Special Topics in Biostatistics and Bioinformatics
Week III

Ege Ülgen, M.D.

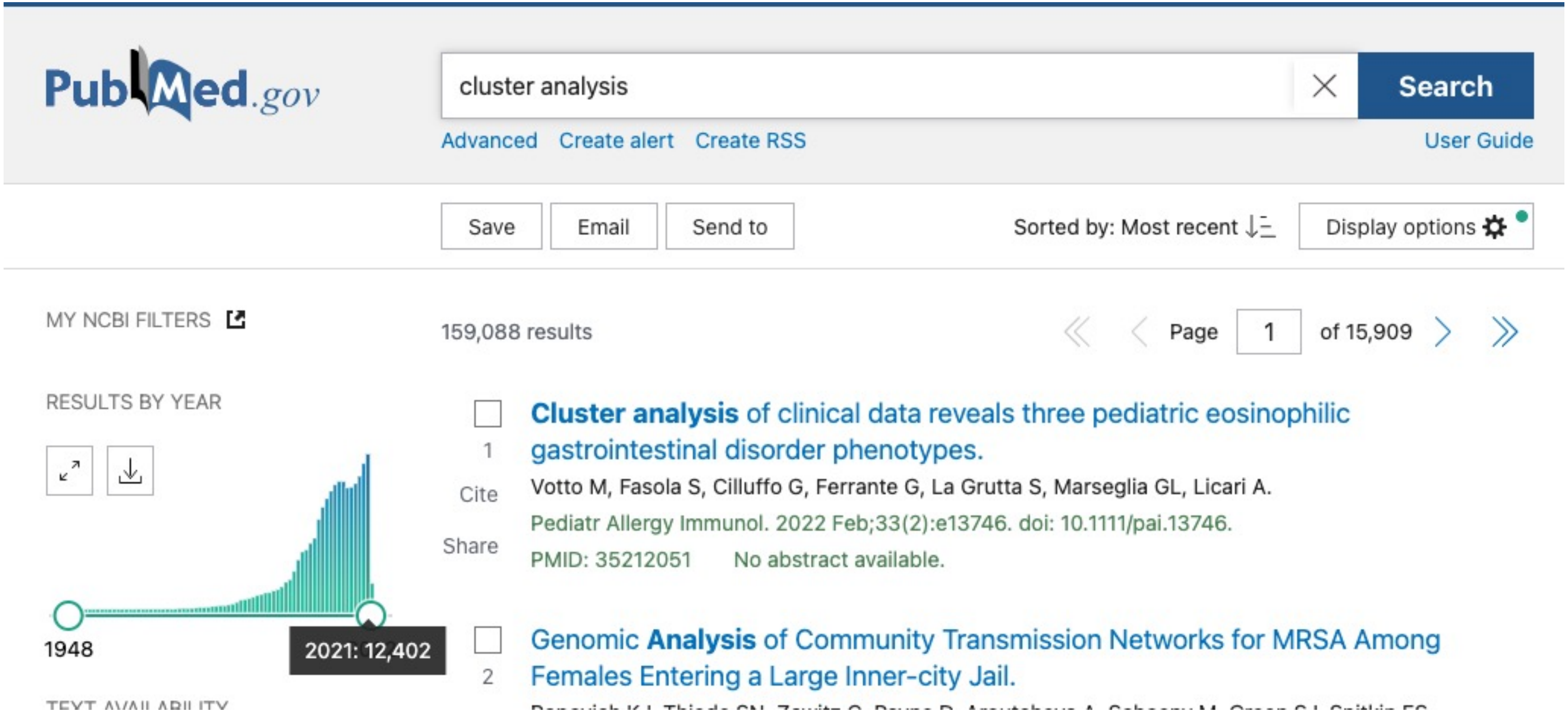
17 March 2022

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Clustering

find “close” samples or genes etc.

put them into groups

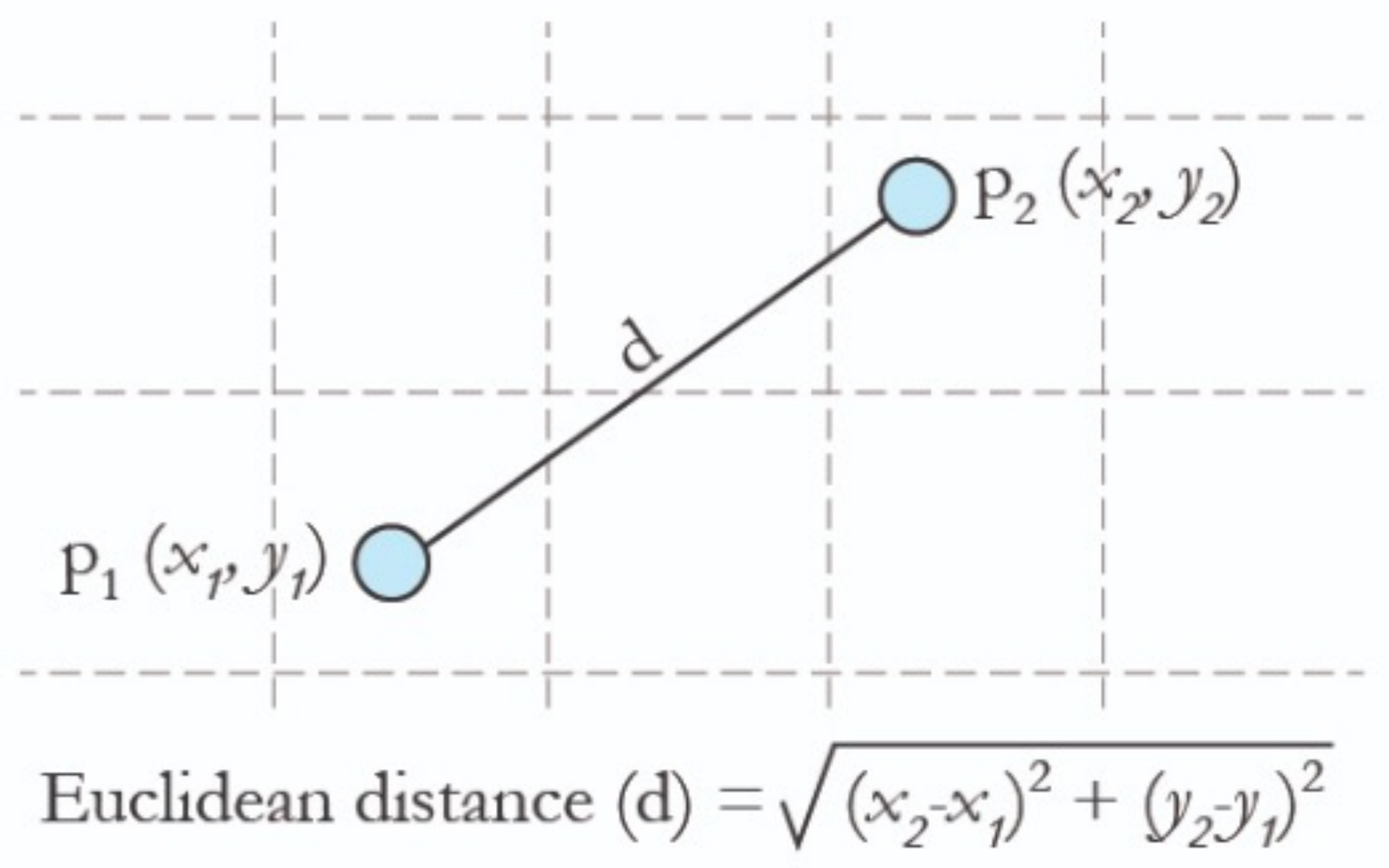

PubMed.gov
cluster analysis
Search
Advanced
Create alert
Create RSS
User Guide
Save
Email
Send to
Sorted by: Most recent
Display options
MY NCBI FILTERS
159,088 results
Page
1
of 15,909
RESULTS BY YEAR
1948
2021: 12,402
Cluster analysis of clinical data reveals three pediatric eosinophilic gastrointestinal disorder phenotypes.
1
Cite
Share
Votto M, Fasola S, Cilluffo G, Ferrante G, La Grutta S, Marseglia GL, Licari A.
Pediatr Allergy Immunol. 2022 Feb;33(2):e13746. doi: 10.1111/pai.13746.
PMID: 35212051
No abstract available.
Genomic Analysis of Community Transmission Networks for MRSA Among Females Entering a Large Inner-city Jail.
2

https://www.tutorialexample.com/calculate-euclidean-distance-in-tensorflow-a-step-guide-tensorflow-tutorial/

Defining "distance"

Eucledian distance

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

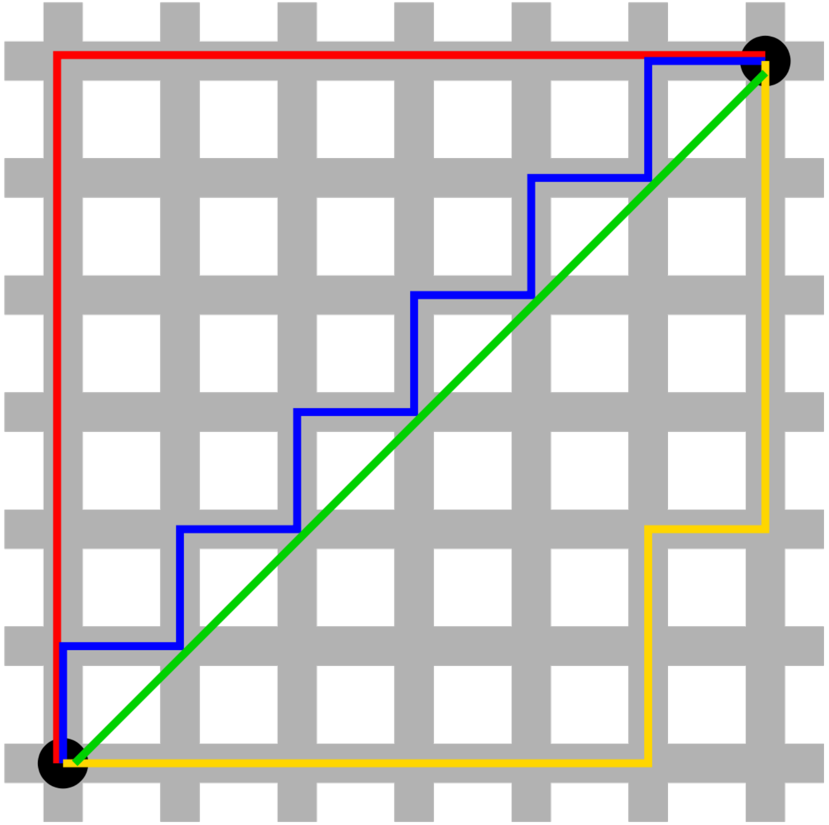


https://en.wikipedia.org/wiki/Taxicab_geometry

Defining "distance"

Manhattan distance

$$d_{man}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$



https://en.wikipedia.org/wiki/Taxicab_geometry

Defining “distance”

Minkowski distance

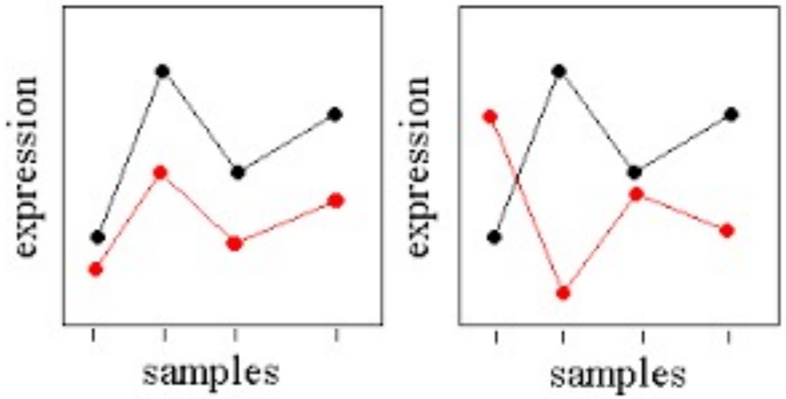
$$d_{mink}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Pearson_Correlation_and_Pearson_Squared_Distance_Metric.htm

Other "dissimilarity" measures

Pearson correlation distance

$$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



https://www.datanovia.com/en/lessons/clustering-distance-measures/

Other "dissimilarity" measures

Eisen cosine correlation distance

$$d_{eisen}(x,y) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2 \sum_{i=1}^{n} {y_i}^2}}$$

Spearman correlation distance

$$d_{spear}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i' - \bar{x'})(y_i' - \bar{y'})}{\sqrt{\sum_{i=1}^{n}(x_i' - \bar{x'})^2 \sum_{i=1}^{n}(y_i' - \bar{y'})^2}}$$

Kendall correlation distance

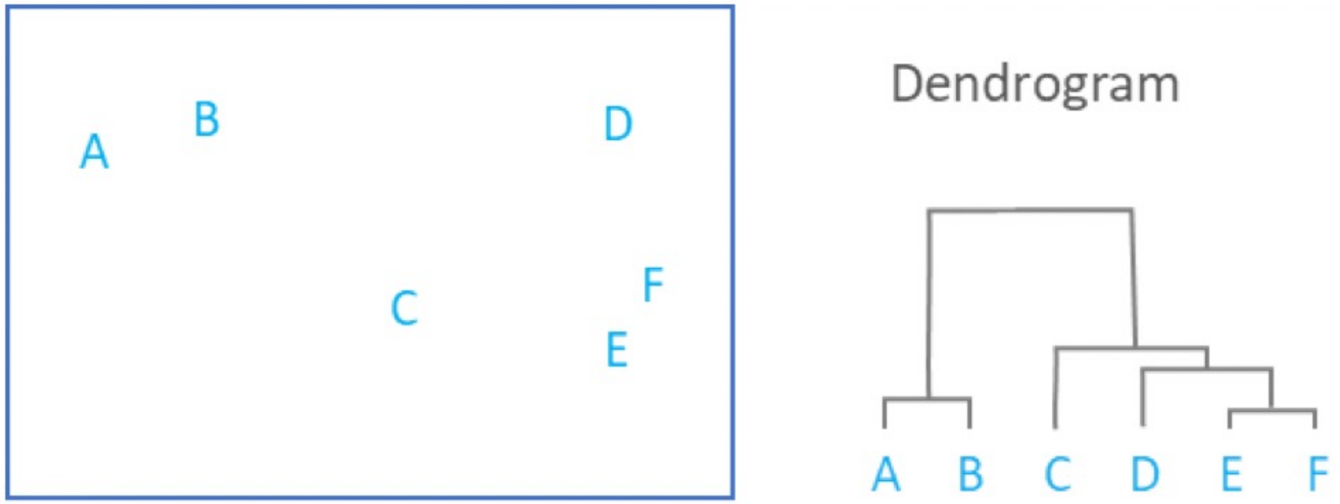
$$d_{kend}(x,y) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Clustering Methods

- Hierarchical clustering
- Centroid-based clustering
- Distribution-based clustering
- Density-based clustering
- Grid-based clustering
- Fuzzy clustering

…

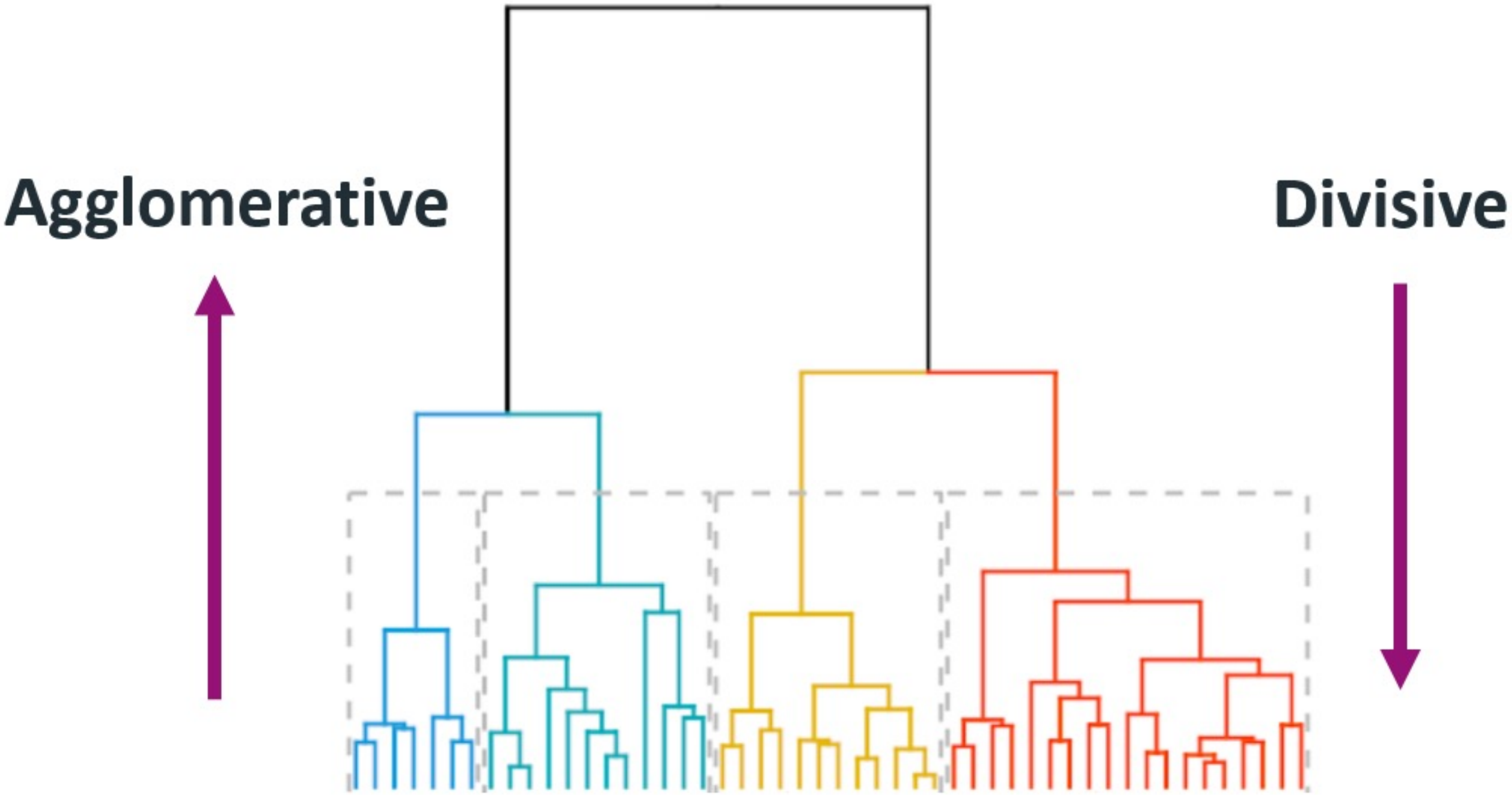
Hierarchical clustering



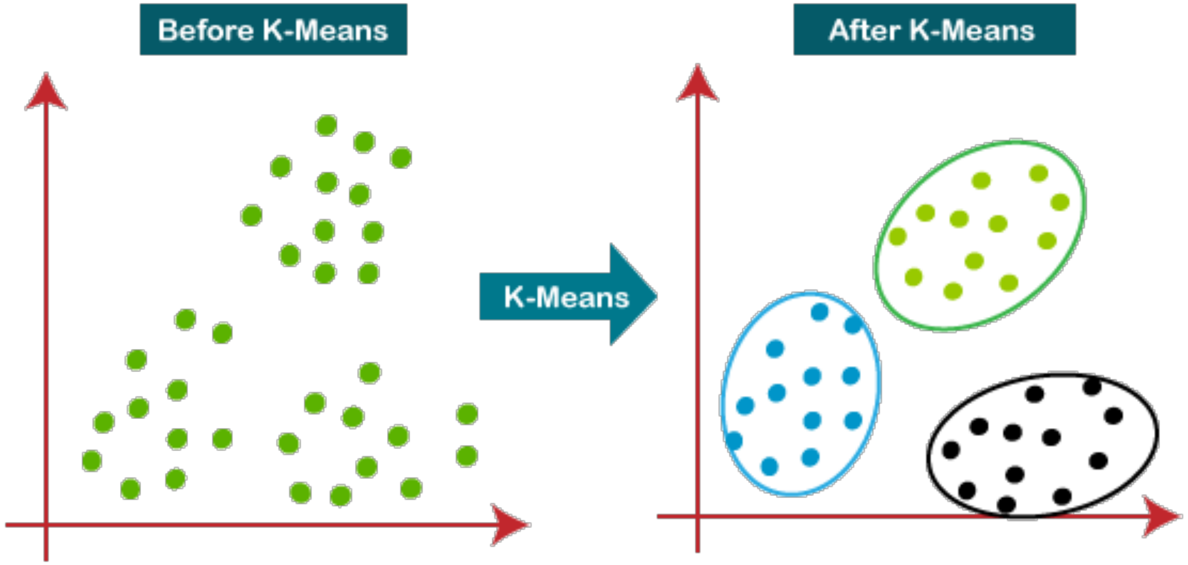
- Find the “closest” points
- Merge
- Repeat

https://harshsharma1091996.medium.com/hierarchical-clustering-996745fe656b

Hierarchical clustering



K-means clustering



- Initialize k cluster “centers” (random seeds)
- Assign each value to the closest center
- Update centers
- Reassign values
- Repeat

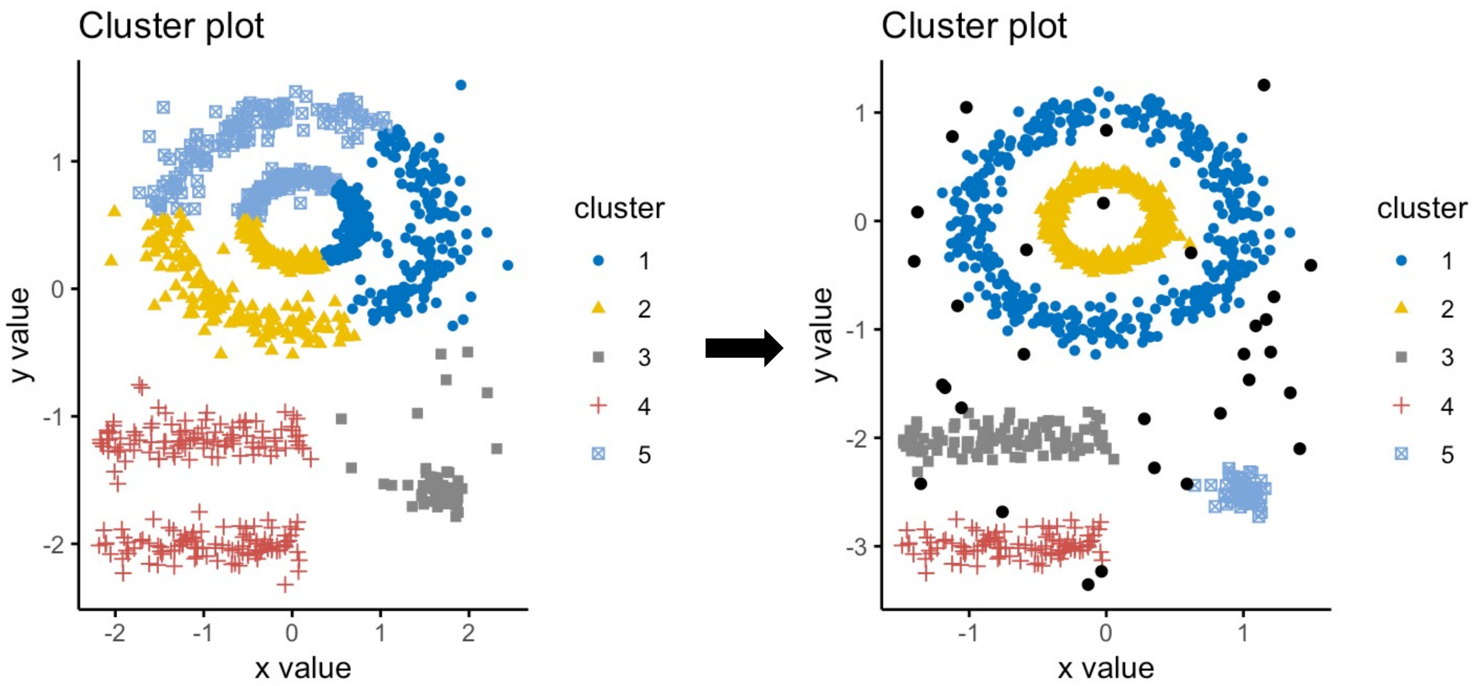
https://www.datanovia.com/en/lessons/dbscan-density-based-clustering-essentials/

Other Approaches

- Hierarchical K-means Clustering
- Hierarchical clustering on principal components

- Fuzzy Clustering
  - Fuzzy c-means

- DBSCAN



https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Determining the optimal number of clusters

- Visualization
- Elbow method
  - location of a bend in total within-cluster sum of square
- Average silhouette method
  - Maximum average silhouette
- Gap statistic method
  - the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap(k) \geq Gap(k+1) - s_{k+1}$.

Dimensionality Reduction

Overcoming the curse of dimensionality

https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e

Curse of dimensionality

- Cannot visualize observations

- Overfitting – terrible out of sample performance.
- Observations become harder to cluster — too many dimensions causes every observation to appear equidistant from all the others

https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e

Curse of dimensionality

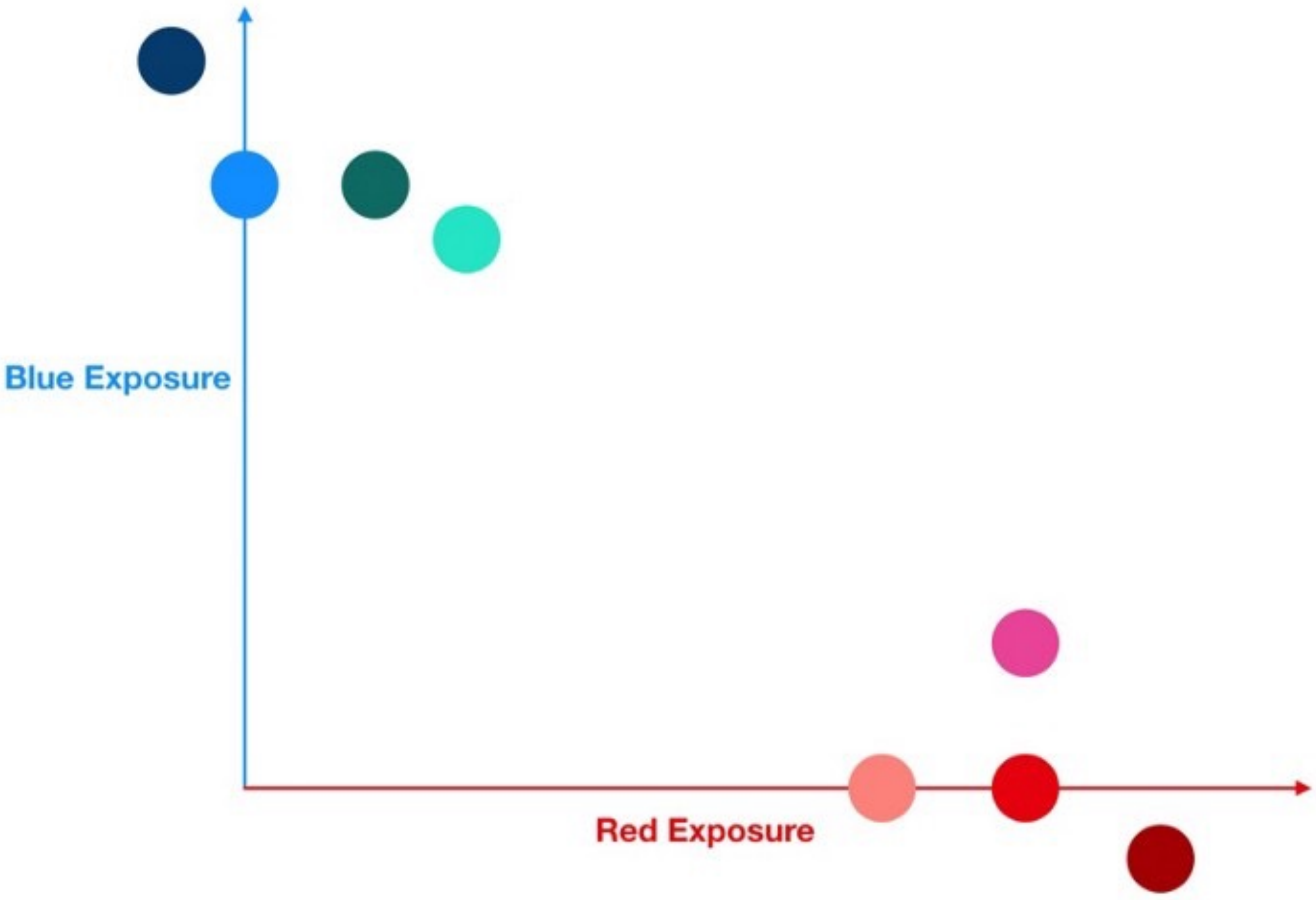
	Red	Maroon	Pink	Flamingo	Blue	Turquoise	Seaweed	Ocean
	1	0	0	0	0	0	0	0
	0	**1**	0	0	0	0	0	0
	0	0	**1**	0	0	0	0	0
	0	0	0	**1**	0	0	0	0
	0	0	0	0	**1**	0	0	0
	0	0	0	0	0	**1**	0	0
	0	0	0	0	0	0	**1**	0
	0	0	0	0	0	0	0	**1**

https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e

Dimensionality Reduction

Latent Features

	Red	Blue
Red	1.00	0
Maroon	1.20	-0.10
Pink	1.00	0.20
Flamingo	0.80	0
Blue	0	1.00
Turquoise	0.25	0.90
Seaweed	0.15	1.00
Ocean	-0.10	1.20



Dimensionality Reduction

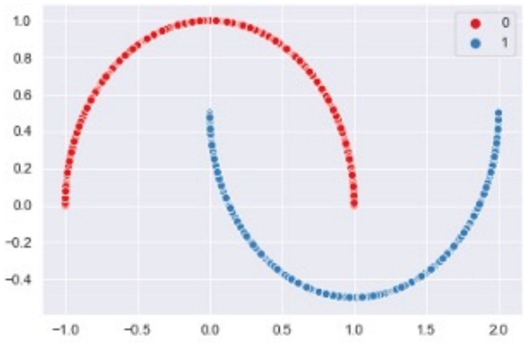
- Projection
  - Projecting every data point in high dimension onto a suitable lower-dimensional space, approximately preserving the distances
- Manifold Learning
  - Modelling the manifold on which the training instance lie
  - Assumption: most real-world high-dimensional datasets lie close to a much lower-dimensional manifold

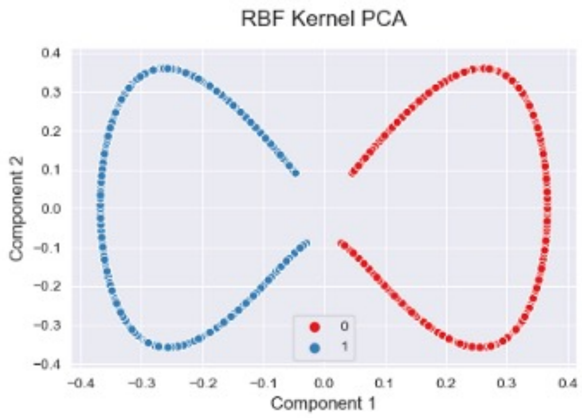
https://towardsdatascience.com/11-different-uses-of-dimensionality-reduction-4325d62b4fa6

Use Cases

- Visualization to observe relationships – 2D/3D
- Batch effect identification
- …

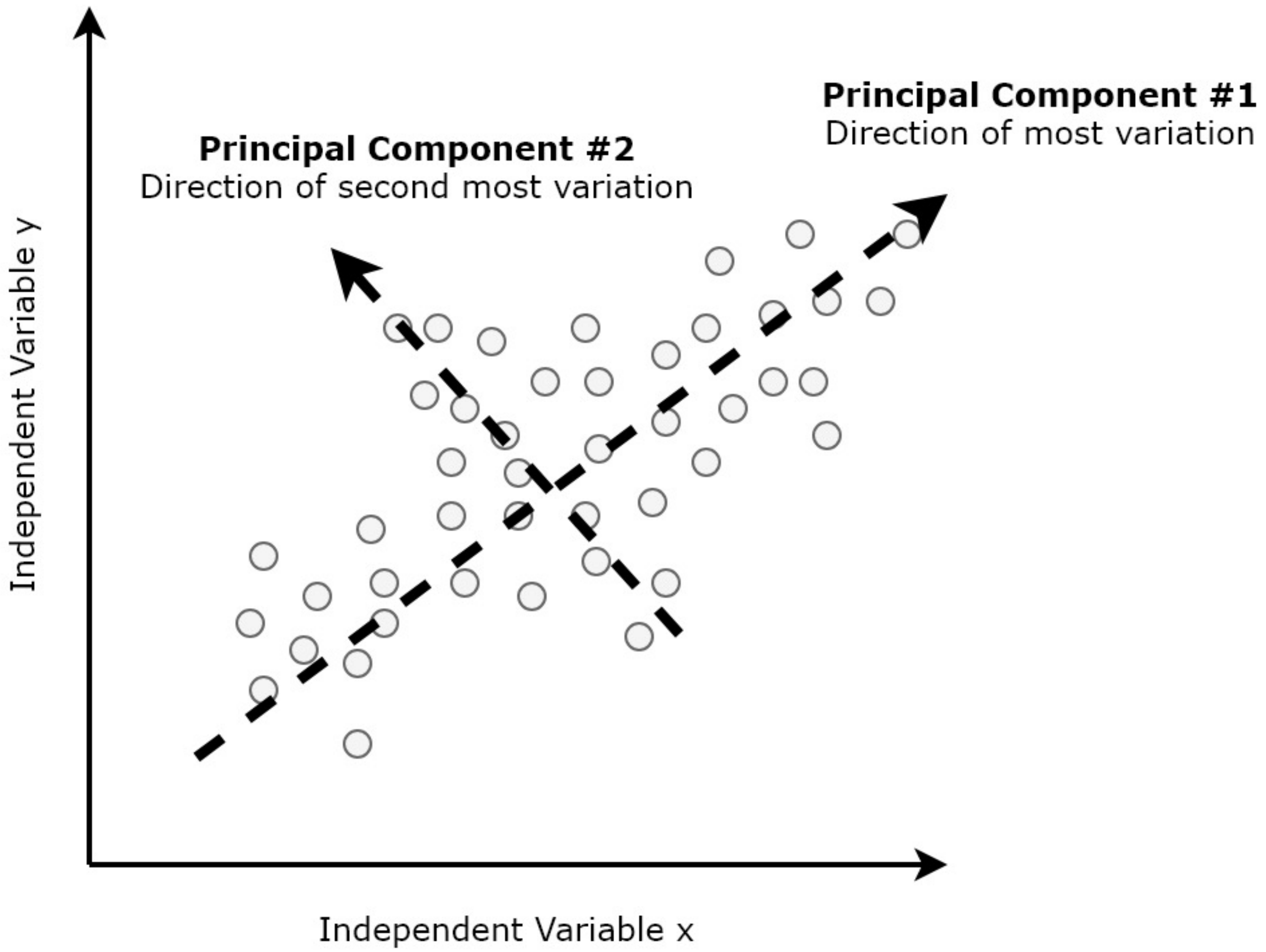
- transforms non-linear data into a linearly-separable form
- removes multicollinearity
- reduces the training time of models
- improves the accuracy of models





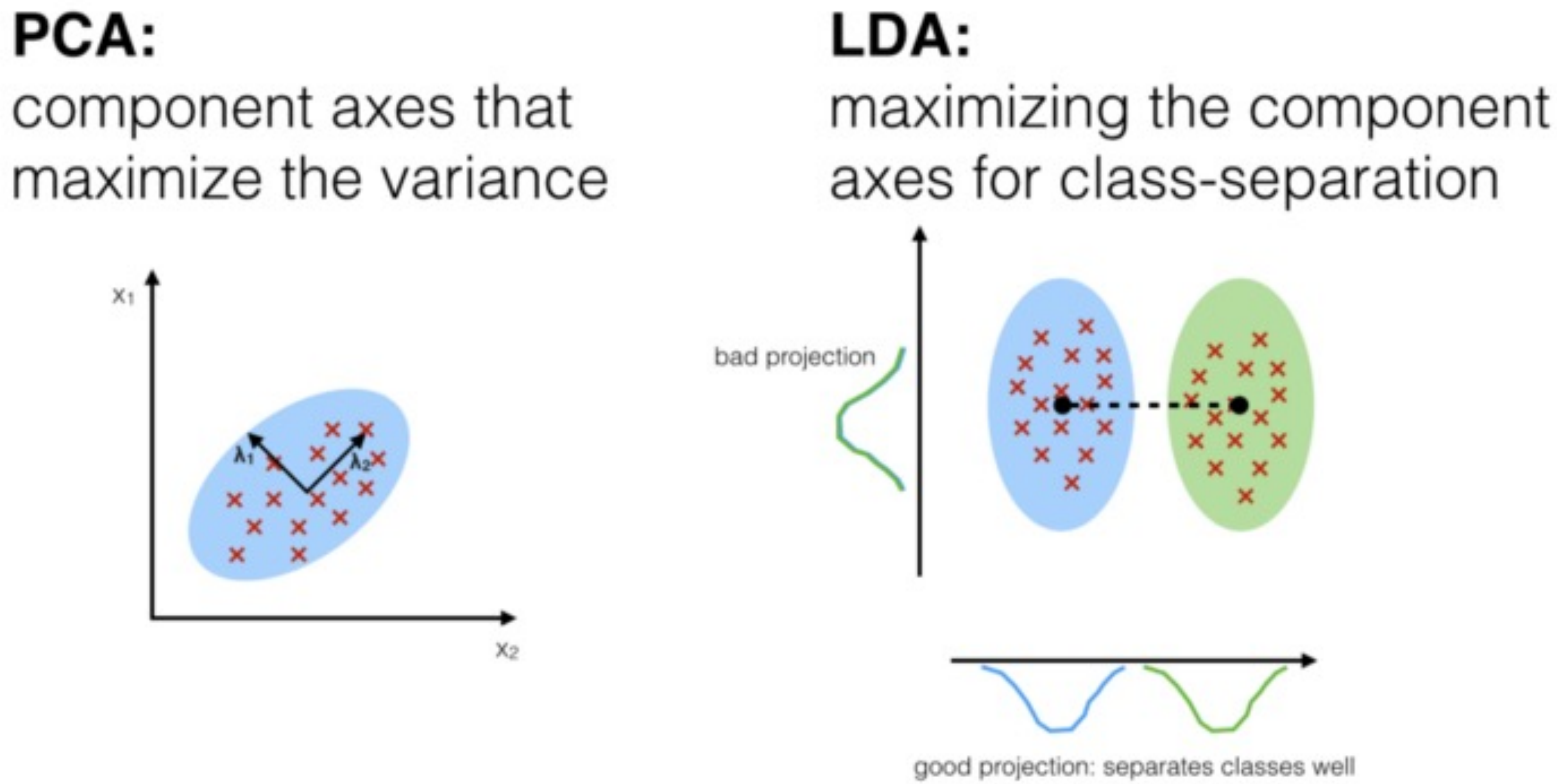
https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

Principal Component Analysis (PCA)



https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

Linear Discriminant Analysis (LDA)



https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

T-distributed stochastic neighbor embedding (t-SNE)

t-SNE reduces dimensionality while trying to keep similar instances close and dissimilar instances apart

van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE" (PDF). Journal of Machine Learning Research. 9: 2579–2605.

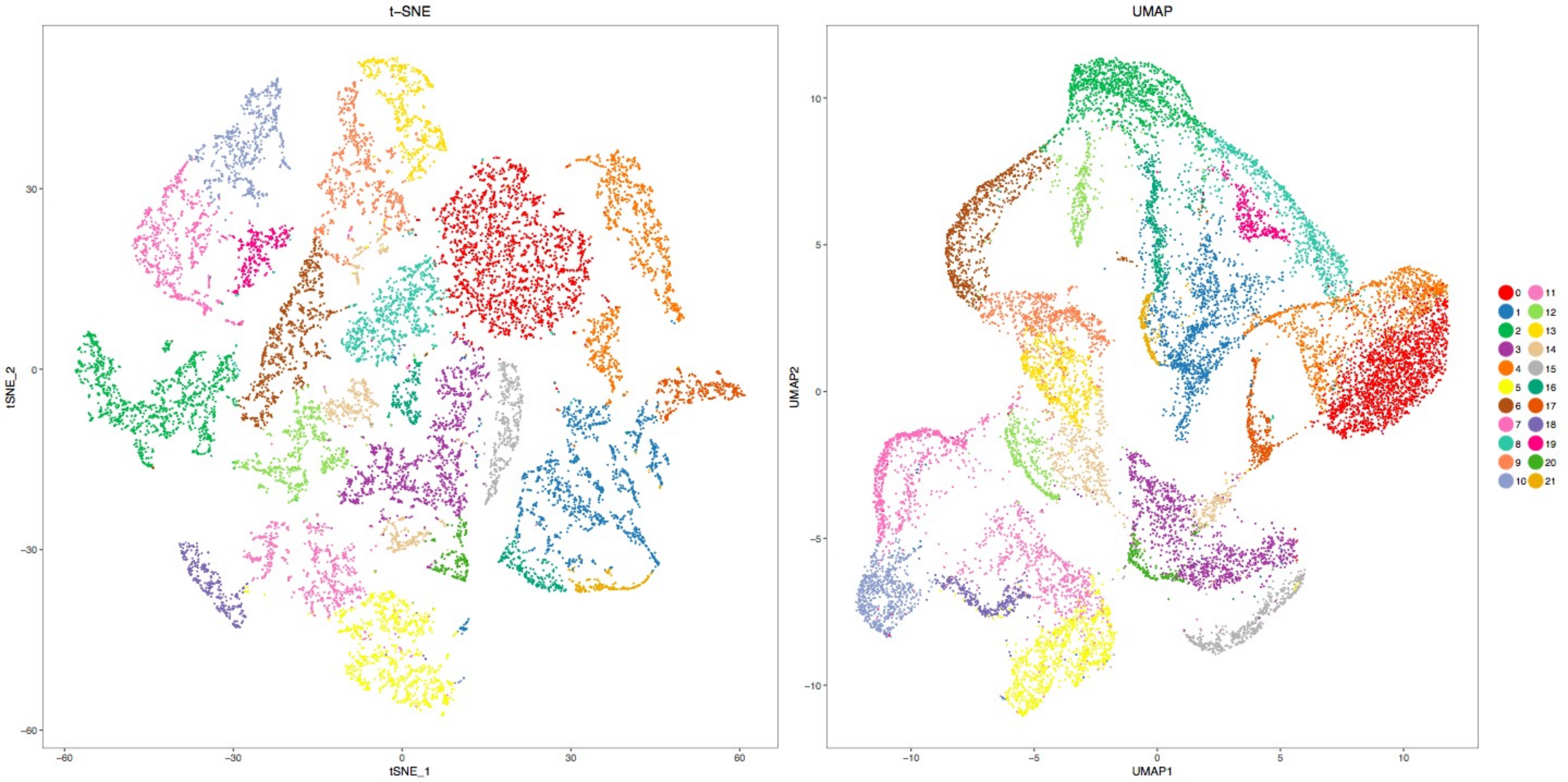
https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

Uniform Manifold Approximation and Projection (UMAP)

McInnes, L., & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints.

- Similar to tSNE
- a number of advantages over tSNE
  - increased speed
  - scalability
  - better preservation of the data's global structure

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



t–SNE
UMAP
tSNE_1
tSNE_2
UMAP1
UMAP2
−60
−30
0
30
60
−10
−5
5
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

More Techniques

- Multidimensional Scaling (MDS)
- Independent Component Analysis (ICA)
- Non-negative Matrix Factorization (NMF)
- …