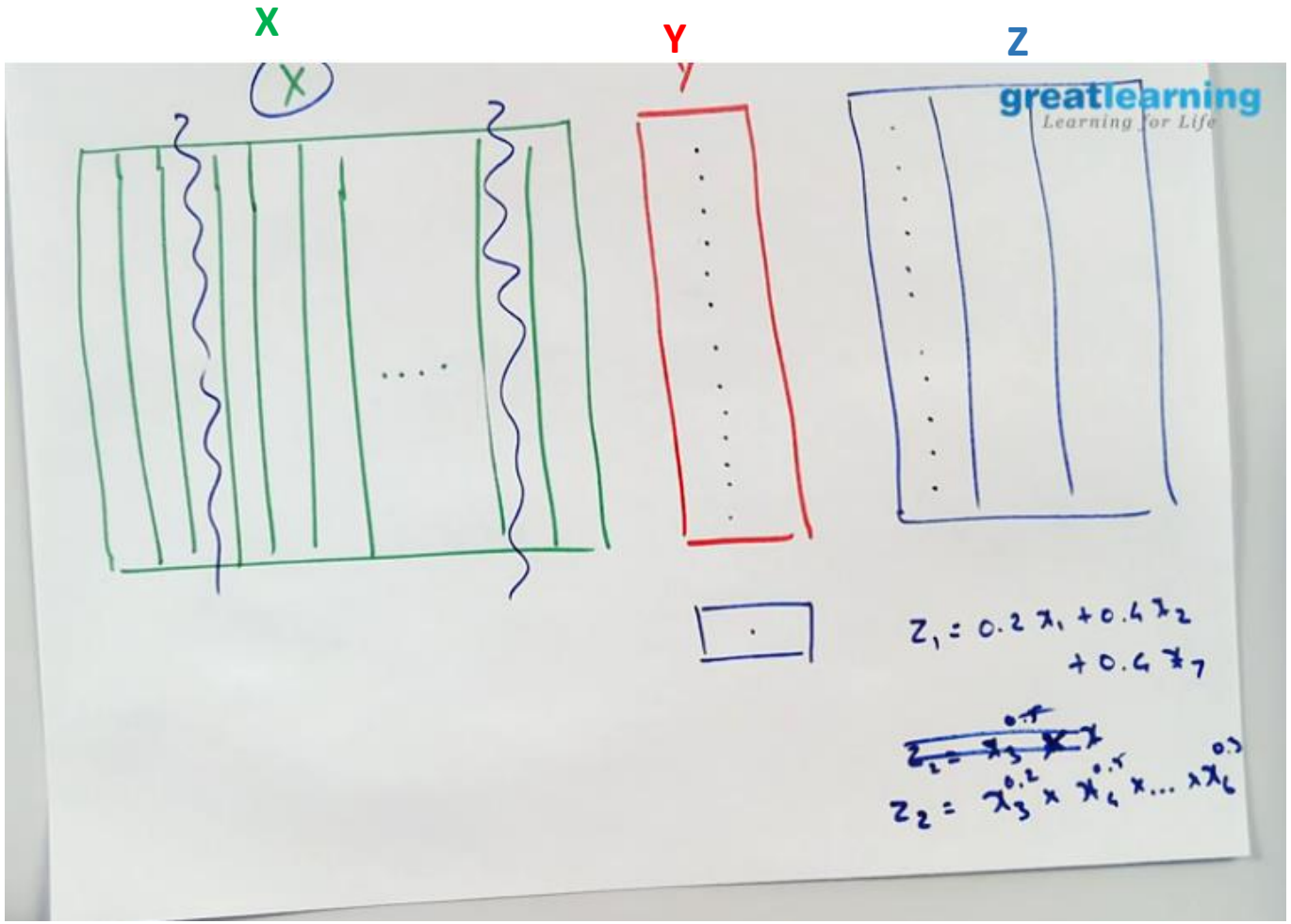
DIMENSIONALITY REDUCTION

- The process of reducing the number of independent variables
- Reducing dimensionality of independent variable helps in many ways
  - Removes multi-collinearity to improve ML model performance
  - Decreases computational times for fitting models
  - Makes visualization easier
  - Decreases storage requirements
  - Avoids curse of dimensionality
  - Helps reduce over fitting
- Hence dimensionality reduction plays a significant role in analyzing data

DIMENSIONALITY REDUCTION TECHNIQUES

- Feature Elimination
  - Simply identify and remove variables (columns) that are not important
  - The disadvantage is that we would gain no insight from those dropped variables and loose any information they contain
- Feature extraction
  - Create a few new variables from the old variables
  - **PCA** Principal Component Analysis: is the most popular feature extraction technique

DIMENSIONALITY REDUCTION

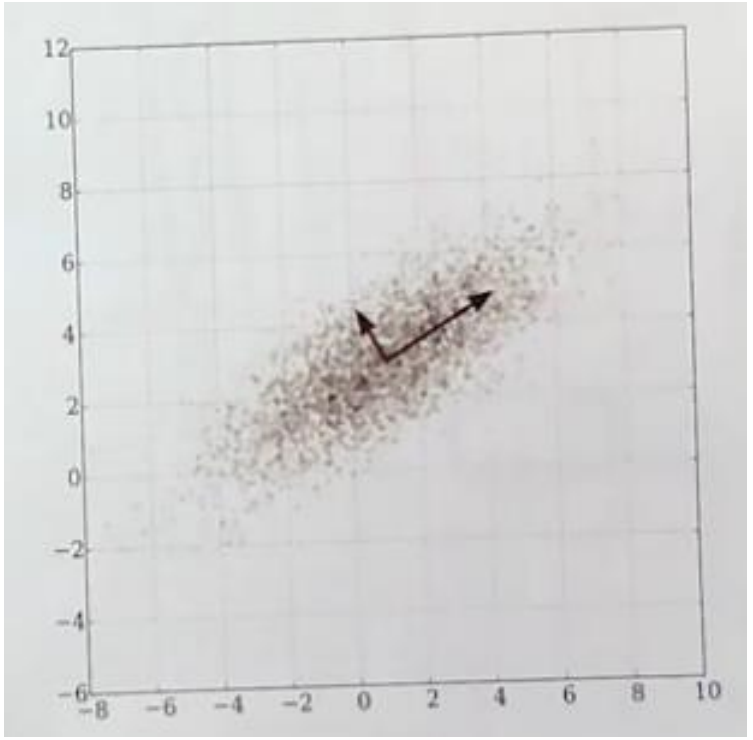


$Z1 = 0.2\ x1 + 0.4\ x2 + 0.4\ x3$

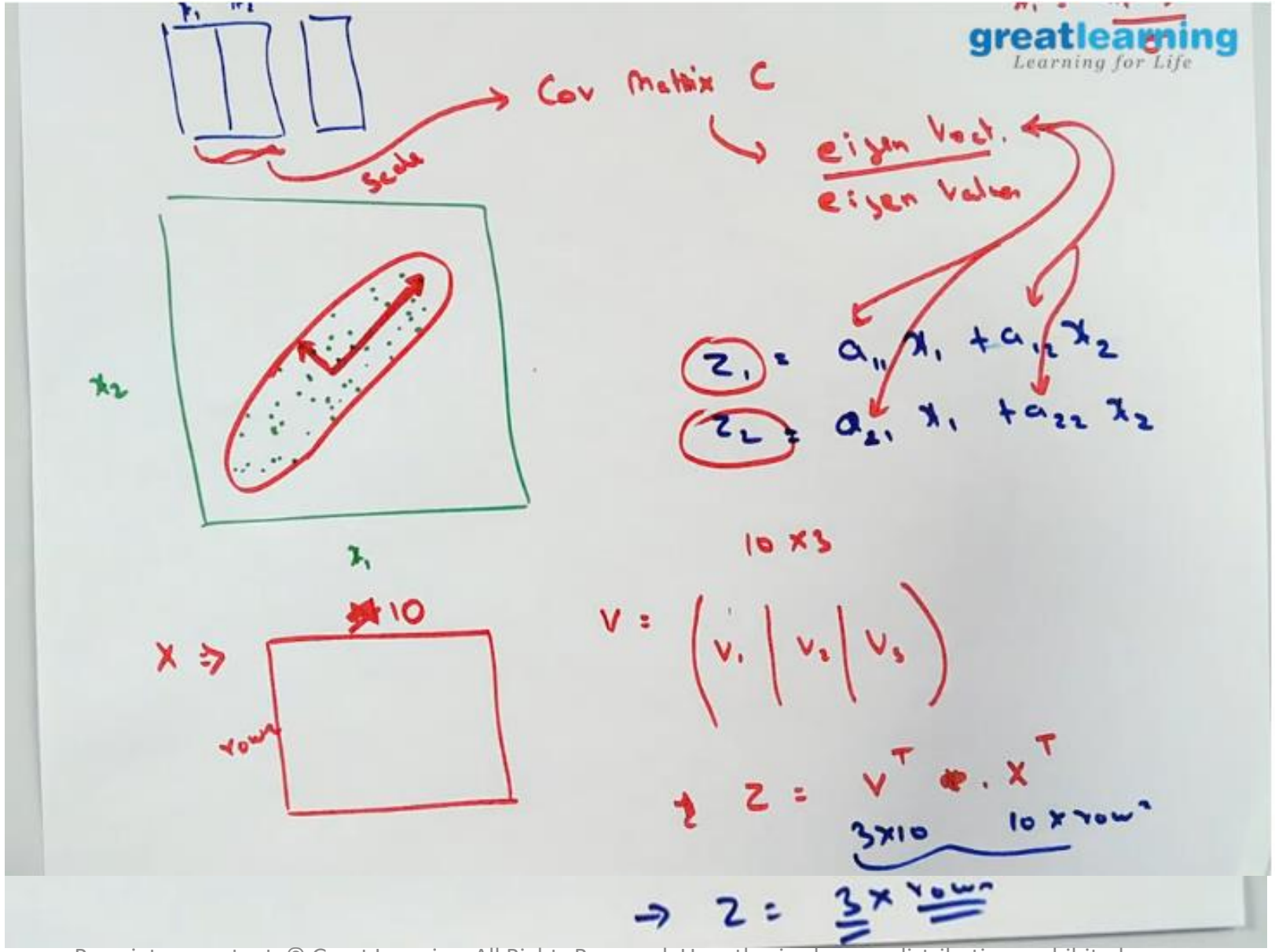
$Z2 = 0.2\ x3 + 0.5\ x4 + 0.5\ x6$

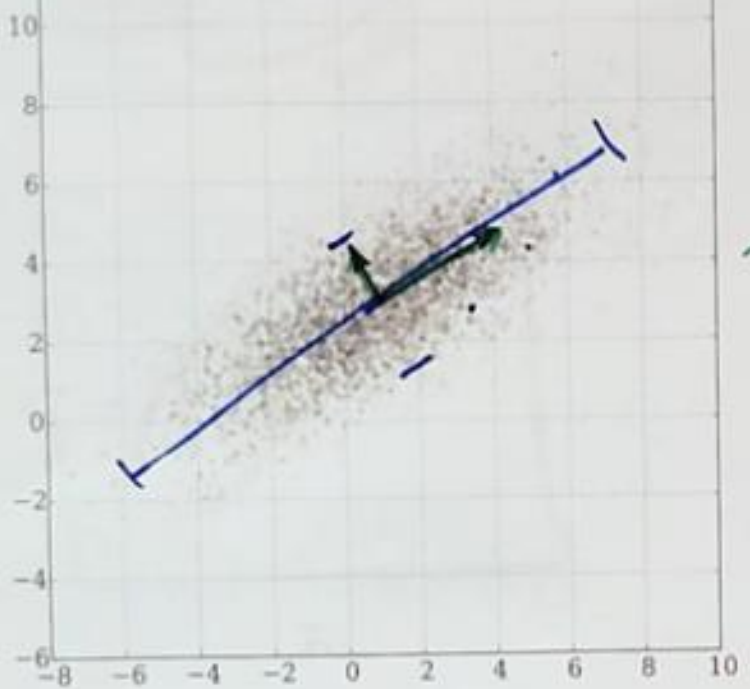
PCA

- Creates new variables using linear combinations of old variables
- Is designed to create variables that are independent of one another
- Also manages to tell us how important each of these new variables are
- This “importance”, helps is to choose how many variables we will use.



Eigen Vectors and values





$z_1 = a_{11} x_1 + a_{12} x_2$

$z_2 = a_{21} x_1 + a_{22} x_2$

2

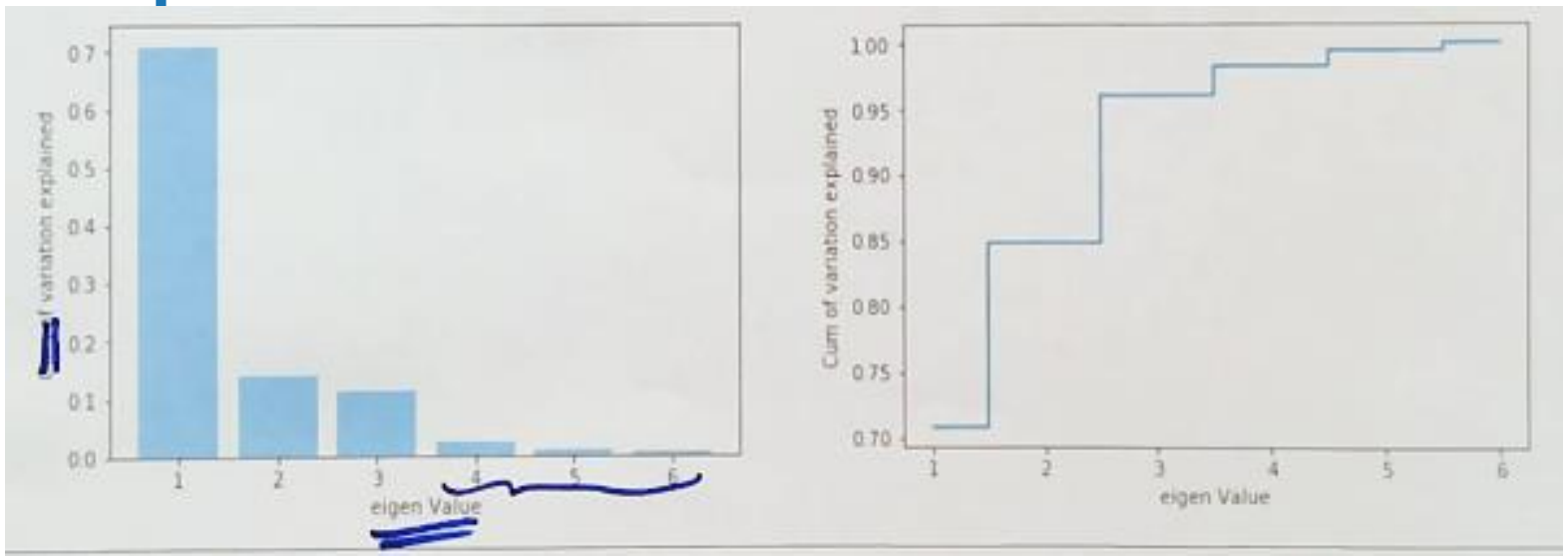
2D

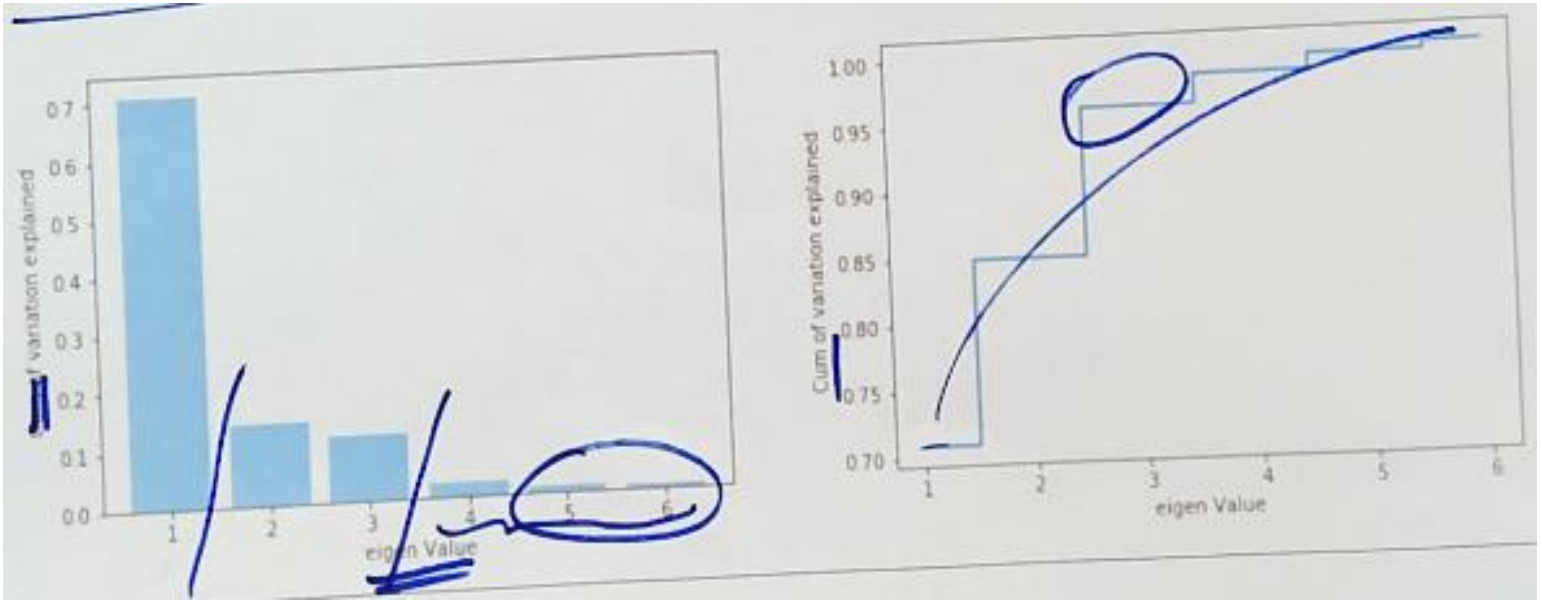
z_1

Overall Steps -PCA

- Scale the data and compute the covariance matrix
- Break the covariance matrix into magnitude and direction. Eigen Vectors and the Eigen values of the covariance matrix can be thought of as the natural axis and magnitudes along those axis
  - Eigen Values of the covariance matrix are the principal components
  - They are all orthogonal to each other – independent
  - The Eigen Values can be used to calculate the percentage of the variation explained by each direction
- Sort in the Eigen values in defending order and calculate the cumulative percentage of variation explained
- Pick the number of principal components you will use
- Transform to new variables

Overall Steps -PCA





Dimensionality Reduction Technique - Summary