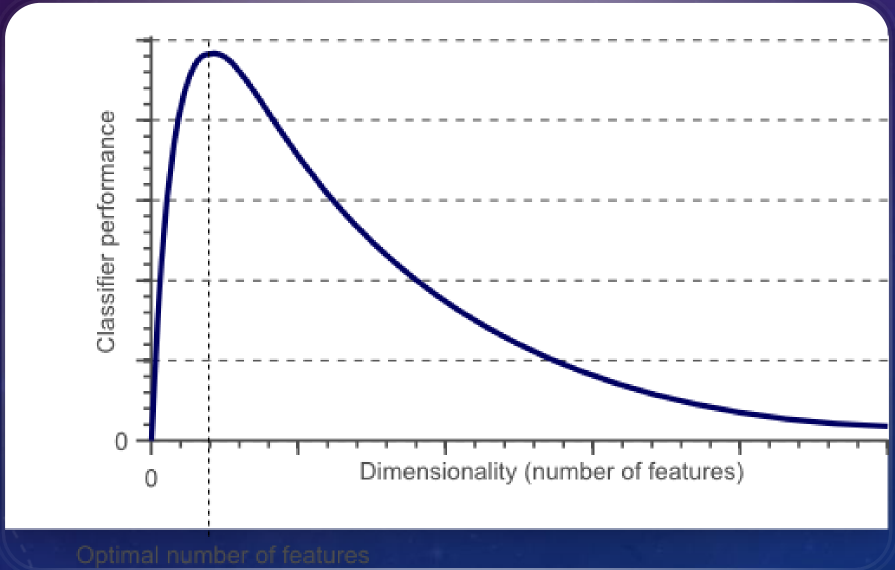
DIMENSIONALITY REDUCTION

CHAPTER 8 – HANDS-ON MACHINE LEARNING WITH SCIKIT-LEARN, KERAS, AND TENSORFLOW

DIMENSIONALITY REDUCTION – PROS AND CONS

(1 Without PCA
Accuracy : 0.7466666666666667
Homogeneity score : 0.7514854021988338
Completeness score : 0.7649861514489815
V-Measure : 0.7581756800057784

(2 With PCA
Accuracy : 0.74
Homogeneity score : 0.736419288125285
Completeness score : 0.7474865805095325
V-Measure : 0.7419116631817838



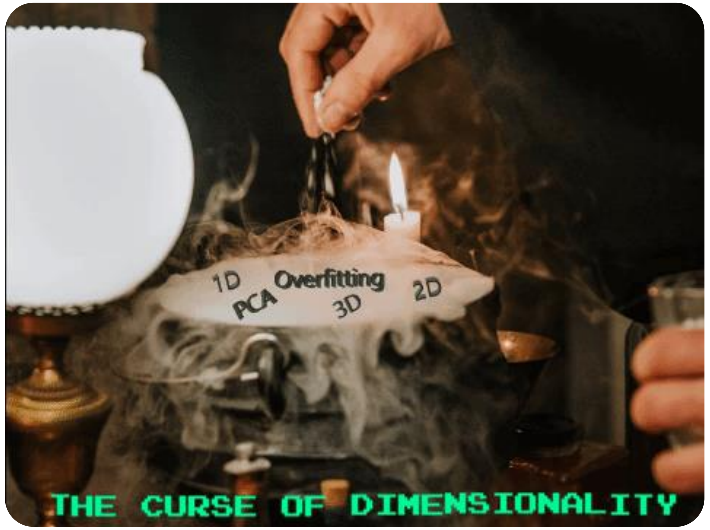
- Pros
  - Removes correlated features
  - Improves model efficiency
  - Reduces overfitting
  - Improves visualization
- Cons
  - PCA is a linear algorithm and does not work well for polynomial or other complex functions
  - Can lead to inefficiencies after reduction if we don’t choose the right number of dimensions to eliminate
  - Less interpretability
  - Preserves global shapes rather than local shapes

(Chaitanya Narava, 2020, “A Complete Guide on Dimensionality Reduction”, [A Complete Guide On Dimensionality Reduction | by Chaitanyanarava | Analytics Vidhya | Medium])

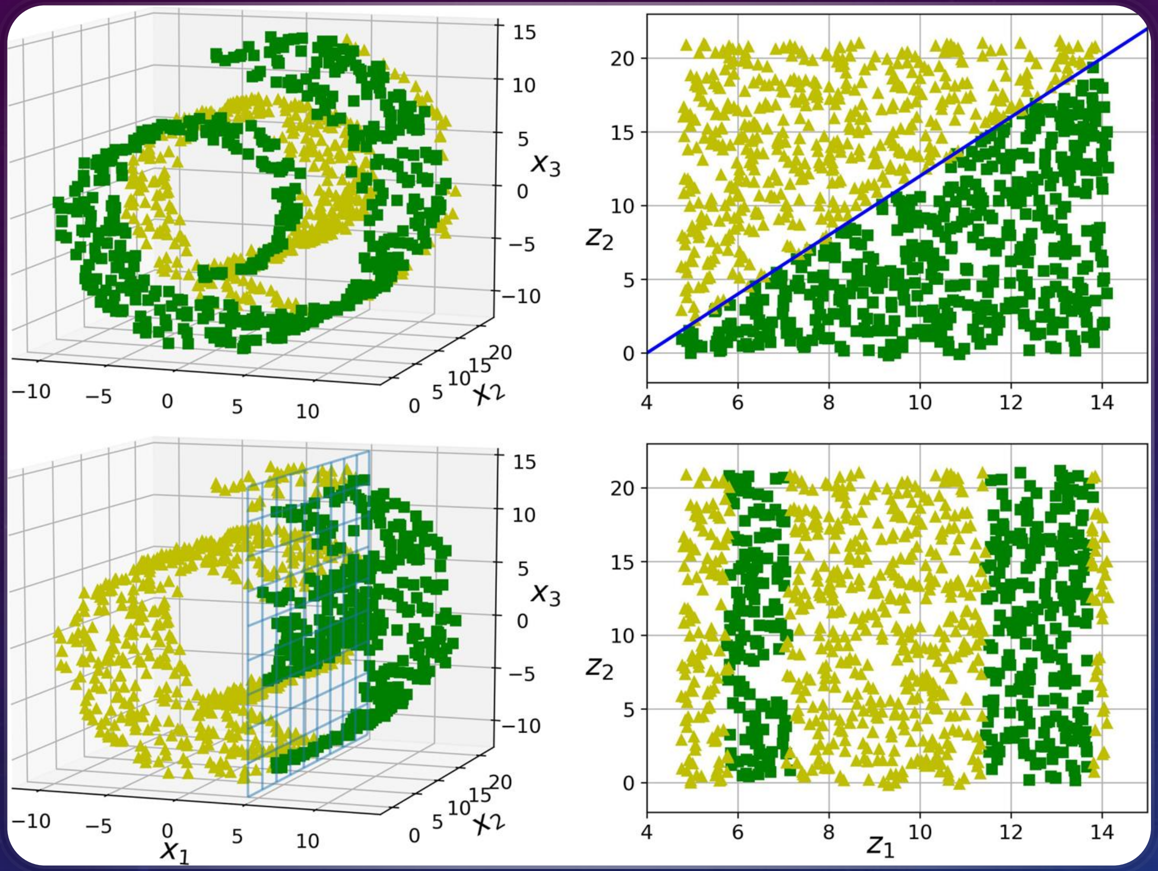
CURSE OF DIMENSIONALITY

- As it relates to Copy-Move Forgery project
    - - (Anuja Dixit and R. K. Gupta, 2016, “Copy-Move Image Forgery Detection a Review”, areviewIJIGSP.pdf)

- Has anyone else run into the curse of dimensionality in other types of projects?



Author	Method	Advantage	Disadvantage
Popescu 2004[5]	PCA	Small variations due to noise and lossy compression can be detected accurately.	For low quality image, as size of block decreases so does efficiency.
Ting 2009[6]	SVD	Less computation complexity and robust against post processing operations.	Cannot deal with JPEG compression.
Bashar 2010[8]	KPCA	Forgeries with additive noise and JPEG compression can be detected.	Average accuracy is less than other methods which are based on wavelet.
Zimba 2011[9]	PCA-EVD	False matches are less and duplication with varying degree of rotations can be detected.	Unable to detect forgeries with scaling and heavy JPEG compression.

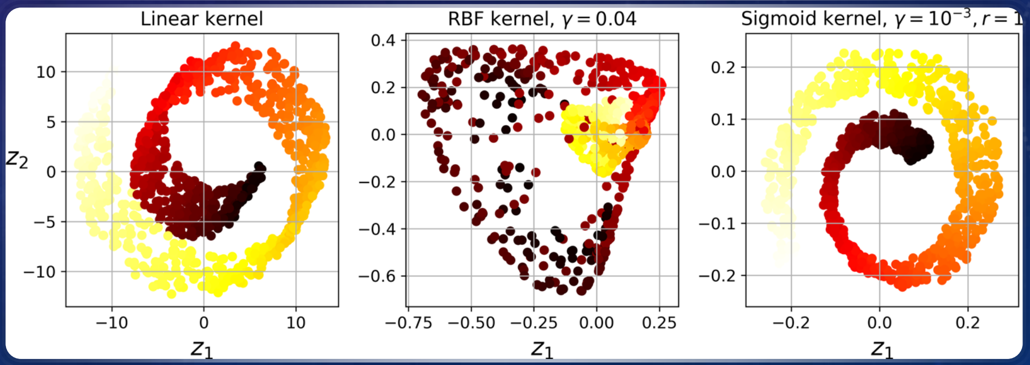


MAIN APPROACHES FOR DIMENSIONALITY REDUCTION

- Projection – this approach works well when many features are almost constant with many highly correlated
- Manifold Learning – the Swiss Roll dataset
  - Assumptions
    - Most real-world, high-dimensional datasets lie close to a much lower-dimensional manifold
    - Task will be simpler if expressed in this lower-dimensional space

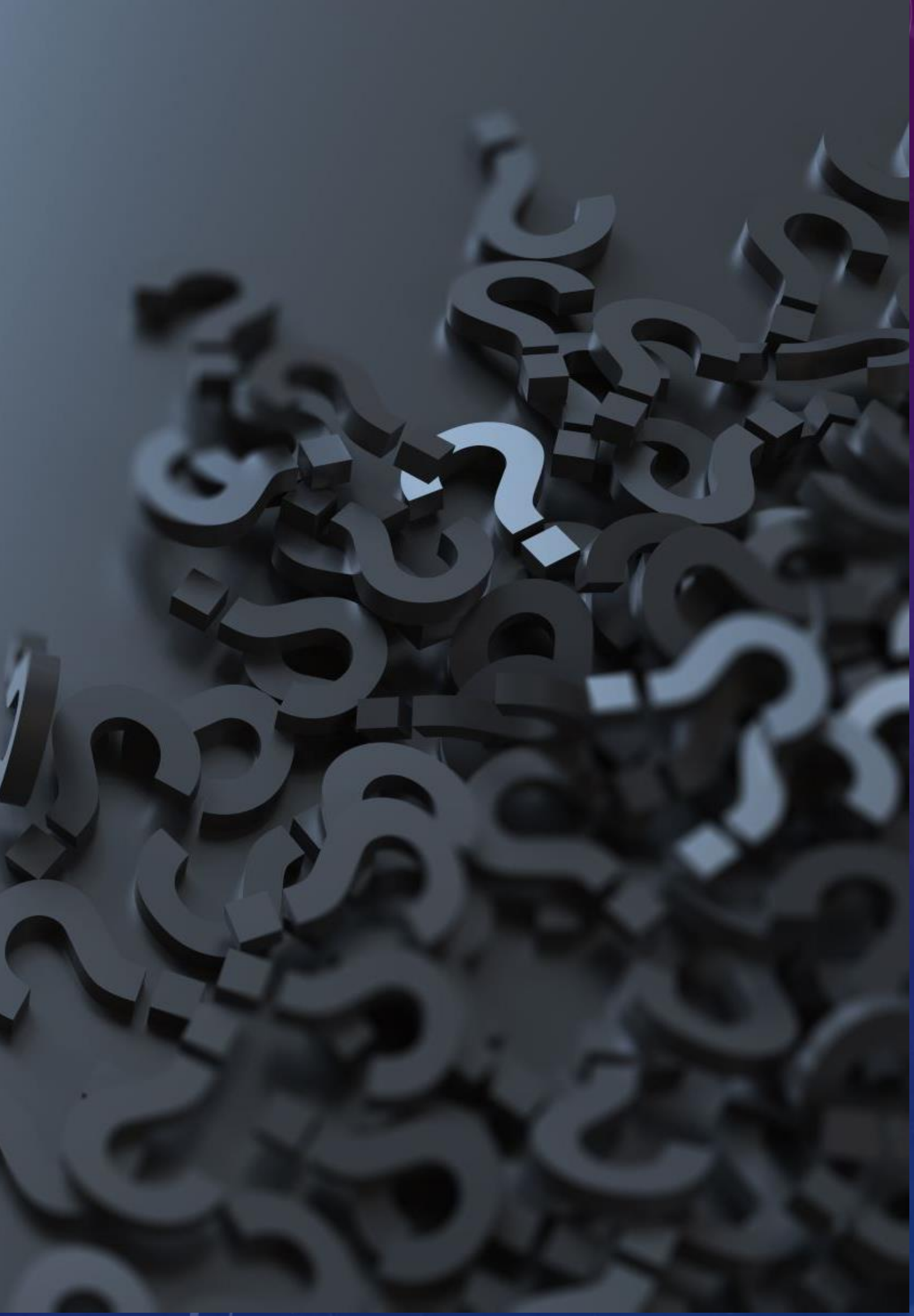
PRINCIPAL COMPONENT ANALYSIS HYPERPARAMETERS

- Randomized PCA – Faster way to tune your model than using "full" when *d* is much smaller than *n*
- Incremental PCA – does not require the full training set to fit in memory for the algorithm to run
- Kernel PCA (kPCA) – allows you to perform nonlinear projections



DIFFERENT TYPES OF DIMENSIONALITY REDUCTION TECHNIQUES

- Principal Component Analysis (or PCA) – the most popular
  - Hyperparameters allow you to alter the PCA to:
    - Randomized PCA
    - Incremental PCA
    - Kernel PCA (or kPCA)
- Locally Linear Embedding (or LLE) – an unsupervised Manifold Learning method that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional data (NON-LINEAR?)
- Random Projections – projects the data to a lower-dimensional space using a random linear projection
- Multidimensional Scaling (or MDS) – a linear method that transforms the given matrix into a low-dimensional matrix based on the distance each element
- Isomap – Manifold Learning method that is non-linear and is better than linear methods when dealing with all types of real image and motion tracking
- T-Distributed Stochastic Neighbor Embedding (t-SNE) – non-linear and more robust towards outlier
- Linear Discriminant Analysis (LDA) – linear technique similar to ANOVA which builds the feature combinations based on differences rather than similarities



QUESTIONS