3050571 Practical Clin Data Sci

Session 7: Dimensionality reduction

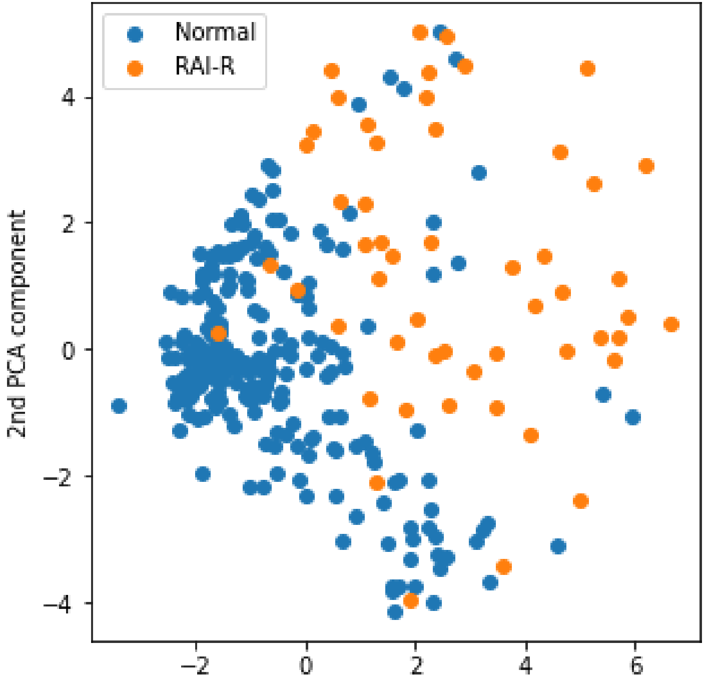
February 13, 2024

Sira Sriswasdi, PhD

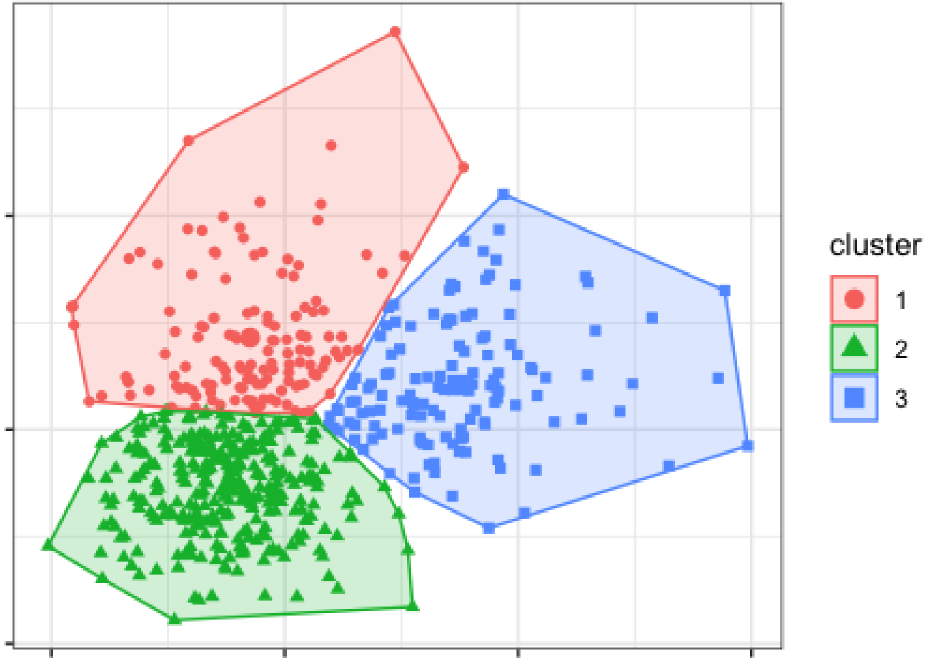
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Two primary branches of unsupervised learning

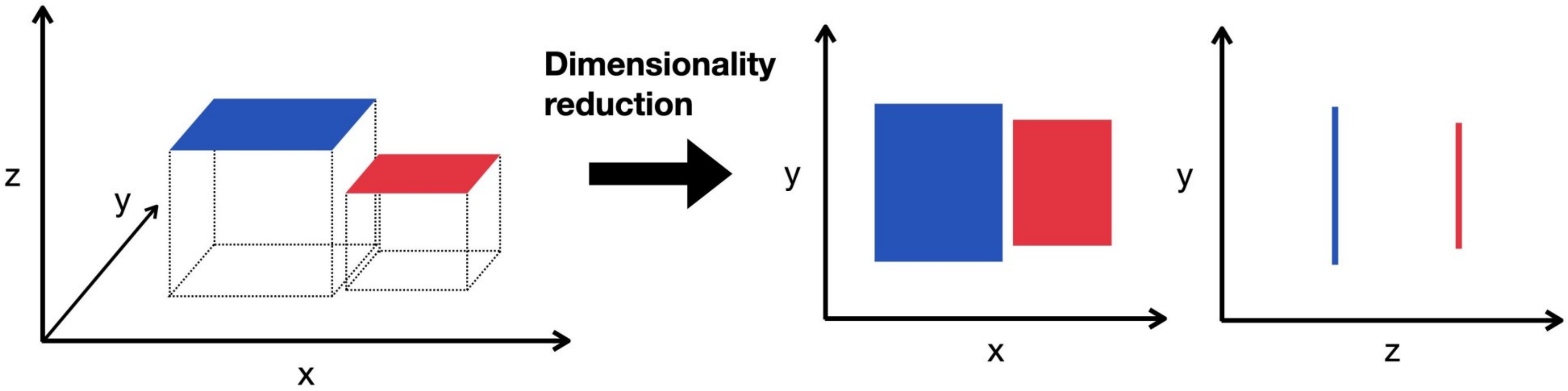
Dimensionality Reduction



Clustering



Dimensionality reduction

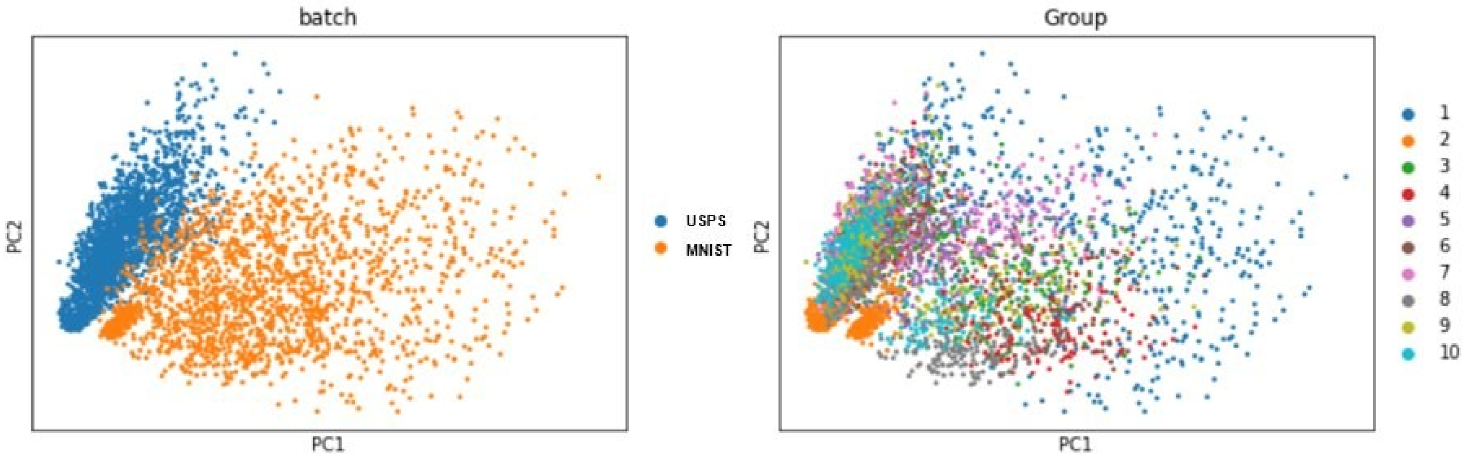


https://www.sc-best-practices.org/preprocessing_visualization/dimensionality_reduction.html

- Reduce dimension (number of features) while maintaining information
- Patient with similar symptoms also exhibit similar lab tests or have similar demographics or similar medical history

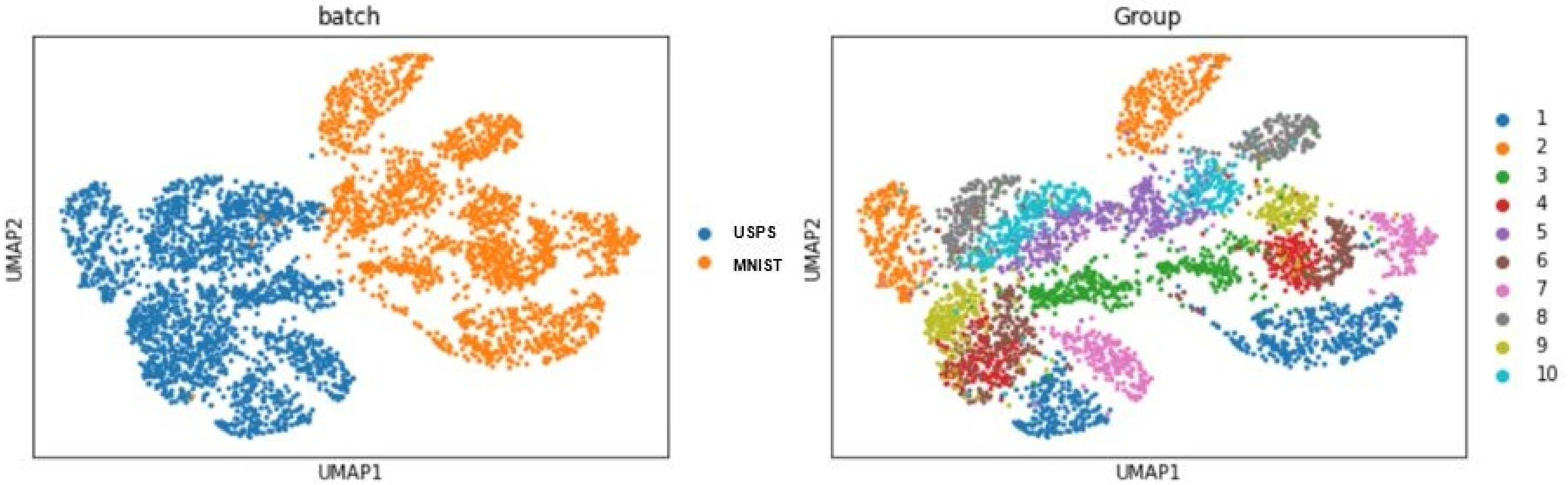
Digit datasets

*adapted from doi:10.1109/TKDE.2017.2669193



https://twitter.com/lkmklsmn/status/1436357177887895555

UMAP captures every group in 2D

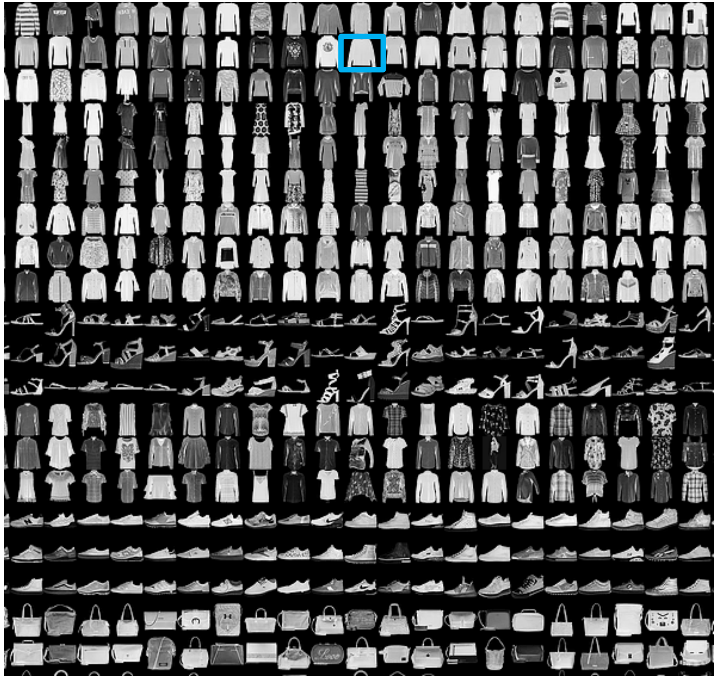


https://twitter.com/lkmklsmn/status/1436357177887895555

- Both data source and digit identity can be distinguished

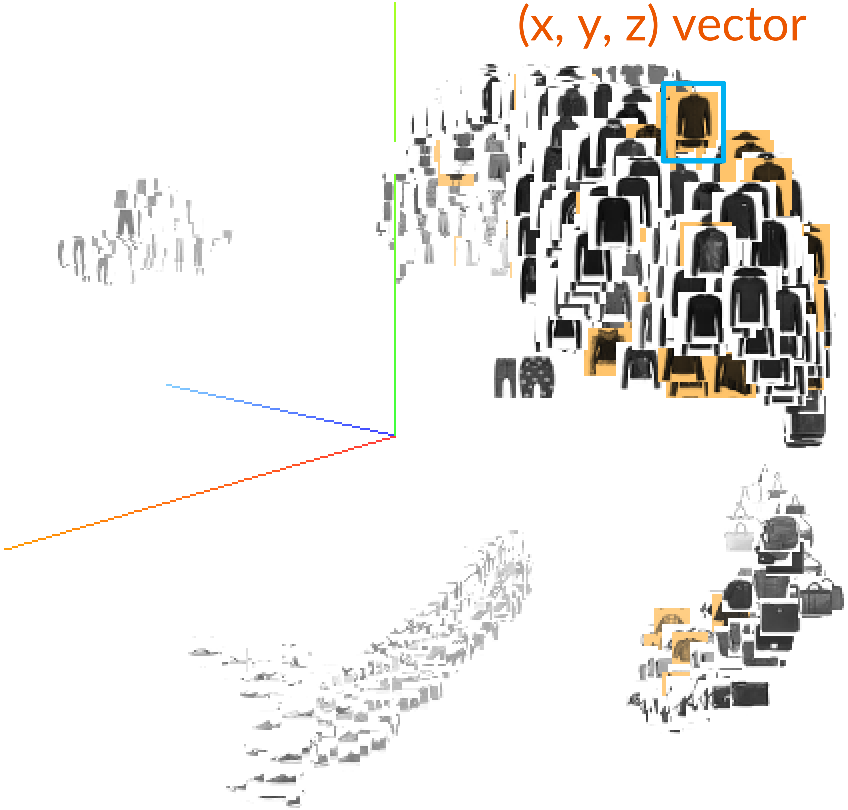
Fashion MNIST

28x28 pixel image



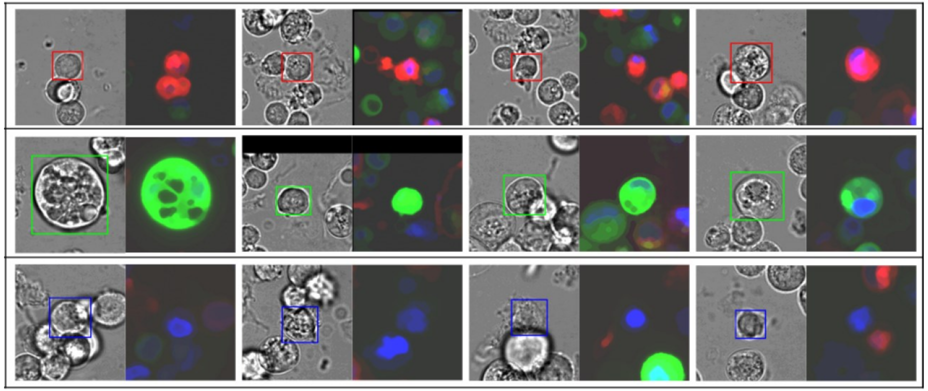
https://github.com/zalandoresearch/fashion-mnist

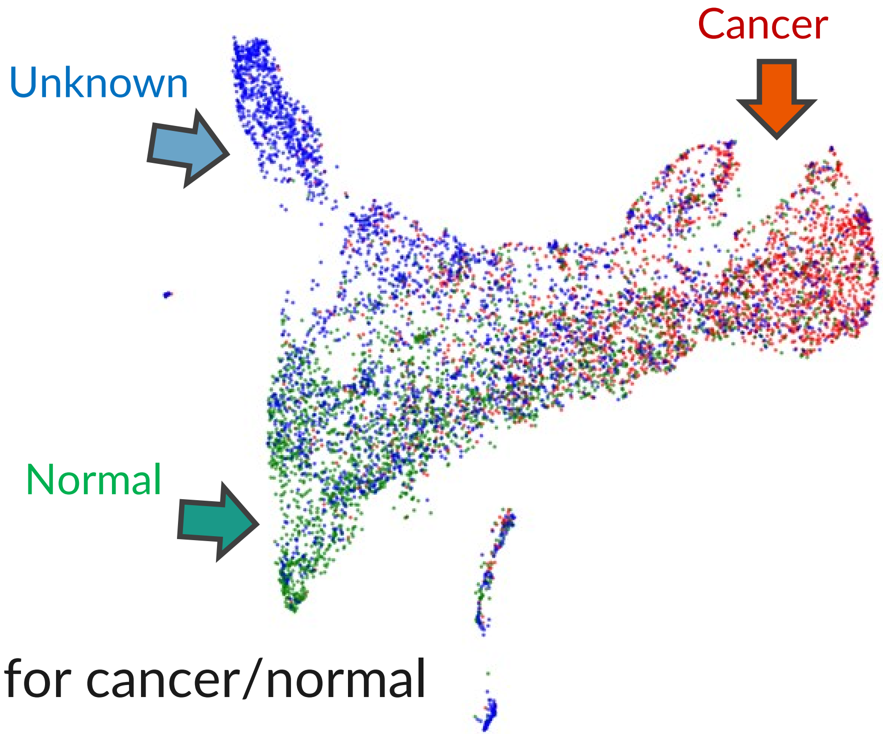
(x, y, z) vector



https://pair-code.github.io/understanding-umap/

2D visualization for cell images

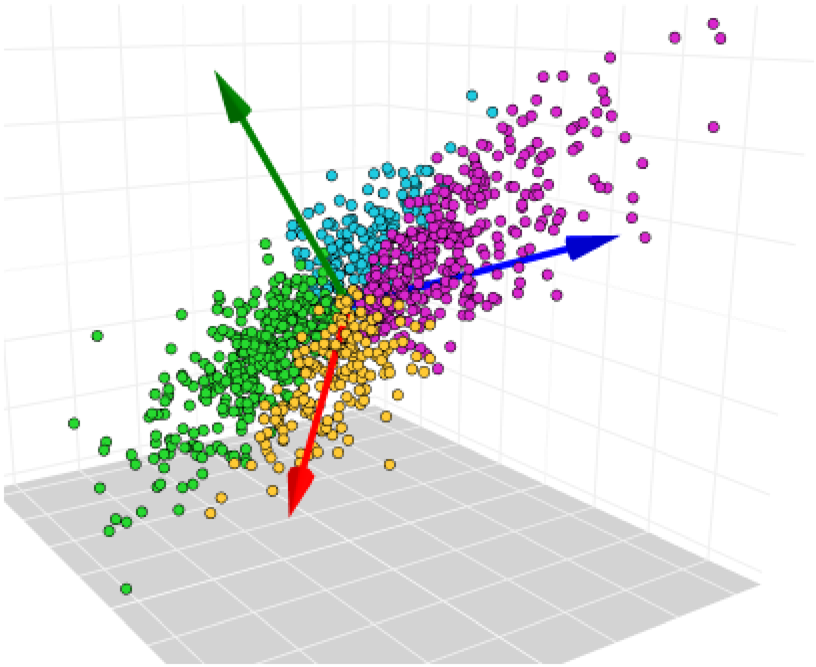




- Extracted from deep learning model for cancer/normal cell type classification

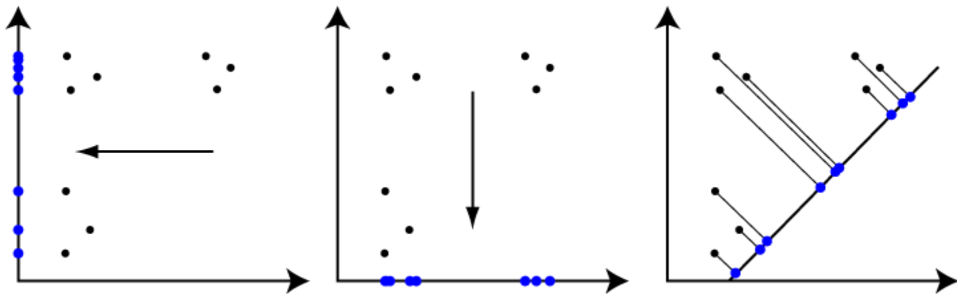
Principal component analysis (PCA)

Variance is information



https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

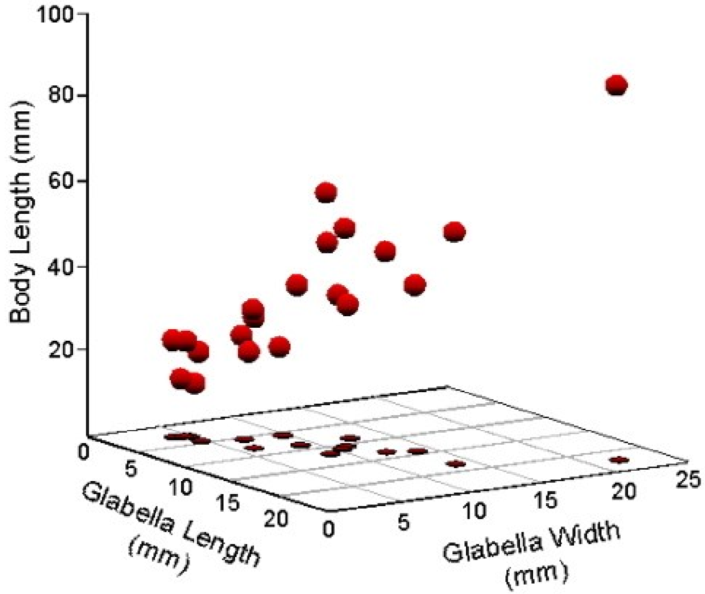
Projection

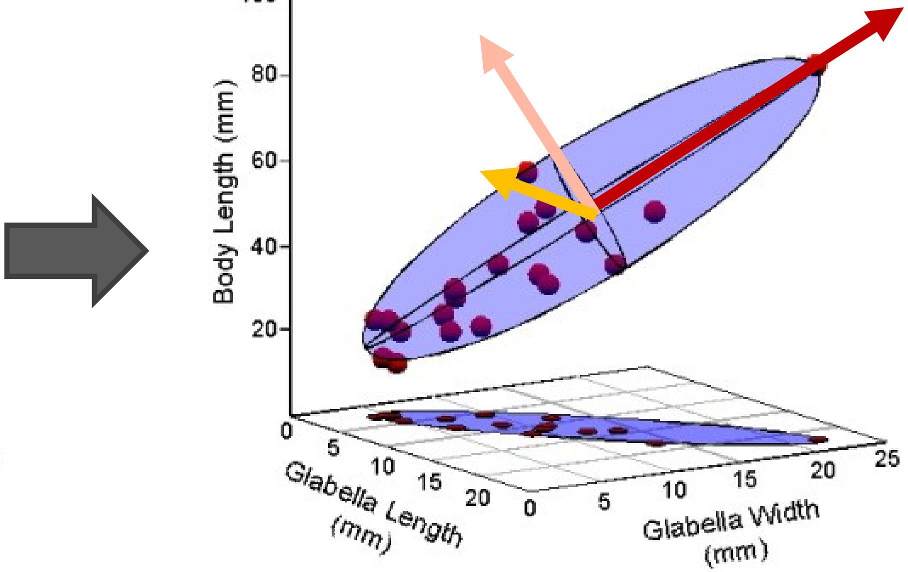


https://shapeofdata.wordpress.com/2013/04/16/visualization-and-projection/

- High variances = more power to distinguish groups of data points

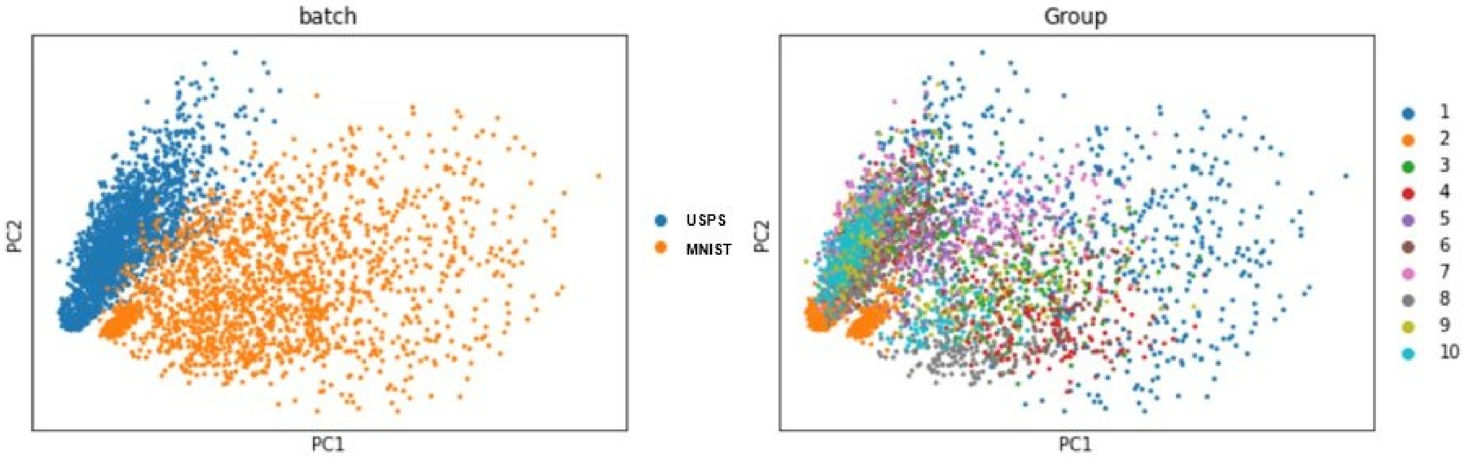
PCA prioritizes directions with high variances





Source: the paleontological association

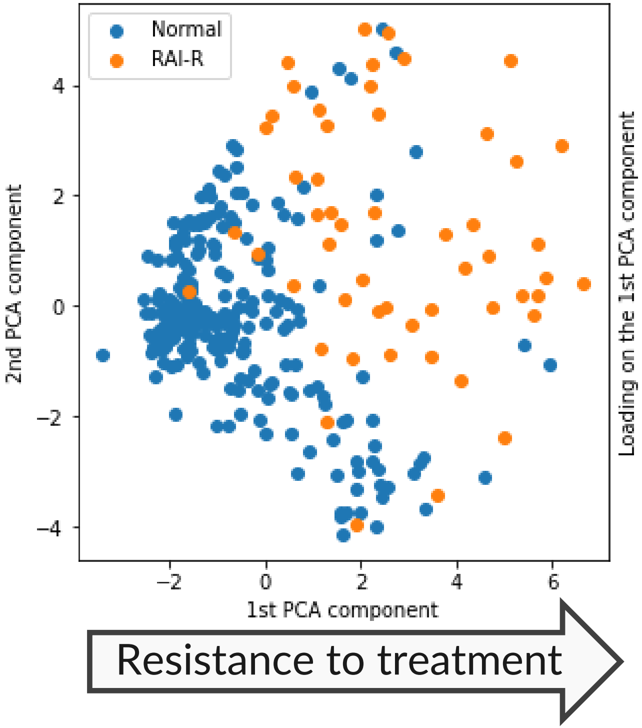
Interpretation of PCA result

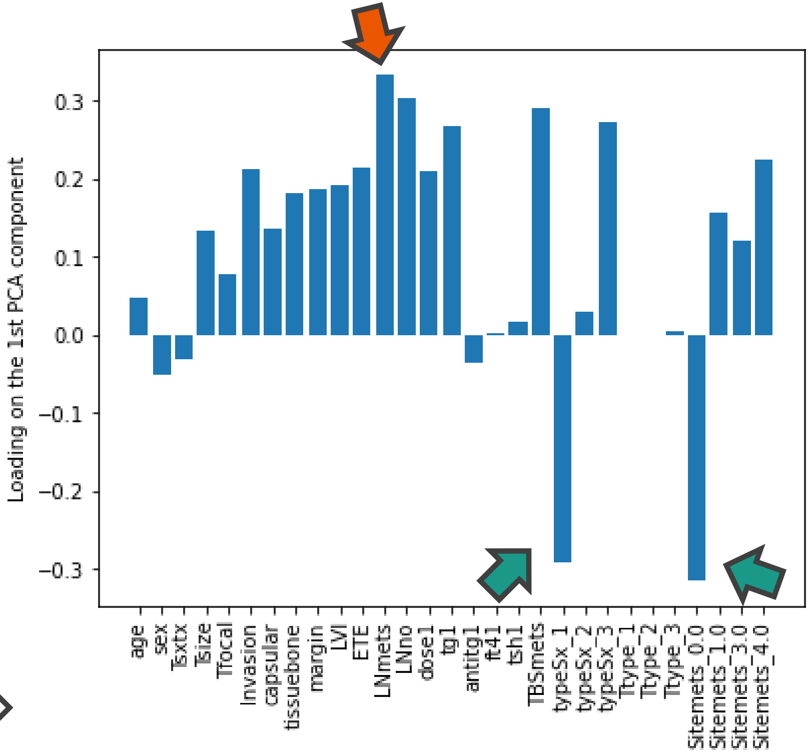


https://twitter.com/lkmklsmn/status/1436357177887895555

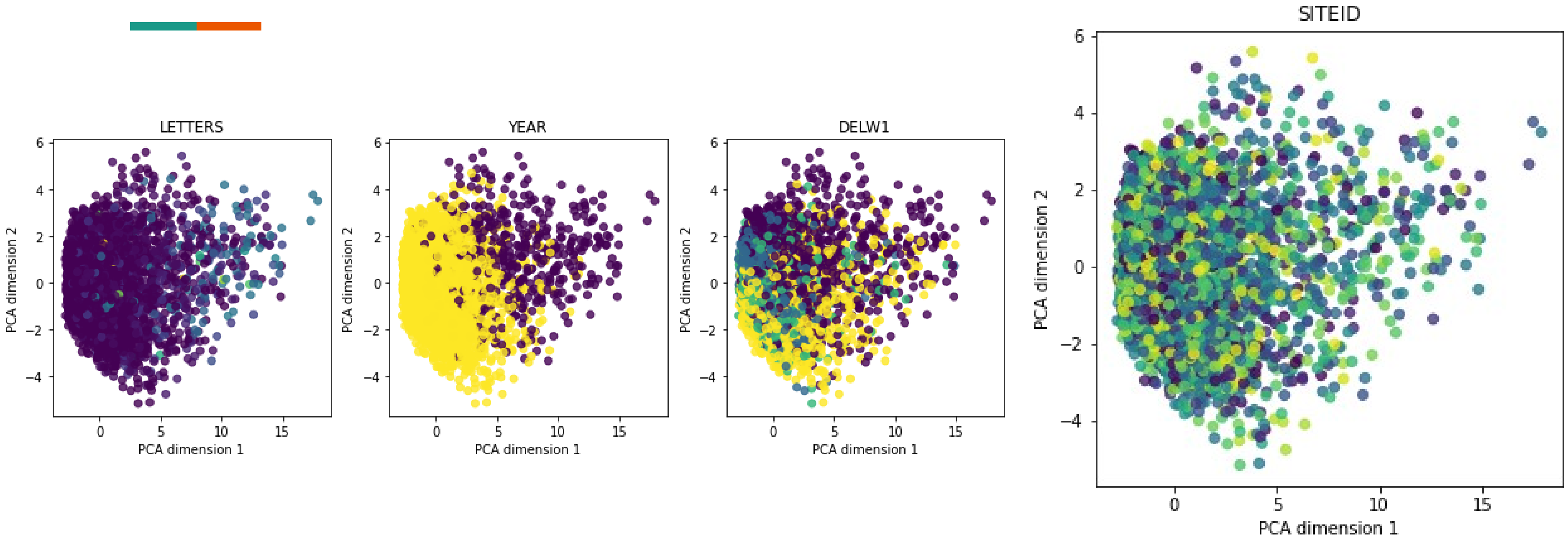
- PC1 captures the variance between data sources
- PC2 somewhat captures the variance between digit identity

Interpreting loadings on individual PC



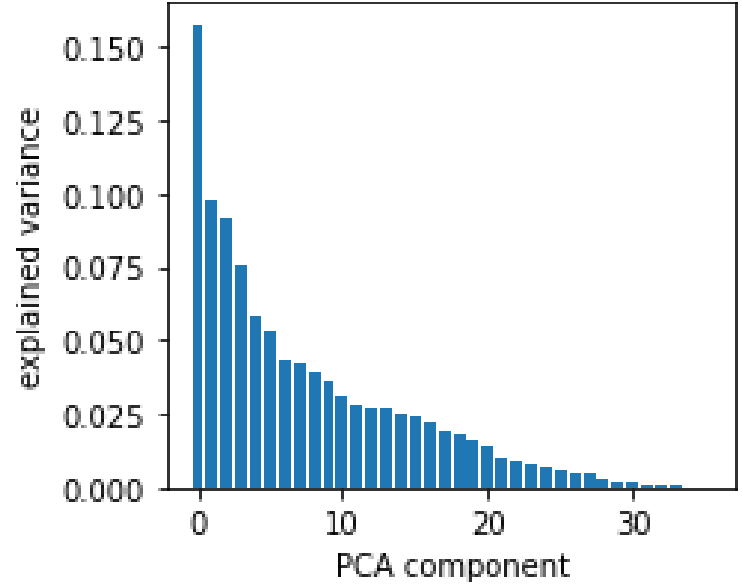


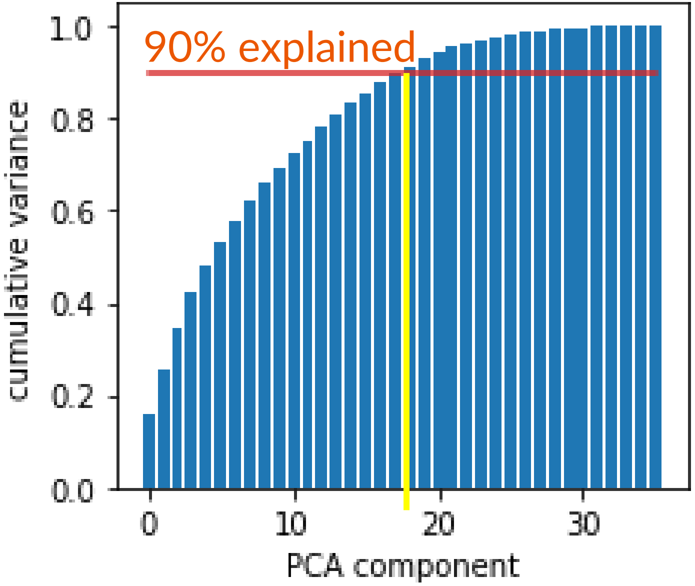
Exploring PCA results



- Color by feature values to understand how PCA group data points
- Color by potential confounding factors

PCA for dimensionality reduction



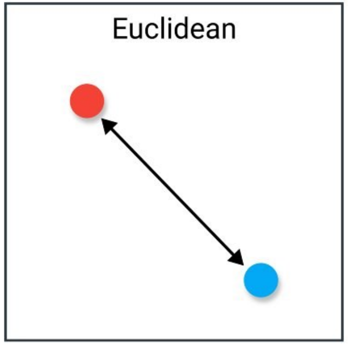


- By default, PCA retains the number of dimensions
- We can select only the first *k* PC for downstream analyses

Pros and cons of PCA

- Each PC can be interpreted from the loadings
- Highly correlated features tend to be grouped into the same PC
- PCA is a good initial dimensionality reduction step

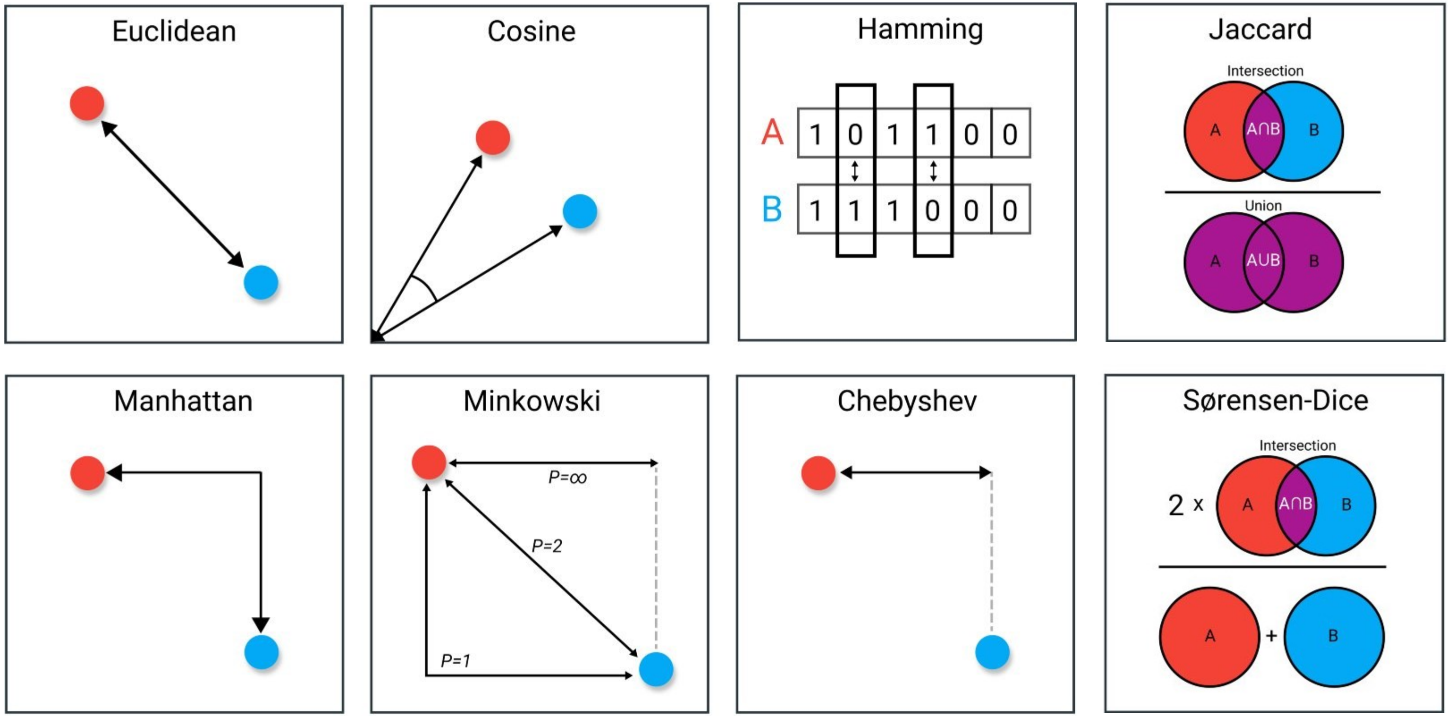
- PCA strictly preserves Euclidean distance
  - But some datasets require different distance metric!



https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Multidimensional Scaling (MDS)

Distances



https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Pairwise distance matrix

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

http://www.slimsuite.unsw.edu.au/teaching/upgma/

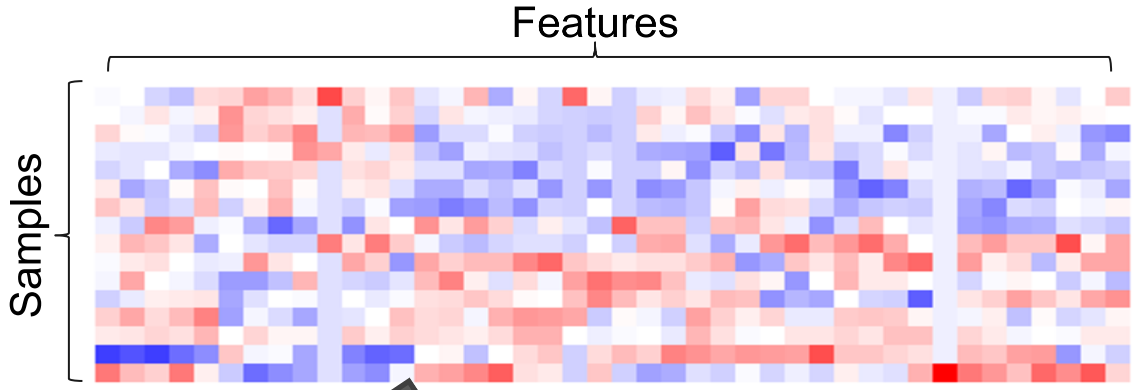
Metric properties

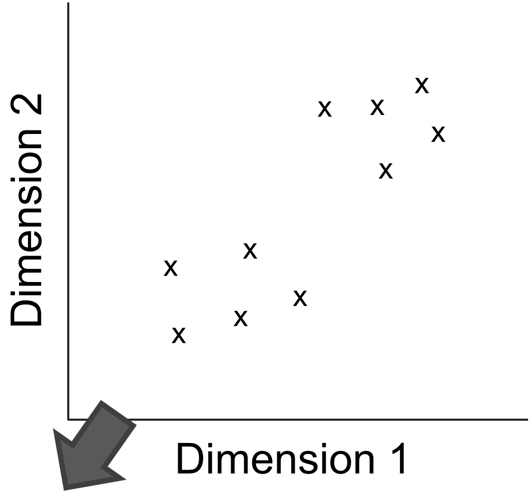
- $D(i, j)$ = distance between sample i and sample j
- $D(i, i) = 0$
- $D(i, j) = 0$ iff $i = j$
- $D(i, j) = D(j, i)$
- $D(i, j) + D(j, k) \geq D(i, k)$

Non-metric

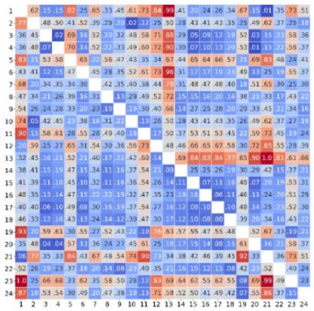
- Any user-defined dissimilarity
- $D(i, j) \mathrel{!=} D(j, i)$

Principal Coordinate Analysis (PCoA)

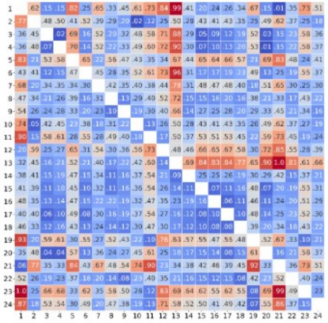




Non-Euclidean distance

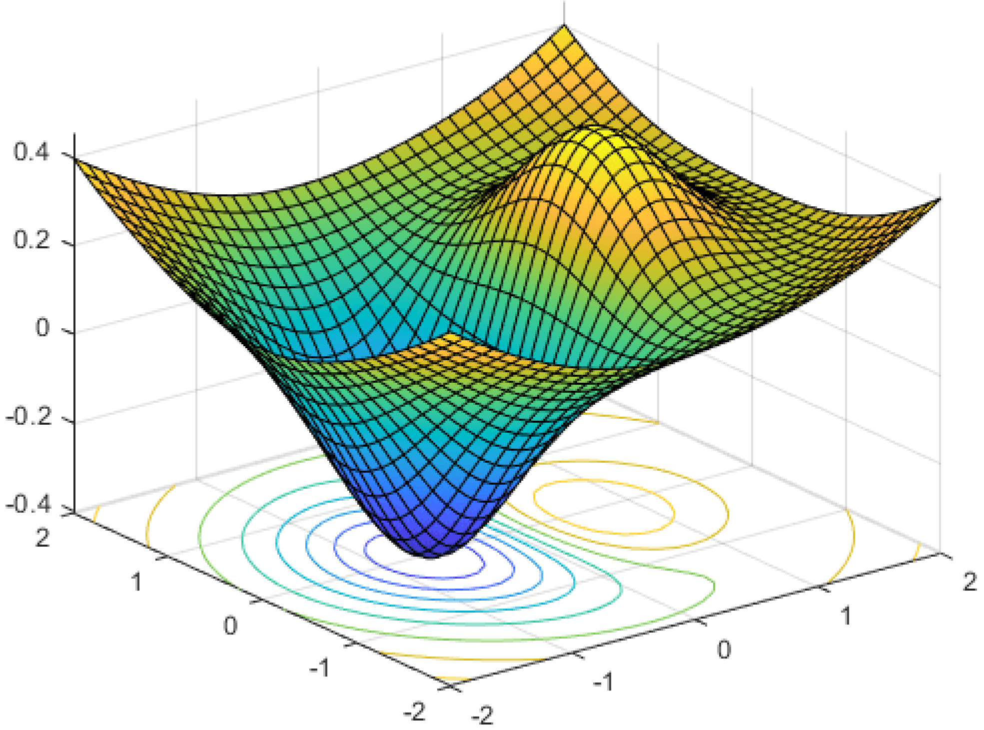


Pearson correlation



Euclidean distance

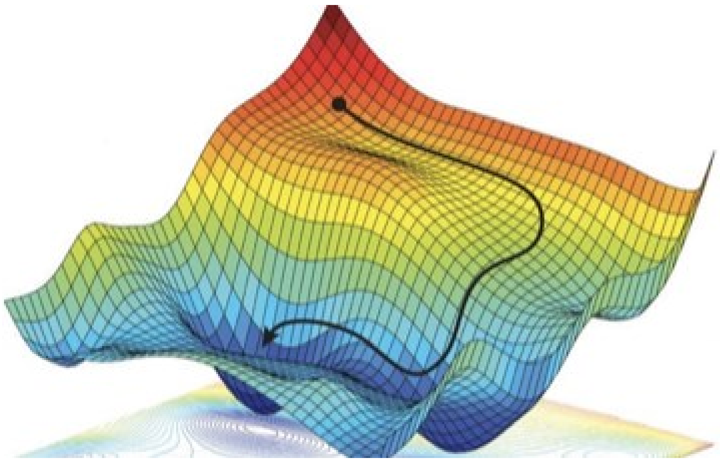
How to optimize a function?



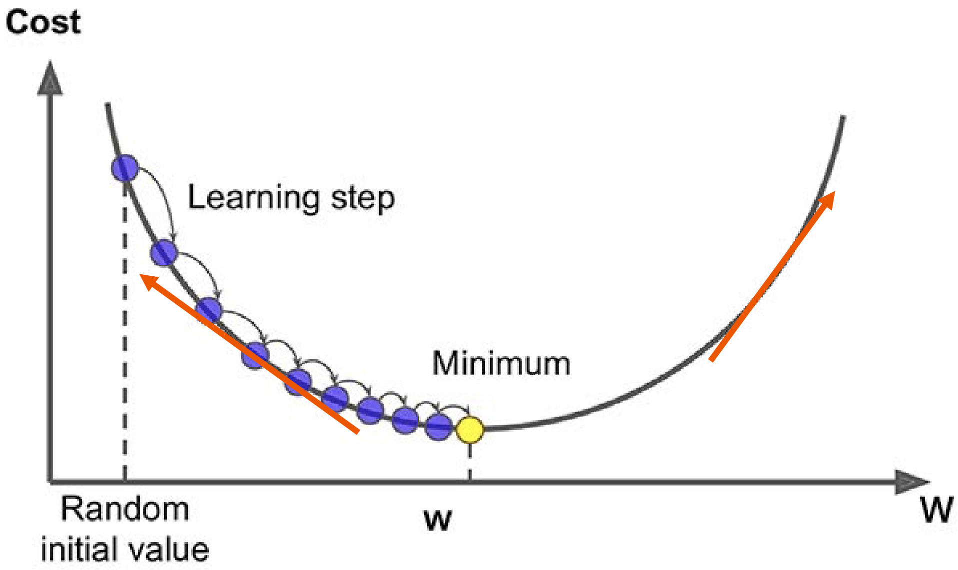
https://es.mathworks.com/help/optim/ug/optimization-toolbox-tutorial.html

- Find $(x_1, x_2, \ldots, x_n)$ that minimize $f(x_1, x_2, \ldots, x_n)$
- At minimum, the slope is zero in all directions
- Take derivative of each variable and set to zero
    - $\frac{\delta f}{\delta x_1} = 0$
    - $\frac{\delta f}{\delta x_2} = 0$
    - *n* equations with *n* variables

Gradient descent

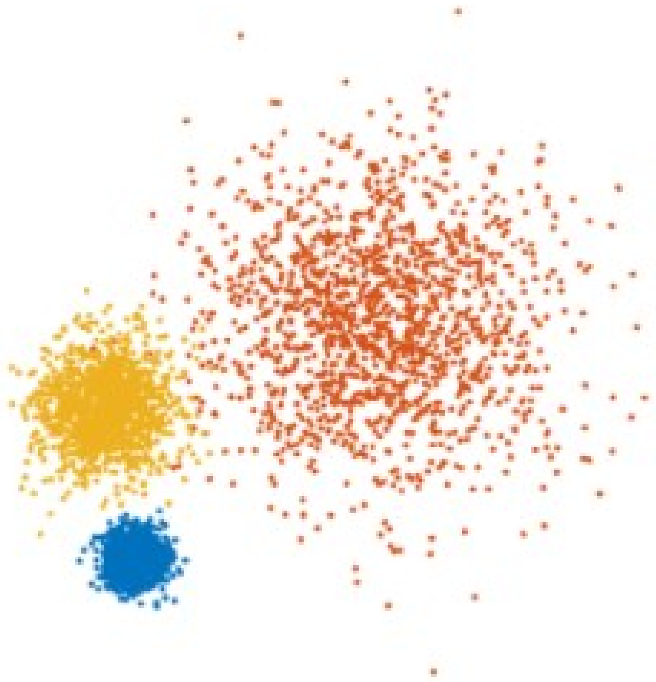


https://medium.com/analytics-vidhya/gradient-descent-b0dc1af33517



- Slope tells us if the function is increasing or decreasing if we increase x_i
  - So, we can update x_i accordingly

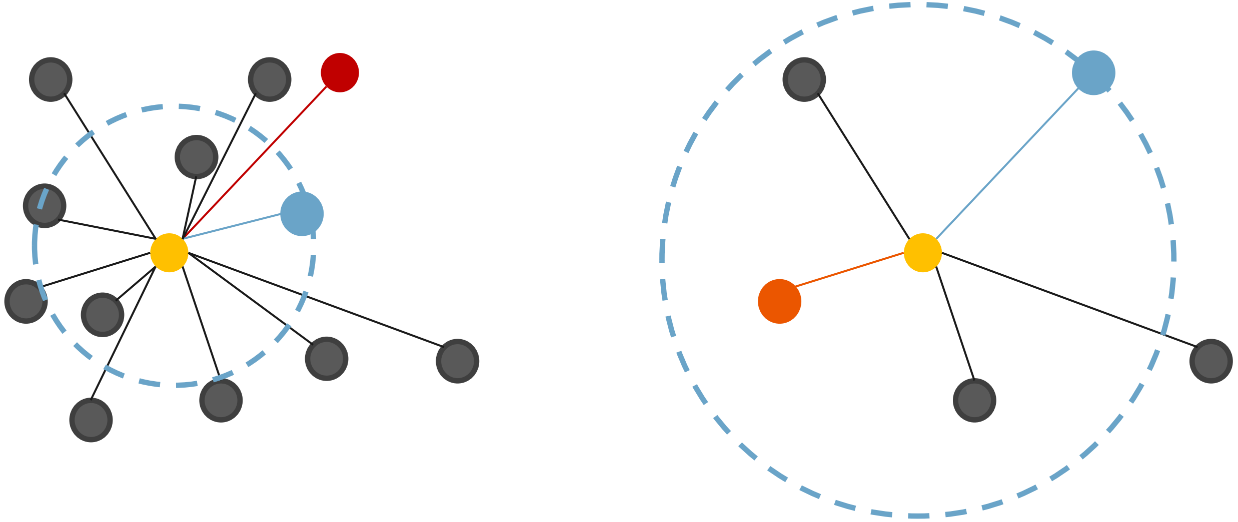
Limitation of PCA and MDS



- A single definition of distance is used throughout the data space
- What if some data groups are noisier than the others?
  - Difference in data density

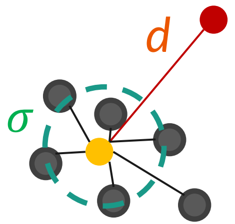
t-distributed stochastic neighbor embedding (*t*-SNE)

Measuring data density



- Distance to the *k*-th nearest neighbor reflects data density
  - Small distance in dense area
  - Large distance in sparse area

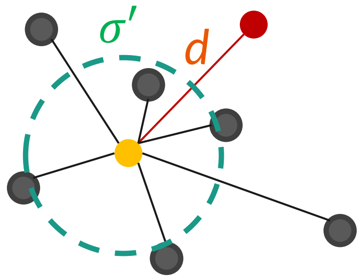
Probability of being a neighbor



score(o | o) = probability that o would pick o as neighbor
under a **normal distribution** center at o

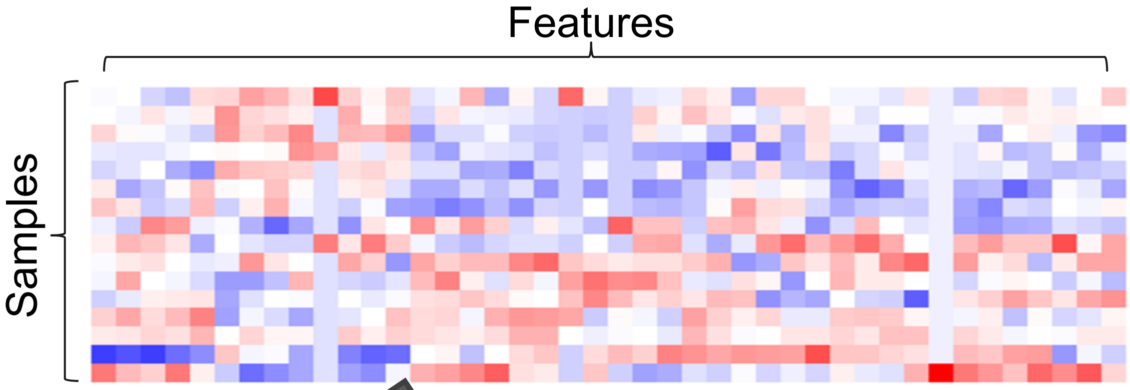
$$= \frac{e^{-\frac{d^2}{2\sigma^2}}/\sigma}{\sum e^{-\frac{(\mathrm{dist}(o,o))^2}{2\sigma^2}}/\sigma}$$

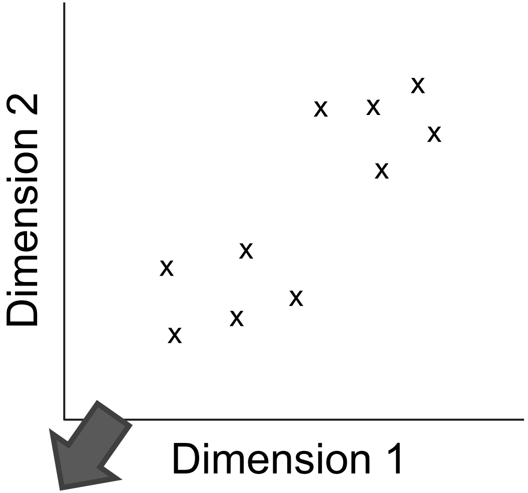
o = other data points



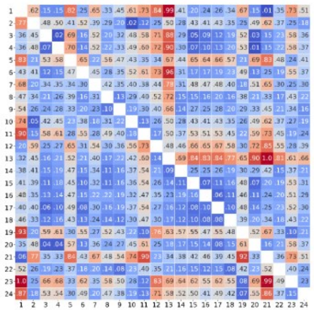
- Same distance d normalized against density σ and distances to other nearby data points o

Finding the optimal projection for *t*-SNE

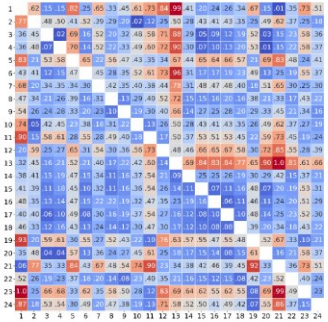




Probability of being a neighbor (Normal) (σ depend on density)



Maximize similarity



Probability of being a neighbor (*t*-distribution) (σ = 1)

Why *t*-distribution for the projection?

- *t*-distribution has fatter tails

- Allow data points to be projected far away from each other

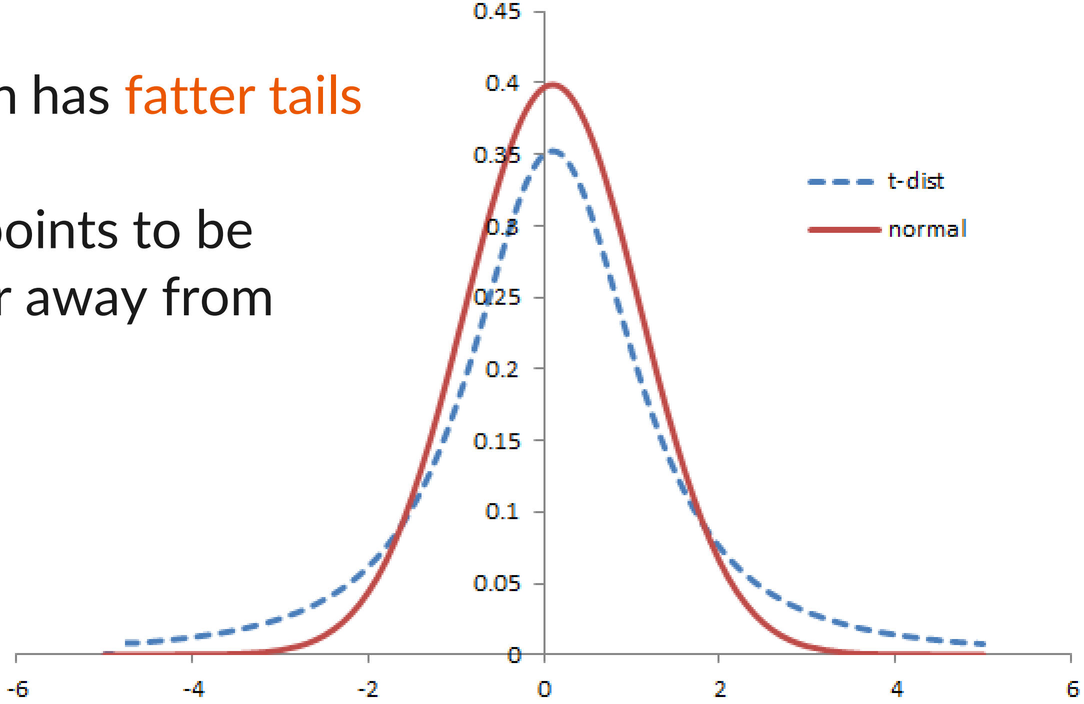
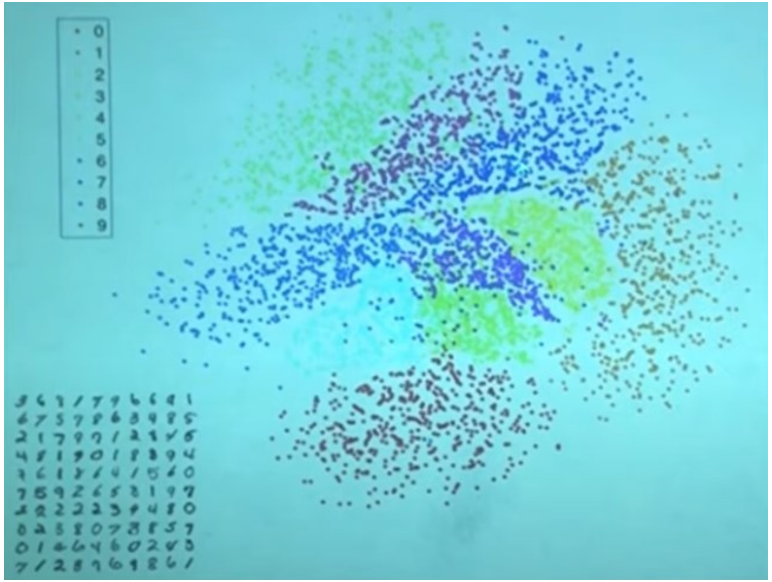
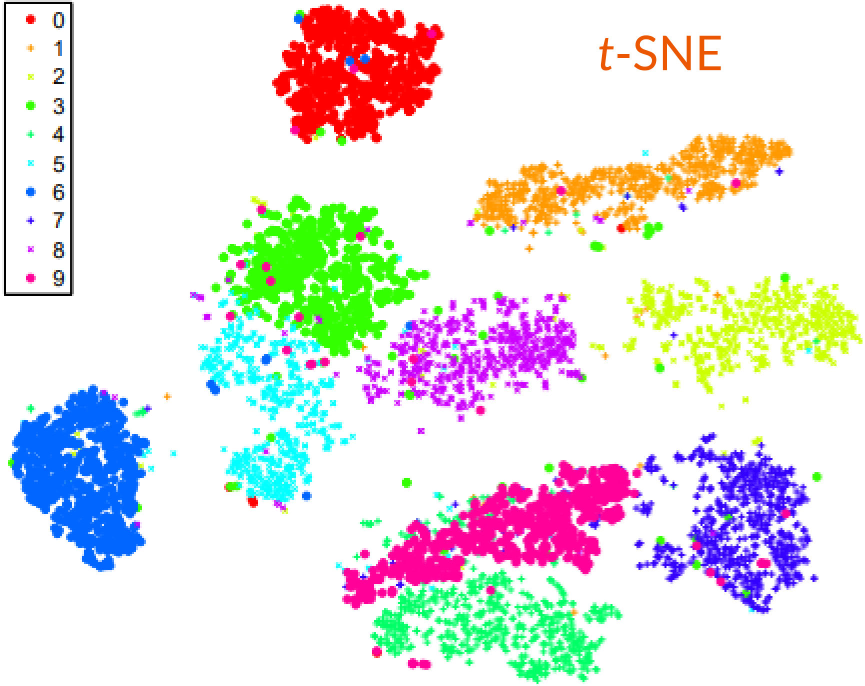


Image from riskprep.com

Impact of *t*-distribution



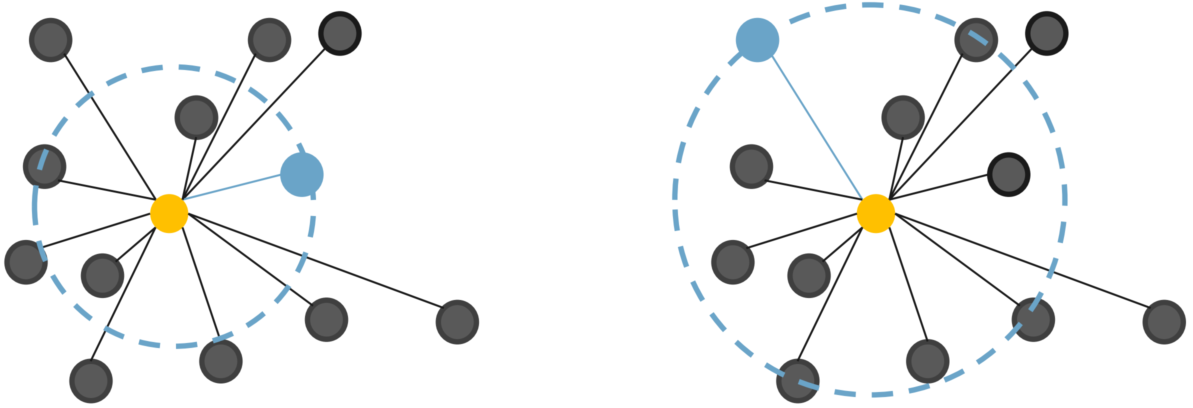
SNE (Normal → Normal)



t-SNE

Maaten, L. and Hinton, G. J of Machine Learning Research 9:2579-2605 (2008)

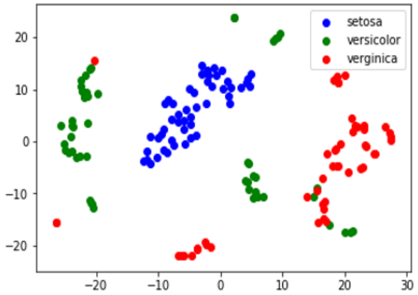
Perplexity



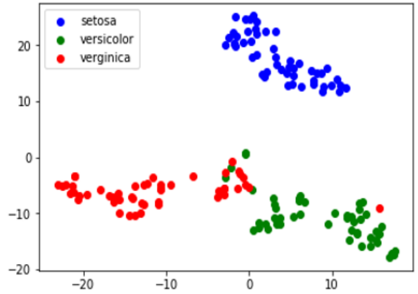
- How many nearest neighbors to consider to normalize data density?
    - Perplexity parameters

Impact of perplexity

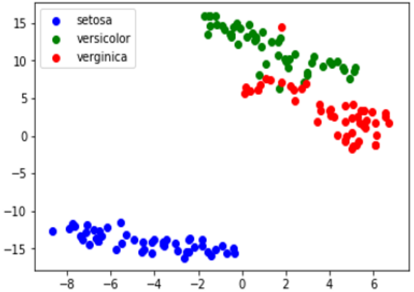
K = 5



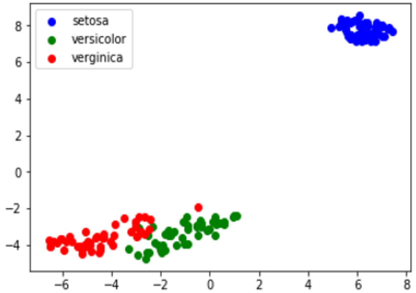
K = 10



K = 25



K = 50



- Too small perplexity = a lot of scatted data groups

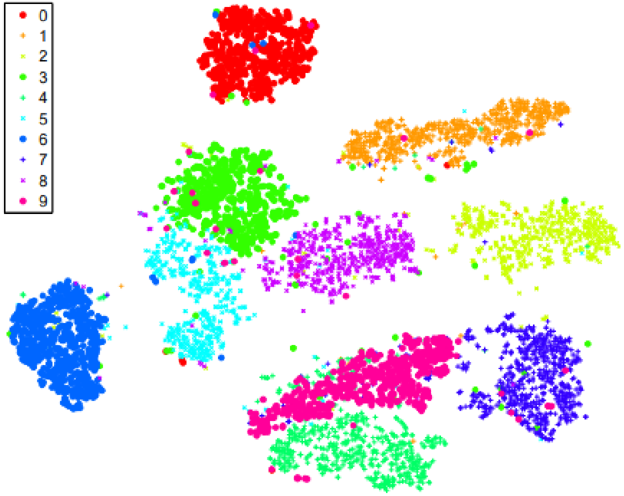
- Try varying the perplexity and identify patterns that consistently appear

Source: blog.paperspace.com/dimension-reduction-with-t-sne/

Pros and cons of *t*-SNE

- Capture neighbor relationship
- Normalize data density

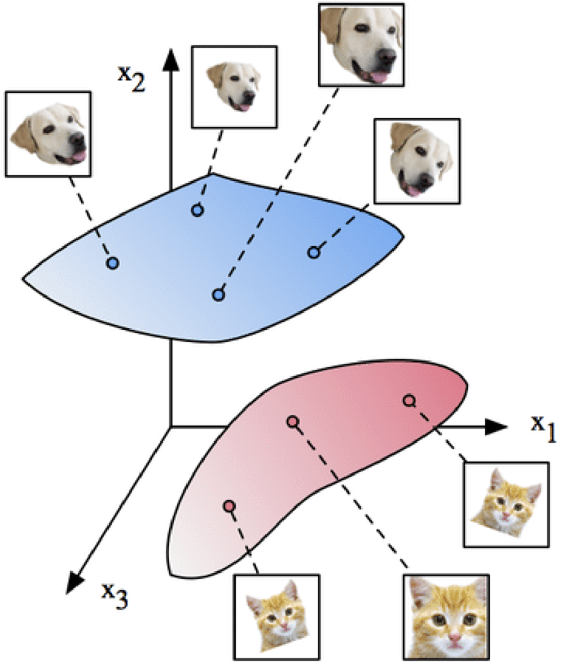
- Recompute every time new data is added
- Lose long-range relationship
- Axes of the resulting projection have no meaning
    - Don’t use *t*-SNE coordinates for clustering or interpretation



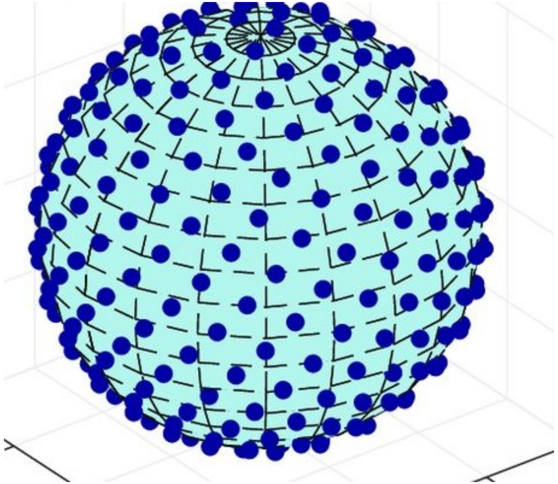
Maaten, L. and Hinton, G. J of Machine Learning Research 9:2579-2605 (2008)

Uniform manifold approximation and projection (UMAP)

Two key assumptions



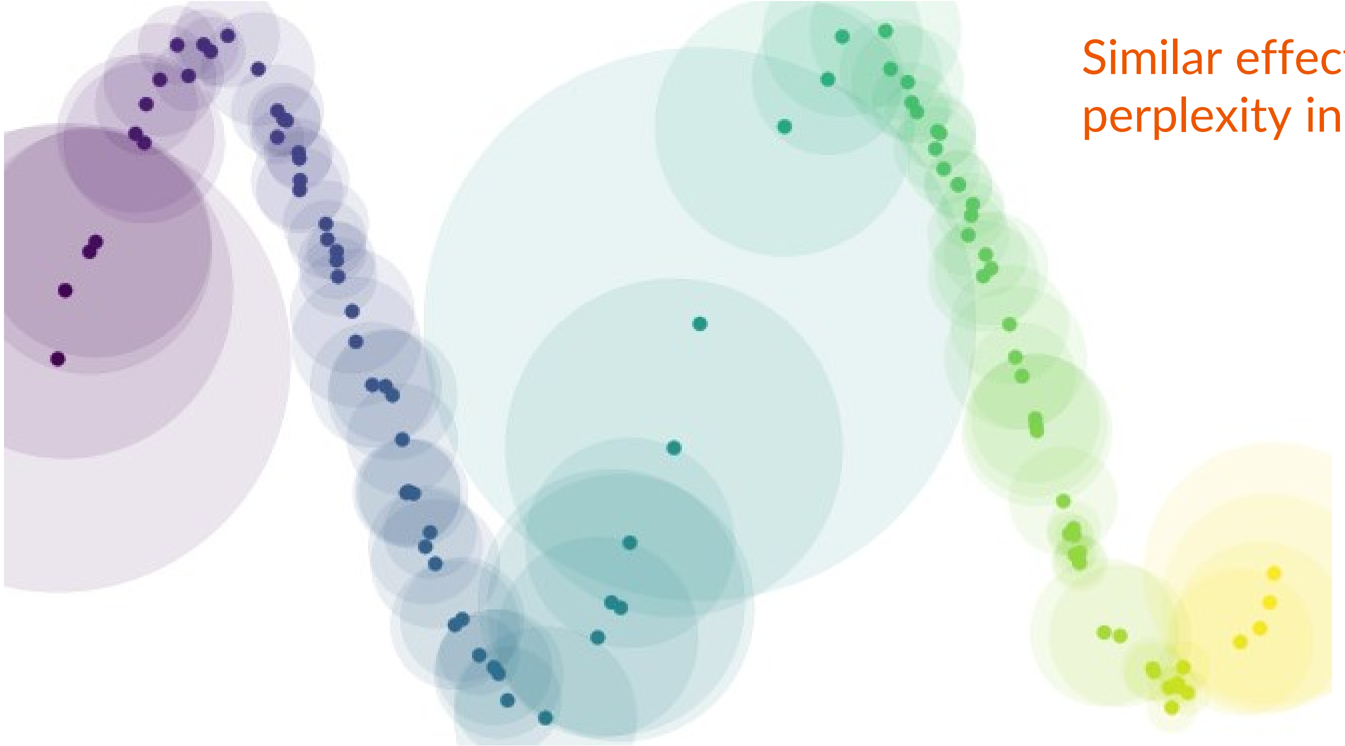
Chung, S. et al. “Classification and Geometry of General Perceptual Manifolds”



Ali, A. et al. IEEE Access PP(99):1 (2021)

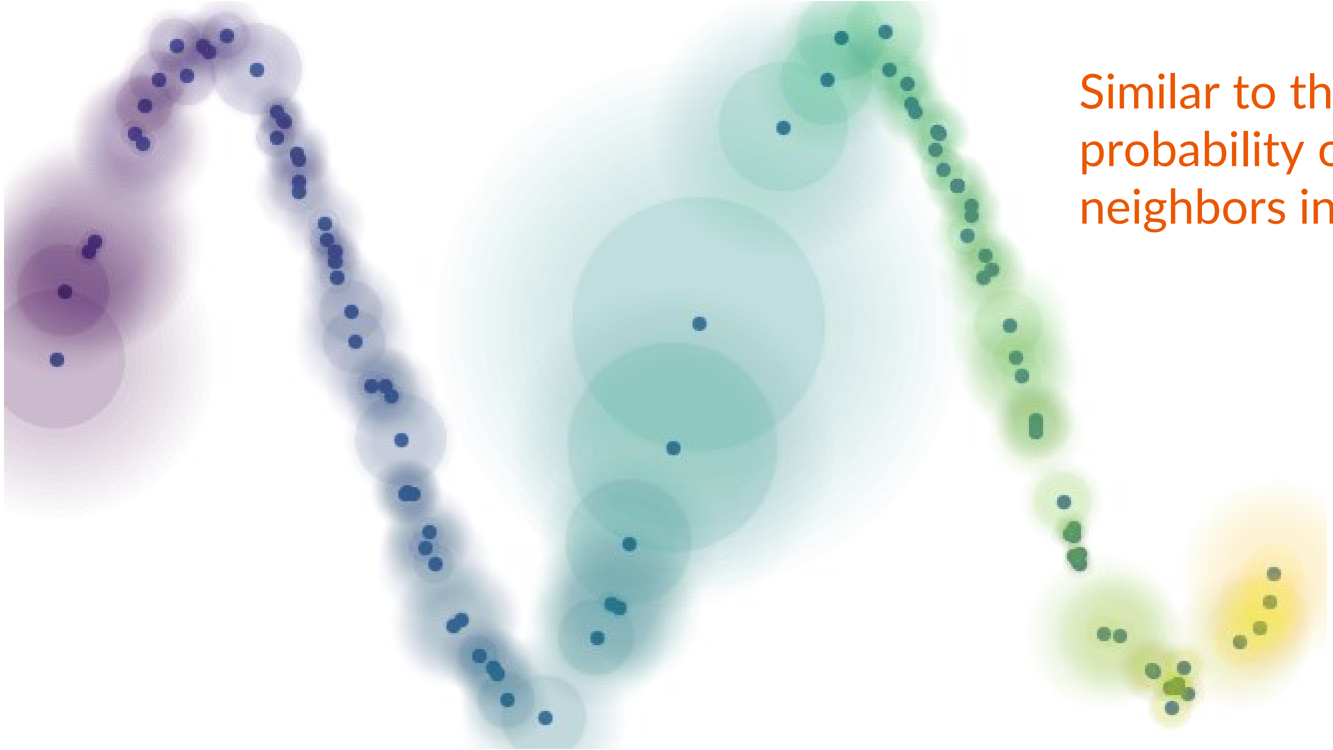
- Data came from multiple manifolds
- Data points were sampled uniformly

Uniform sampling = similar distance to *k*-th neighbor



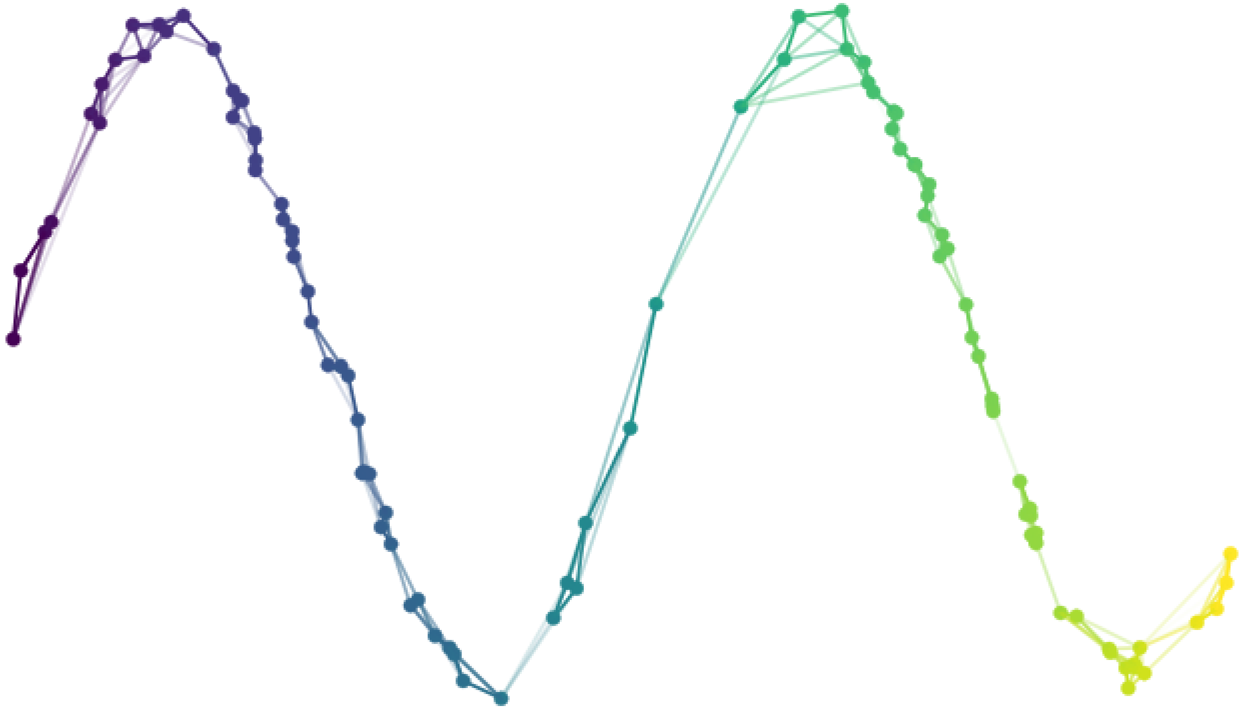
Similar effect as perplexity in t-SNE

Adding uncertainty between faraway data points

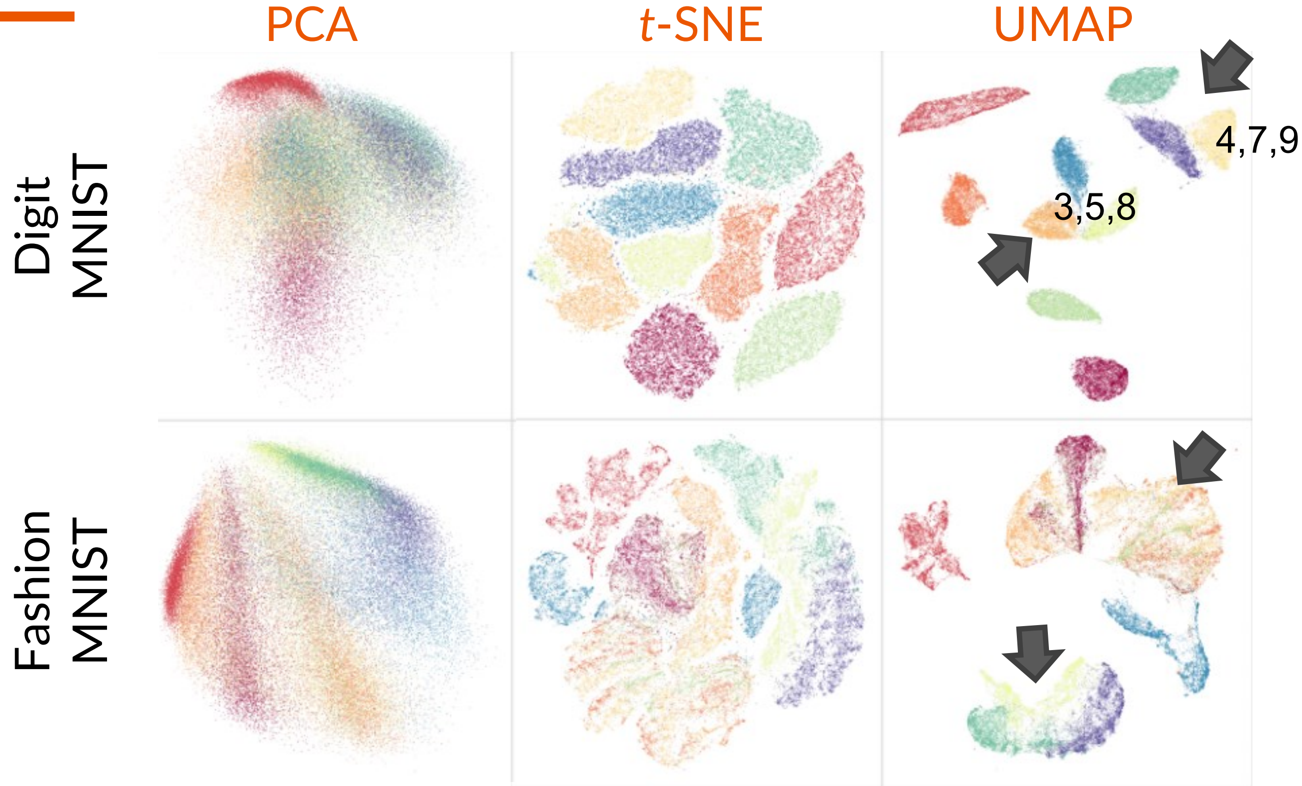


Similar to the probability of being neighbors in *t*-SNE

Network representation of neighbor relationship

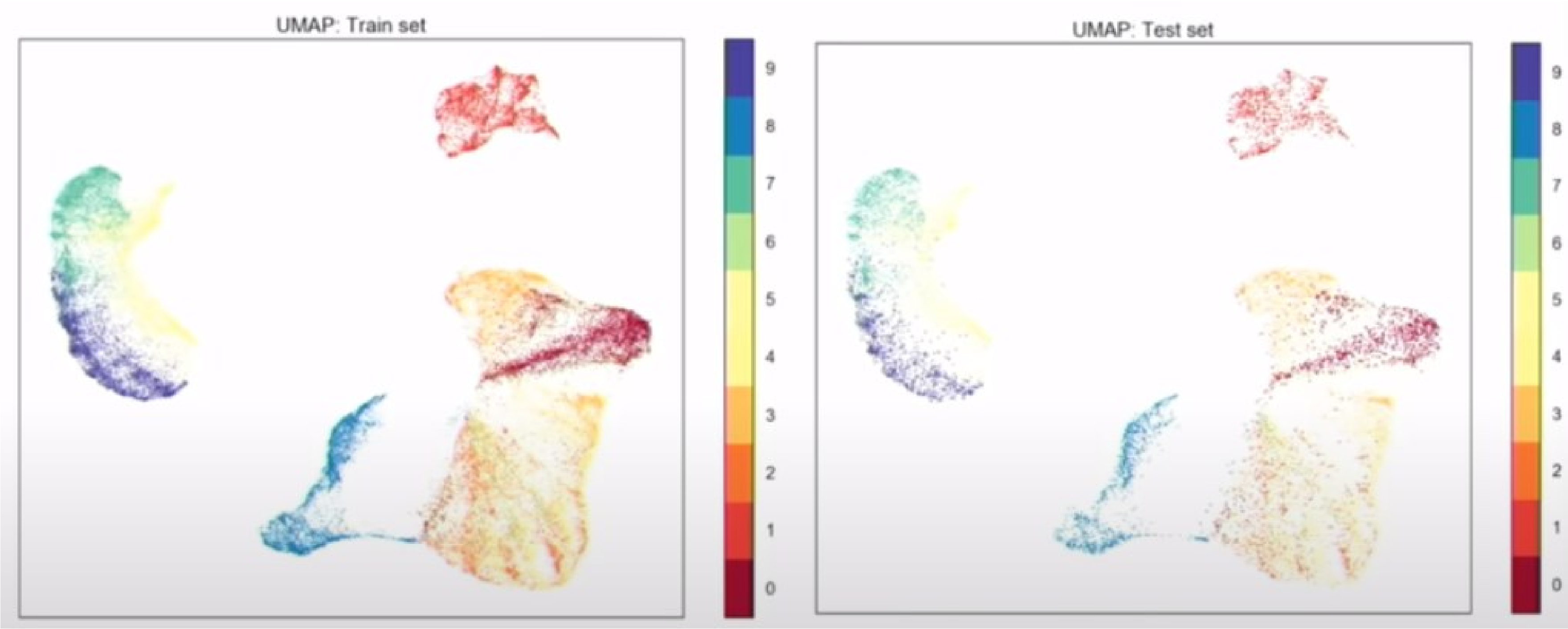


UMAP can capture long-range relationship



McInnes, L., Healy, J. and Melville, J. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”

UMAP can transform new data points



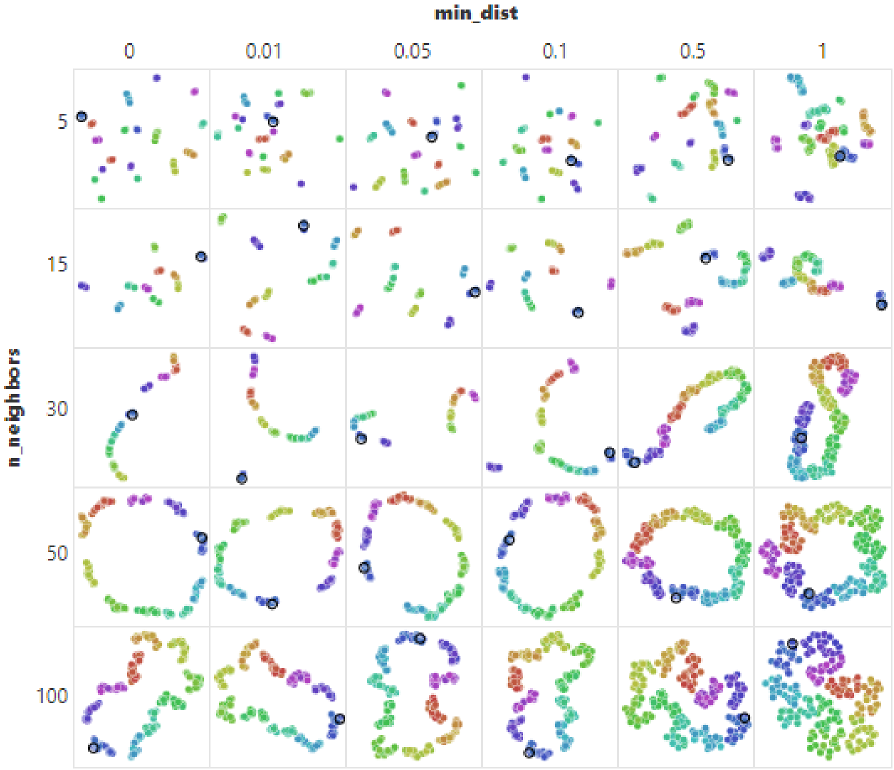
UMAP presentation by Dr. McInnes

Pros and cons of UMAP

- Can capture long-range relationship
- Can be applied to new data points without recomputing

- Require strong assumptions

Customizing UMAP outputs



- Number of neighbors (n_neighbors) is perplexity

- Minimum distant for placing similar data point (min_dist) is for adjusting the scale of visualization

Source: https://pair-code.github.io/understanding-umap/

Any questions?

See you on February 15th