Dimensionality Reduction in Linear Data

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Dimensionality reduction using Principal Components Analysis (PCA)

Dimensionality reduction the Singular Value Decomposition method in Factor Analysis

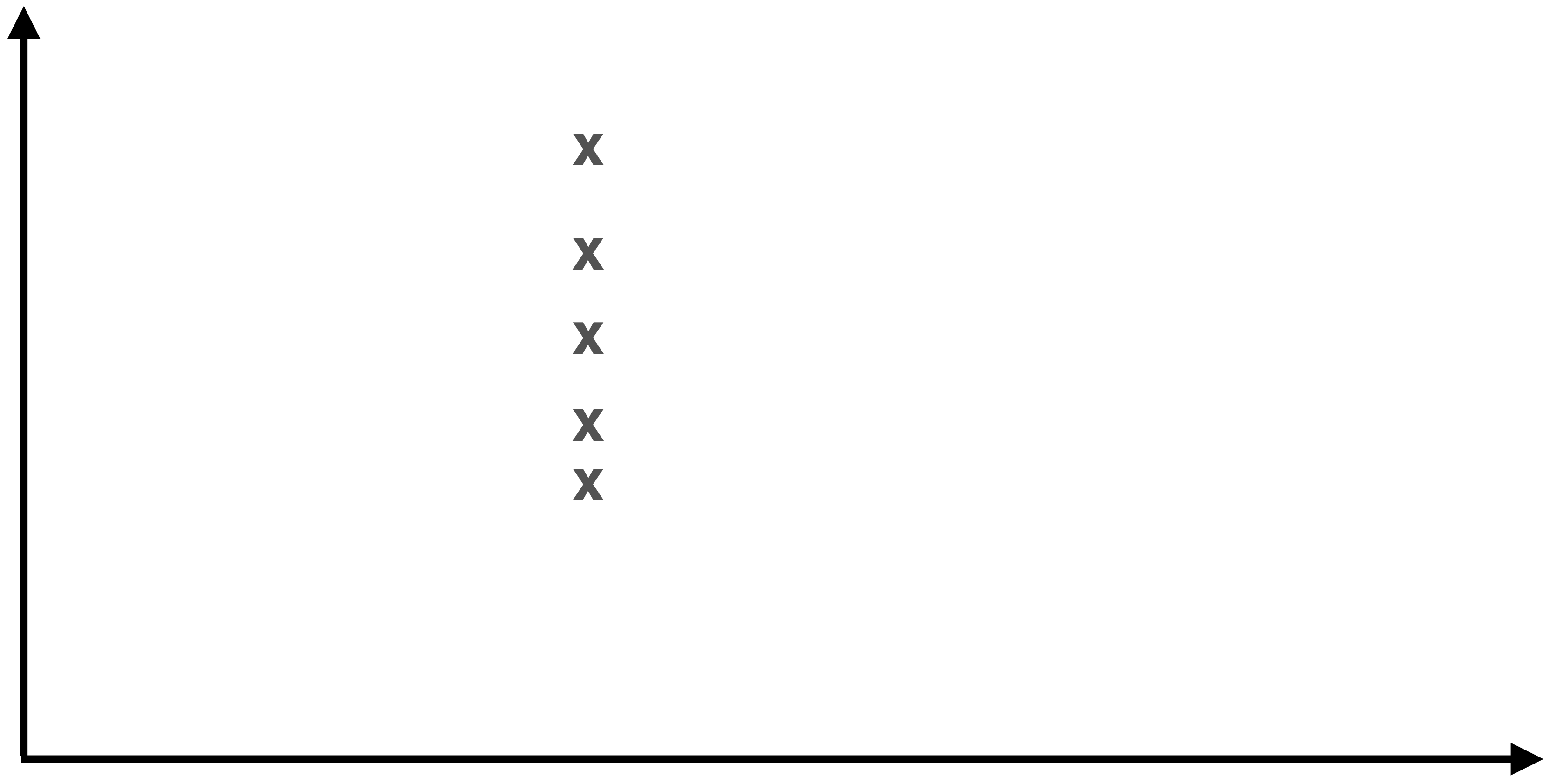
Dimensionality reduction using Linear Discriminant Analysis

Principal Components Analysis

Choosing PCA and Factor Analysis

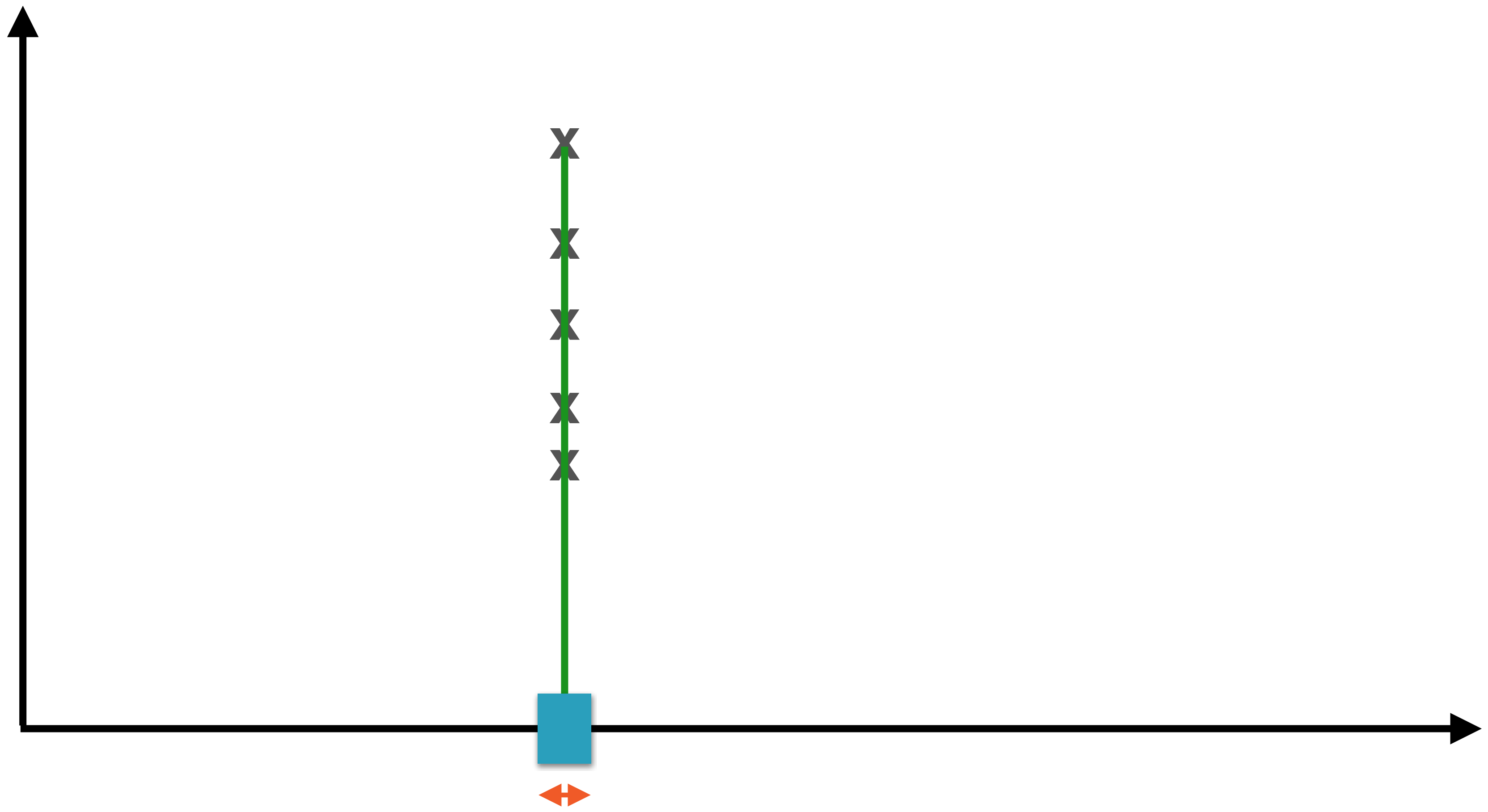
Use Case	Possible Solution
Large number of X-variables	Principal Components Analysis (PCA) or Factor Analysis
Most of which are meaningful	
Highly correlated to each other	
Linearly related to each other	
For use in regression	

A Question of Dimensionality



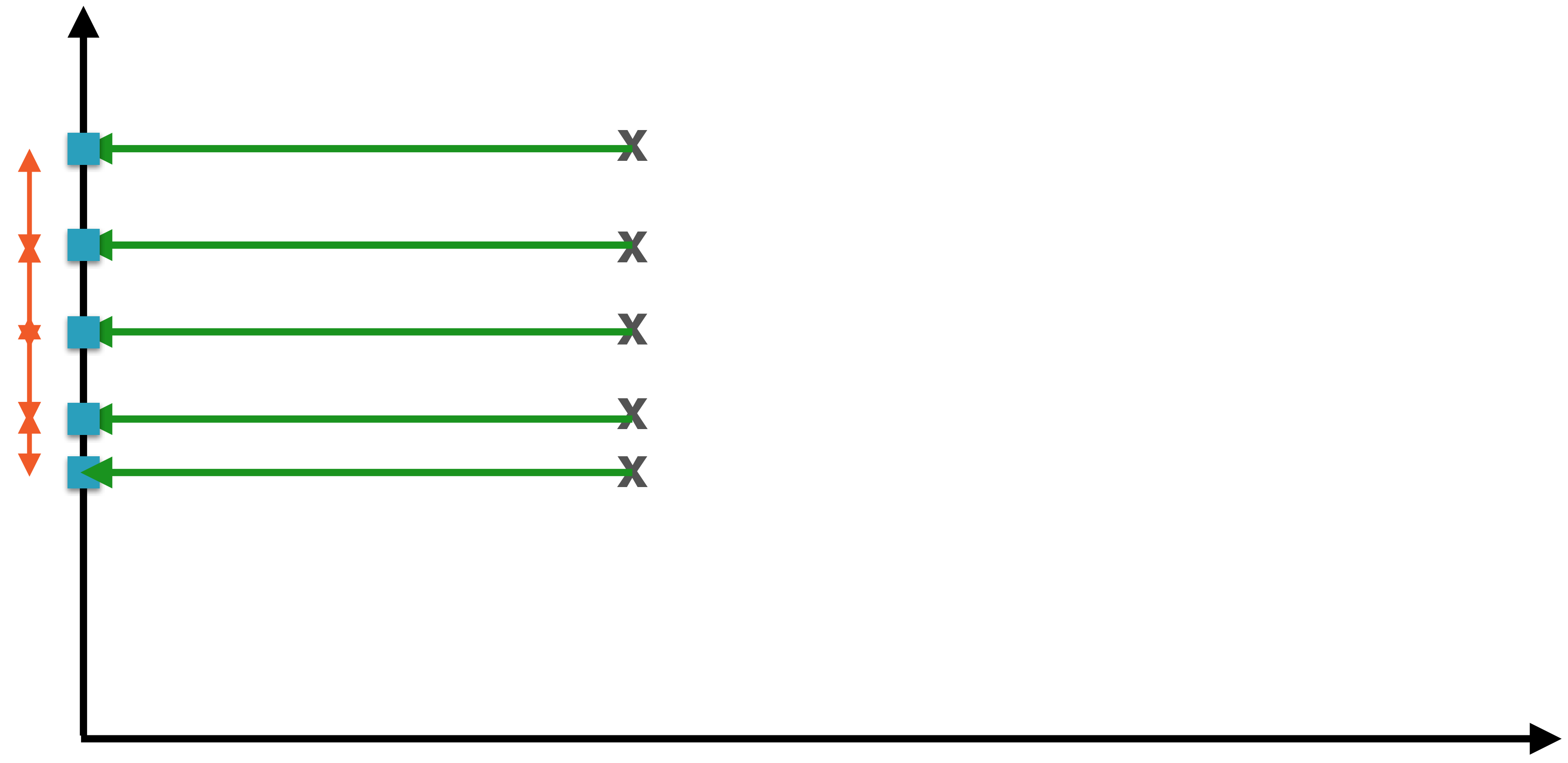
Pop quiz: Do we really need two dimensions to represent this data?

Bad Choice of Dimensions



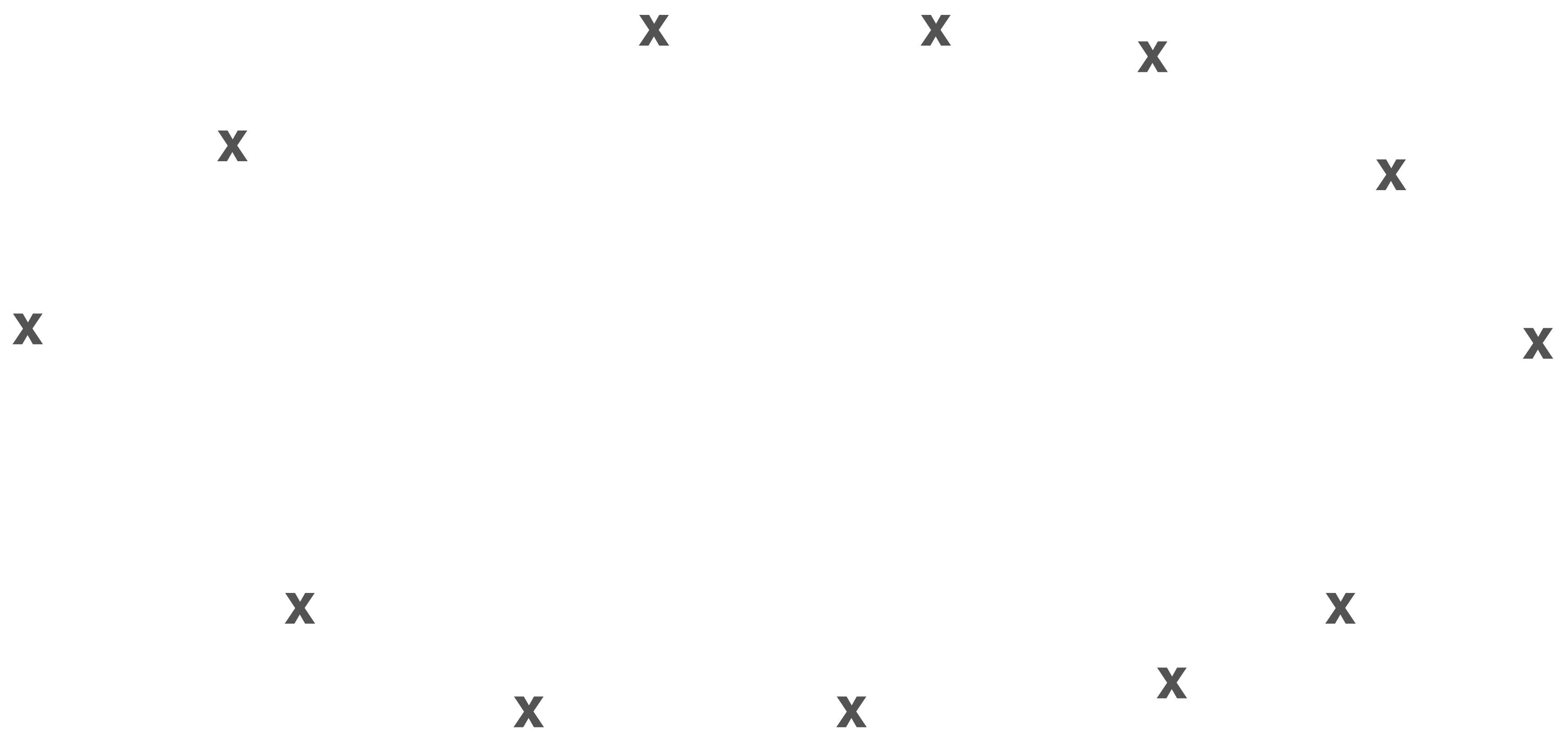
If we choose our axes (dimensions) poorly then we do need two dimensions

Good Choice of Dimensions



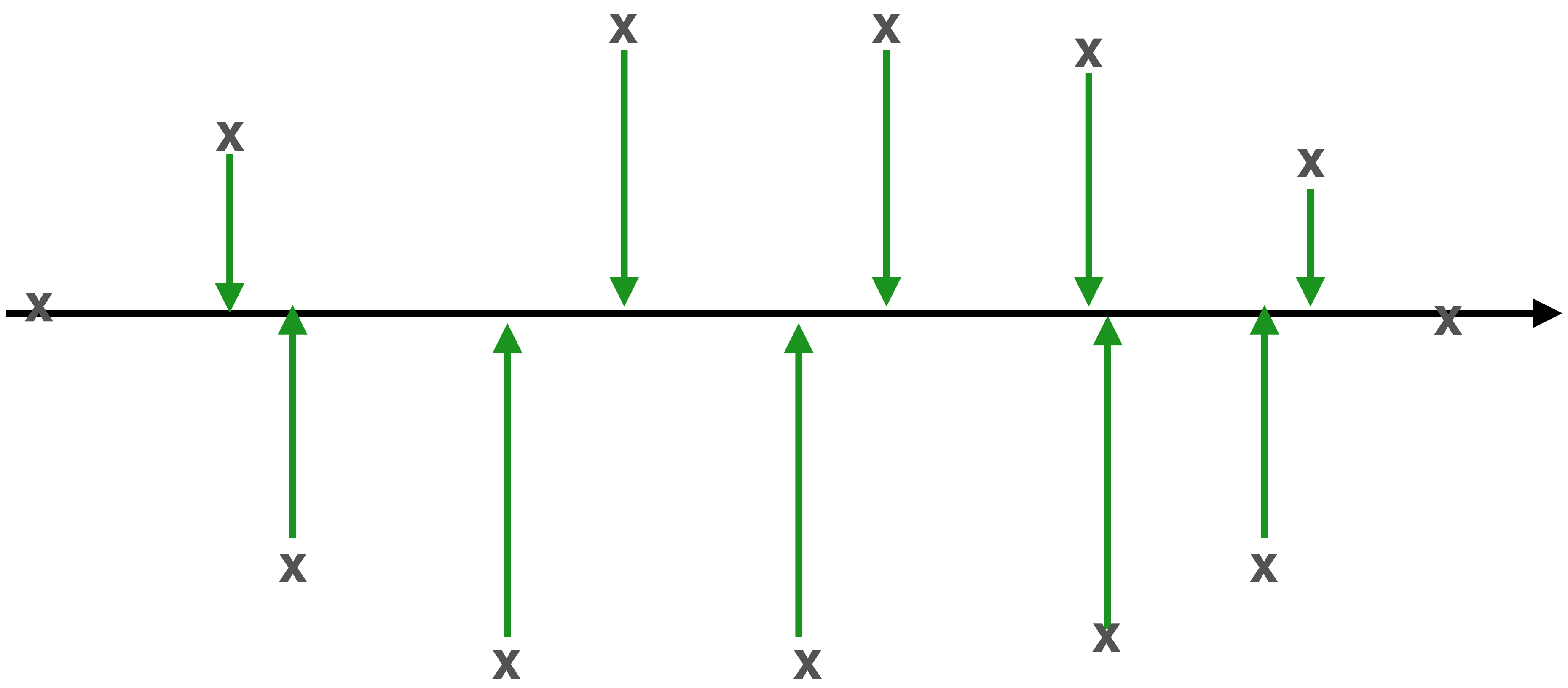
If we choose our axes (dimensions) well then one dimension is sufficient

Intuition Behind PCA



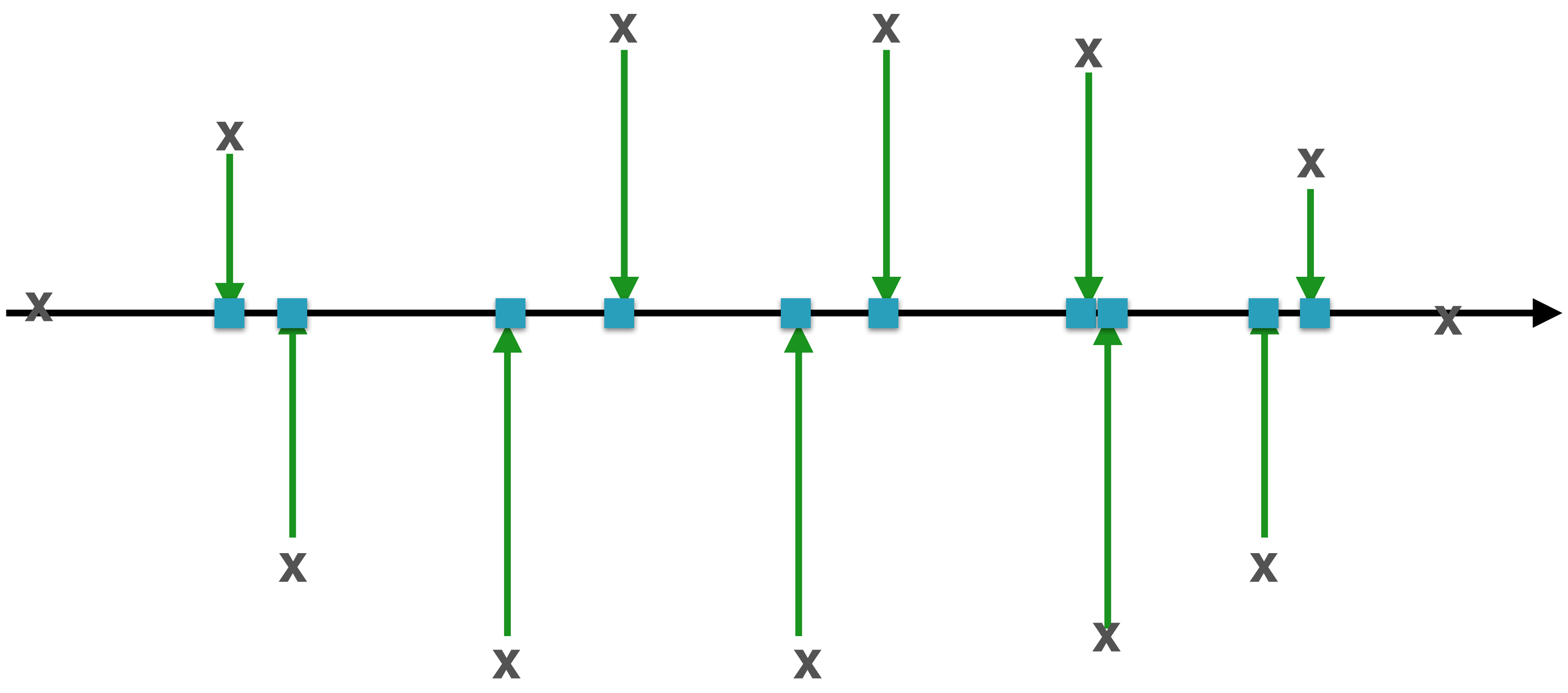
Objective: Find the “best” directions to represent this data

Intuition Behind PCA



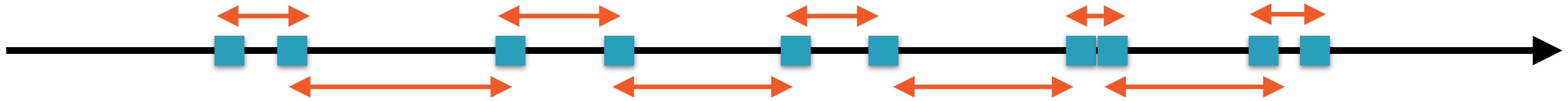
Start by "projecting" the data onto a line in some direction

Intuition Behind PCA



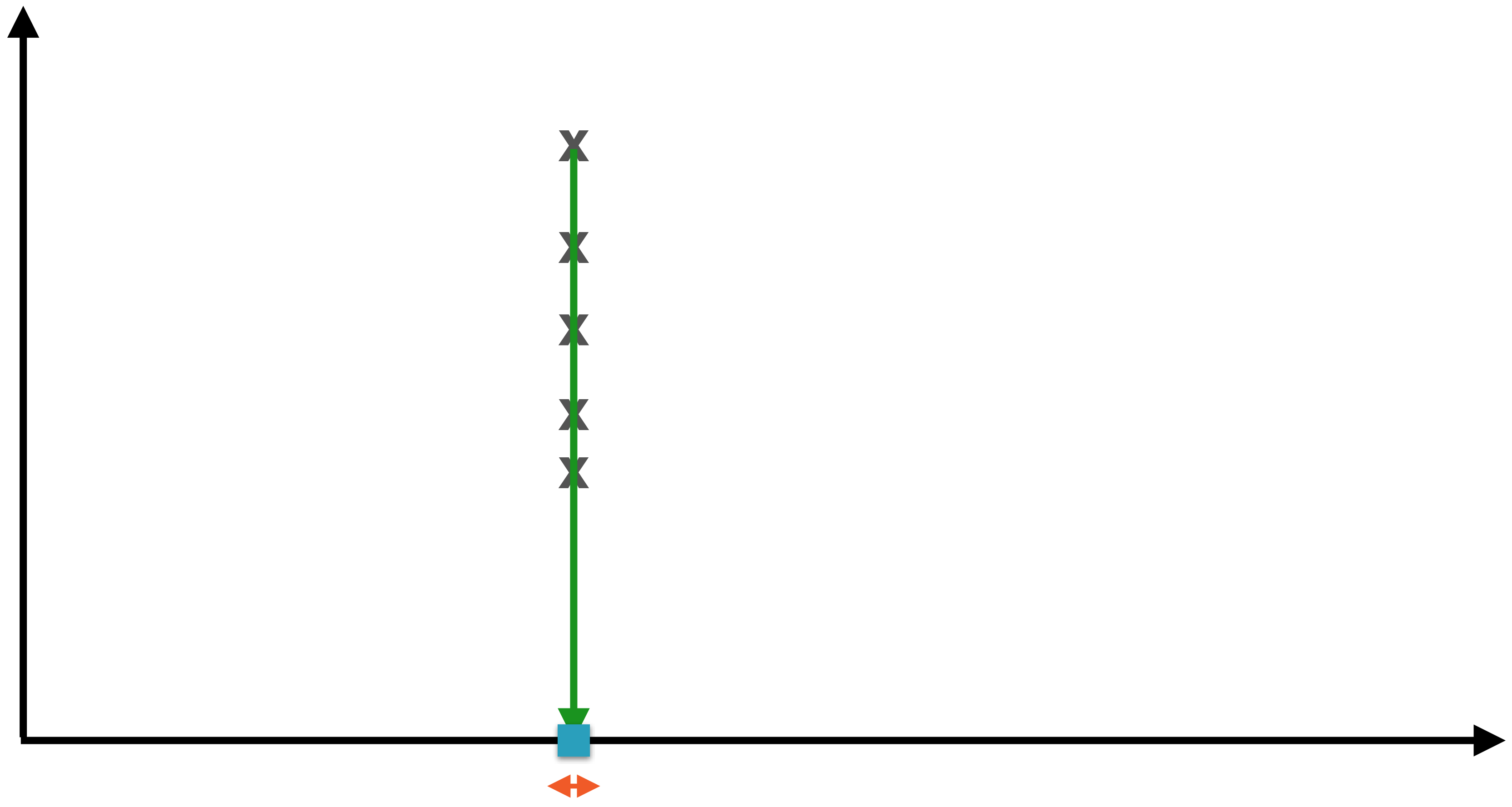
Start by "projecting" the data onto a line in some direction

Intuition Behind PCA



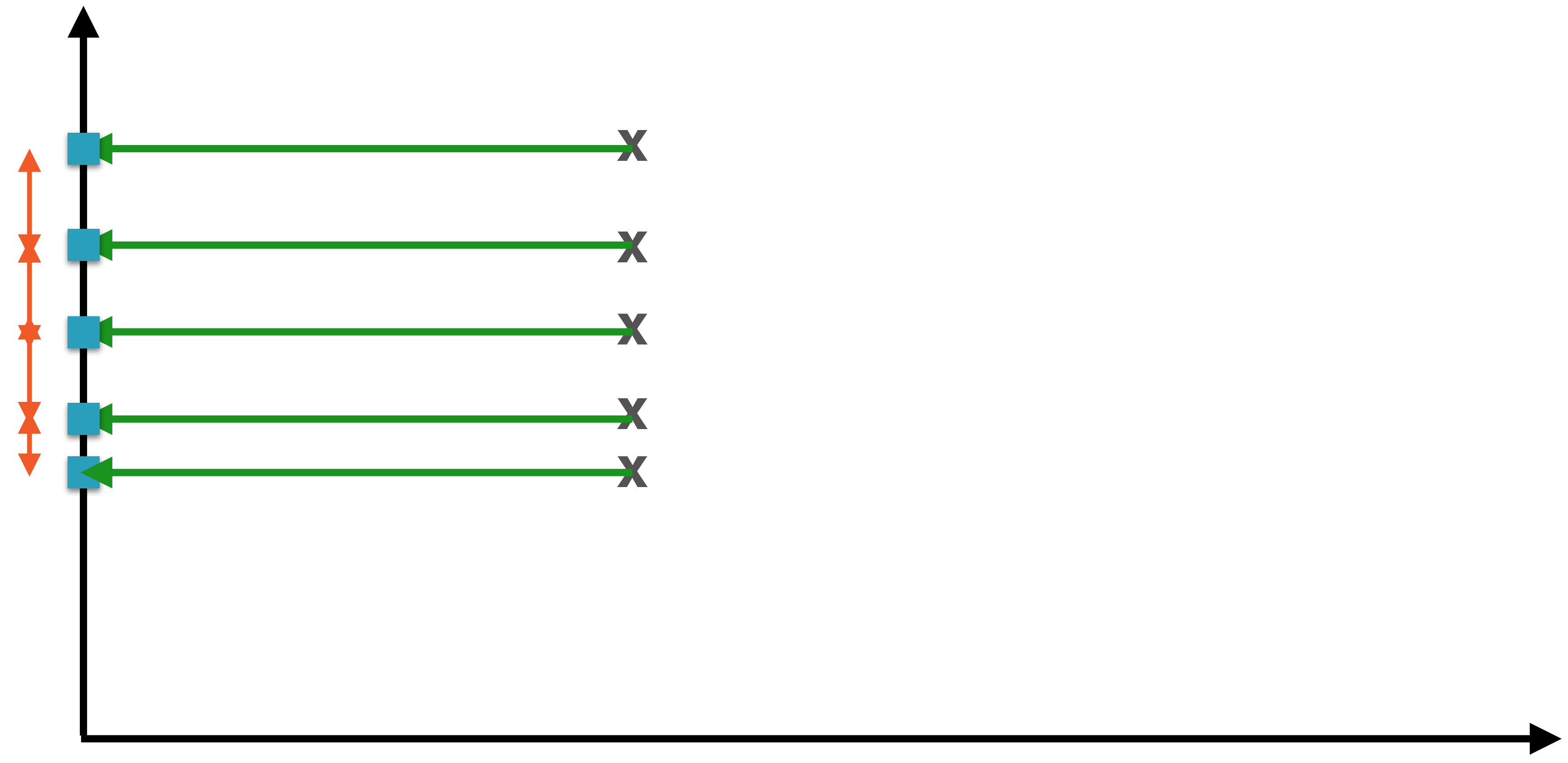
The greater the distances between these projections, the “better” the direction

Bad Projection



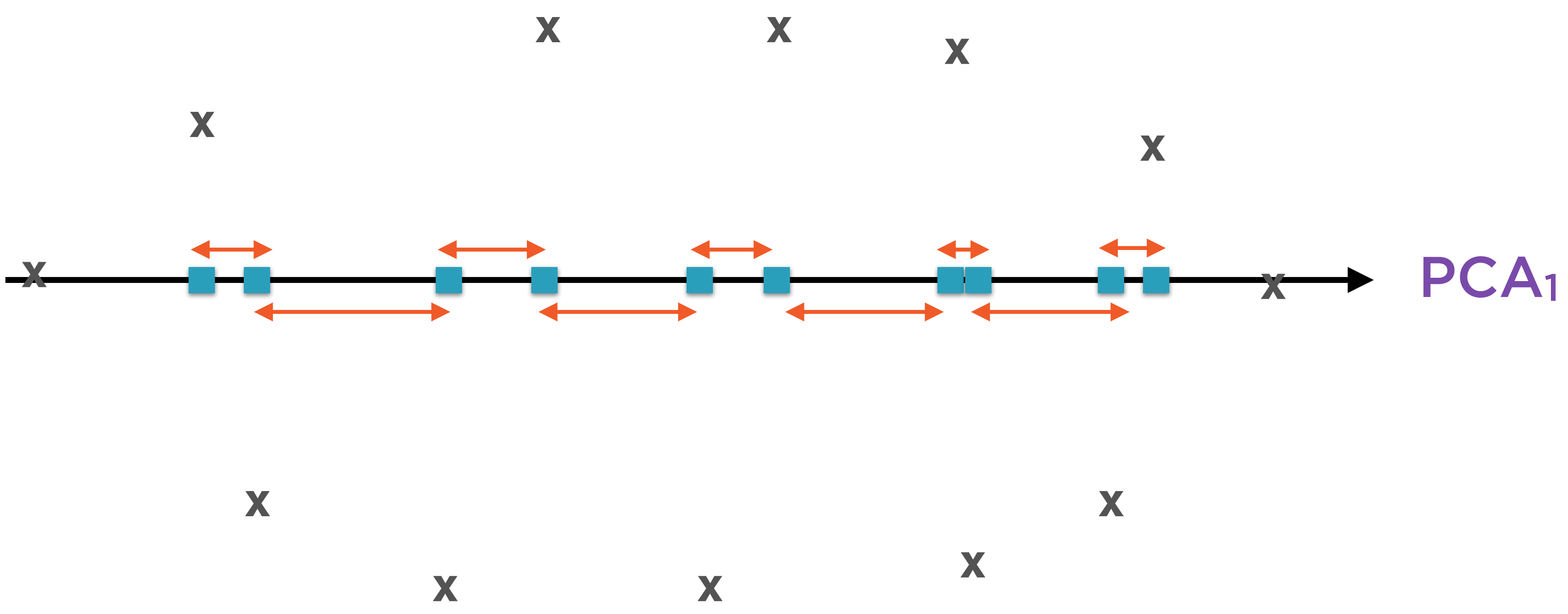
A projection where the distances are minimized is a bad one - information is lost

Good Projection



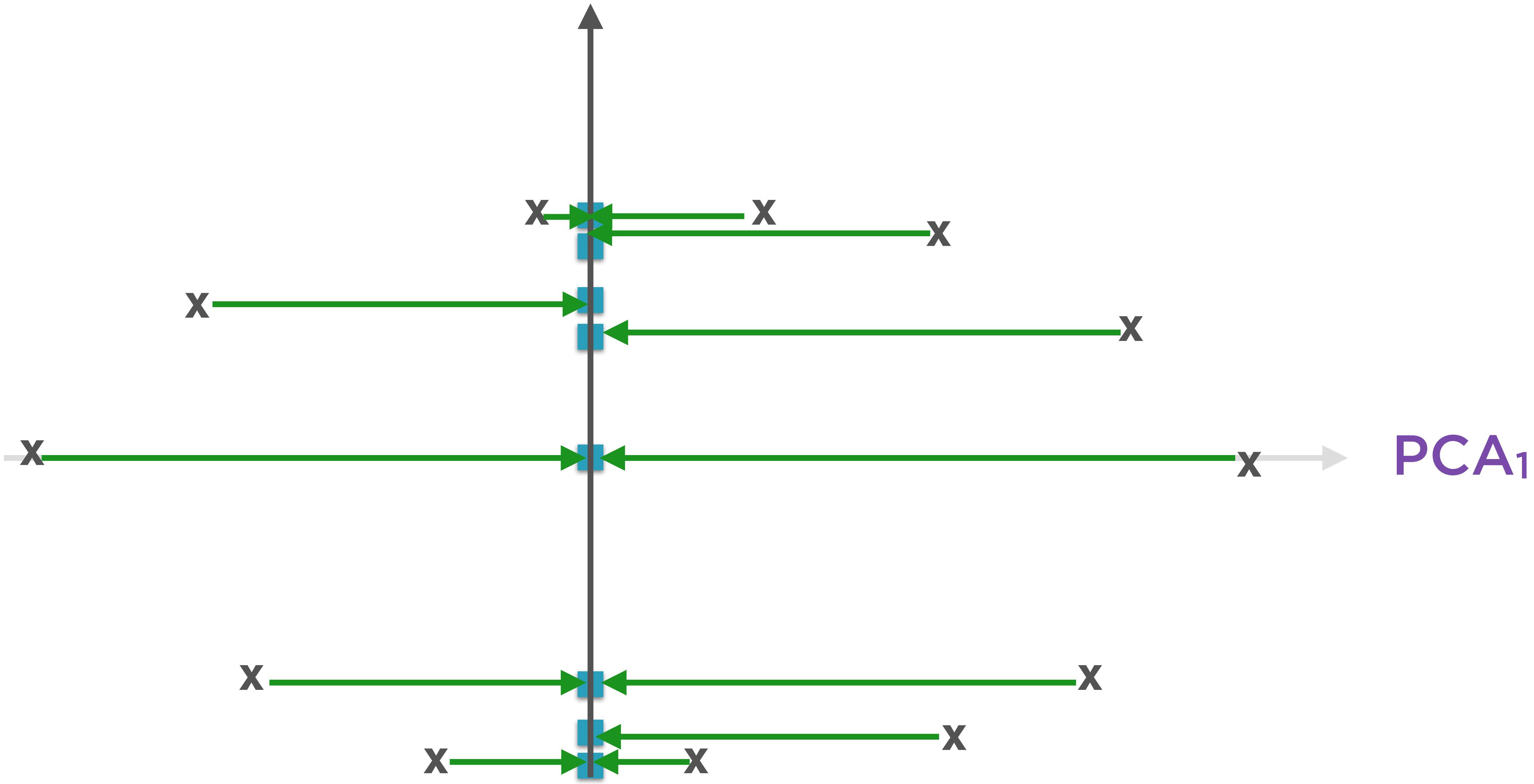
A projection where the distances are maximised is a good one - information is preserved

Intuition Behind PCA



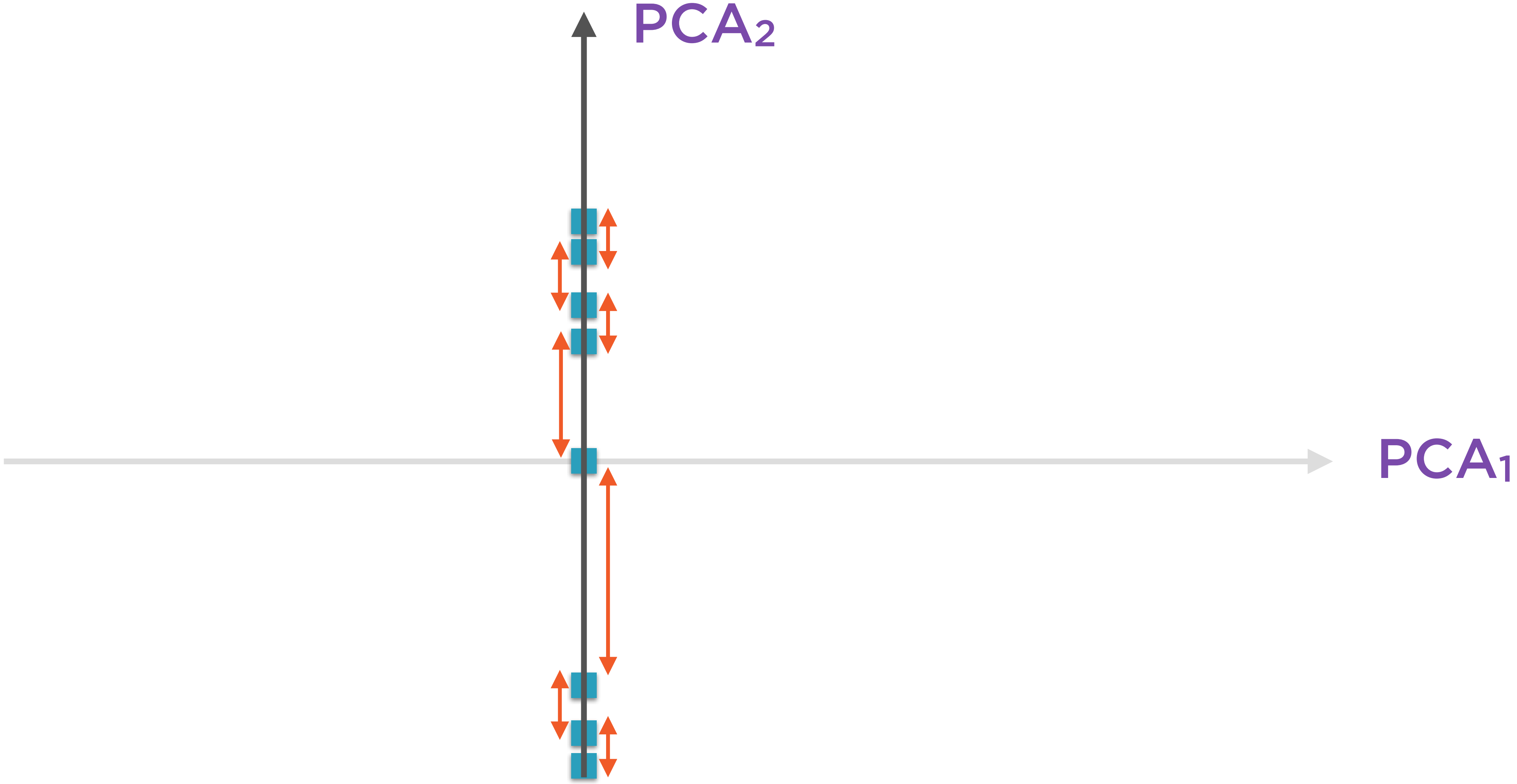
The direction along which this variance is maximised is the first principal component of the original data

Intuition Behind PCA



Find the next best direction, the **second principal component**, which must be at right angles to the first

Intuition Behind PCA

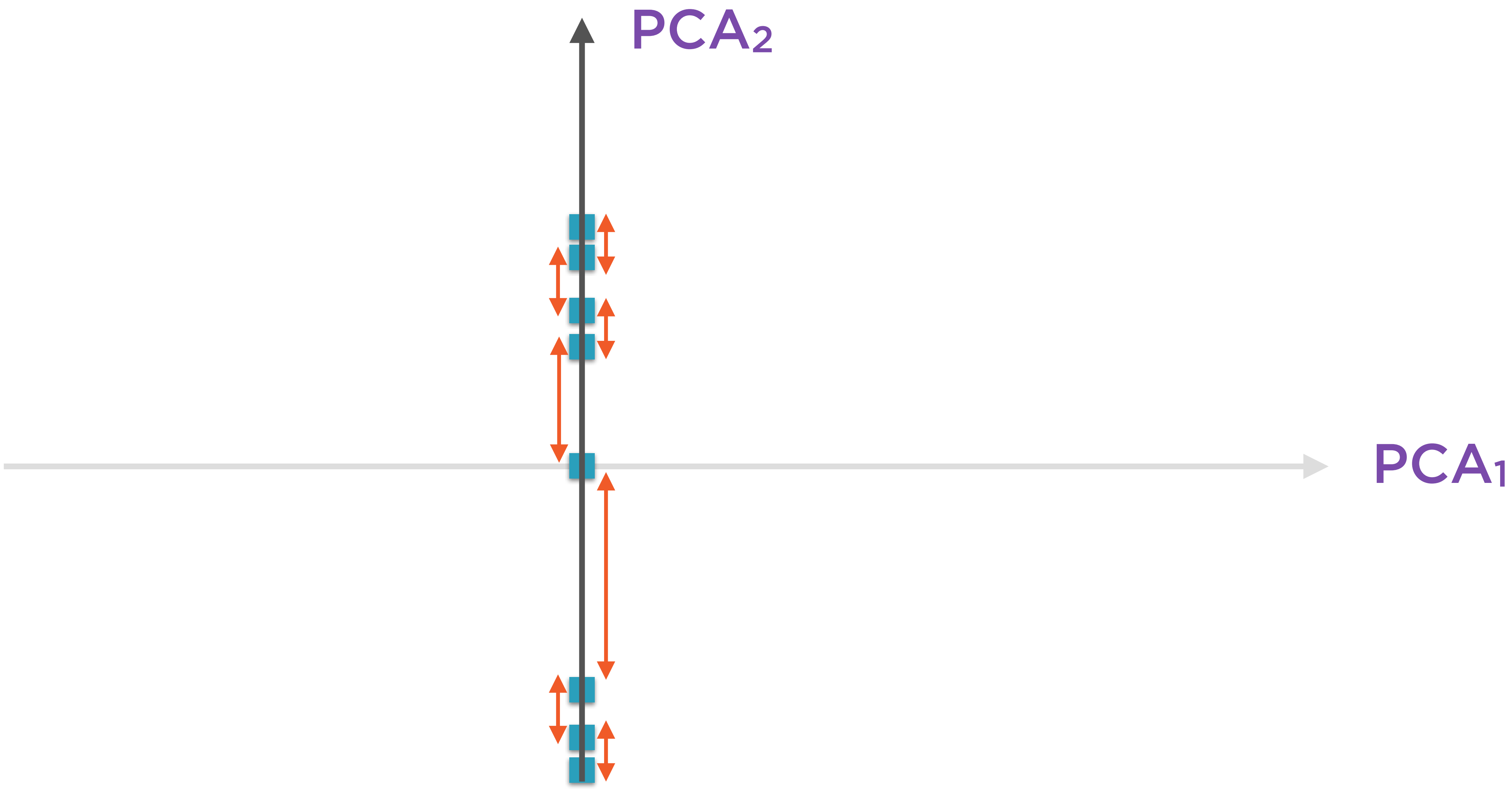


Find the next best direction, the **second principal component**, which must be at right angles to the first

Principal Components at Right Angles

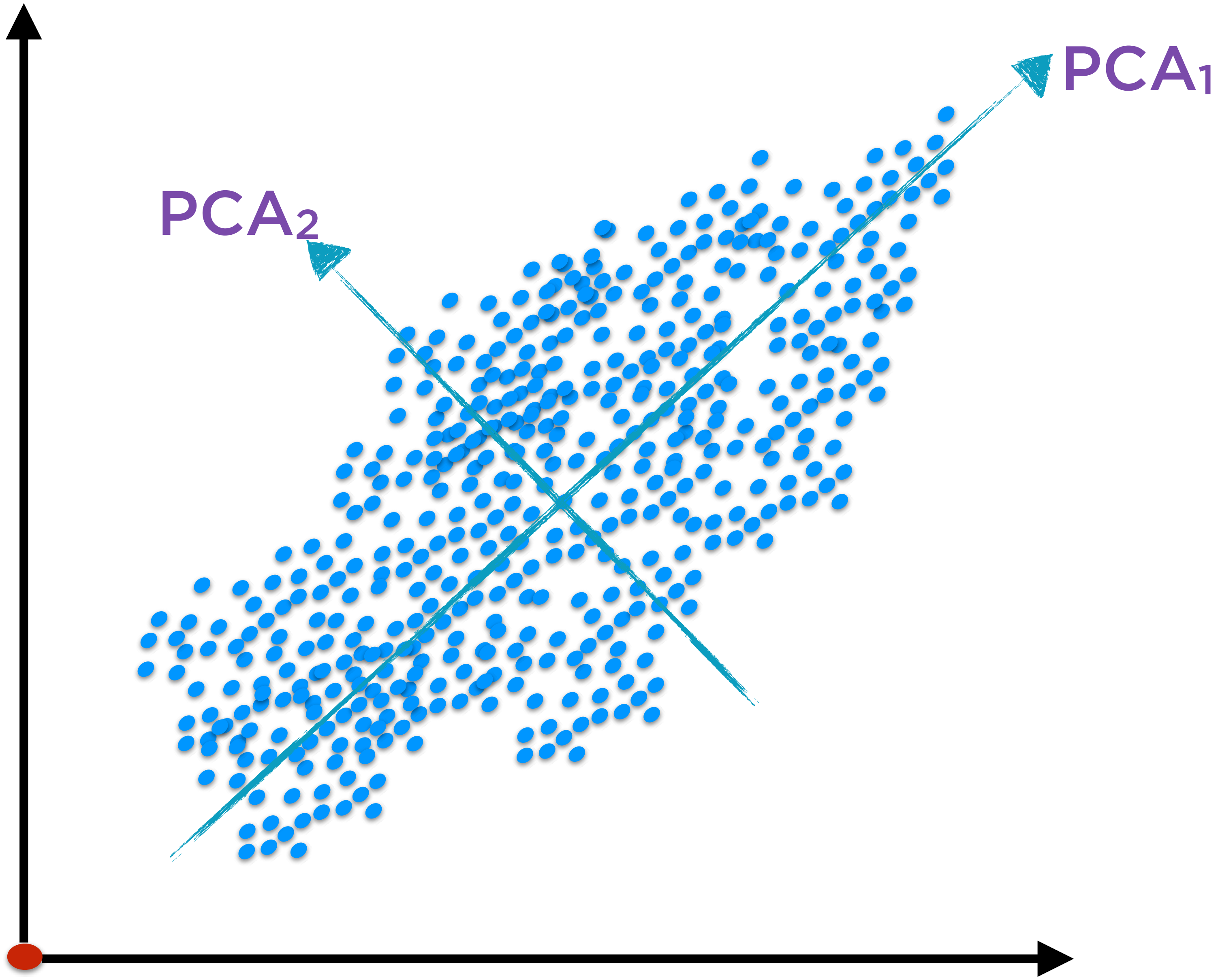
Directions at right angles help express the most variation with the smallest number of directions

Intuition Behind PCA



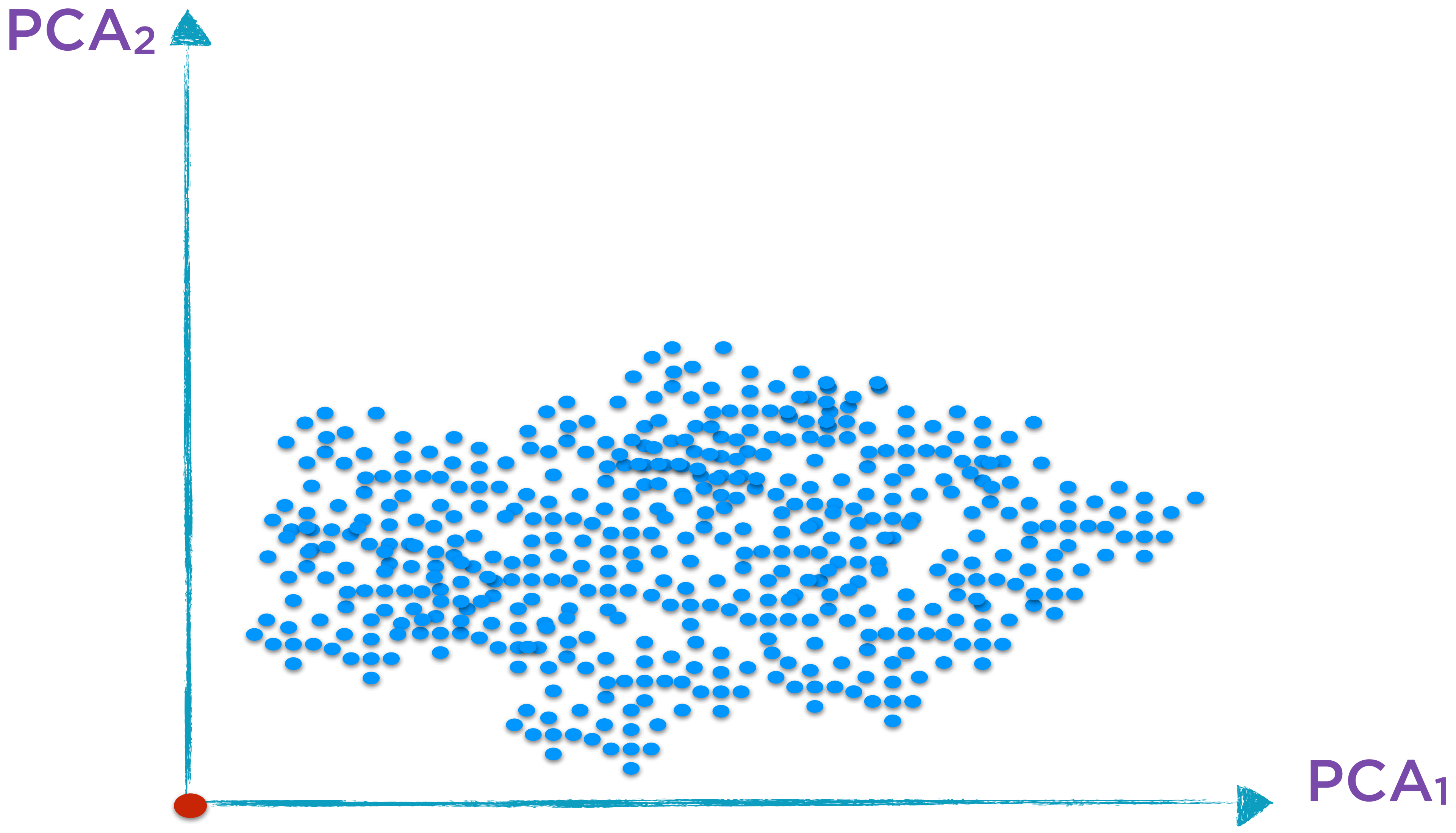
The variances are clearly smaller along this **second principal component** than along the first

Intuition Behind PCA



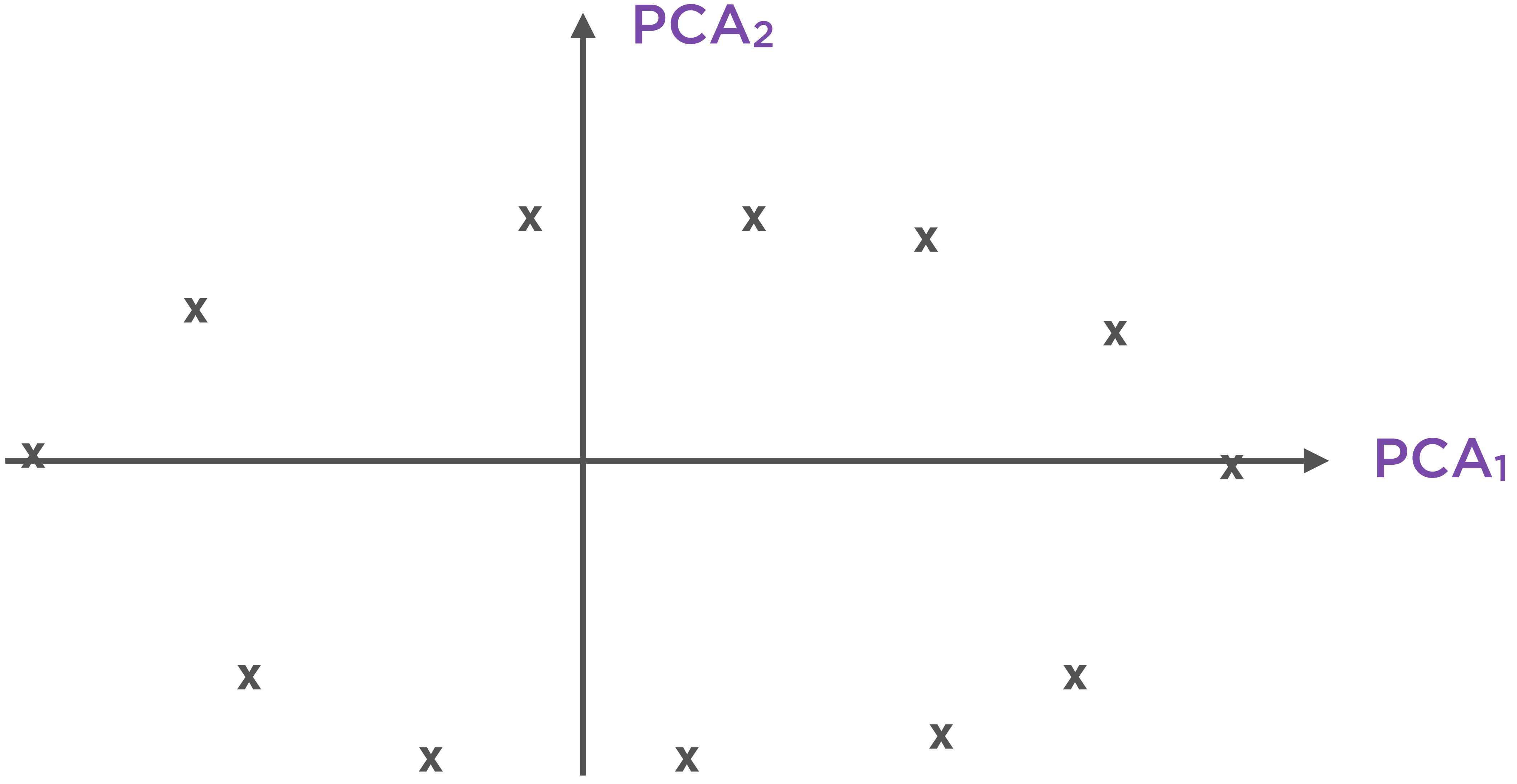
In general, there are as many principal components as there are dimensions in the original data

Intuition Behind PCA



Re-orient the data along these new axes

Dimensionality Reduction



If the variance along the second principal component is small enough, we can just ignore it and use just 1 dimension to represent the data

Demo

Implement Principal Components Analysis (PCA) with linear regression

Factor Analysis

SVD Factor Analysis

Apply Singular Value Decomposition (SVD) to re-express highly correlated X-variables in terms of new, unrelated components

Correlated Random Variables

$$[\ X_1 \quad X_2 \quad X_3 \quad \ldots \quad X_k\ ]$$

n rows

k columns

SVD, like PCA is used when the elements X_i of this matrix are highly correlated with each other

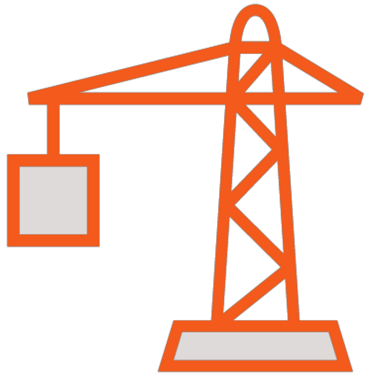
Factor Analysis

k columns

$$[\ X_1 \quad X_2 \quad X_3 \quad \ldots \quad X_k\ ]$$

n rows

X_i are highly correlated with each other



Factor Analysis:
SVD, PCA, etc

F_i are completely uncorrelated with each other

$$[\ F_1 \quad F_2 \quad F_3 \quad \ldots \quad F_k\ ]$$

n rows

k columns

Factor Analysis

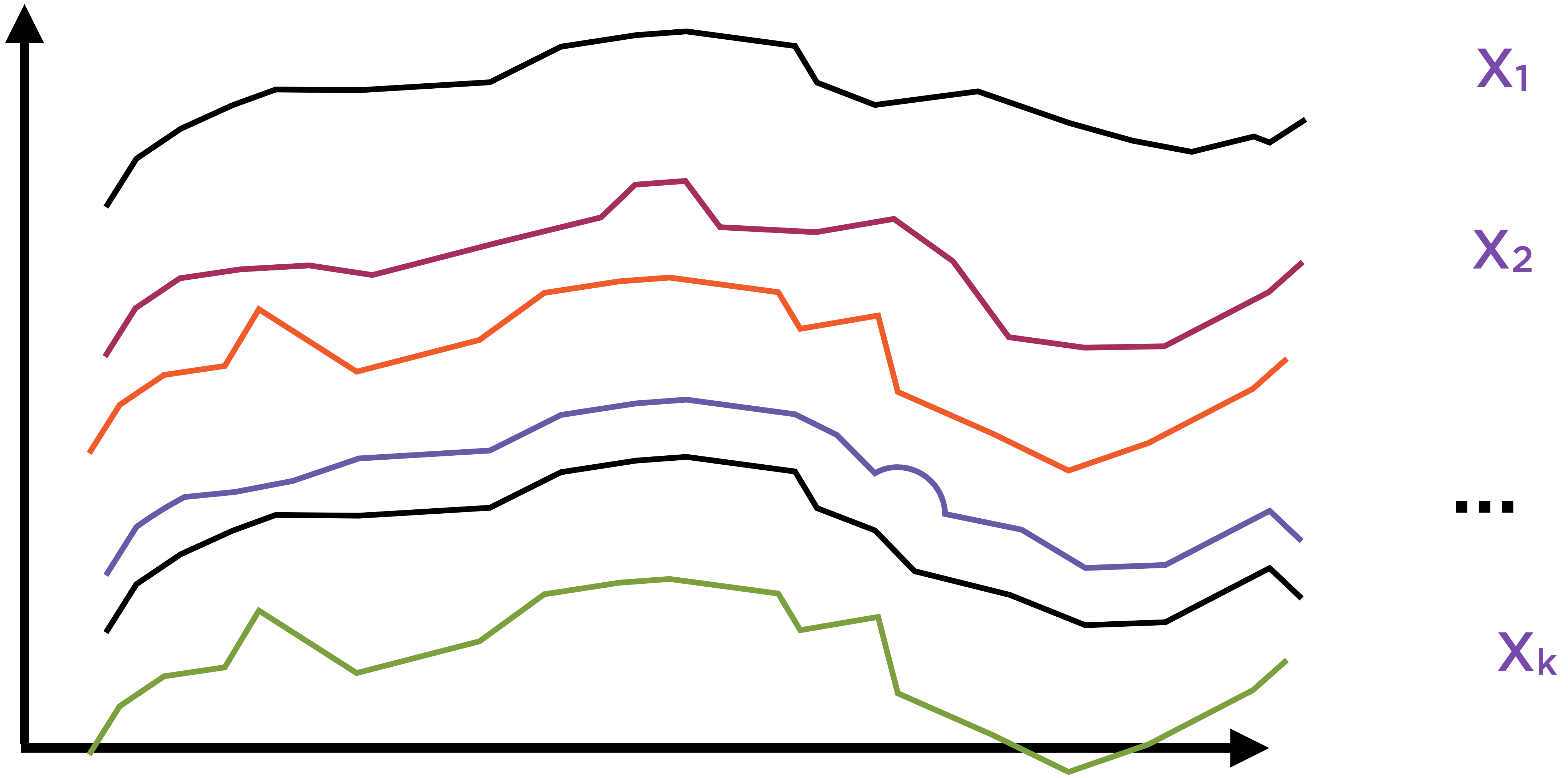
$$[\ F_1 \quad F_2 \quad F_3 \quad \ldots \quad F_k\ ]$$

n rows

k columns

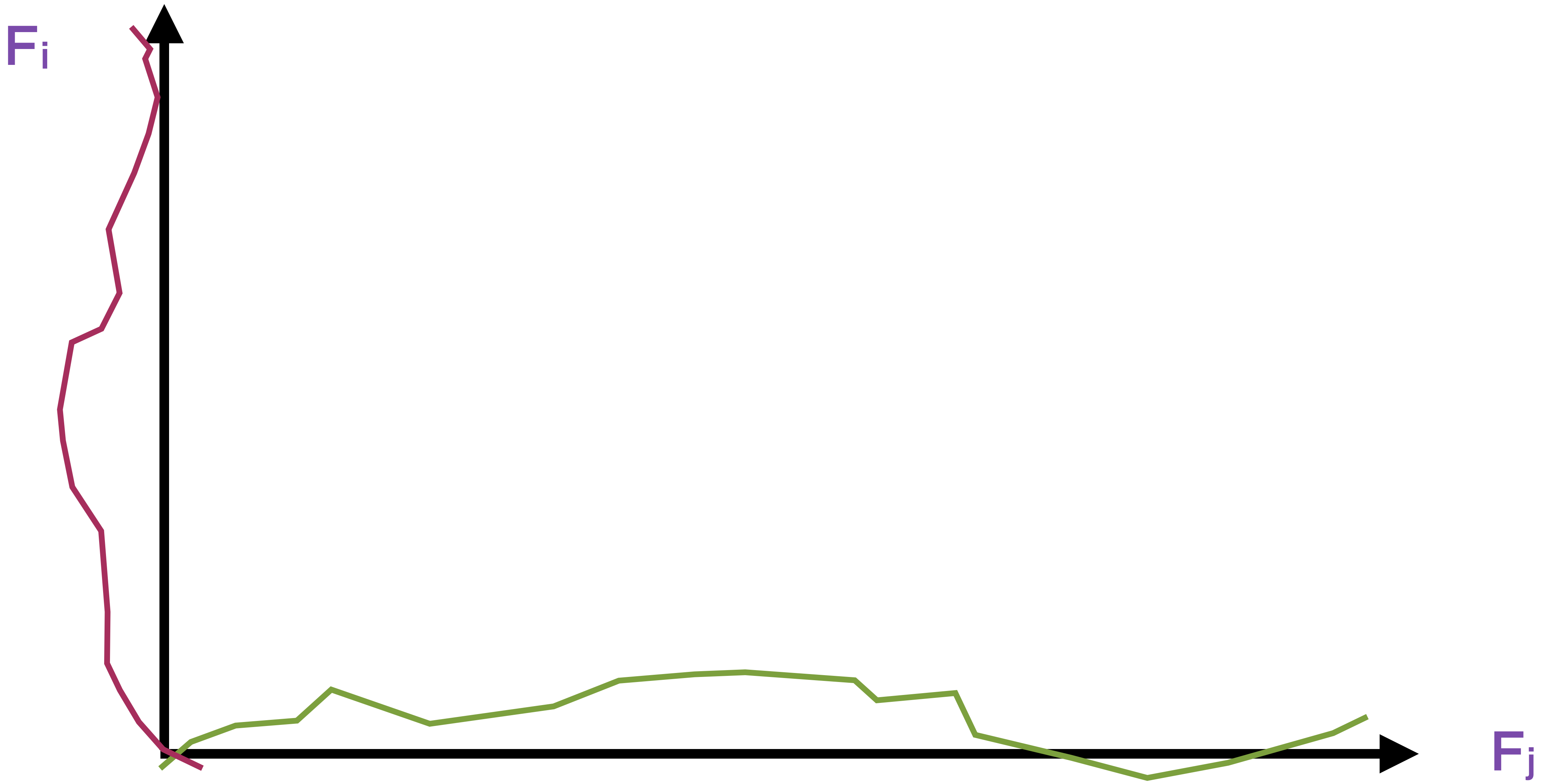
These vectors F_i are the factor representations of the original vectors X_i

Correlated Random Variables



Highly correlated variables are not suitable for use in regression

Uncorrelated F_i



Factors generated by SVD, like those from PCA, are perfectly uncorrelated to each other

Demo

Implement Factor Analysis with classification

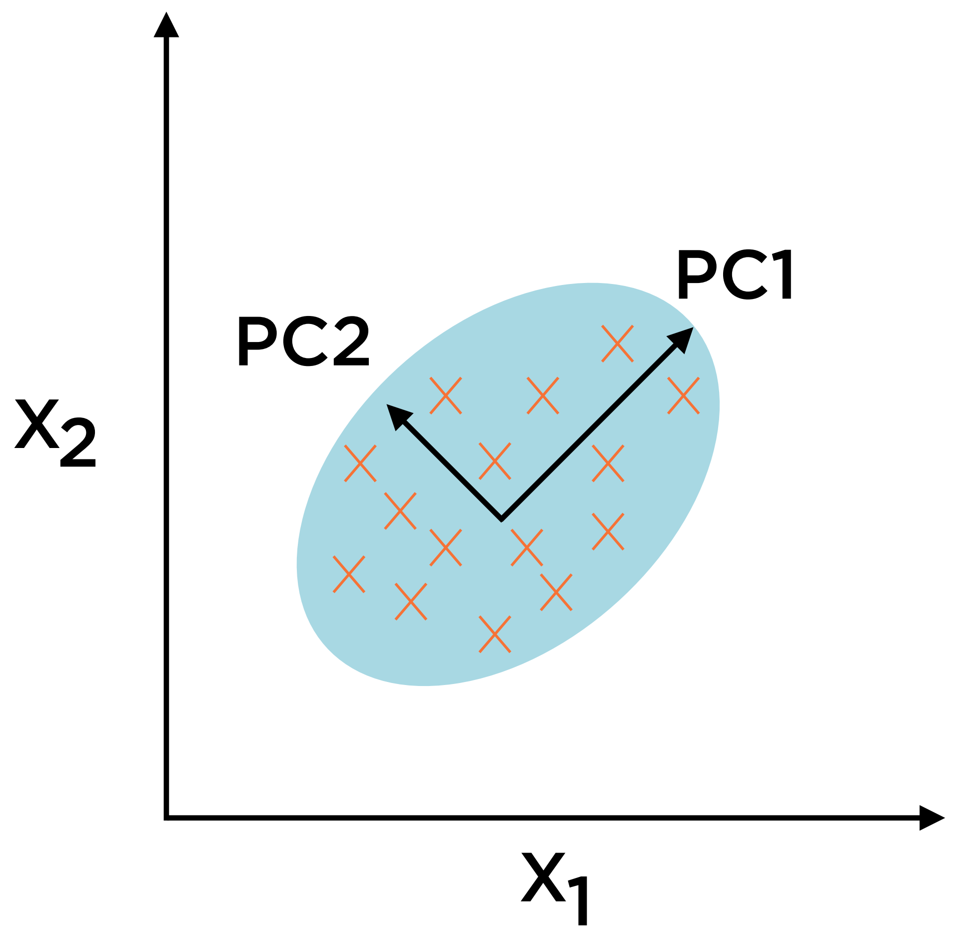
Linear Discriminant Analysis

Choosing Linear Discriminant Analysis

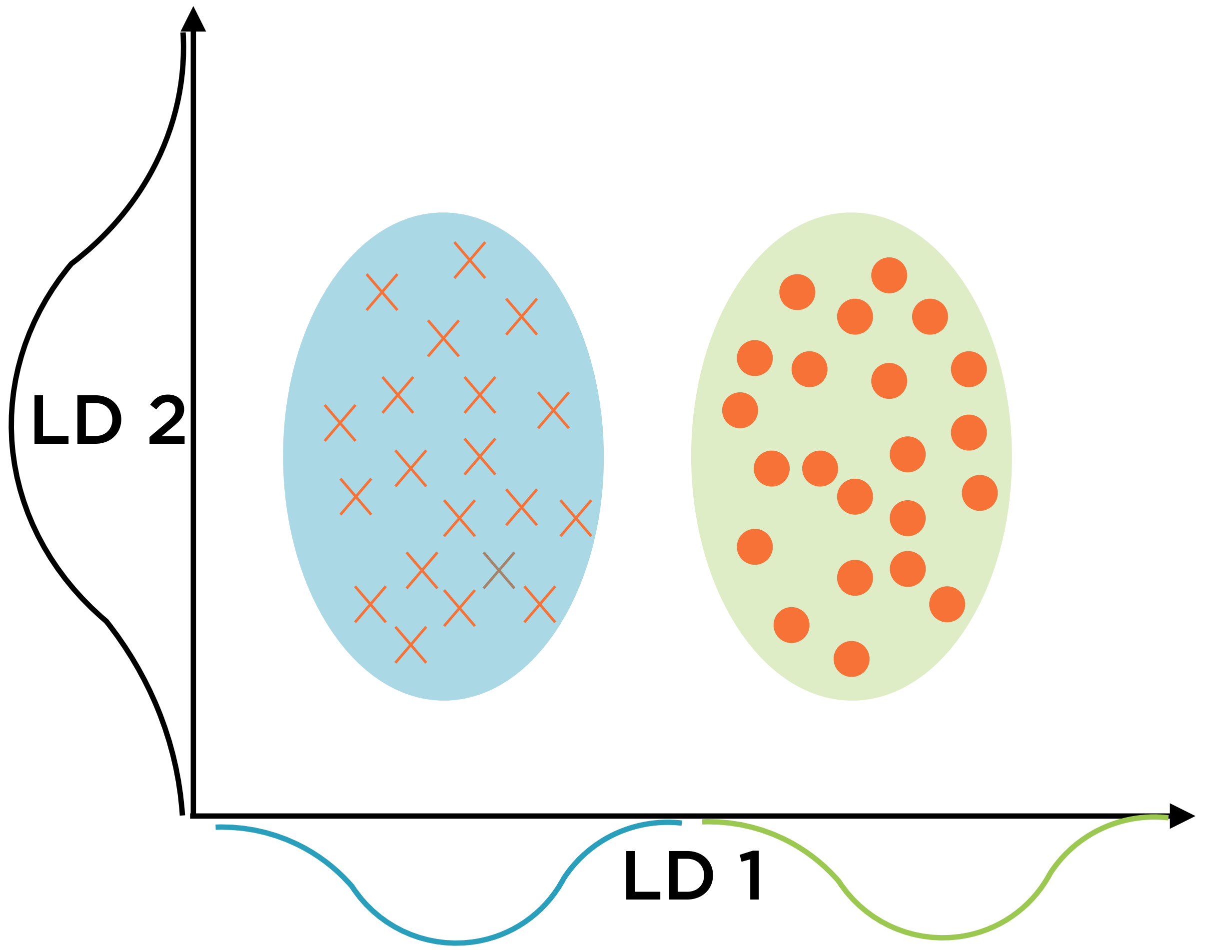
Use Case	Possible Solution
Large number of X-variables	
Most of which are meaningful	
Highly correlated to each other	Linear Discriminant Analysis
Linearly related to each other	
For use in classification	

PCA vs. LDA

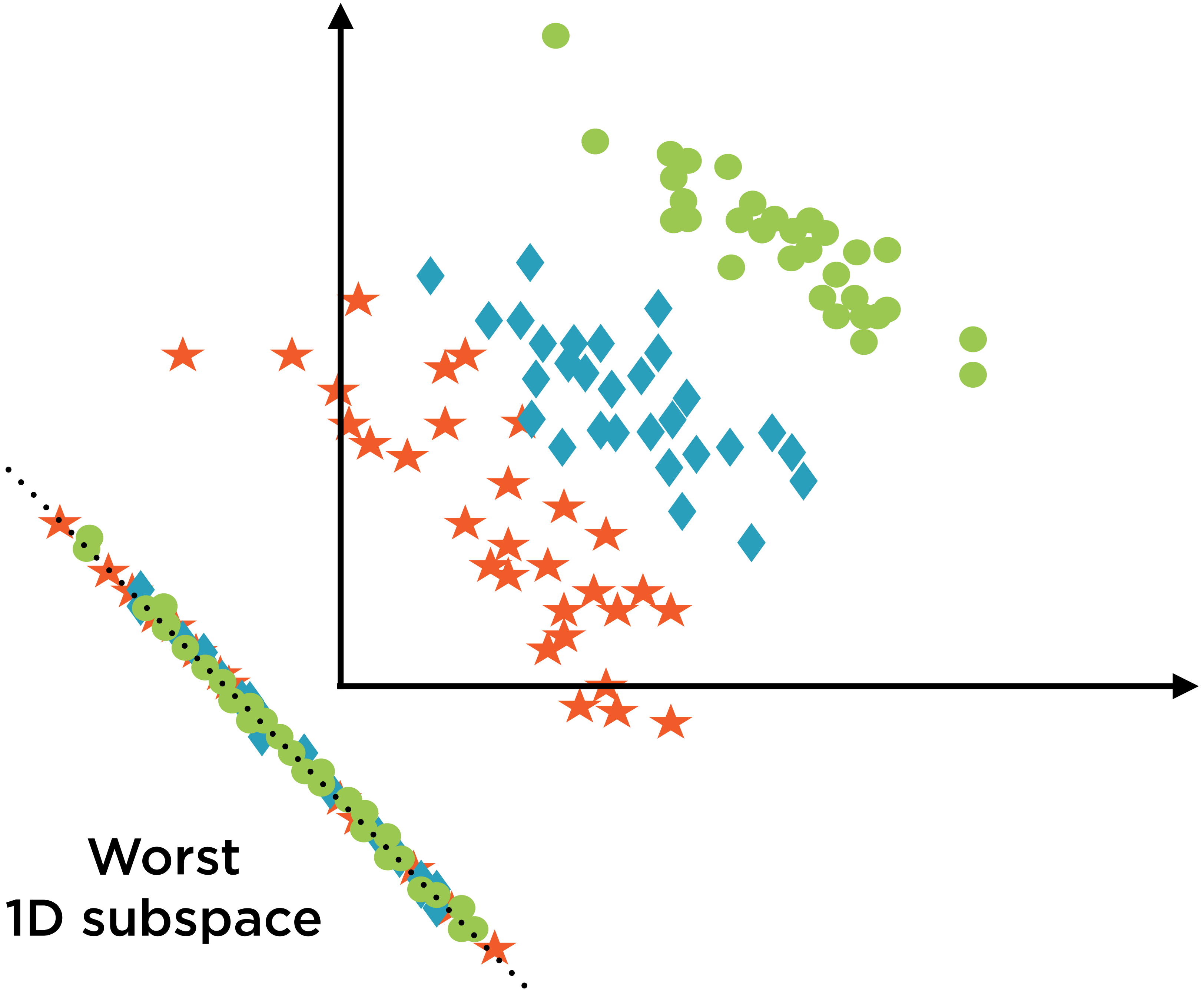
PCA



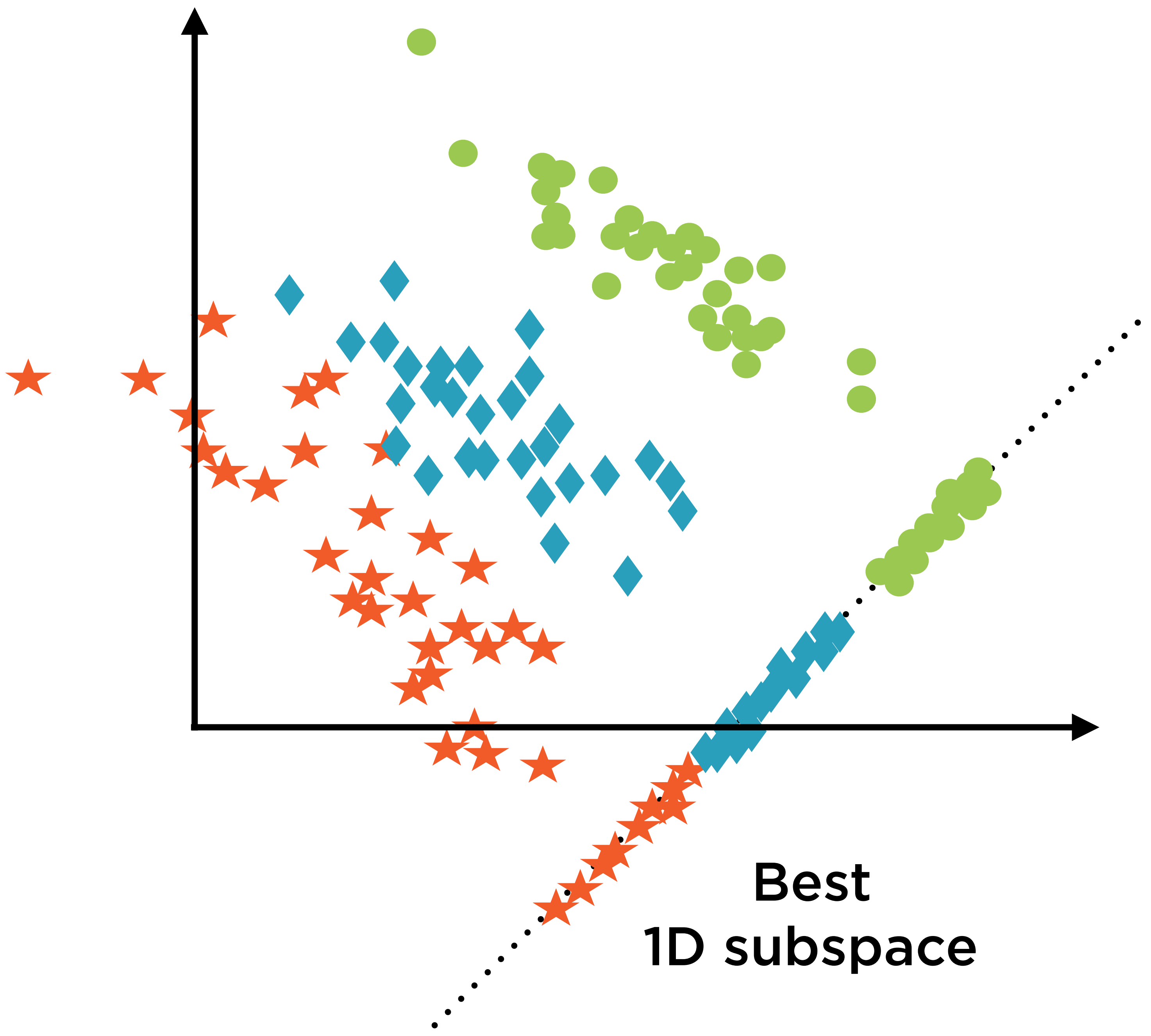
LDA



Choosing Axes



Choosing Axes



The scikit-learn LDA estimator object can be used for both dimensionality reduction as well as classification

Demo

Implement Linear Discriminant Analysis (LDA) to reduce dimensions

Summary

Dimensionality reduction using Principal Components Analysis (PCA)

Dimensionality reduction the Singular Value Decomposition method in Factor Analysis

Dimensionality reduction using Linear Discriminant Analysis