi Akshar Barakhadi Digital Dataset (SHABD Dataset)



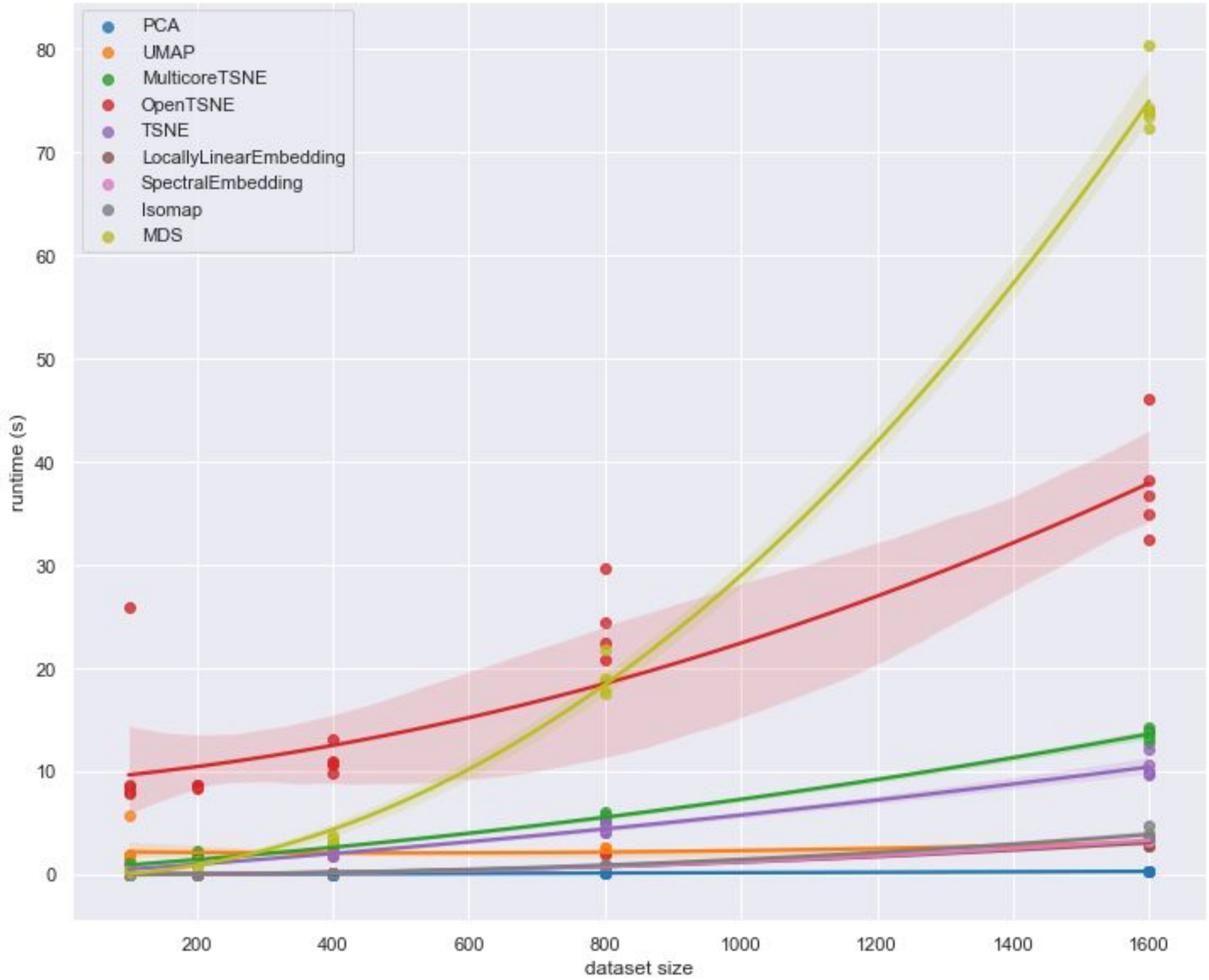


Appendix A: (src Salinas Valley Dataset)

Groundtruth classes for the Salinas scene and their respective samples number

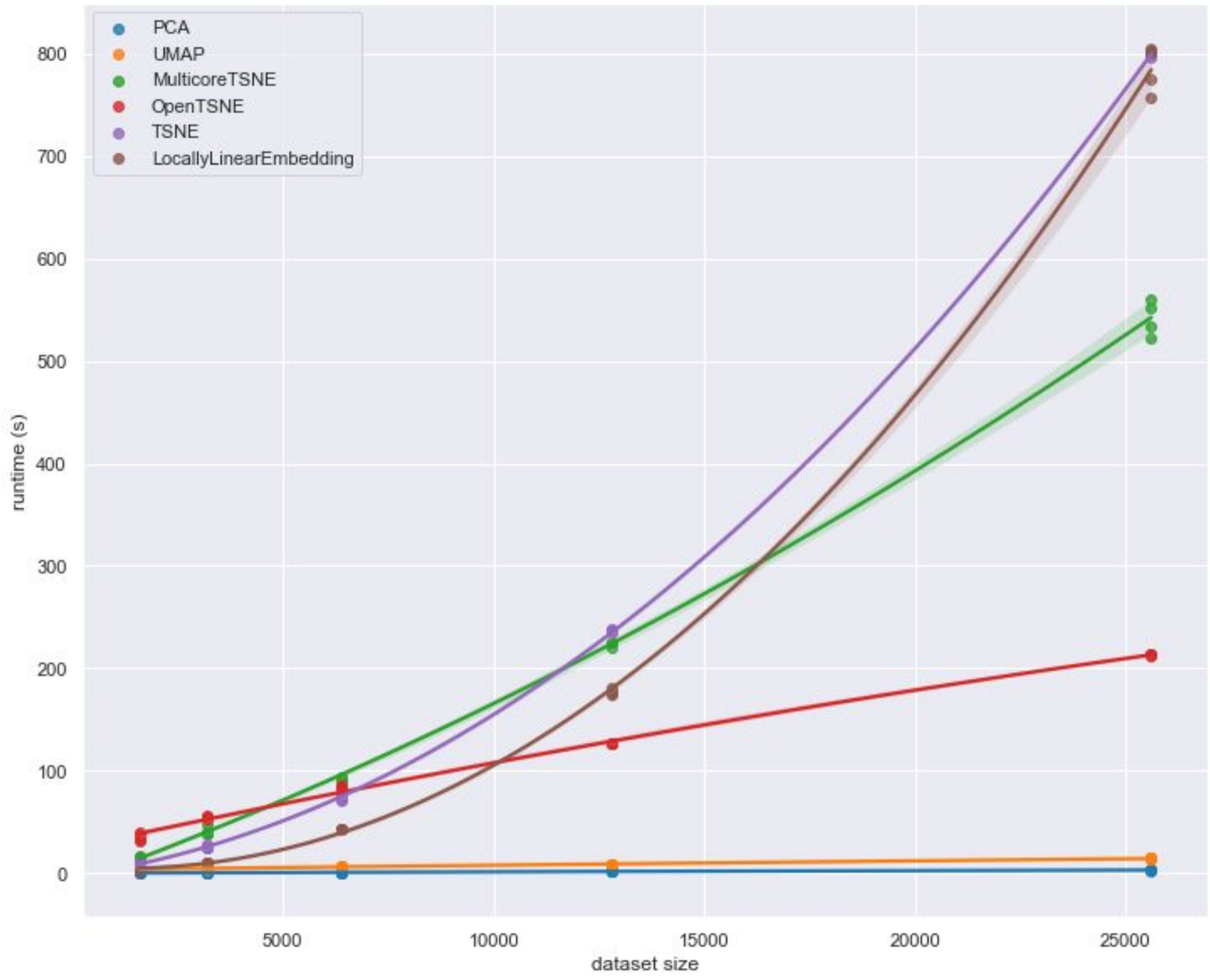
#	Class	Samples
1	Brocoli_green_weeds_1	2009
2	Brocoli_green_weeds_2	3726
3	Fallow	1976
4	Fallow_rough_plow	1394
5	Fallow_smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes_untrained	11271
9	Soil_vinyard_develop	6203
10	Corn_senesced_green_weeds	3278
11	Lettuce_romaine_4wk	1068
12	Lettuce_romaine_5wk	1927
13	Lettuce_romaine_6wk	916
14	Lettuce_romaine_7wk	1070
15	Vinyard_untrained	7268
16	Vinyard_vertical_trellis	1807

Appendix B: UMAP Performance Comparison



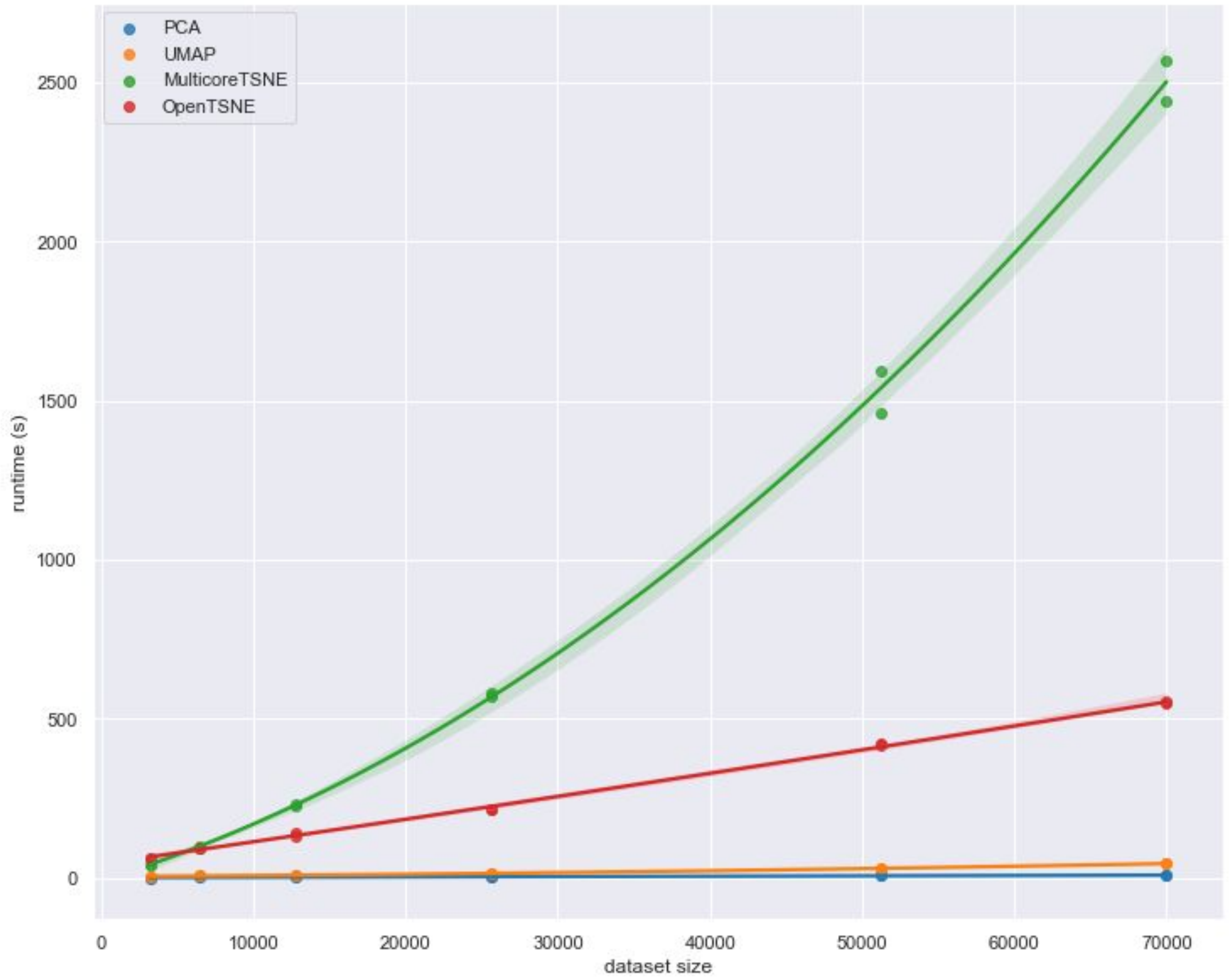
UMAP-Learn Documentation
https://umap-learn.readthedocs.io/en/latest/performance.html

Appendix B: UMAP Performance Comparison (cont.)



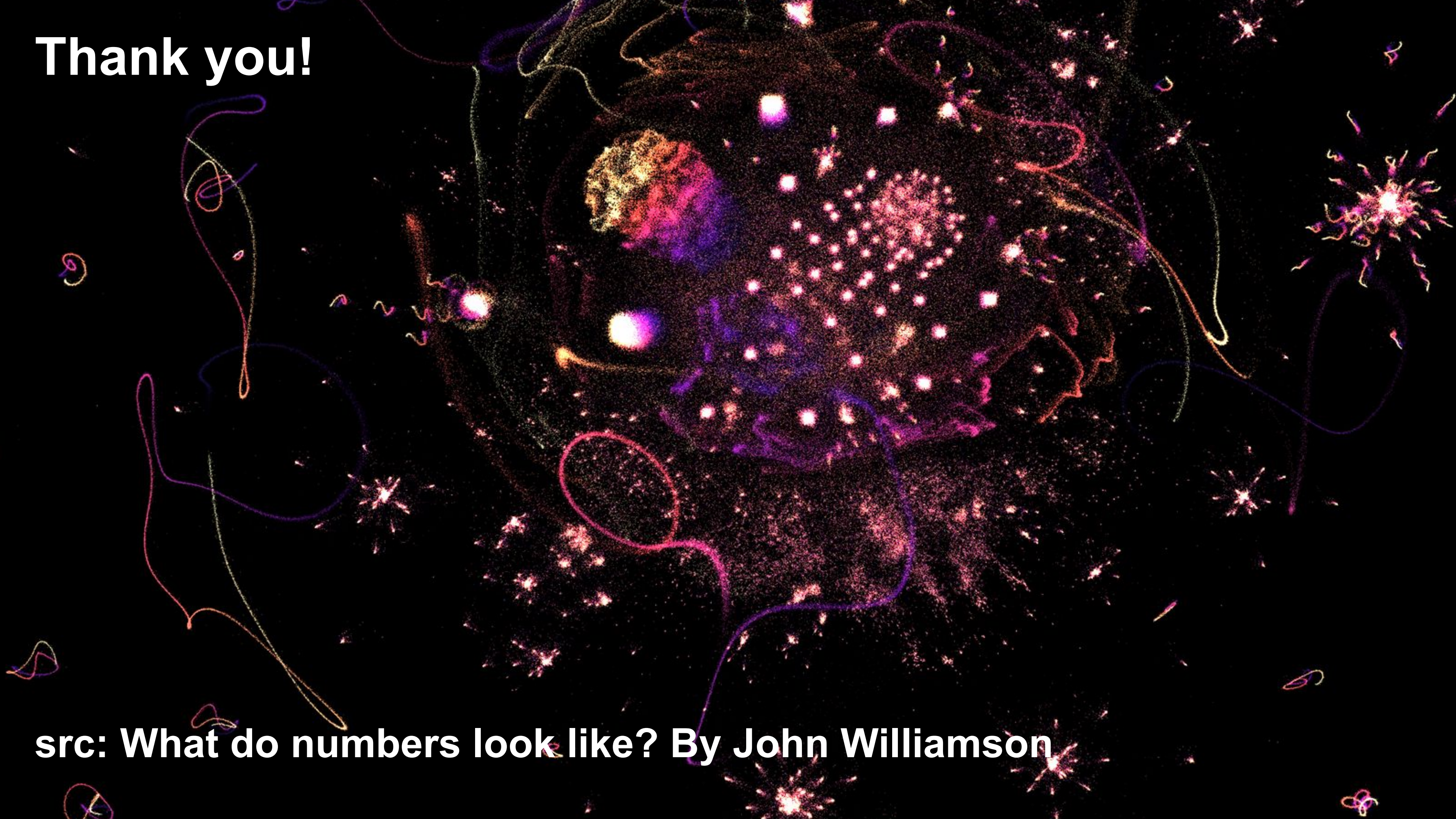
UMAP-Learn Documentation
https://umap-learn.readthedocs.io/en/latest/performance.html

Appendix B: UMAP Performance Comparison (cont.)



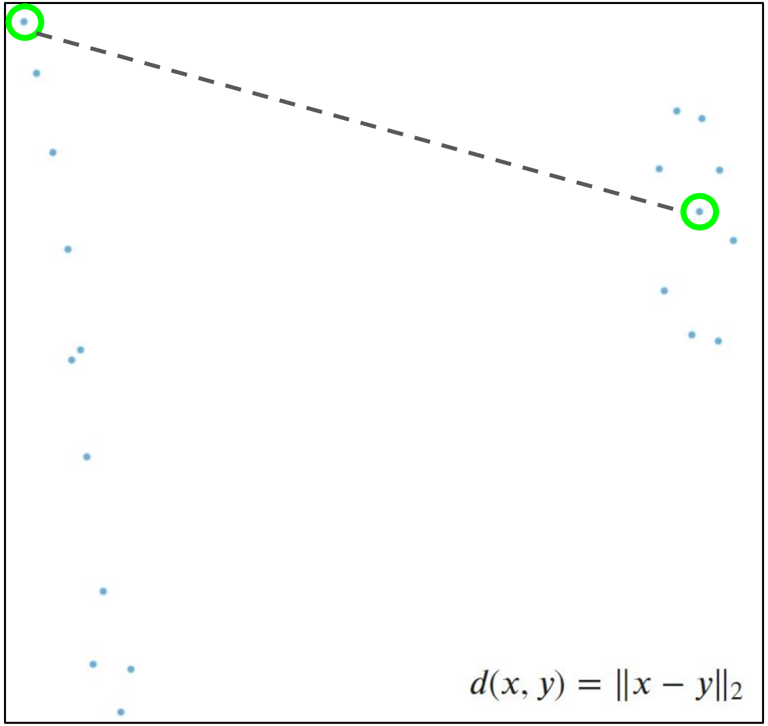
UMAP-Learn Documentation
https://umap-learn.readthedocs.io/en/latest/performance.html

Thank you!



src: What do numbers look like? By John Williamson

Euclidean Metric



Geodesic Metric