Reducing Complexity in Linear Data

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Principal Components Analysis (PCA)

Factor Analysis with Singular Value Decomposition (SVD)

Linear Discriminants Analysis (LDA)

Reducing Complexity

- Feature Selection
  - Statistical Techniques
    - Variance Thresholding
    - ANOVA
    - Mutual Information
- Dimensionality Reduction
  - Projection
  - Manifold Learning
  - Autoencoding

Reducing Complexity

Feature Selection

Dimensionality Reduction

Transform original X variables into new dimensions

Projection

Manifold Learning

Autoencoding

Reducing Complexity

Feature Selection

Dimensionality Reduction

Projection

Manifold Learning

Autoencoding

Find new, better axes and re-orient data

Reducing Complexity

Feature Selection

Dimensionality Reduction

Projection

Manifold Learning

Autoencoding

e.g. PCA, Factor Analysis, LDA, QDA

Principal Components Analysis

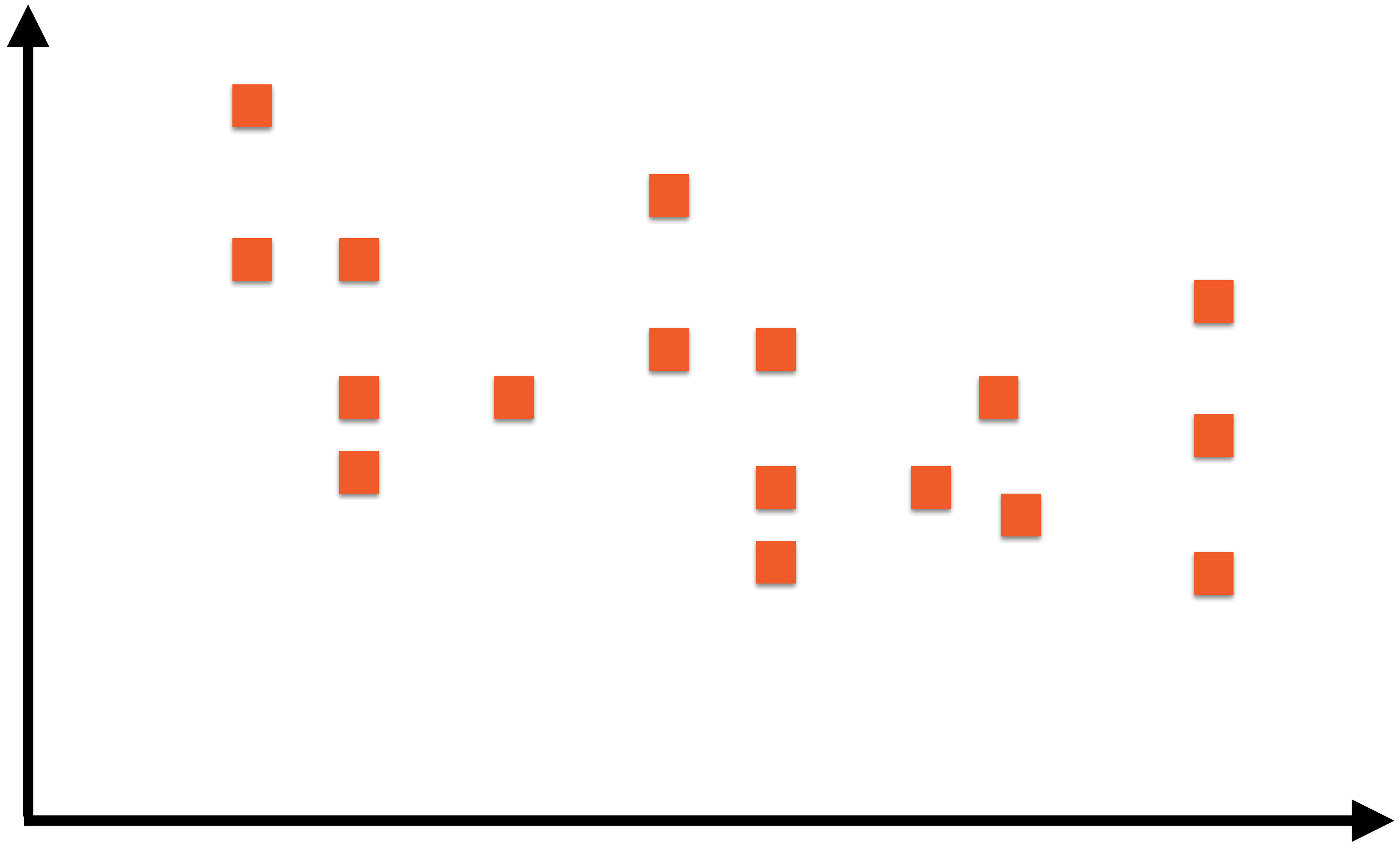
Choosing PCA and Factor Analysis

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	

Data in One Dimension

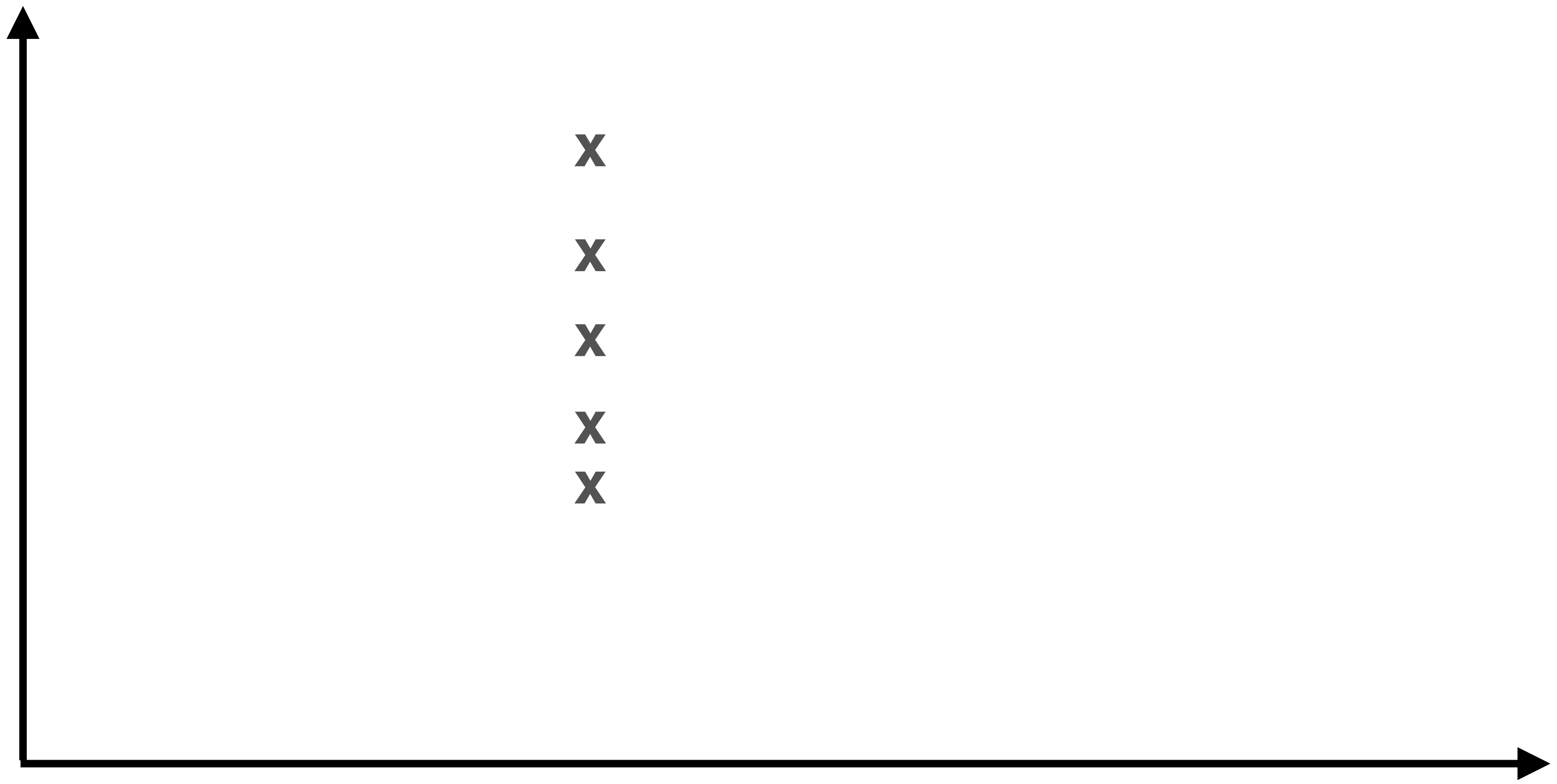
Unidimensional data points can be represented using a line, such as a number line

Data in Two Dimensions



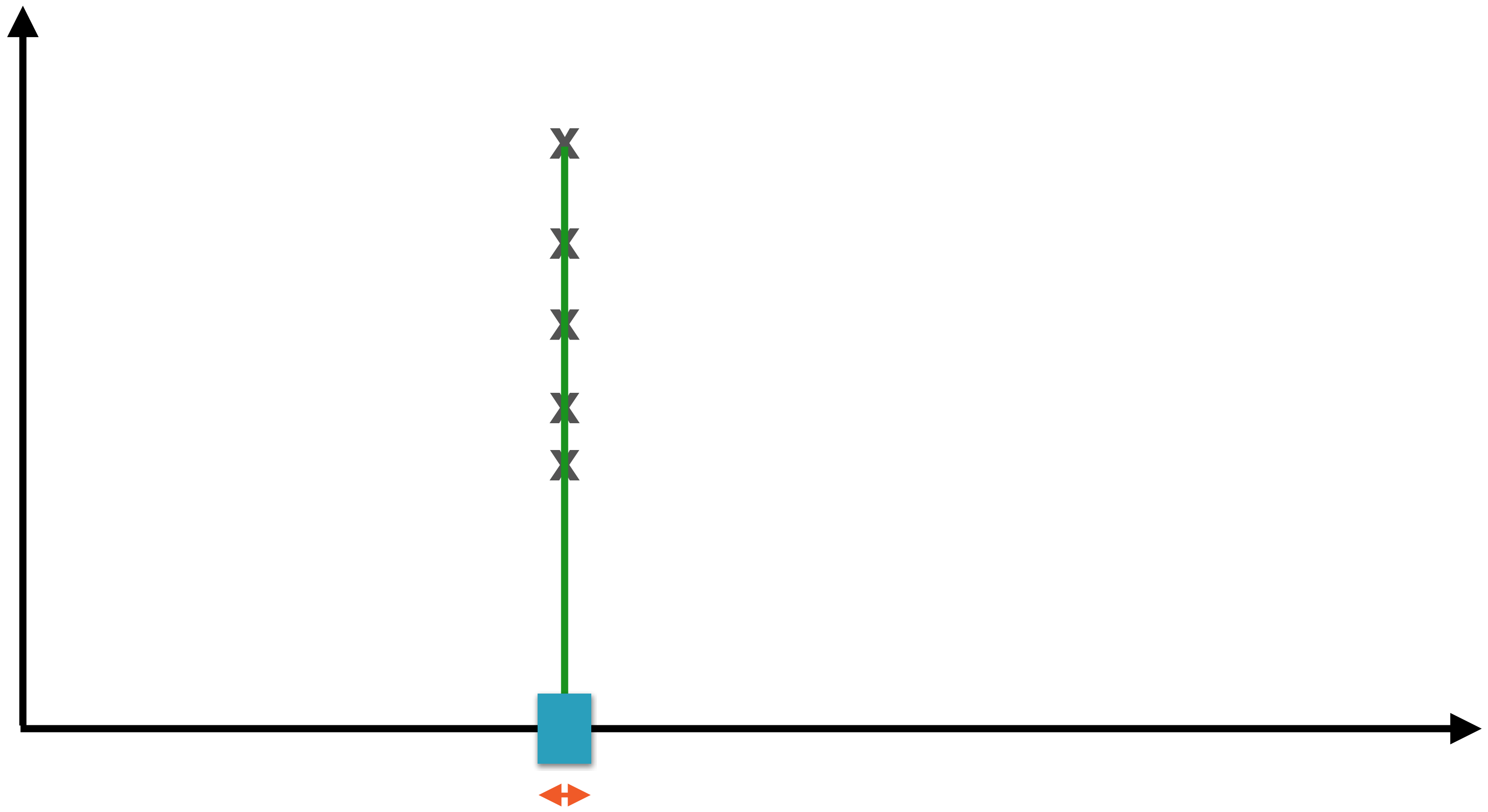
It's often more insightful to view data in relation to some other, related data

A Question of Dimensionality



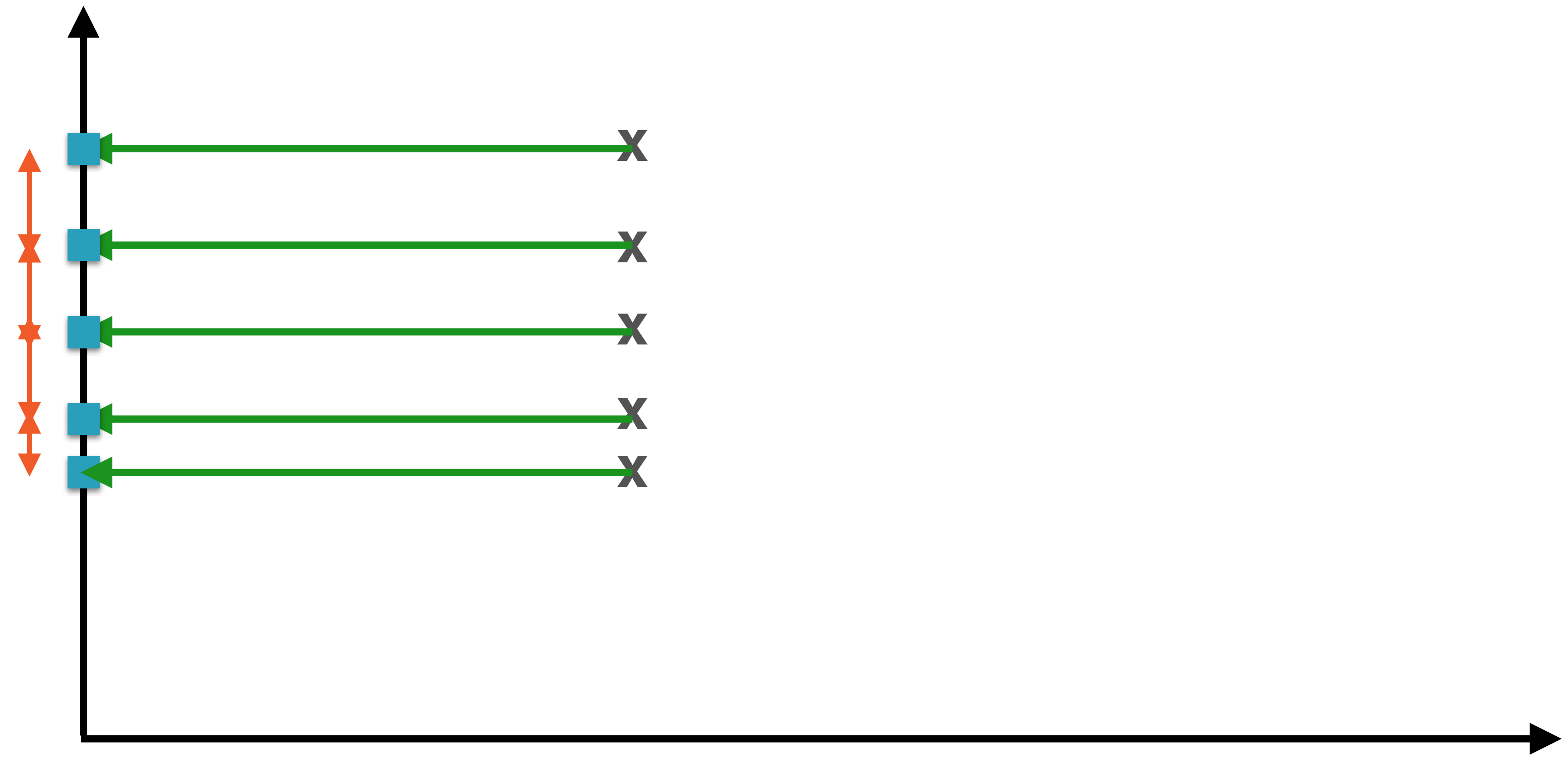
Pop quiz: Do we really need two dimensions to represent this data?

Bad Choice of Dimensions



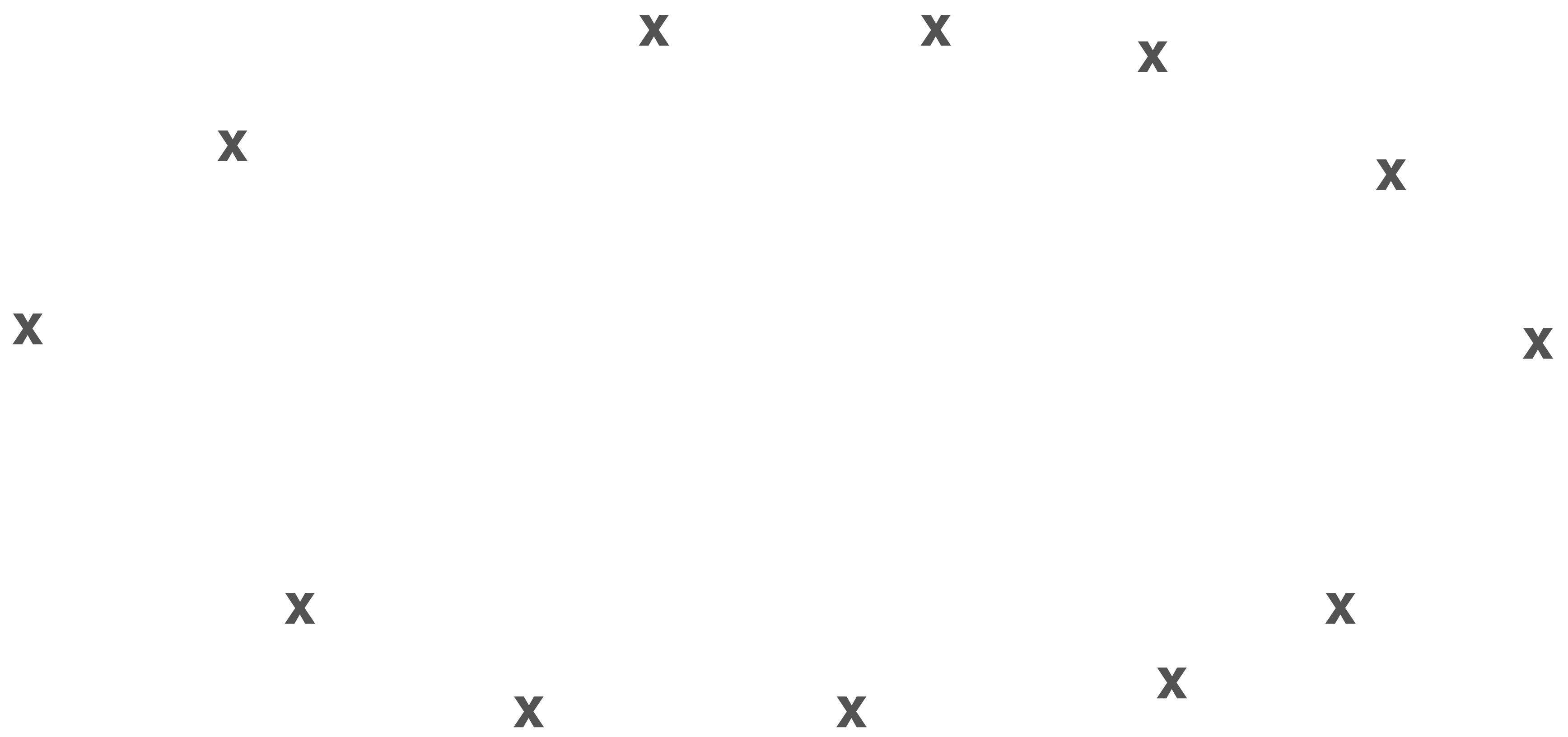
If we choose our axes (dimensions) poorly then we do need two dimensions

Good Choice of Dimensions



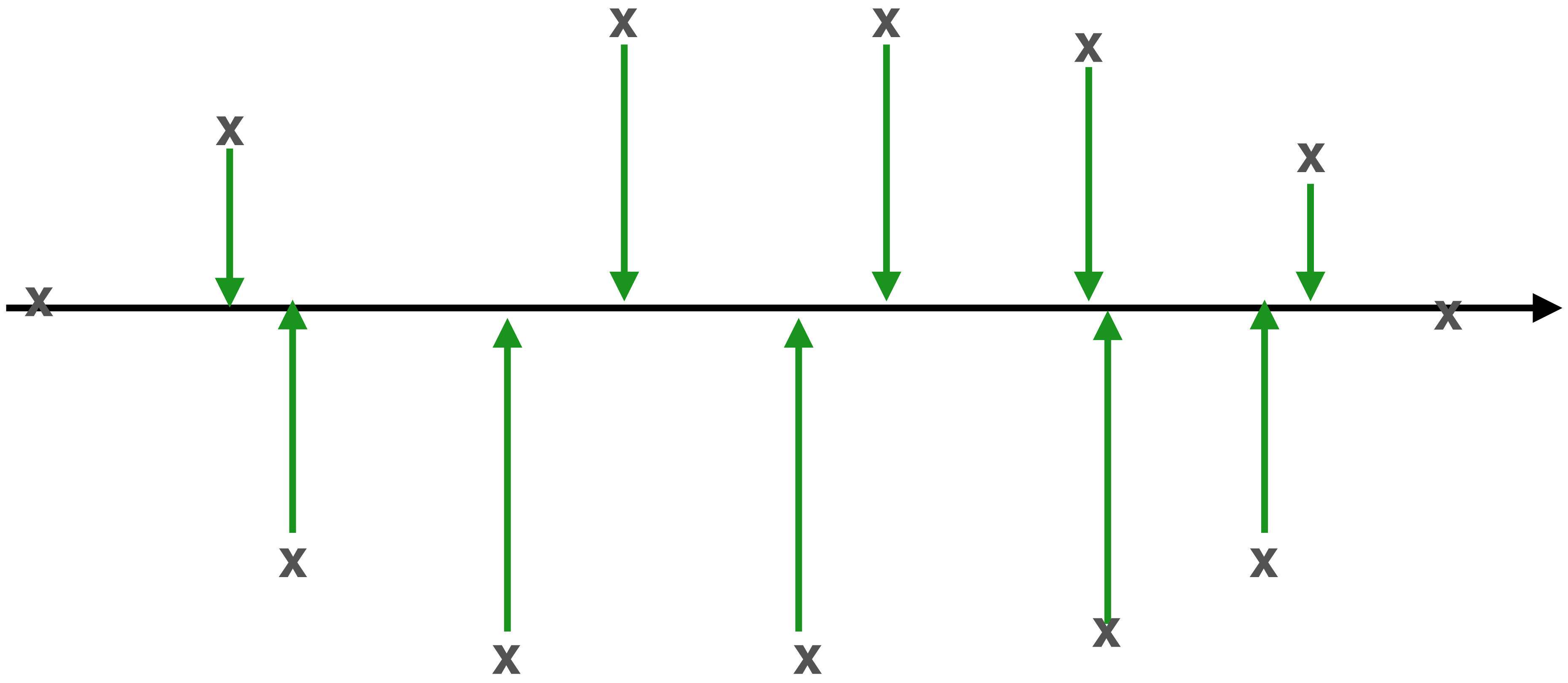
If we choose our axes (dimensions) well then one dimension is sufficient

Intuition Behind PCA



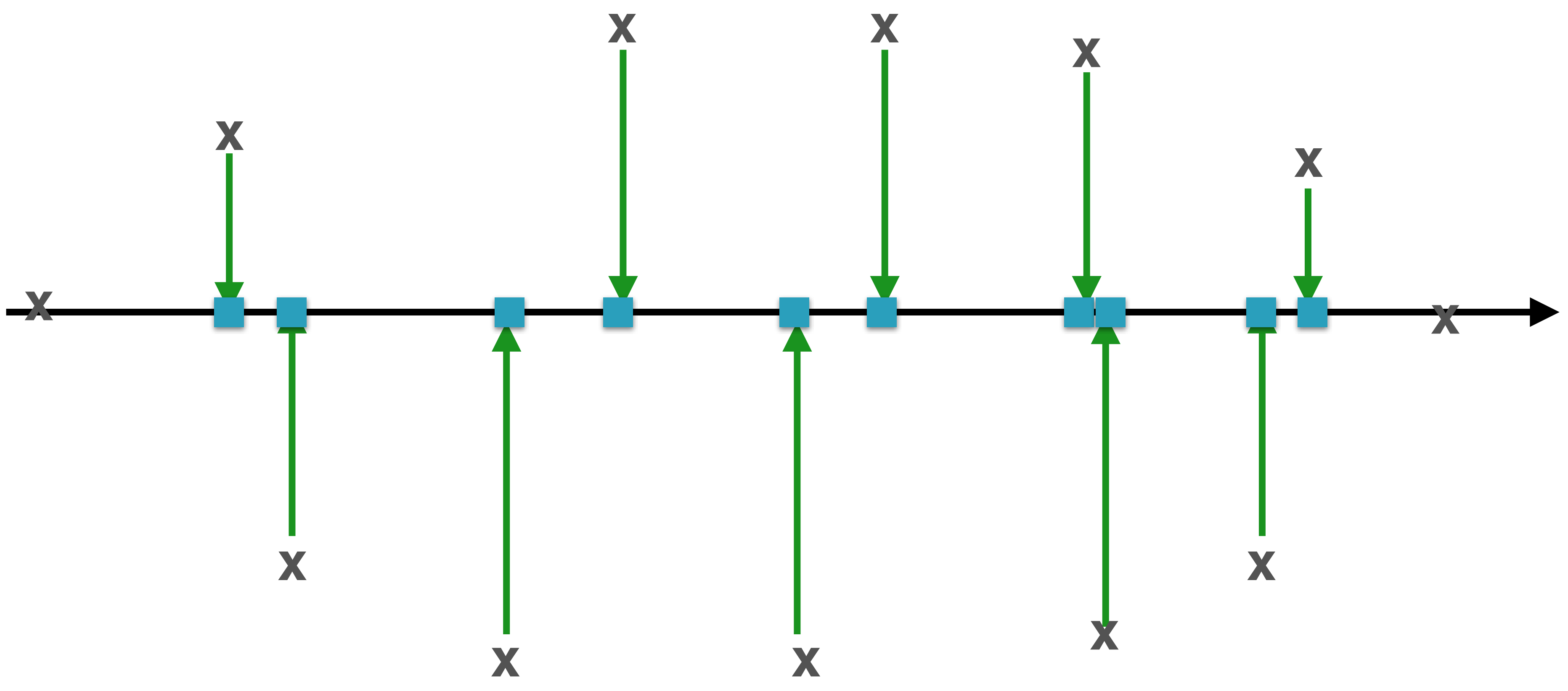
Objective: Find the “best” directions to represent this data

Intuition Behind PCA



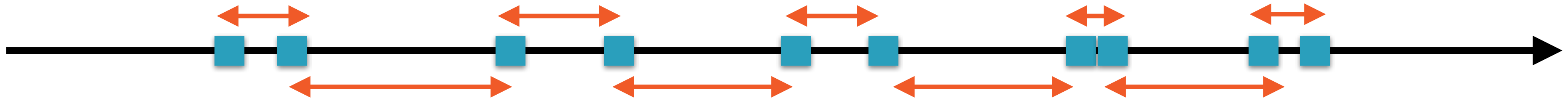
Start by "projecting" the data onto a line in some direction

Intuition Behind PCA



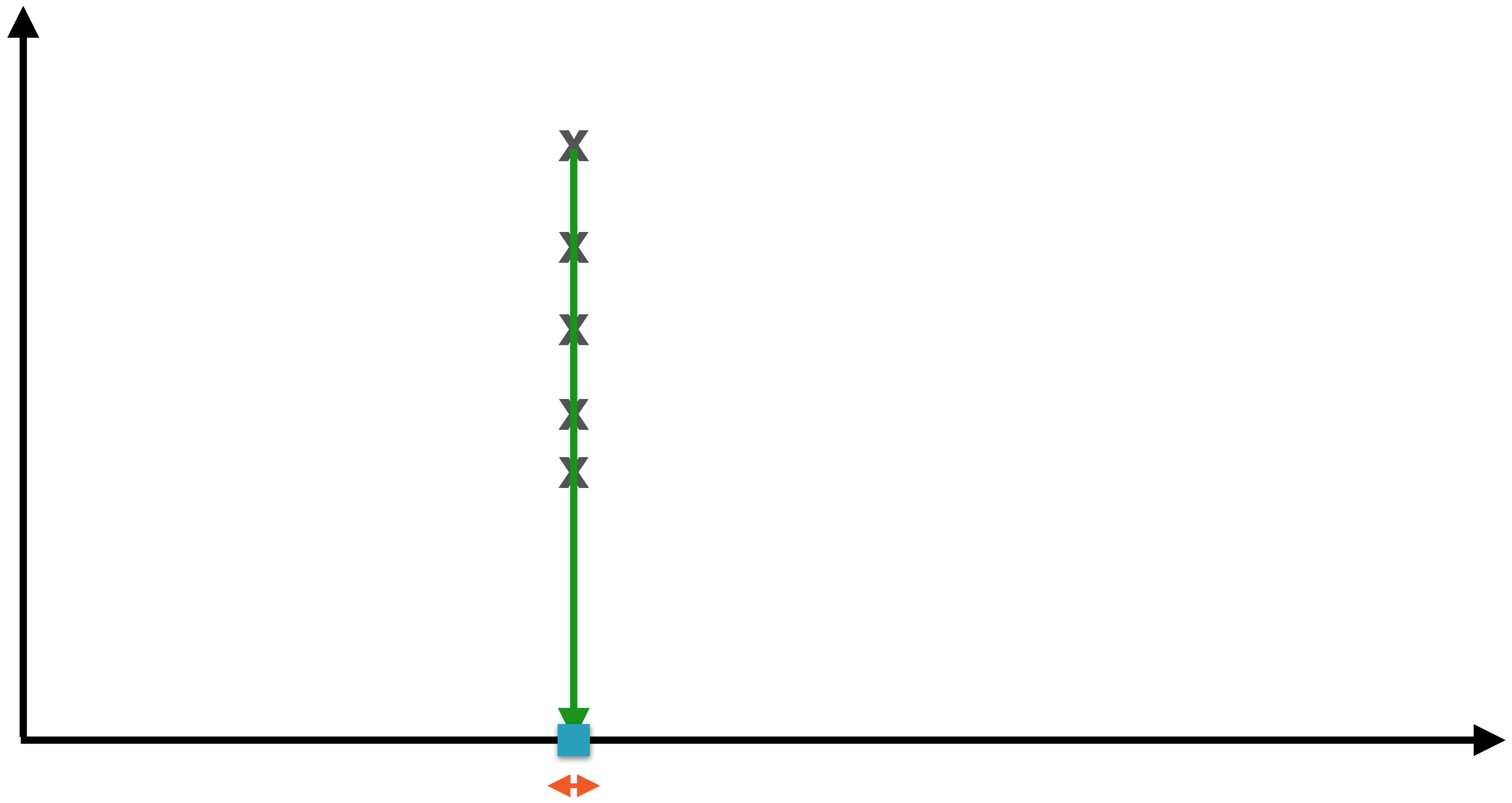
Start by "projecting" the data onto a line in some direction

Intuition Behind PCA



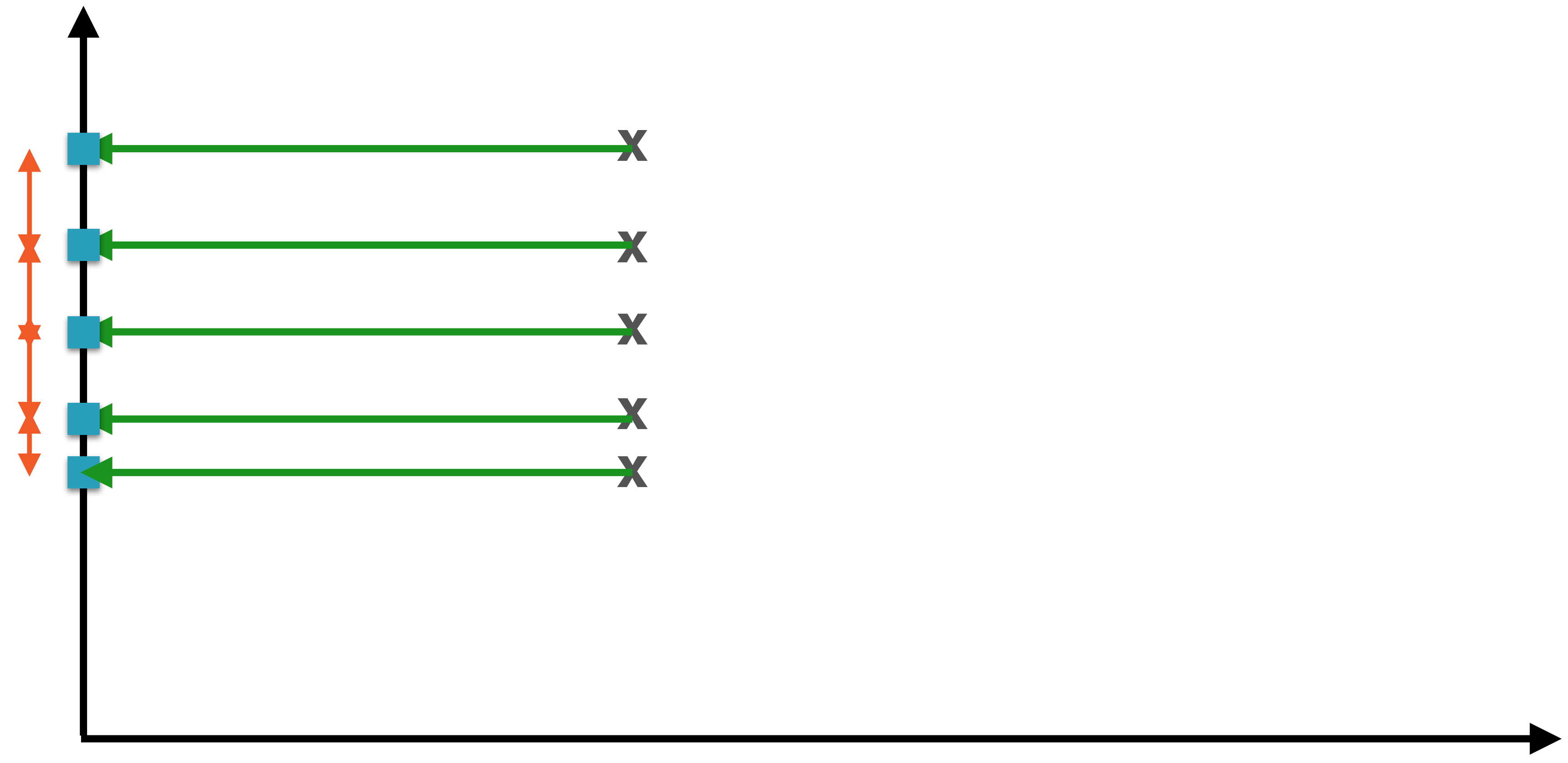
The greater the distances between these projections, the “better” the direction

Bad Projection



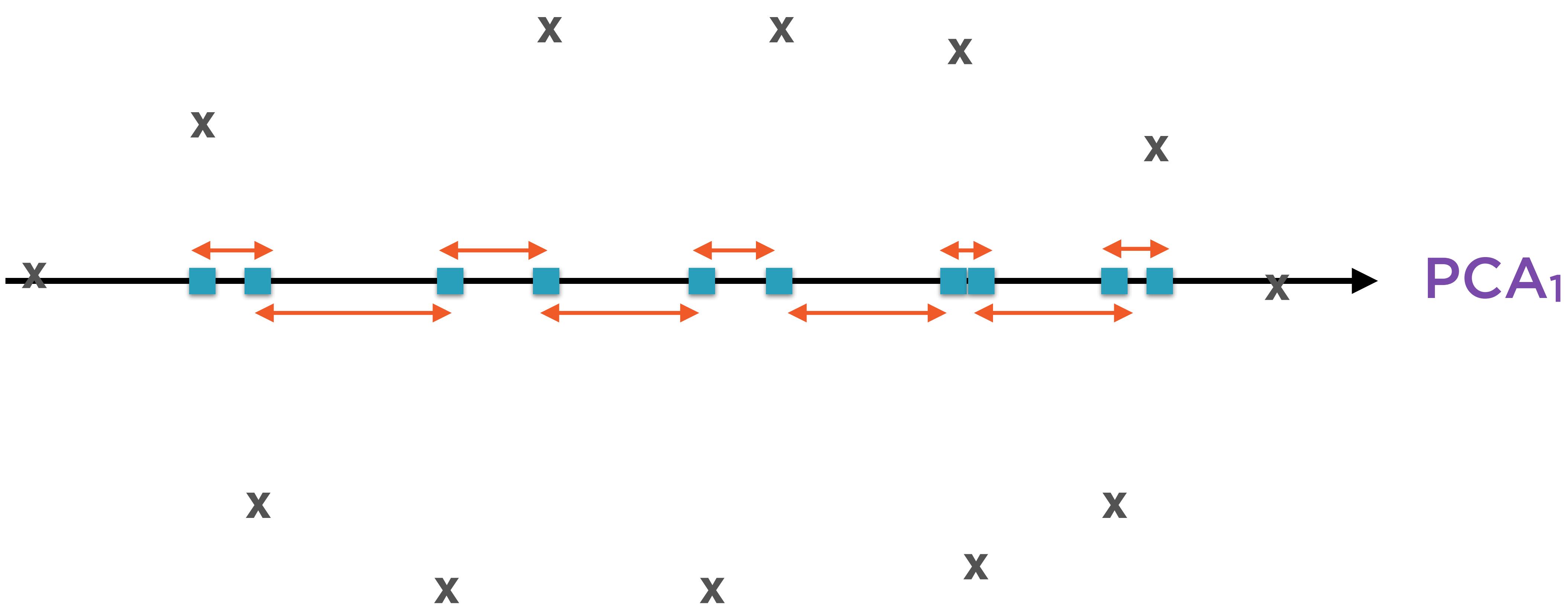
A projection where the distances are minimized is a bad one - information is lost

Good Projection



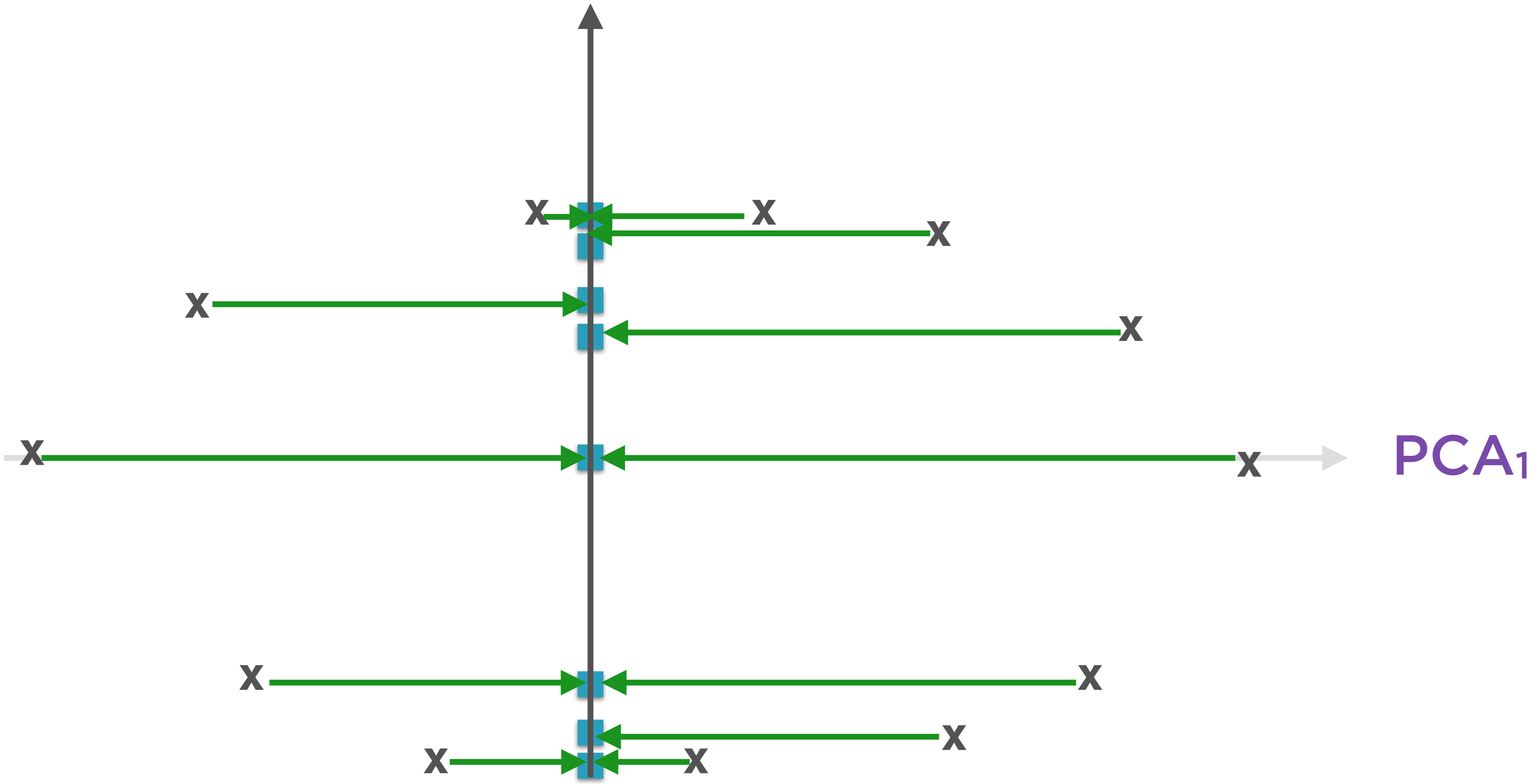
A projection where the distances are maximized is a good one - **information is preserved**

Intuition Behind PCA



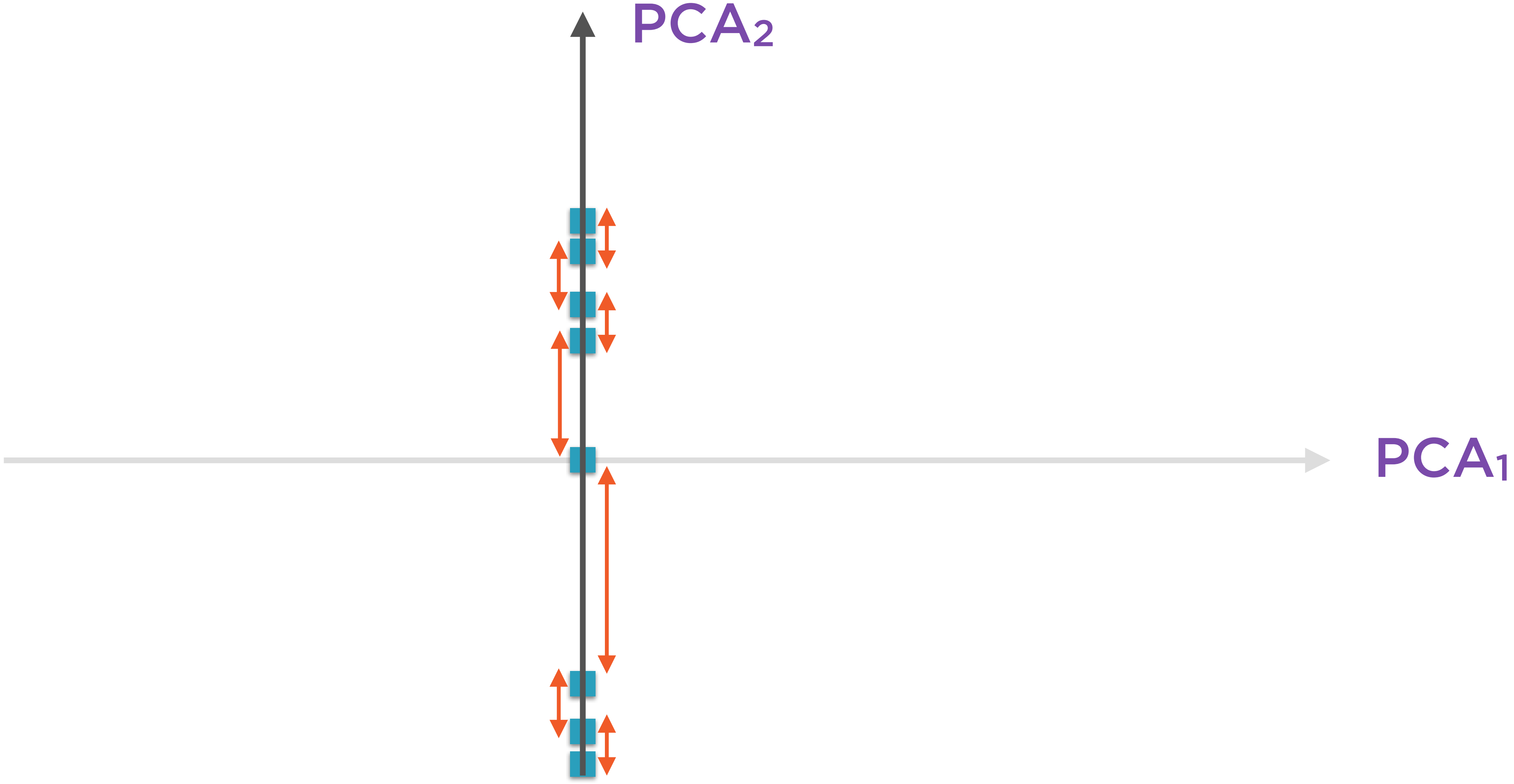
The direction along which this variance is maximized is the **first principal component** of the original data

Intuition Behind PCA



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

Intuition Behind PCA

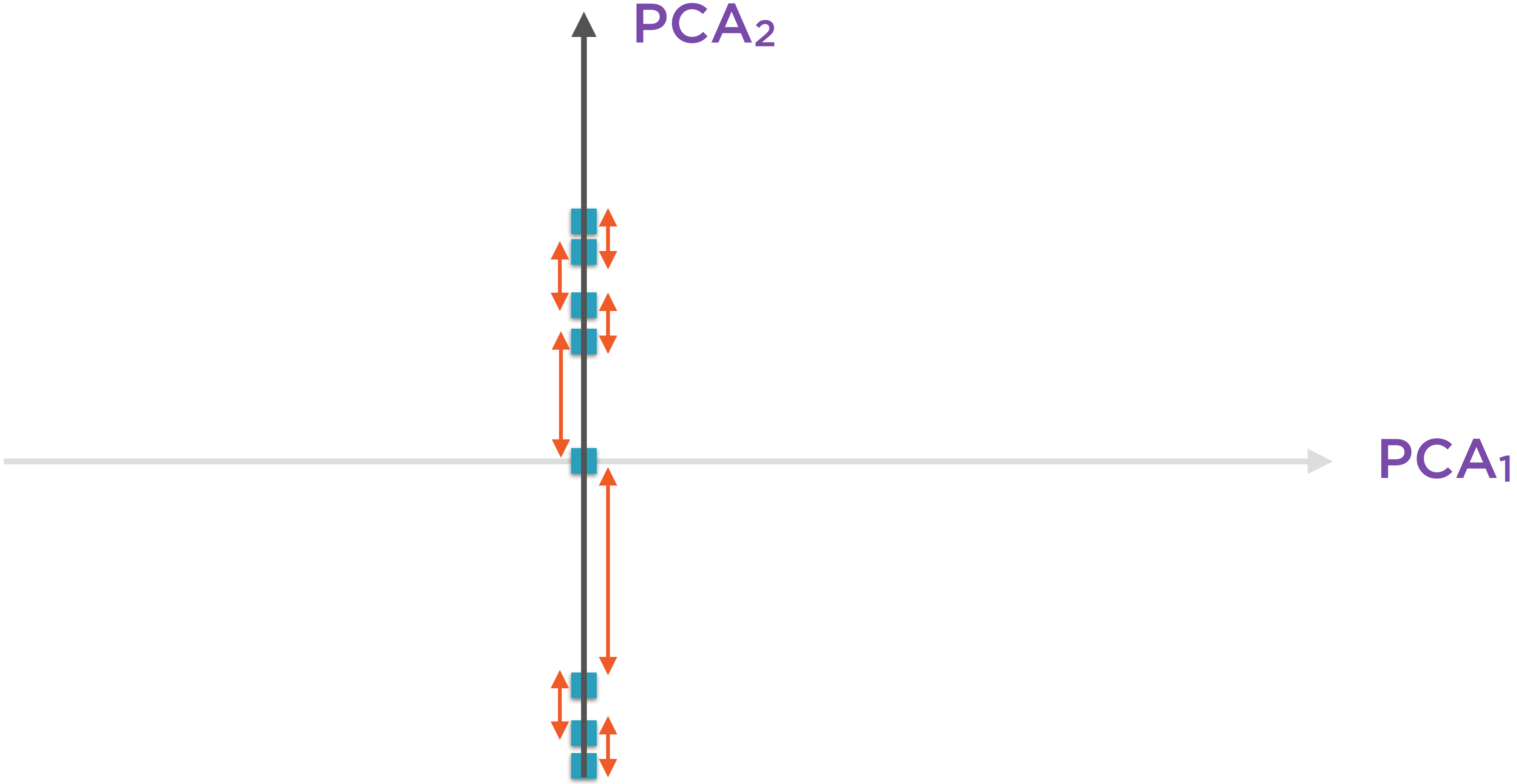


Find the next best direction, the **second principal component**, which must be at right angles to the first

Principal Components at Right Angles

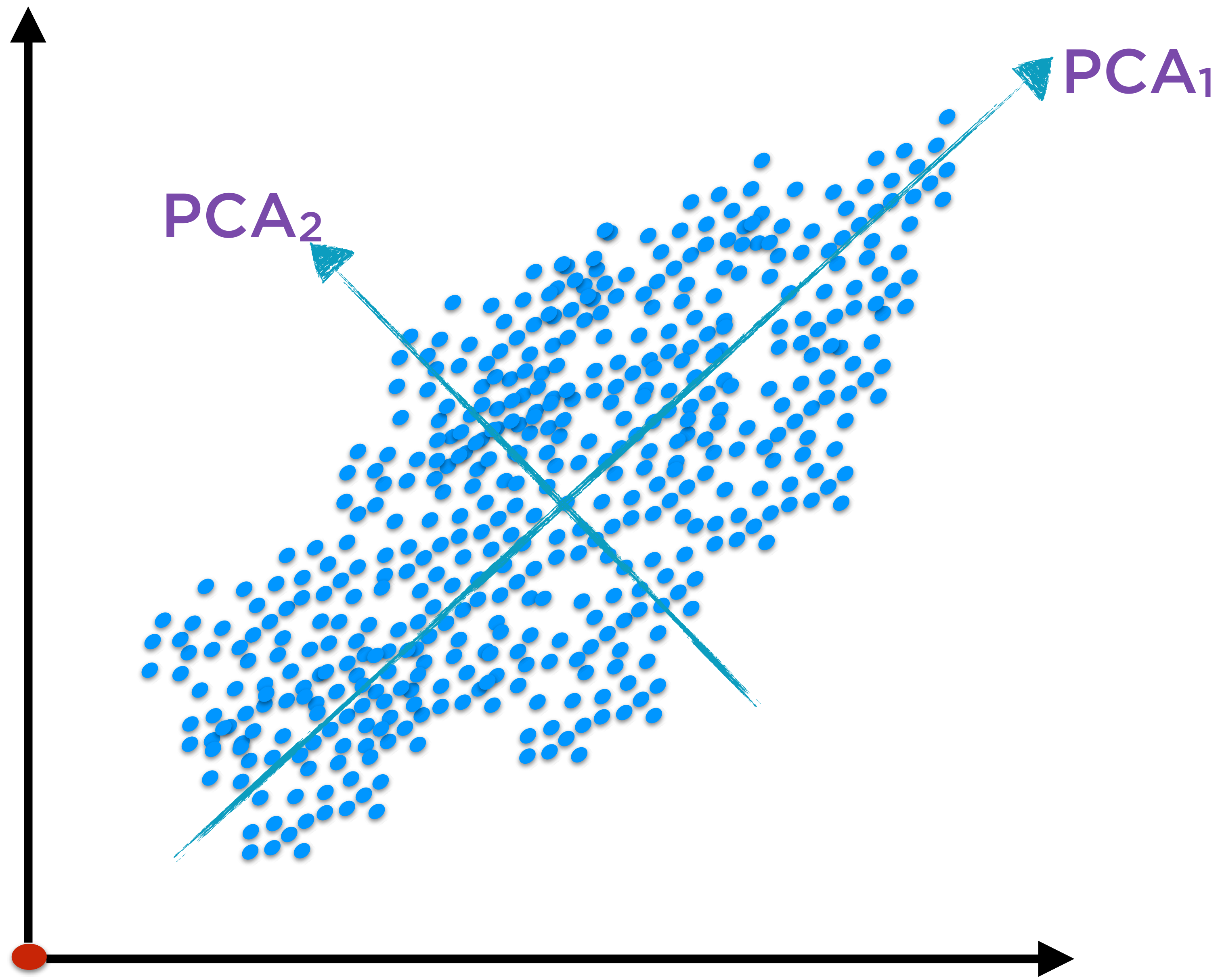
Directions at right angles help express the most variation with the smallest number of directions

Intuition Behind PCA



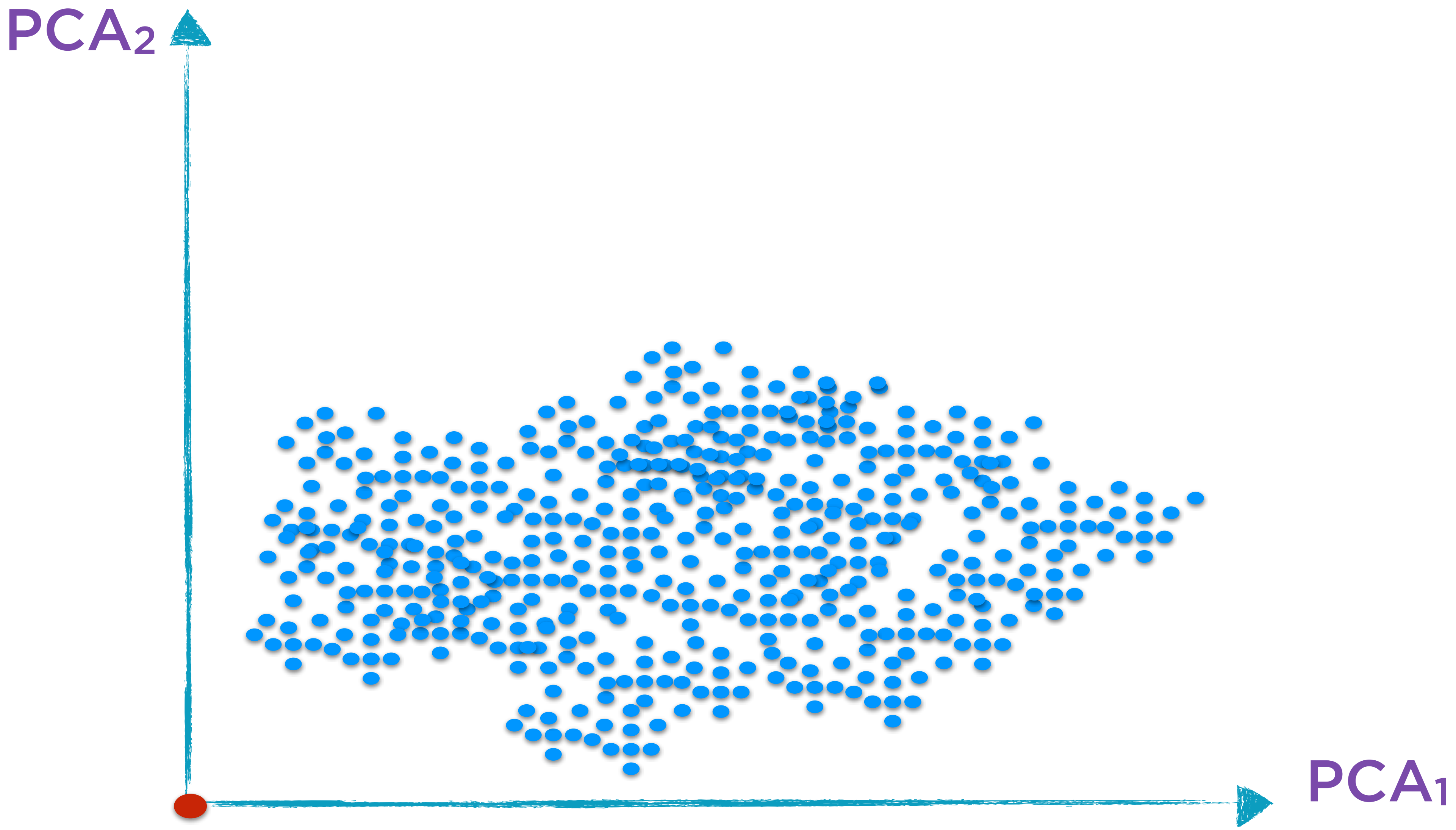
The variances are clearly smaller along this **second principal component** than along the first

Intuition Behind PCA



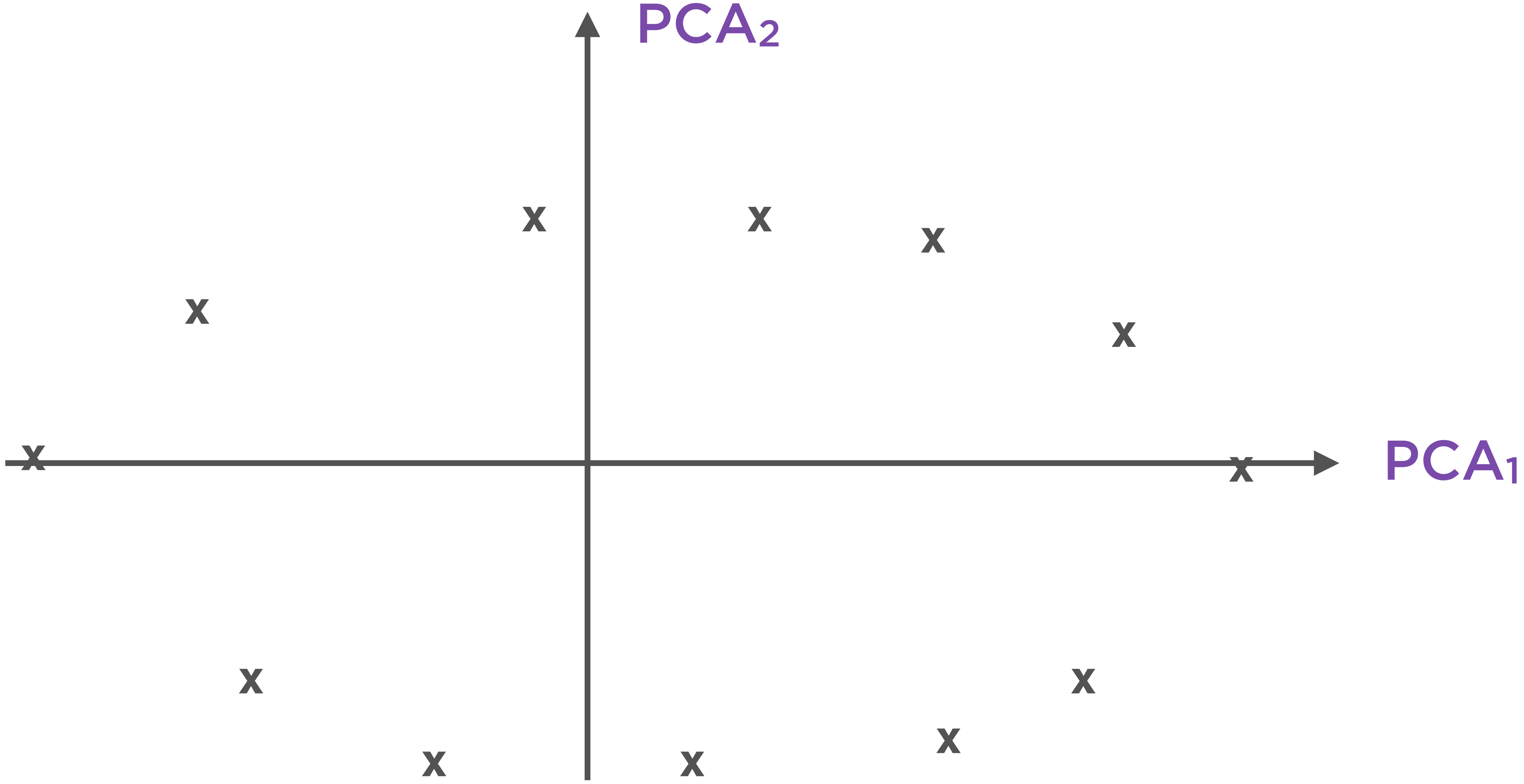
In general, there are as many principal components as there are dimensions in the original data

Intuition Behind PCA



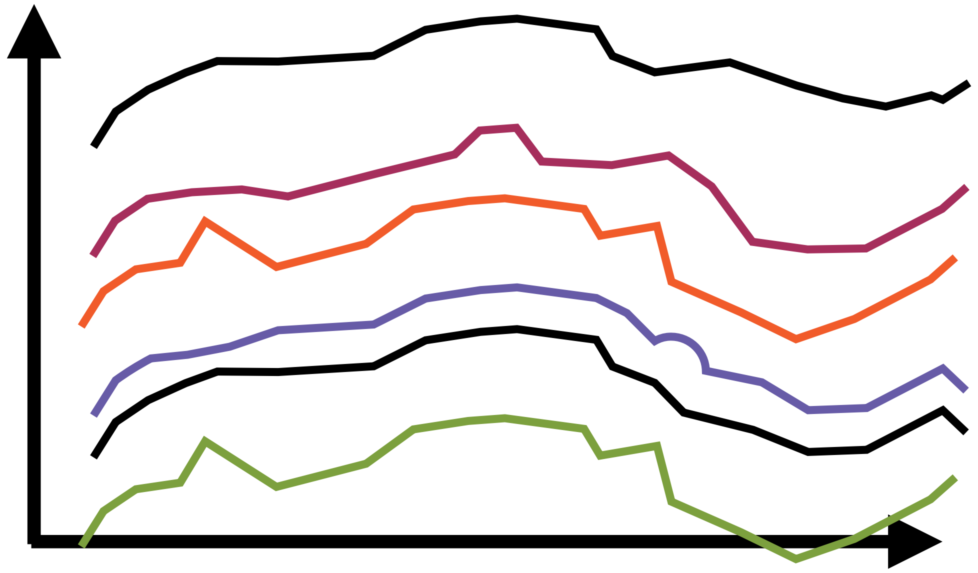
Re-orient the data along these new axes

Dimensionality Reduction

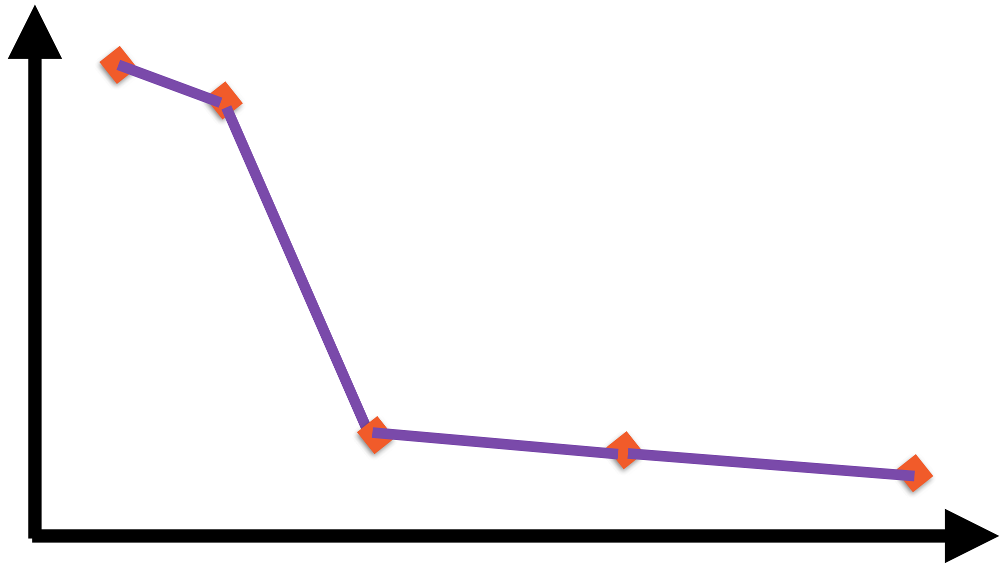


If the **variance** along the second principal component is small enough, we can just **ignore** it and use just 1 dimension to represent the data

PCA's Forte

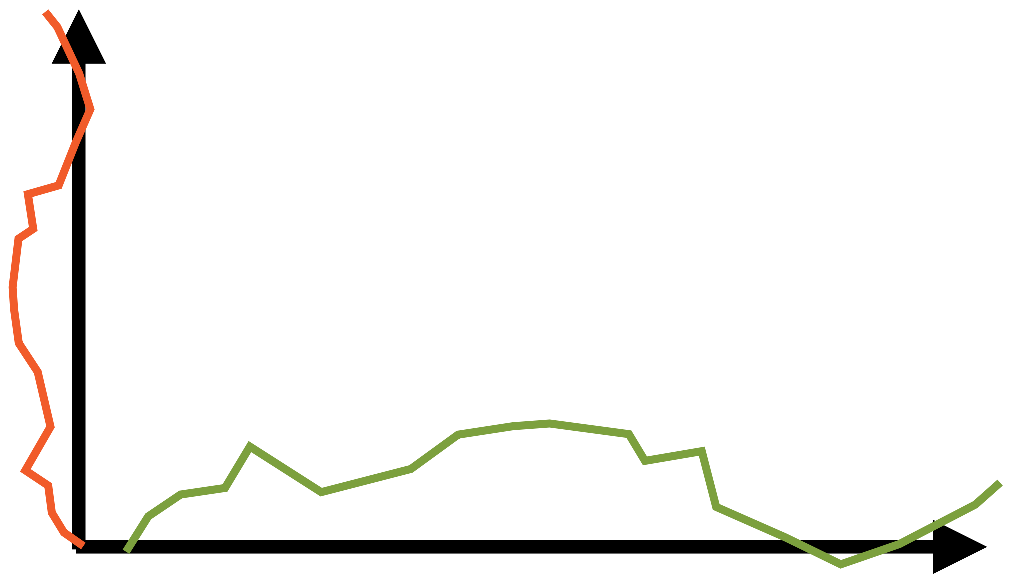


Many, highly correlated X variables

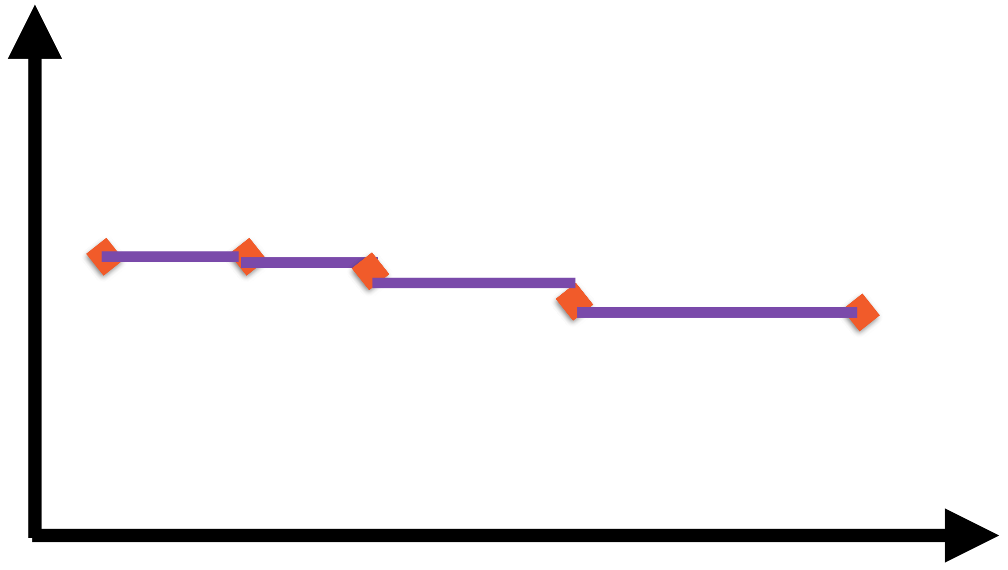


Unequal explained variance ratios

PCA's Weak Spots



Few, uncorrelated X variables



Almost equal explained variance ratios

Demo

Implement PCA for dimensionality reduction in linear regression

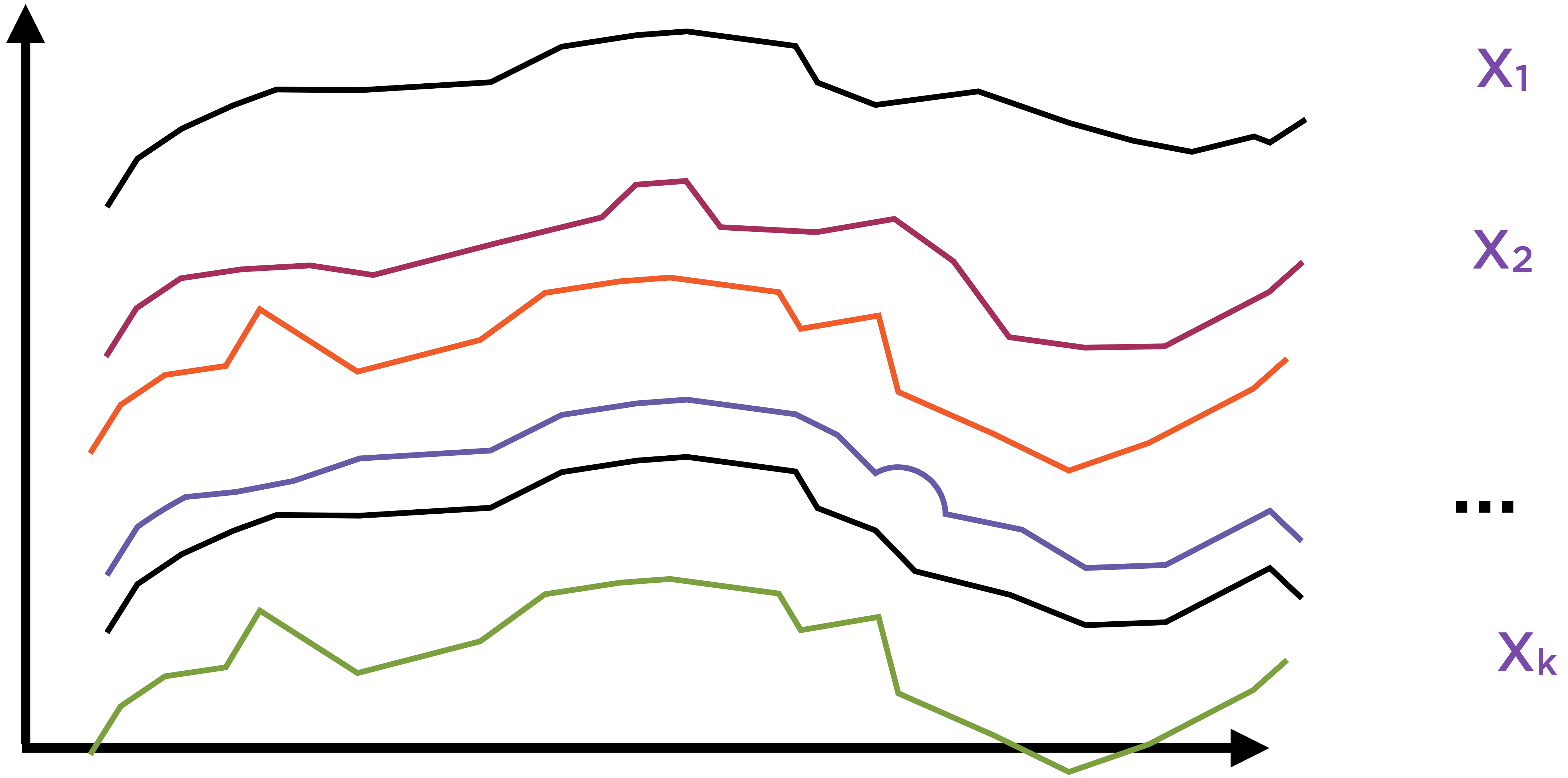
Factor Analysis

PCA is one specific implementation of Factor Analysis; a common alternative is to use a procedure named SVD

SVD Factor Analysis

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

Correlated Random Variables



Highly correlated variables are not suitable for use in regression

Correlated Random Variables

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

n rows

k columns

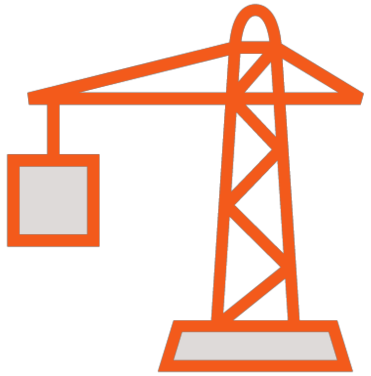
SVD, like PCA is used when the elements X_i of this matrix are highly correlated with each other

Factor Analysis

k columns

$[\ X_1 \quad X_2 \quad X_3 \quad \dots \quad X_k\ ]$ n rows

X_i are highly correlated with each other



Factor Analysis: SVD, PCA, etc

F_i are completely uncorrelated with each other

$[\ F_1 \quad F_2 \quad F_3 \quad \dots \quad F_k\ ]$ n rows

k columns

Factor Analysis

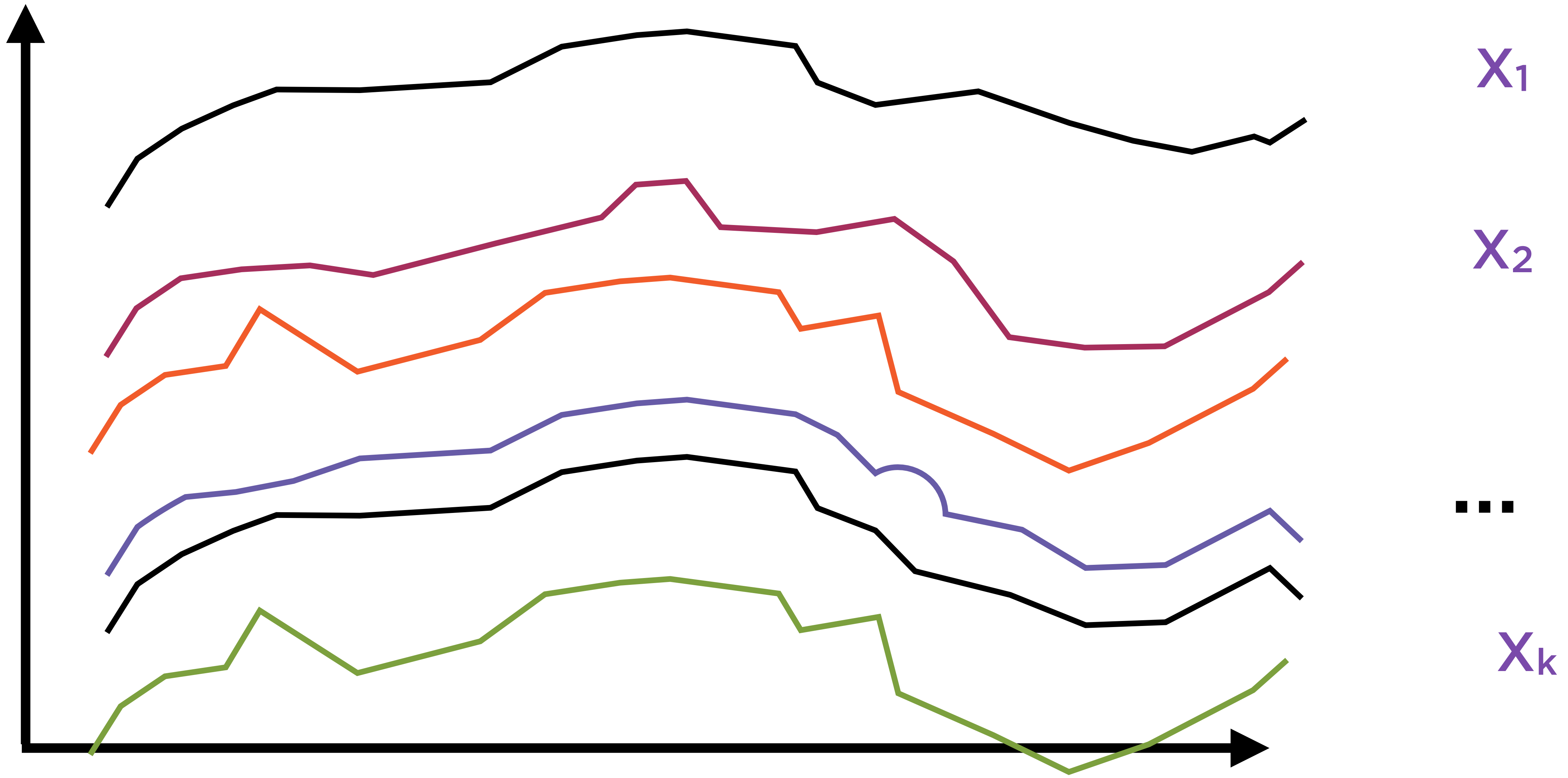
$$[\ F_1 \quad F_2 \quad F_3 \quad \cdots \quad F_k\ ]$$

n rows

k columns

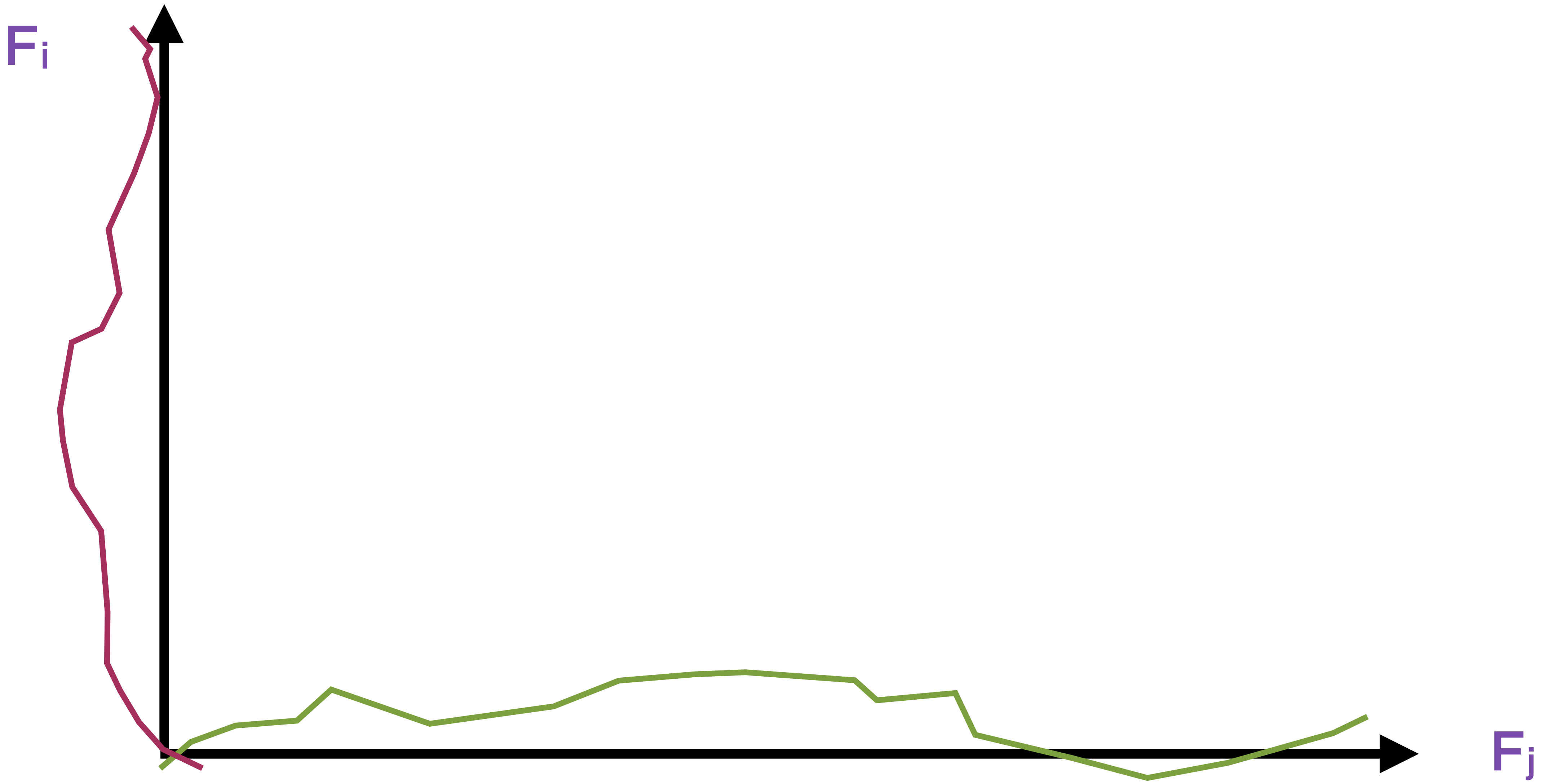
These vectors F_i are the factor representations of the original vectors X_i

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 for dimensionality reduction in classification

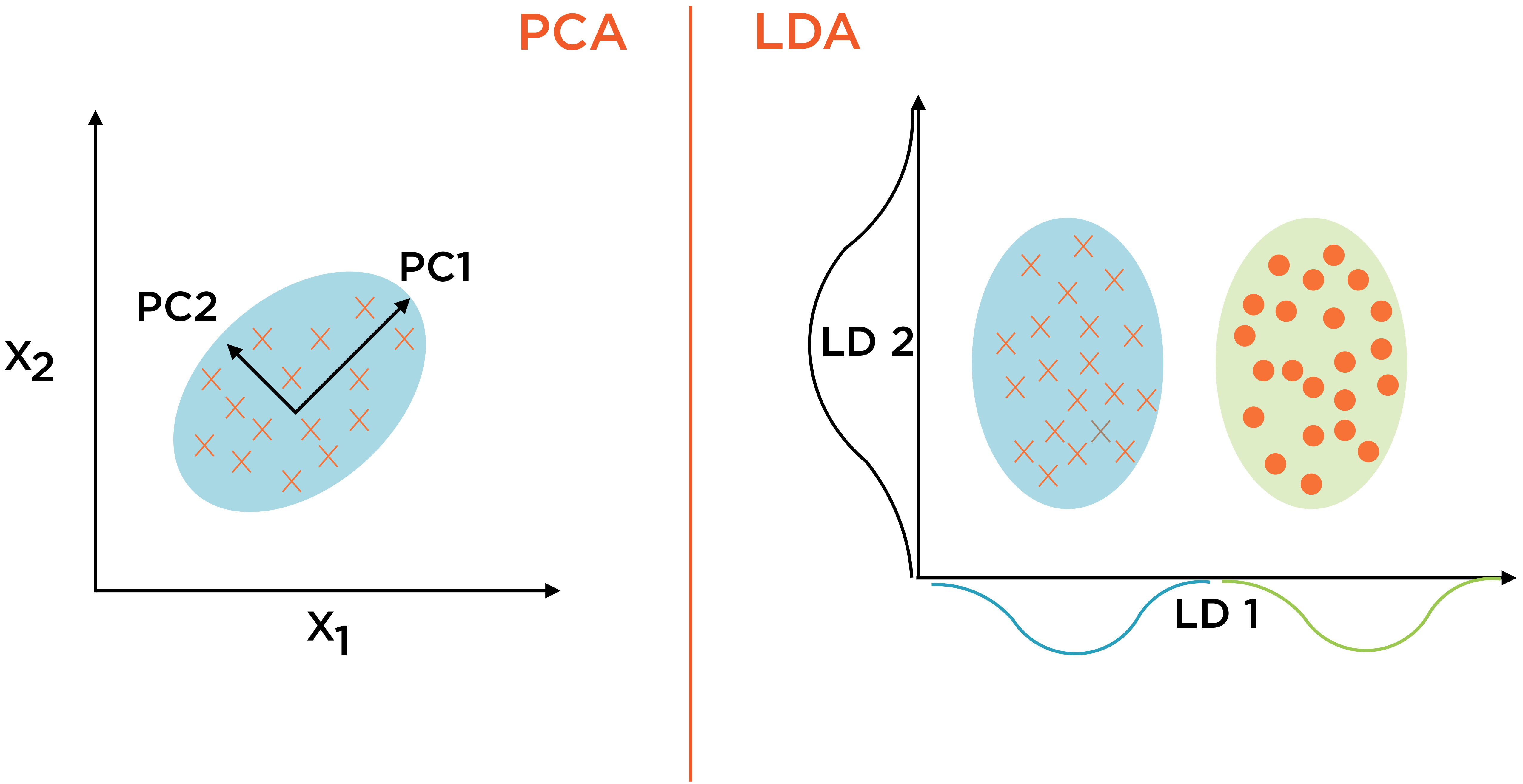
Linear Discriminant Analysis

Choosing PCA and Factor Analysis

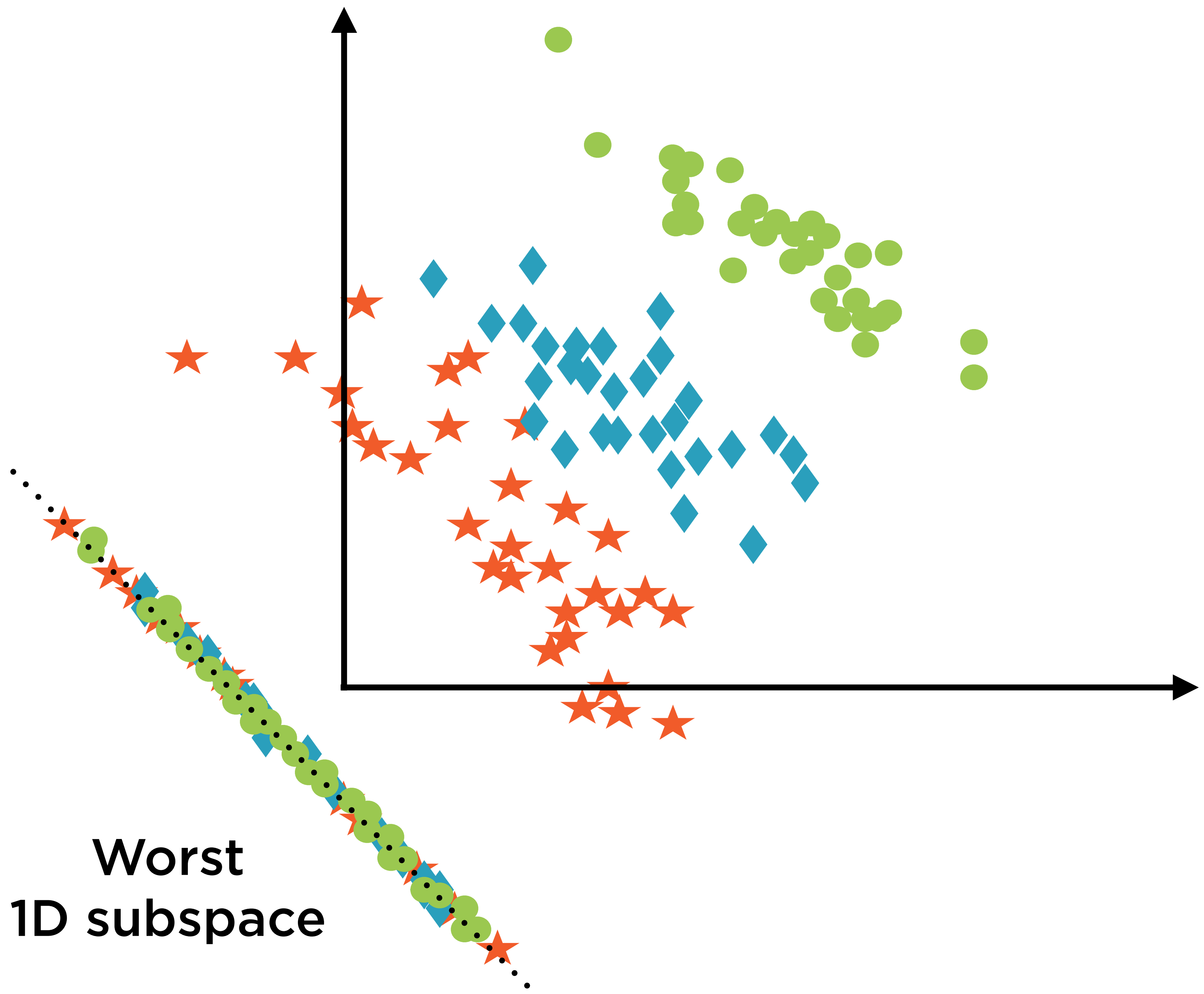
Use Case	Possible Solution
Large number of X-variables	
Most of which are meaningful	
Highly correlated to each other	Linear Discriminant Analysis (LDA) or Dictionary Learning
Linearly related to each other	
For use in classification	

LDA is similar to PCA, but chooses axis to maximize distance between points of different categories

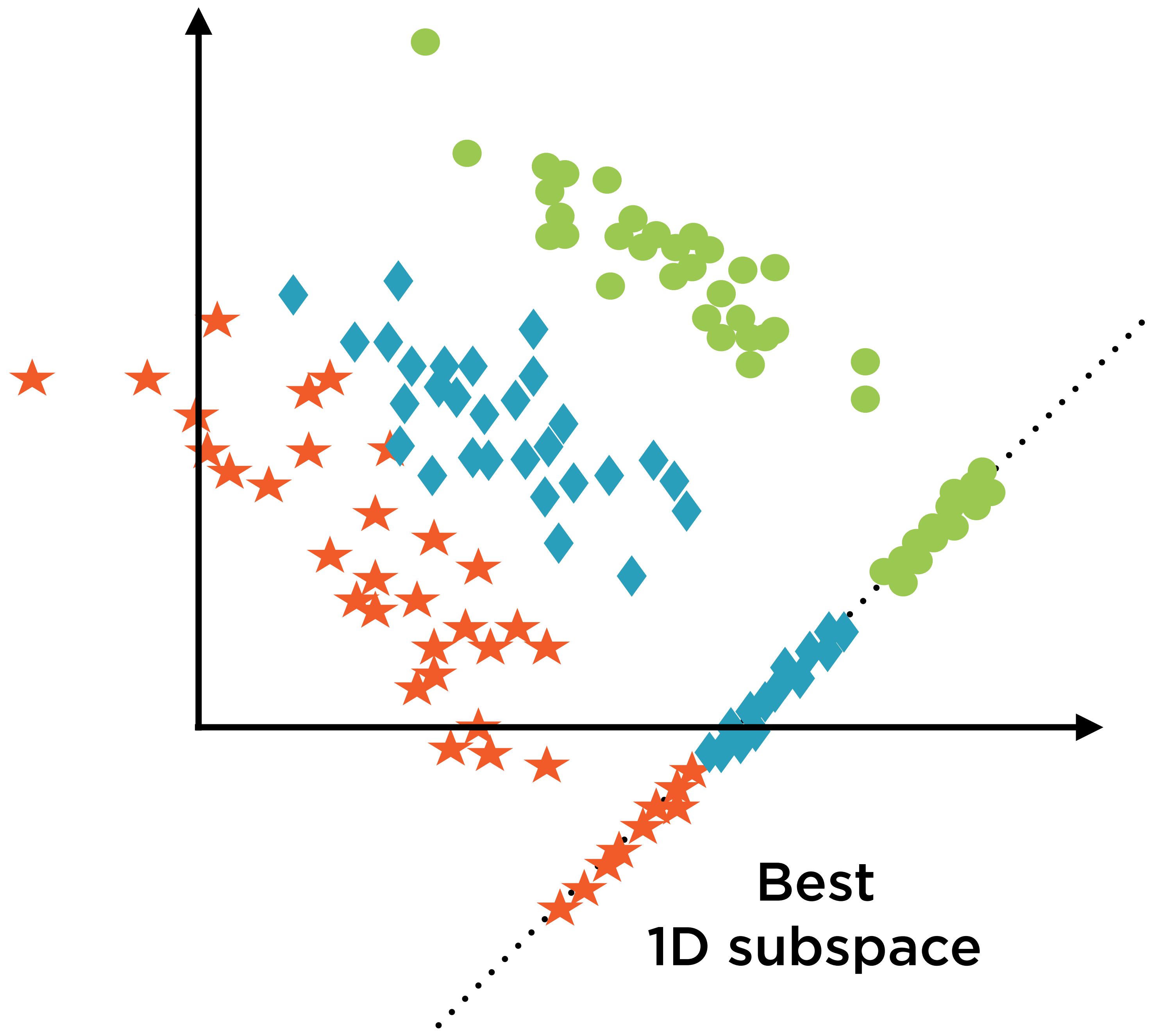
PCA vs. LDA



Choosing Axes



Choosing Axes



Demo

Implement Linear Discriminant Analysis (LDA)

Summary

Principal Components Analysis (PCA)

Factor Analysis with Singular Value Decomposition (SVD)

Linear Discriminants Analysis (LDA)