# Welcome to the CoGrammar Revision: Principal Component Analysis

The session will start shortly...

Questions? Drop them in the chat. We'll have dedicated moderators answering questions.

CoGrammar

# Data Science Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

# **Data Science Session Housekeeping** cont.

- For all **non-academic questions**, please submit a query:

  [www.hyperiondev.com/support](www.hyperiondev.com/support)

- Report a **safeguarding** incident:

  [www.hyperiondev.com/safeguardreporting](www.hyperiondev.com/safeguardreporting)

- We would love your **feedback** on lectures: **Feedback on Lectures**

CoGrammar

# Skills Bootcamp
# 8-Week Progression Overview

## Fulfil 4 Criteria to Graduation

### ✅ Criterion 1: Initial Requirements

Timeframe: First 2 Weeks
Guided Learning Hours (GLH): Minimum of 15 hours
Task Completion: First four tasks

**Due Date: 24 March 2024**

### ✅ Criterion 2: Mid-Course Progress

**60** Guided Learning Hours

Data Science - **13 tasks**
Software Engineering - **13 tasks**
Web Development - **13 tasks**

**Due Date: 28 April 2024**

CoGrammar

# Skills Bootcamp Progression Overview

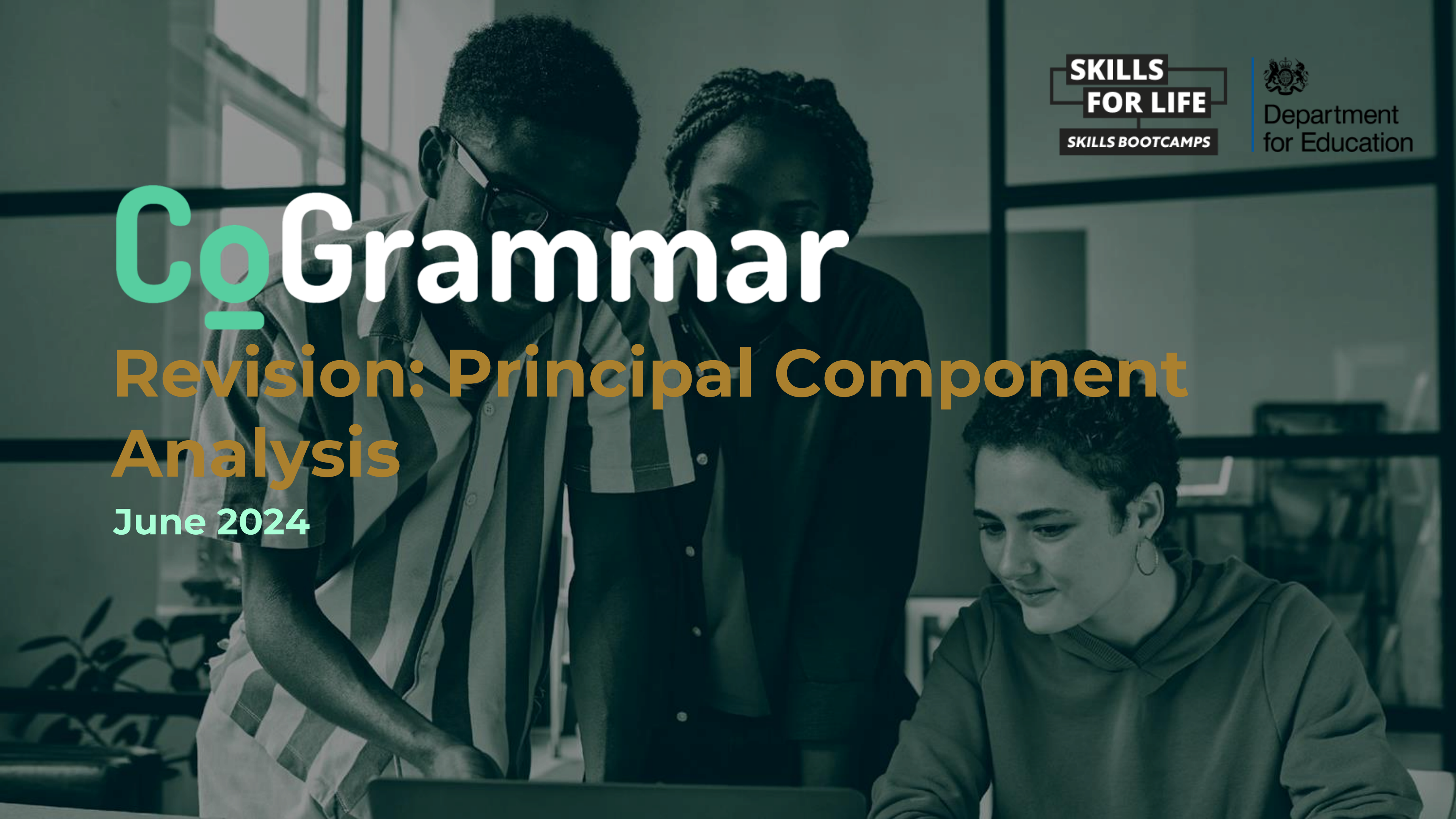
## ✅ Criterion 3: Course Progress

Completion: All mandatory tasks, including Build Your Brand and resubmissions by study period end
Interview Invitation: Within 4 weeks post-course
Guided Learning Hours: Minimum of 112 hours by support end date
(10.5 hours average, each week)

## ✅ Criterion 4: Demonstrating Employability

Final Job or Apprenticeship Outcome: Document within 12 weeks post-graduation
Relevance: Progression to employment or related opportunity

CoGrammar



SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar

# Revision: Principal Component Analysis

June 2024

# Learning objectives

- ❖ Understand the fundamental concepts underlying Principal Component Analysis
- ❖ Understand the issues we have with high dimensionality and how PCA allows us to work around these issues
- ❖ Grasp how we use PCA to rank observations
- ❖ Create and interpret a biplot

CoGrammar

# PCA Recap

CoGrammar

# Principal Component Analysis

- **Principal Component Analysis (PCA)** is a statistical technique used for dimensionality reduction, transforming data into a set of linearly uncorrelated variables called principal components.
- In practice, the number of variables in a machine learning task can be high.
- As the number of variables grows, the data becomes harder to work with. Relationships between variables become harder to see, training slows down, and the chance of overfitting increases. It is, therefore, useful to know a bit about how to reduce the number of variables while still retaining enough information about our dataset.

# Principal Components

- Each **Principal Component** is a weighted sum of the original features i.e., PCs are linear combinations of the original variables.

- Components are ordered by the amount of variance they explain, with the first component explaining the most.

- **Formula**: $PC_i = w_{i\,1}\,X_1 + w_{i\,2}\,X_2 + \ldots + w_{i\,p}\,X_p$

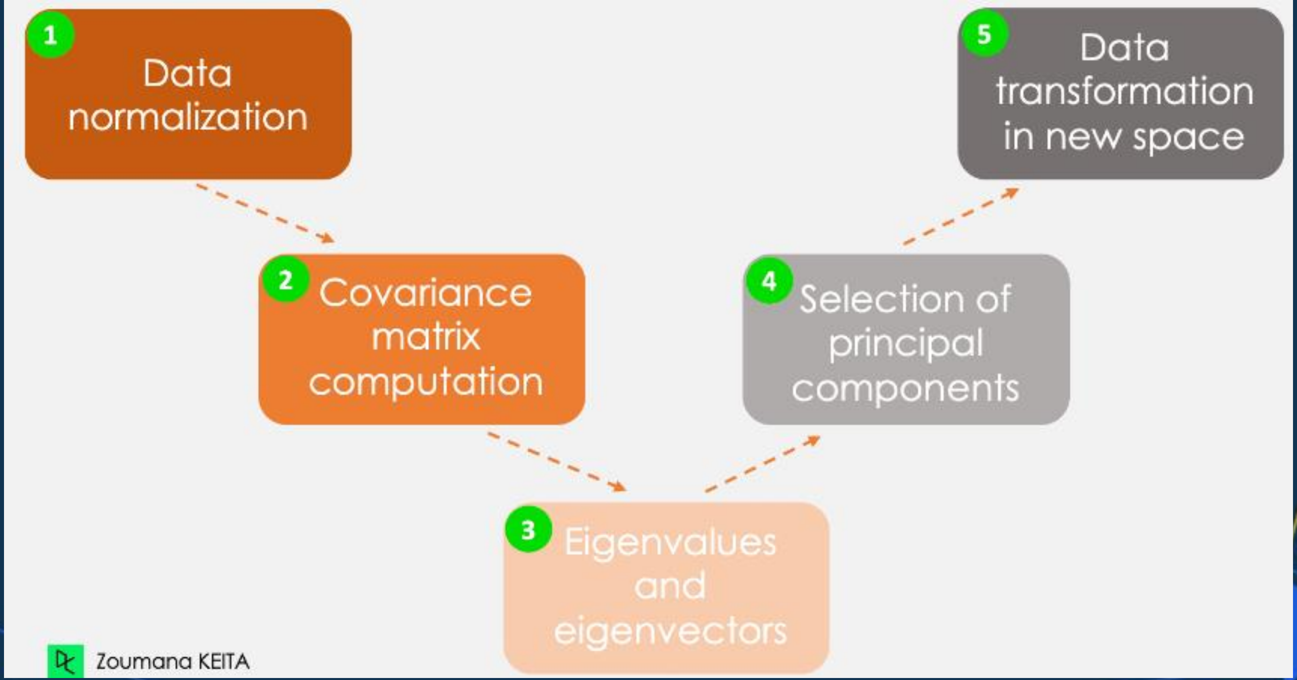
  where $w_{i\;j}$ are the weights (loadings) of the *j*-th feature in the *i*-th principal component. The weights indicate the contribution of each original feature to the principal component.

  **Example**: for a dataset with three features $X_1$, $X_2$, $X_3$, the first 2 PCs are:

  Principal Component 1: $PC_1 = w_{1\,1}\,X_1 + w_{1\,2}\,X_2 + w_{1\,3}\,X_3$

  Principal Component 2: $PC_2 = w_{2\,1}\,X_1 + w_{2\,2}\,X_2 + w_{2\,3}\,X_3$

# Steps to perform PCA


1 Data normalization
2 Covariance matrix computation
3 Eigenvalues and eigenvectors
4 Selection of principal components
5 Data transformation in new space
Zoumana KEITA

Source: [https://www.datacamp.com/tutorial/pca-analysis-r](https://www.datacamp.com/tutorial/pca-analysis-r)

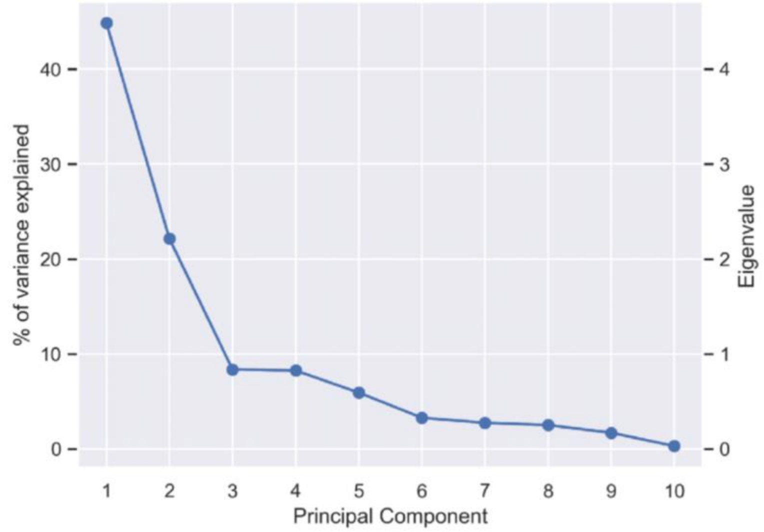
# Scree Plot

- A **Scree plot** is a plot of the eigenvalues or the % of explained variance in descending order. It helps us determine the number of principal components to retain.

- We are interested in the 'elbow' point.

- In the scree plot, the cumulative percentage of variance explained by the first 3 PCs is (approximately) 44% + 22% + 9% = 75%.

- Notice how the proportion of explained variance decreases as we move along the PCs. That is because PCs are ranked, with the first PC explaining the most. This is also why we only choose to work the first few PCs - they explain majority of the variance.



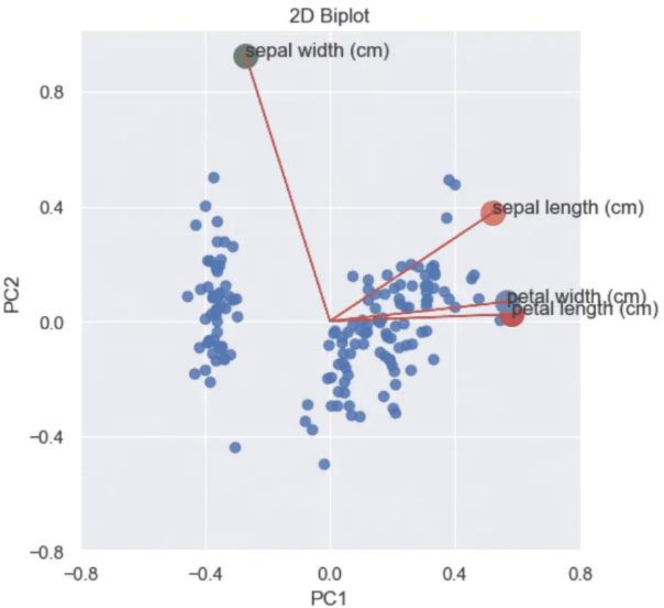
CoGrammar

# Biplots

- A **Biplot** plots the scores of the first two principal components. It is a scatterplot of observations projected onto a 2-dimensional space defined by the first two PCs.

- When interpreting biplots, we look for the following:

  - Clusters of observations - the closer points are to one another, the more similar they are
  - Distance from origin - the farther points are from the origin, the more extreme their values are on the PCs
  - Angles between vectors - small angles indicate strong positive correlations between variables, right angles indicate no correlation, and large angles (close to 180º) indicate strong negative correlations.
  - Projection of points onto vectors - points in the same direction of a vector indicate higher values for that value. Opposite direction indicates lower values.



Source: [https://www.jcchouinard.com/python-pca-biplots-machine-learning/](https://www.jcchouinard.com/python-pca-biplots-machine-learning/)

# Polls

# In PCA, what is the first principal component?

A. The component with the smallest variance
B. The component with the largest variance
C. The component orthogonal to the others
D. The component with the largest mean

# In PCA, what is the first principal component?

A. The component with the smallest variance

**B. The component with the largest variance**

C. The component orthogonal to the others

D. The component with the largest mean

CoGrammar

# What does a scree plot show in PCA?

A. The amount of variance each principal component explains

B. The correlation between variables

C. The distribution of data points

D. The mean and standard deviation of the data

# What does a scree plot show in PCA?

A. **The amount of variance each principal component explains**
B. The correlation between variables
C. The distribution of data points
D. The mean and standard deviation of the data

# How many principal components can be extracted from a dataset with 10 features?

A. 5, half the number of original features

B. 3, we cannot exceed three dimensions in the PC space

C. 20, the number of PCs we can create is double the number of original features

D. 10, same as the number of original features

CoGrammar

# How many principal components can be extracted from a dataset with 10 features?

A. 5, half the number of original features

B. 3, we cannot exceed three dimensions in the PC space

C. 20, the number of PCs we can create is double the number of original features

**D. 10, same as the number of original features**

CoGrammar

# What happens to the principal components if the data is not standardised?

A. They remain unaffected
B. They may be biased towards features with larger scales
C. They become correlated
D. They explain more variance

# What happens to the principal components if the data is not standardised?

A. They remain unaffected

**B. They may be biased towards features with larger scales**

C. They become correlated

D. They explain more variance

# PCA can only be applied to datasets with numerical features.

A. True

B. False

# PCA can only be applied to datasets with numerical features.

A. **True**

B. False

# The first principal component always explains all the variance in the data.

A. True

B. False

# The first principal component always explains all the variance in the data.

A. True

**B. False**

CoGrammar

# In PCA, what is an eigenvector?

A. A vector representing the direction of maximum variance in the data
B. A scalar value representing the magnitude of variance
C. The average of all feature values
D. A measure of central tendency

# In PCA, what is an eigenvector?

A. **A vector representing the direction of maximum variance in the data**

B. A scalar value representing the magnitude of variance

C. The average of all feature values

D. A measure of central tendency

# Questions and Answers

CoGrammar

# Thank you for attending

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar