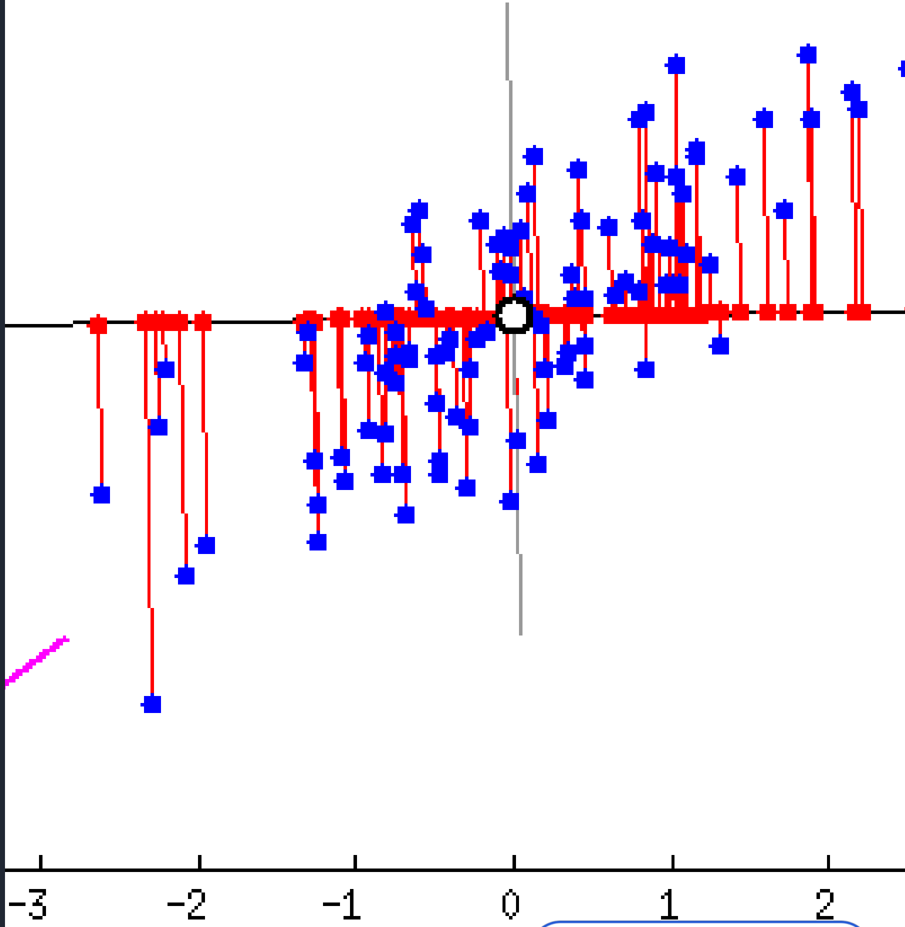
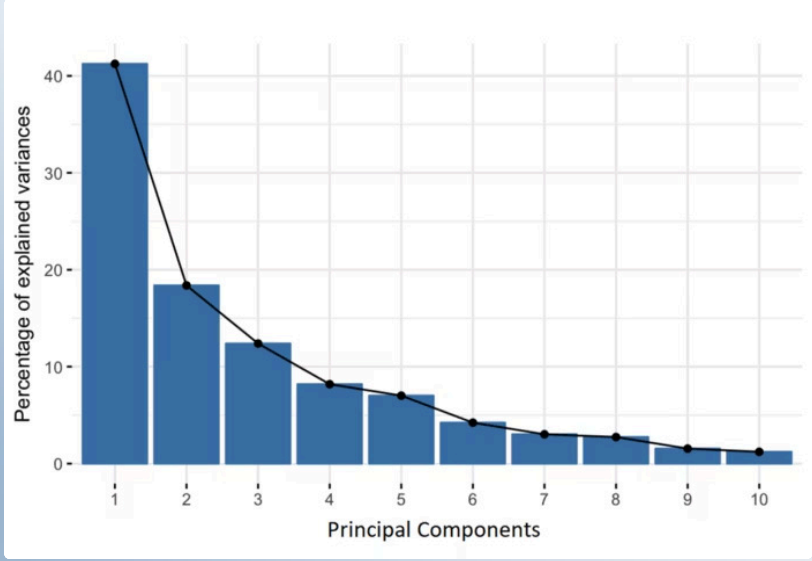
# NDL Lab 4: Dimensionality Reduction Using PCA & t-SNE

11/2/25



# Meeting Name: "Reduction"



# Why Dimensionality Reduction Matters

Real-world datasets often have hundreds or thousands of features. Working with all of them creates challenges we need to solve strategically.

## Curse of Dimensionality

High-dimensional spaces become sparse, making patterns harder to find and models prone to overfitting.

## Visualization

Humans can't perceive beyond 3D. Reduction lets us see meaningful patterns and relationships.

## Noise Reduction

Focuses on signal by removing irrelevant noise and redundant information in the data.

# Palmer Penguins Dataset

## Dataset Overview

The Palmer Penguins dataset contains measurements from three penguin species: Adelie, Chinstrap, and Gentoo. We have physical measurements like bill length, bill depth, flipper length, and body mass.

Our goal: use unsupervised learning to cluster penguins by species and sex, discovering natural groupings without predefined labels.

# Unsupervised Learning: Finding Patterns Without Labels

Unlike supervised learning, unsupervised learning explores data structure independently. It discovers hidden patterns, groupings, and relationships the data naturally contains.

01

## Collect Features

Gather data without knowing the true groups in advance.

02

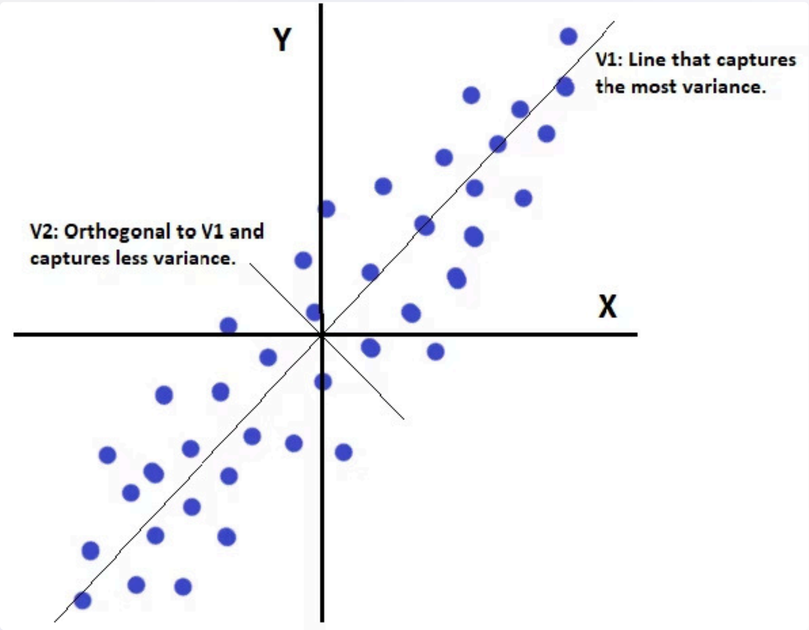
## Reduce Dimensions

Compress high-dimensional features into visualizable form while preserving meaningful structure.

03

## Identify Clusters

Examine the reduced space to find natural groupings and validate data.


Y
V1: Line that captures the most variance.
V2: Orthogonal to V1 and captures less variance.
X

# Principal Component Analysis: The Core Idea

PCA is used on a data set with many variables and reduces them to a smaller set of uncorrelated principal components which retain most of the original information

### Variance Capture

Principal components are ordered by how much variance they explain. The first component explains the most variance, the second component explains the second most variance, and so on.

### Mathematical Foundation

PCA uses eigenvectors and eigenvalues of the covariance matrix. Eigenvectors define the directions of axes; eigenvalues are how much variance they explain

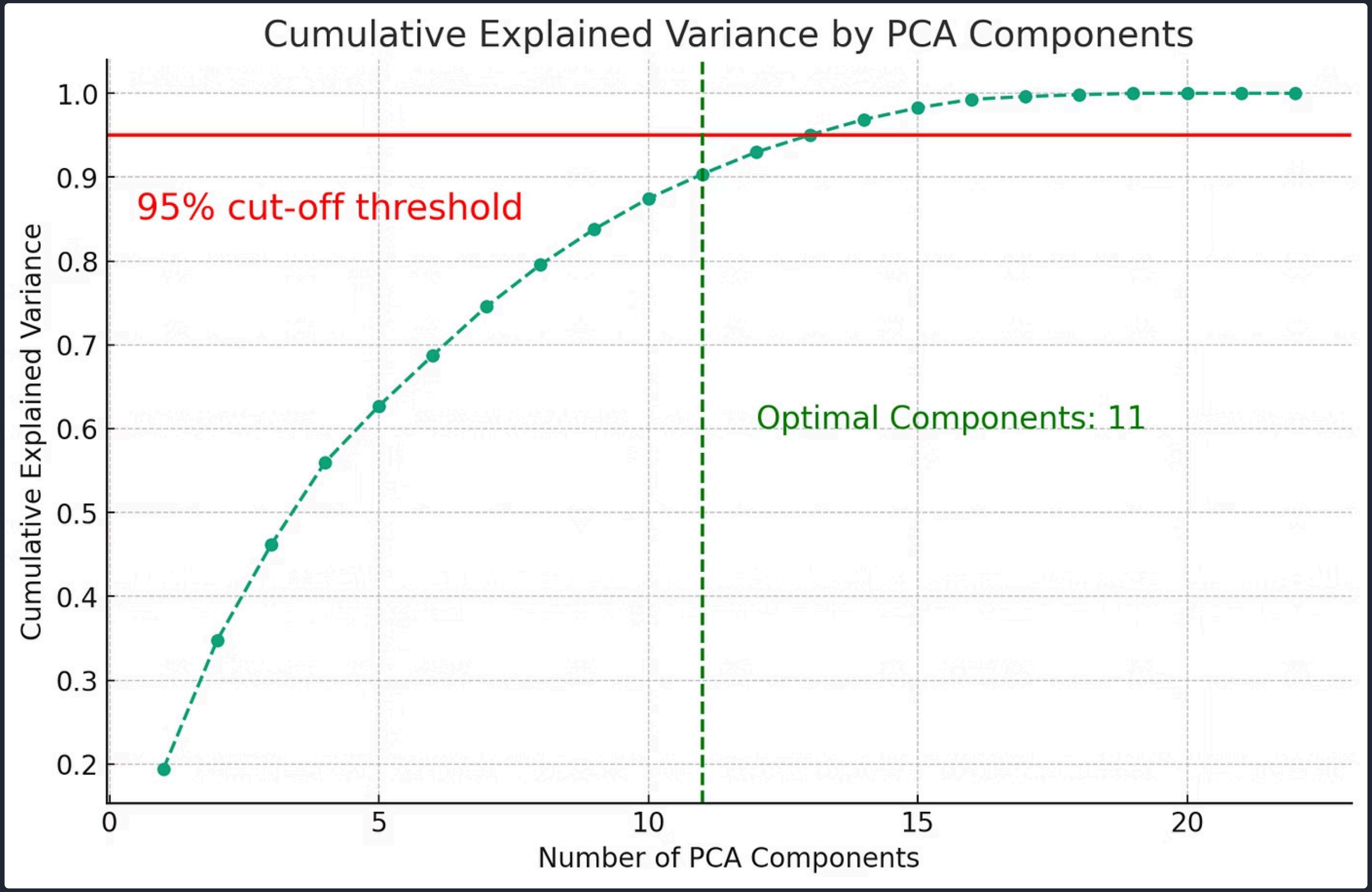
### Linear Transformation

PCA projects your original features onto these principal components

$$FinalDataSet = FeatureVector^T * StandardizedDataSet^T$$

# Choosing # of Components: Explained Variance Plot

Cumulative explained variance shows how much total variance you retain as you include more components. Look for the "elbow"—the point where adding more components yields diminishing returns.



## Reading the Elbow

If 3 components explain 90% of variance, and adding a 4th component only explains an extra 1%,  you've likely found a good balance between dimensionality reduction and information retention.

## Decision Rule

Choose the number of components where the curve flattens noticeably. This is where you stop gaining meaningful information per added component. We use 95% as a benchmark to signify where adding extra components only captures noise.

# t-SNE (t-distributed Stochastic Neighbor Embedding)

t-SNE takes a fundamentally different approach than PCA. Instead of maximizing the amount of variance explained, it focuses on preserving local structure (useful for visualizing clusters)

### Non-Linear Approach

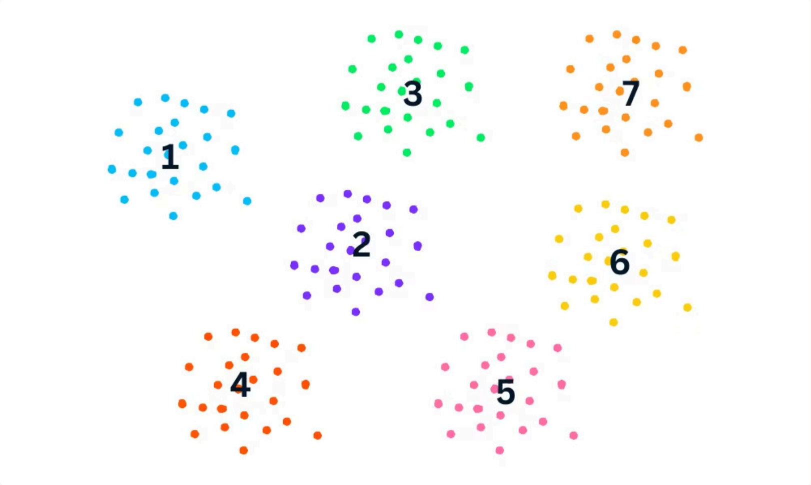
Unlike PCA's linear transformation, t-SNE is based on conditional probabilities. This calculates how likely it is that one point that is near another

### Local vs Global Structure

- Preserves local structure: similar data points remain close
- Does not preserve global structure: distorts global distances

### Trade-Offs

t-SNE excels at visualization but is computationally expensive and non-deterministic. It can be useful for clustering



# PCA vs. t-SNE: Side-by-Side Comparison

Each technique reveals different aspects of data. PCA retains global structure; t-SNE reveals visually distinct clusters.

## PCA Results

- Linear reduction preserves global structure.
- Species clusters visible but less separated.
- Useful for noise filtering and downstream tasks.

## t-SNE Results

- Non-linear reduction emphasizes local structure.
- Forms well-separated clusters.
- Excellent for visualization.

## When to Use Each

- PCA for larger datasets or for downstream tasks
- t-SNE when you need compelling visualizations to explore and communicate patterns.

# Google Colab

# Thanks NDL!

## Next Meeting: 11/9

Topic: Clustering (K-means & DBscan)