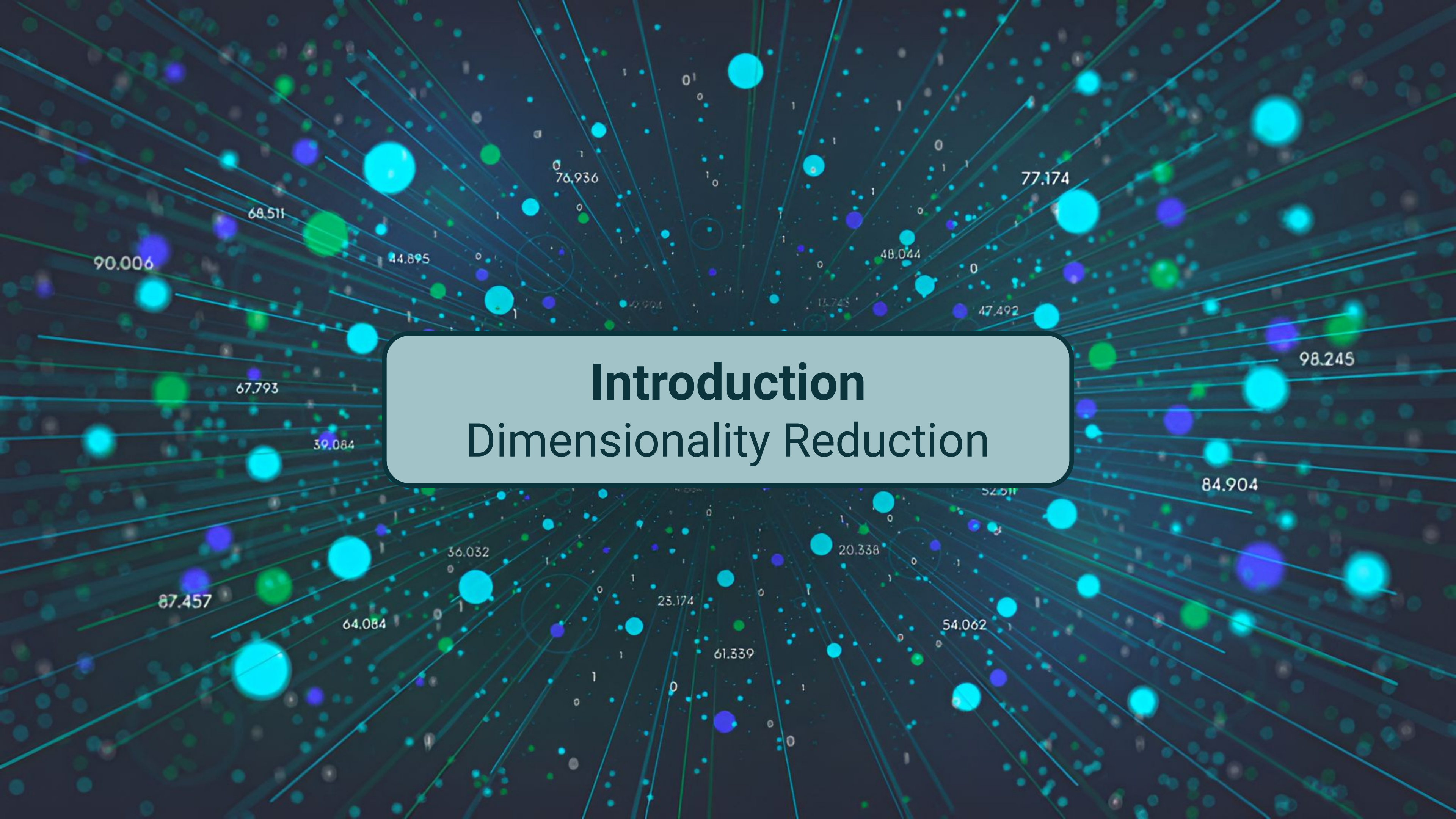
Simplifying Complexity

Extracting key patterns in immunological data with PCA

Team Sydney Brenner
Erica Anton, Annelisa Faché, Ziqing Yu, Alex Norman
November 14, 2024



Introduction

Dimensionality Reduction

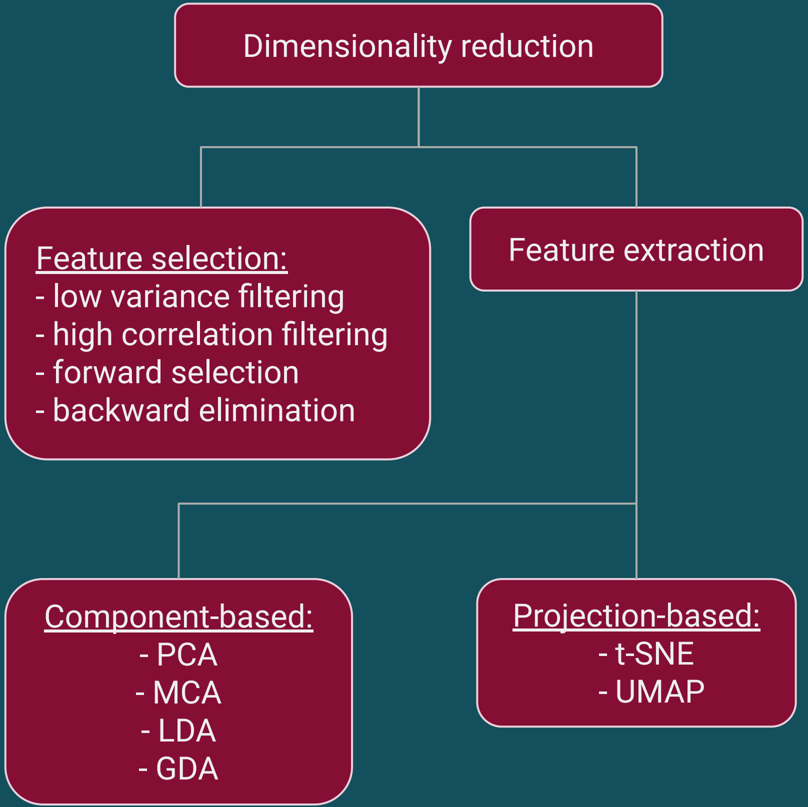
What is Dimensionality Reduction?

- **Preprocessing step** applied in high dimensional data analysis:
  - Extracts transformed features from the raw data
  - Removes noise and redundancy
  - Simplifies data processing

- Utilizes **“feature engineering”**:
  - Classified as suitable, unnecessary, or repeated
  - **Feature selection** = identify essential features from the input dataset
  - **Feature extraction** = create new features from the existing features in the input dataset

Types of Dimensionality Reduction

1. Principal component analysis (PCA)
   a. Multiple correspondence analysis (MCA)
2. Linear discriminant analysis (LDA)
3. Generalized discriminant analysis (GDA)
4. T-distributed stochastic neighbor embedding (t-SNE)
5. Uniform Manifold Approximation and Projection (UMAP)
6. Low variance or high correlation filtering
7. Forward or backward feature selection and elimination



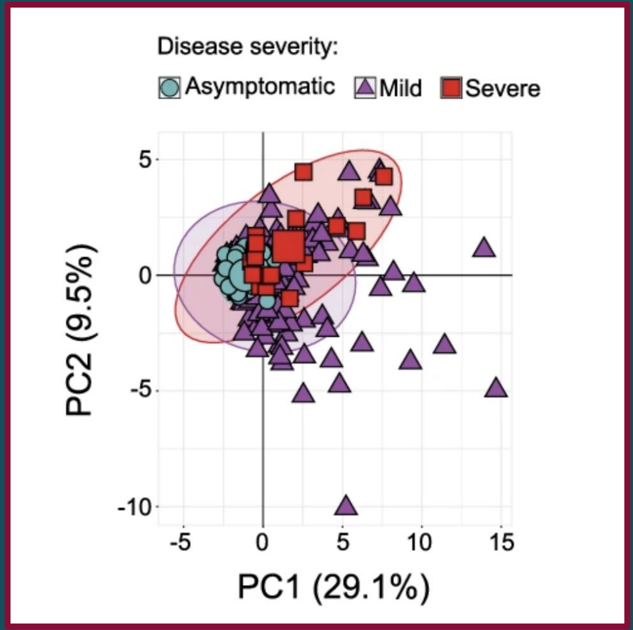
What is Principal Component Analysis (PCA)?

Statistical method reduces data to essential features or “principal components”

- **Principal components** = linear combinations of variables with maximum variance

Given a dataset of **p** numerical variables for **n** individuals → n x p matrix, **X**

Goal: obtain a linear combination of matrix X columns with maximum variance



Tomic, 2022

New coordinate system visualizing maximum variation

What is Multiple Correspondence Analysis (MCA)?

Analysis of **relationship patterns** for **category-based dependent variables**

- Generalization of PCA where variables are analyzed categorically instead of quantitatively

		Expert 1							Expert 2										Expert 3					
Wine	Oak Type	fruity		woody			coffee		red fruit		roasted		vanillin			woody			fruity		butter		woody	
W1	1	1	0	0	0	1	0	1	1	0	0	1	0	0	1	0	1		0	1	0	1	0	1
W2	2	0	1	0	1	0	1	0	0	1	1	0	0	1	0	1	0		0	1	1	0	1	0
W3	2	0	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0		0	1	1	0	1	0
W4	2	0	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0		1	0	1	0	1	0
W5	1	1	0	0	0	1	0	1	1	0	0	1	0	0	1	0	1		1	0	0	1	0	1
W6	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0	0	1		1	0	0	1	0	1
W?	?	0	1	0	1	0	.5	.5	1	0	1	0	0	1	0	.5	.5		1	0	.5	.5	0	1

Abdi, 2007

Nominal variables require conversion to **binary representation**

Ex. Male vs. female represented as 10 and 01 respectively

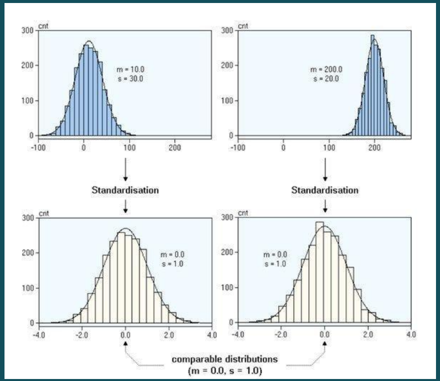
Quantitative variables can be **converted to categorical variables**

Ex. Ranking of 0-10 separated into <5, =5, or >5 and represented as 100, 010, or 001

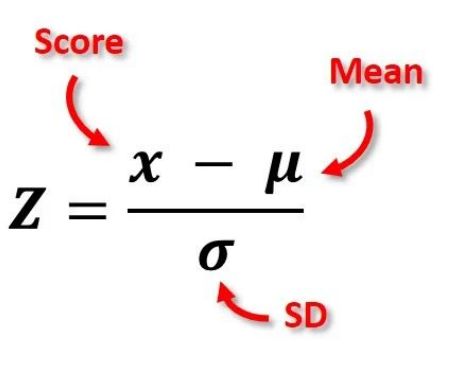
How PCA works

Data Standardized

- If data has different units or scales, must first be standardized
- Brings data points to a scale which can all be compared
- Standardization found by finding the z-score of each data point
- Makes the data set mean = 0 and standard deviation = 1



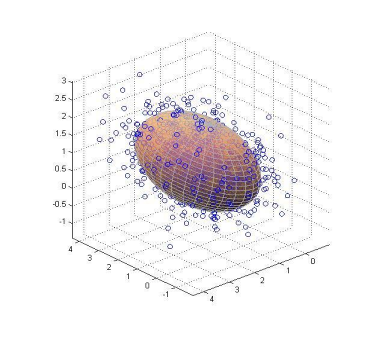
https://www.simplypsychology.org



https://medium.com/@vinodkumargr

Compute the Covariance Matrix

- Covariance matrix must be found
- A matrix composed of x-variance, y-variance, z-variance, and covariance
- Shows the distribution of magnitude + direction data (eigen values/vectors)



https://stackoverflow.com

$$\Sigma = \begin{bmatrix} \sigma_{x_R}^2 & \sigma_{x_R}\sigma_{x_G} & \sigma_{x_R}\sigma_{x_B} \\ \sigma_{x_R}\sigma_{x_G} & \sigma_{x_G}^2 & \sigma_{x_G}\sigma_{x_B} \\ \sigma_{x_R}\sigma_{x_B} & \sigma_{x_G}\sigma_{x_B} & \sigma_{x_B}^2 \end{bmatrix}$$

https://stackoverflow.com

$$\sigma^2 = \frac{\sum(xi - \bar{x})^2}{N}$$

https://www.indeed.com

Population covariance:

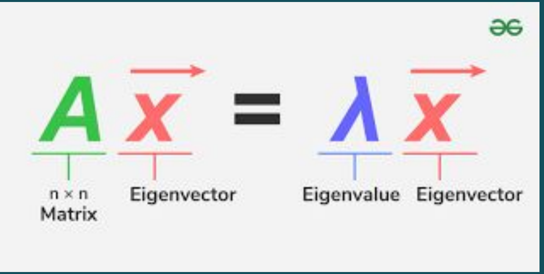
$$\mathrm{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

https://community.microstrategy.com

Calculate the Eigenvalues and Eigenvectors

If a matrix can be multiplied by a vector, v, and the product = λv where λ is a constant, then v is an eigenvector of that matrix and λ is an eigenvalue

Eigenvectors and Eigenvalues of the covariance matrix are found



https://www.geeksforgeeks.org/eigen-values/

$$A = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}$$

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

$$\det(A - \lambda I) = 0$$

$$\det\left(\begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix} 1-\lambda & 4 \\ 3 & 2-\lambda \end{bmatrix}\right) = 0$$

$$(1-\lambda)(2-\lambda) - 12 = 0$$

$$\lambda^2 - 3\lambda - 10 = 0$$

$$(\lambda - 5)(\lambda + 2) = 0$$

$$\lambda = 5, -2$$

https://towardsdatascience.com

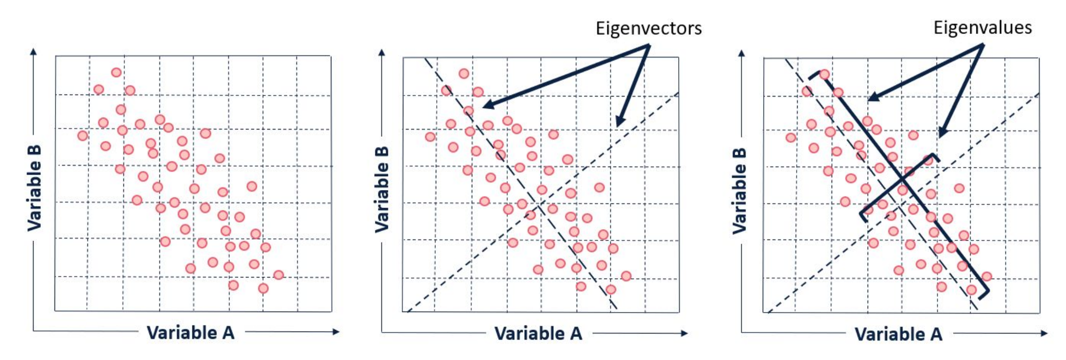
Sort the Eigenvectors by Eigenvalues

Eigen values are sorted in descending order

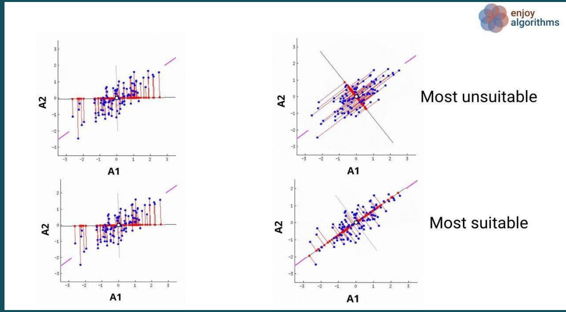
Largest eigen values provide the most information on the variance of the data

Eigen vectors sorted into corresponding order of their respective eigenvalues

The eigenvector with largest eigenvalue = first principal component, the vector with the second largest value = second principal component etc.



https://community.alteryx.com



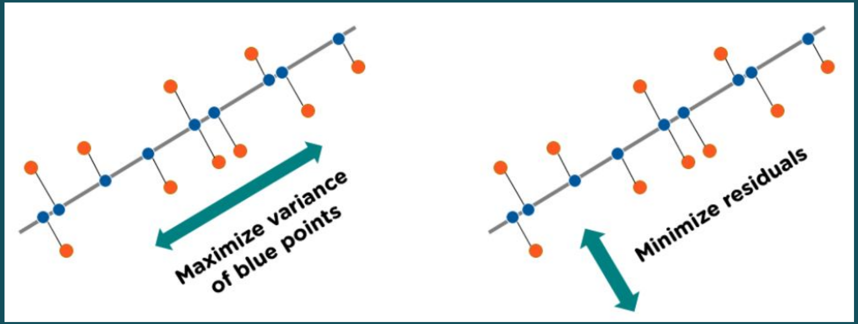
https://medium.com

Forming a Feature Vector

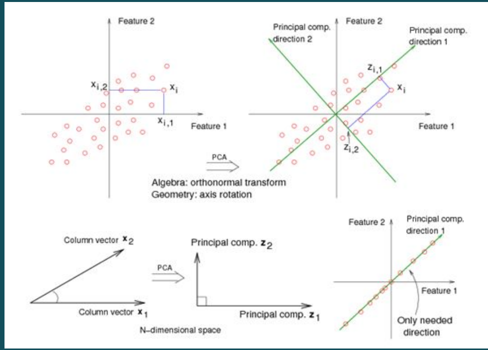
Select percentage of larger eigenvalues desired to be kept in the data

Cut-off point selected based on the amount of variance desired in the data set

Eigenvectors make up the columns of a new matrix called a feature vector



https://www.bogotobogo.com

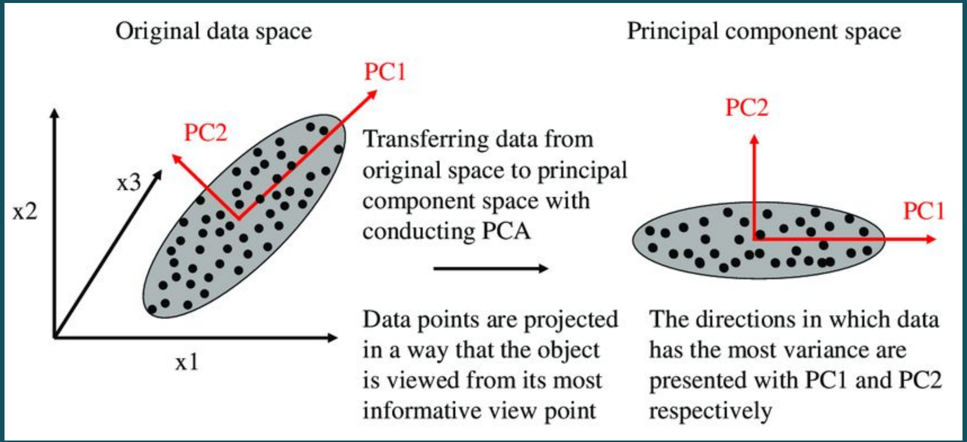


https://www.bogotobogo.com

Plotting the Data

- Data is plotted from the original axes to the axes of the principal components

- the transpose of the original data set is multiplied by the transpose of the feature vector.

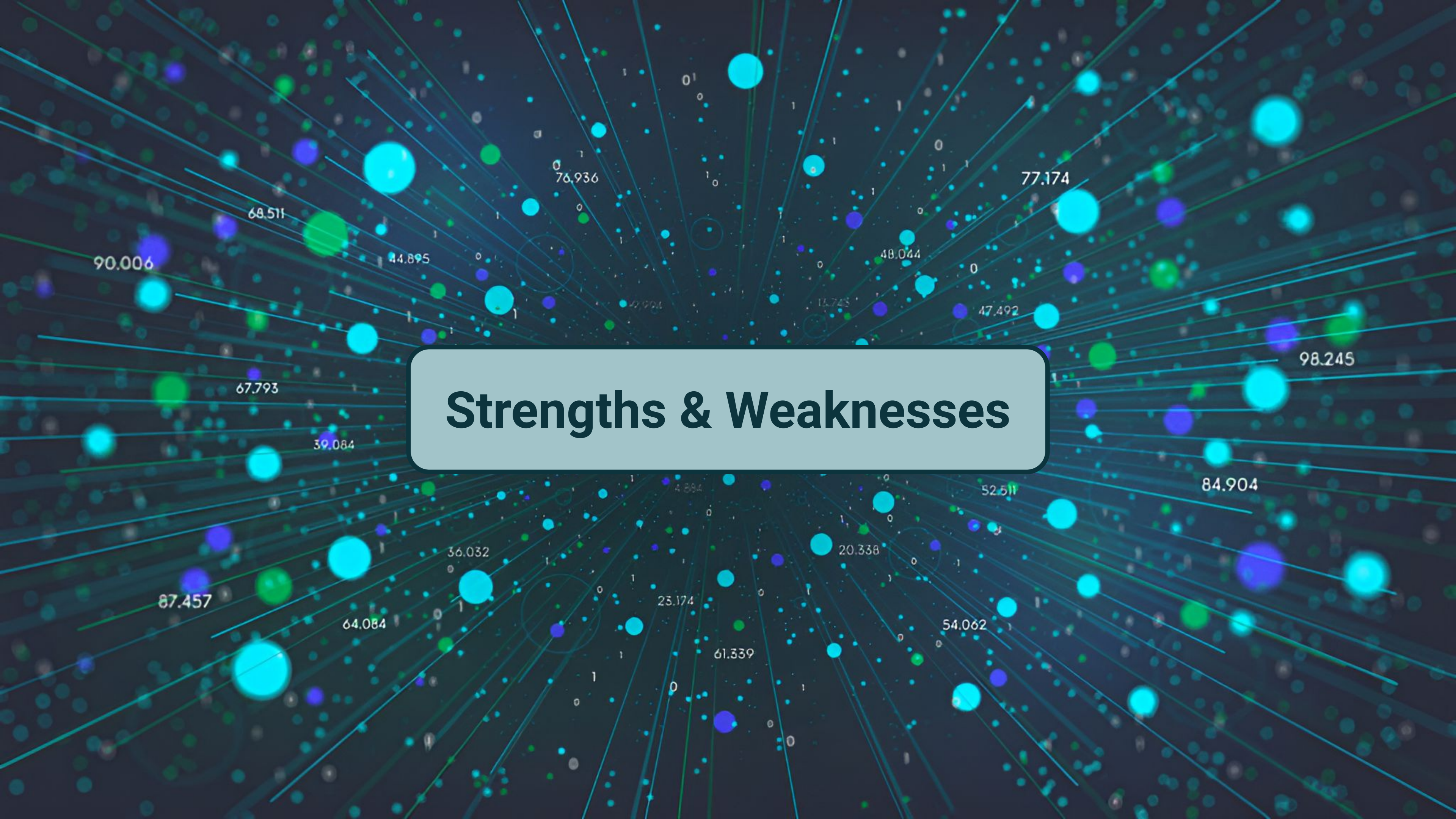


https://www.analyticsvidhya.com

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

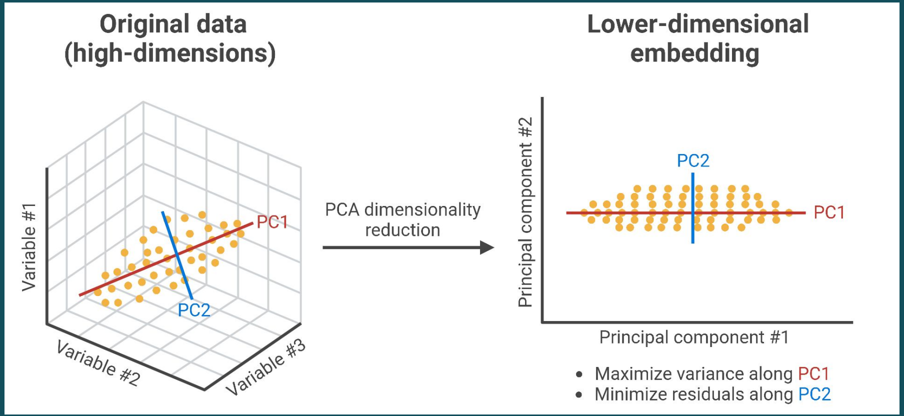
https://builtin.com

Strengths & Weaknesses



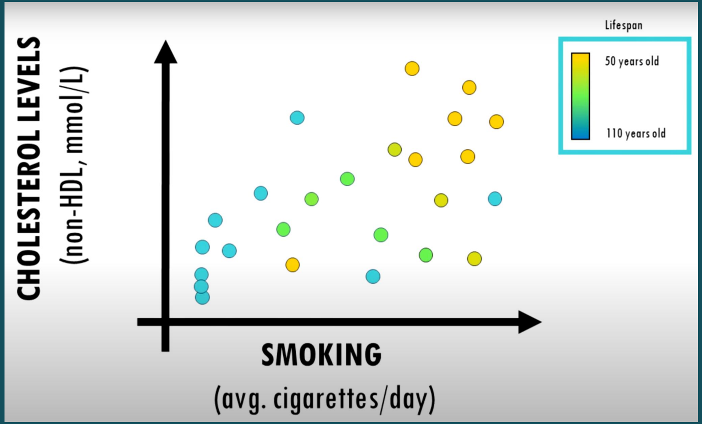
PCA Strengths

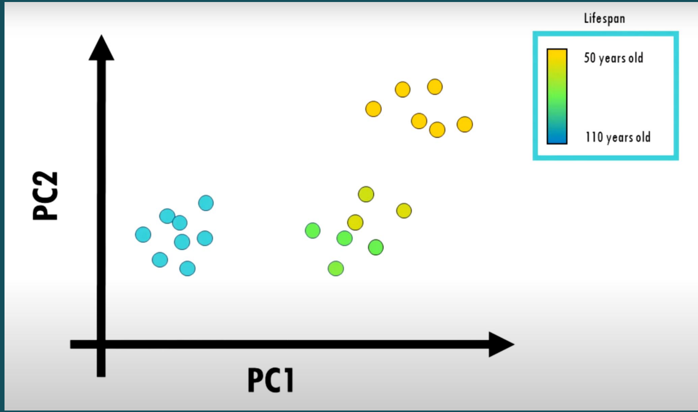
- **Improved visualization** when using two dimensional data
- **Reduced multicollinearity and noise**
  - Uncorrelated, orthogonal axes
  - Eliminates correlated features
- **Simplified and faster training**
  - Fewer dimensions simplifies model calculations, leading to faster results



PCA Example

		1	2	3	4	5	6	7		200
	Lifespan	Height	Weight	Average blood pressure	Average heart rate	BMI	Cholesterol levels	Average cigarettes/day	...	Sugar levels
Person 1	82	150	80	140/90	63	36	5.0	0		99
Person 2	73	174	90	90/60	100	32	4.1	0		95
Person 3	95	183	109	120/80	95	29	3.6	1		92
Person 4	92	186	95	123/75	84	28	4.8	5		89
Person 5	87	170	67	95/60	76	23	2.7	10		100
Person 6	65	180	82	92/60	78	25	3.7	10		112
Person 7	93	165	71	124/80	81	26	3.8	0		113
Person 8	80	172	70	97/70	90	24	3.4	0		100
...										
Person 20	72	190	75	90/60	78	21	4.2	0		82



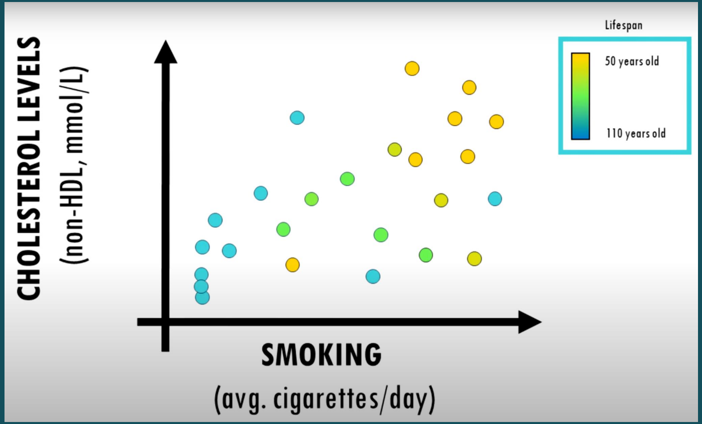


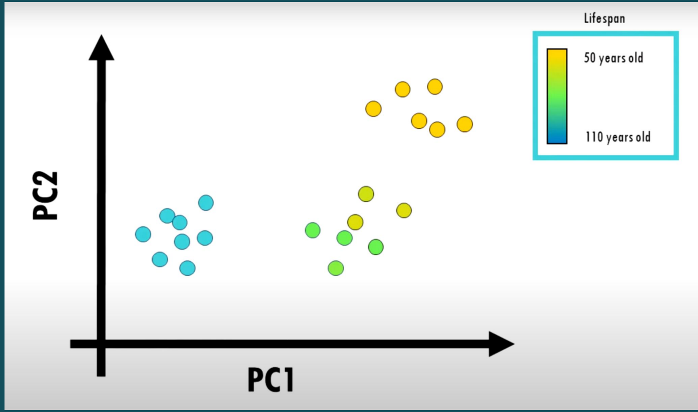
PCA Weaknesses

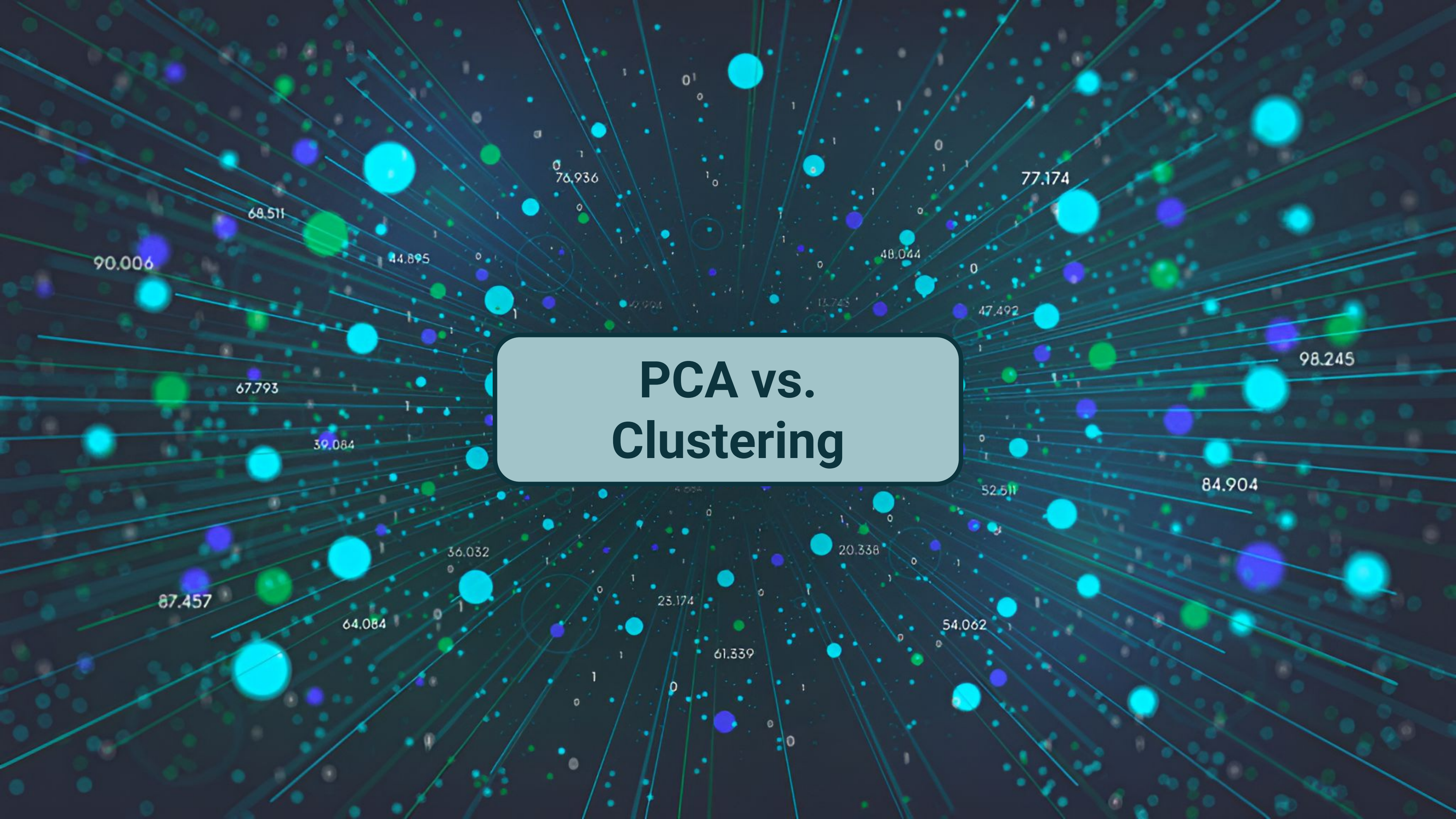
- **Loss of information**
  - Dimensionality reduction can lead to details being distorted or lost
- **Impact of outliers**
  - Outliers may distort principal components, impacting accuracy of results
- **Principal components are not interpretable**
  - Linear combinations of initial variables
  - Don’t themselves have real-world meaning

PCA Example

	Lifespan	1 Height	2 Weight	3 Average blood pressure	4 Average heart rate	5 BMI	6 Cholesterol levels	7 Average cigarettes/day	...	200 Sugar levels
Person 1	82	150	80	140/90	63	36	5.0	0		99
Person 2	73	174	90	90/60	100	32	4.1	0		95
Person 3	95	183	109	120/80	95	29	3.6	1		92
Person 4	92	186	95	123/75	84	28	4.8	5		89
Person 5	87	170	67	95/60	76	23	2.7	10		100
Person 6	65	180	82	92/60	78	25	3.7	10		112
Person 7	93	165	71	124/80	81	26	3.8	0		113
Person 8	80	172	70	97/70	90	24	3.4	0		100
...										
Person 20	72	190	75	90/60	78	21	4.2	0		82

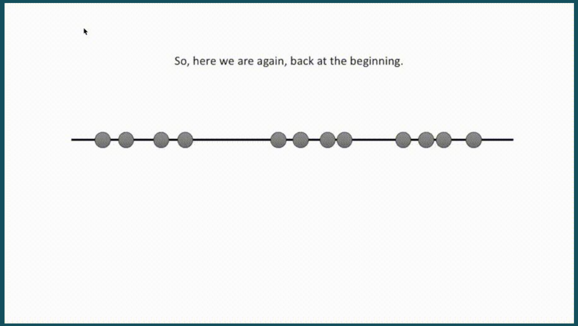






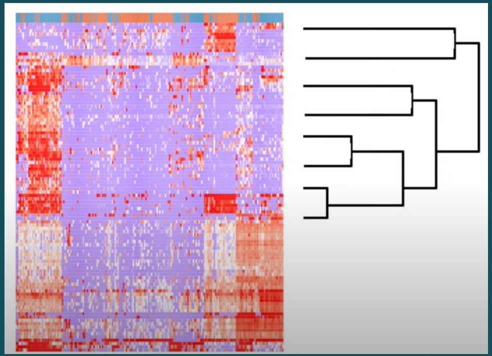
PCA vs. Clustering

CLUSTERING



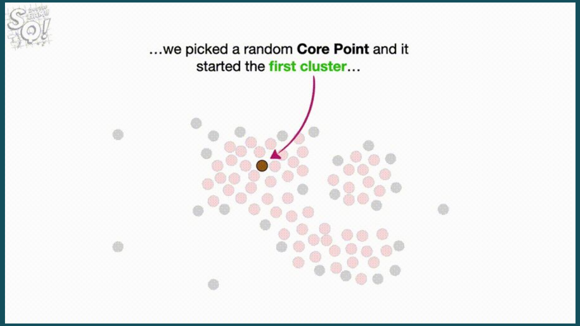
K-means

- High efficiency, easy to implement and flexible
- Use centroid to represent the entire group
- Limitations with cluster shapes, better for linear patterns
- Sensitive to noise



Hierarchical clustering

- Easy to implement, could show hierarchical relationships between clusters (taxonomy)
- Computationally expensive for large datasets
- Sensitive to noise



DBSCAN

- Density based clustering algorithm, can discover arbitrarily clusters
- Robust to outlier detection (noise)

PCA vs. Clustering

	PCA	Clustering
Purpose	**Reduces dimension,** keeping key information.	Forms **clusters** based on data similarity, without reducing dimensions
Type of learning	Unsupervised learning	Unsupervised learning
Output	Projected data onto principal components	Groups of similar data points
Method	Identifies **main factors** explaining data variance.	Finds **groups based on similarity or distance**
Strengths	Reduces variables, helps reveal important patterns.	Works well on large datasets; useful for exploratory analysis.
Weaknesses	Needs a correlation matrix; may lose information. Principal components may not interpretable	Groups may not be meaningful

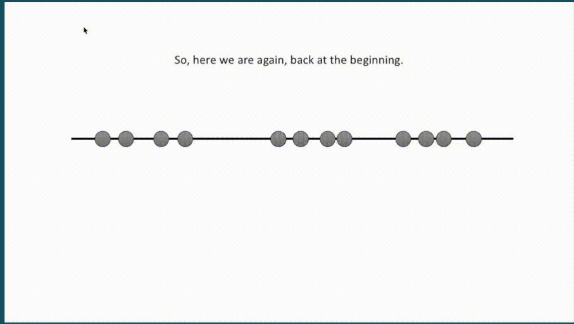
PCA vs. Clustering

	PCA	Clustering
Noise and Outliers	**Sensitive** to noise and outliers.	DBSCAN is robust to noise and outliers, K-means and hierarchical are more sensitive.
Computational Efficiency	Generally **efficient** but can be costly when have large datasets and high dimension	Hierarchical clustering and DBSCAN are **computationally expensive for large datasets**
Link	PCA could provide **pre-processing method for clustering**. For instance, combining PCA with K-means clustering can improve clustering performance by reducing dimensionality, removing noise, and preserving data structure	

References

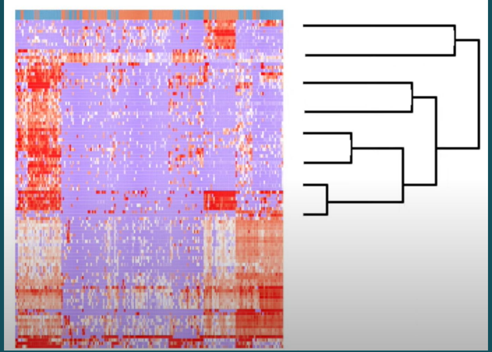
1. https://www.sciencedirect.com/science/article/pii/S1877050920300879
2. https://ieeexplore.ieee.org/abstract/document/9036908
3. https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/#h-4-brief-nbsp-summary-of-when-to-use-each-dimensionality-reduction-techniques
4. https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202
5. https://www.nature.com/articles/s41467-022-28898-1
6. https://personal.utdallas.edu/~herve/Abdi-MCA2007-pretty.pdf
7. https://www.researchgate.net/figure/PCA-biplot-graph-representing-genotypes-in-two-main-principal-components-for-traits_fig3_359097887
8. https://blog.dailydoseofds.com/p/the-advantages-and-disadvantages
9. https://elitedatascience.com/dimensionality-reduction-algorithms
10. https://builtin.com/data-science/step-step-explanation-principal-component-analysis
11. https://www.biorender.com/template/principal-component-analysis-pca-transformation
12. https://www.youtube.com/watch?v=5vgP05YpKdE
13. https://datarundown.com/k-means-clustering-pros-cons/#:~:text=Pros%20of%20K-Means%20clustering%20include%20its%20ease%20of,the%20risk%20of%20getting%20stuck%20in%20local%20minima.
14. https://datarundown.com/hierarchical-clustering/
15. https://datarundown.com/dbscan-clustering/
16. https://www.researchgate.net/figure/Advantages-Disadvantages-and-Applications-of-DBSCAN_tbl2_271520302
17. https://datarundown.com/cluster-vs-factor-analysis/
18. K-means Clustering via Principal Component Analysis, Chris Ding, Xiaofeng He, 2004.
19. https://www.youtube.com/watch?v=5vgP05YpKdE

CLUSTERING



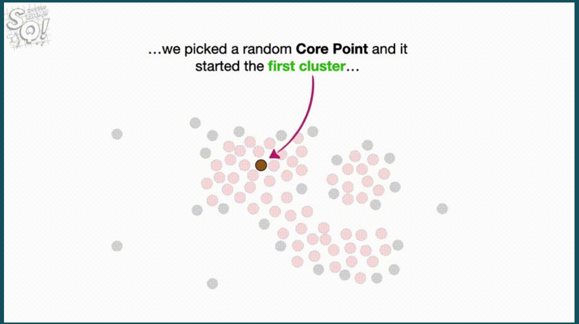
K-means

- High efficiency, easy to implement and flexible
- Use centroid to represent the entire group (carefully choose the initial centroids)
- Need to define the number of clusters
- Limitations with cluster shapes, better for linear patterns
- Sensitive to noise



Hierarchical clustering

- Easy to implement, could show hierarchical relationships between clusters (taxonomy)
- No need to define the number of clusters
- Better for non-linear patterns
- Computationally expensive for large datasets
- Sensitive to noise



DBSCAN

- Density based clustering algorithm, can discover arbitrarily clusters
- Robust to outlier detection (noise)
- No need to define the number of clusters
- Computationally expensive for large datasets
- Sensitive to distance threshold and minimum number of points