Understanding Factor Analysis and PCA

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Understand eigenvalue decomposition, a technique that underpins PCA

Calculate the principal components which explain all the variance in data

Apply PCA to dimensionality reduction and latent factor identification

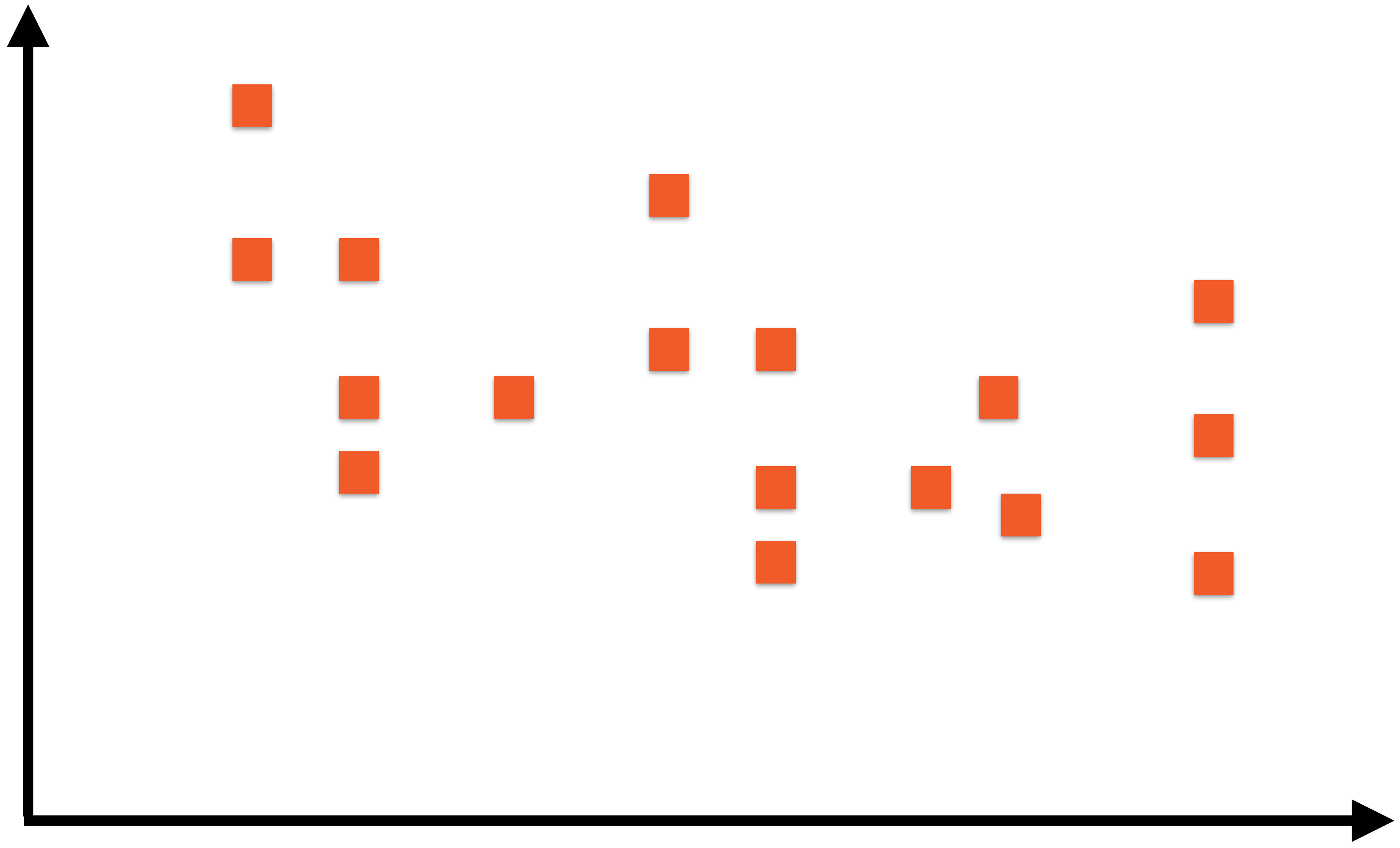
Introduce and contrast exploratory and confirmatory factor analysis

The Intuition Behind Principal Components

Data in One Dimension

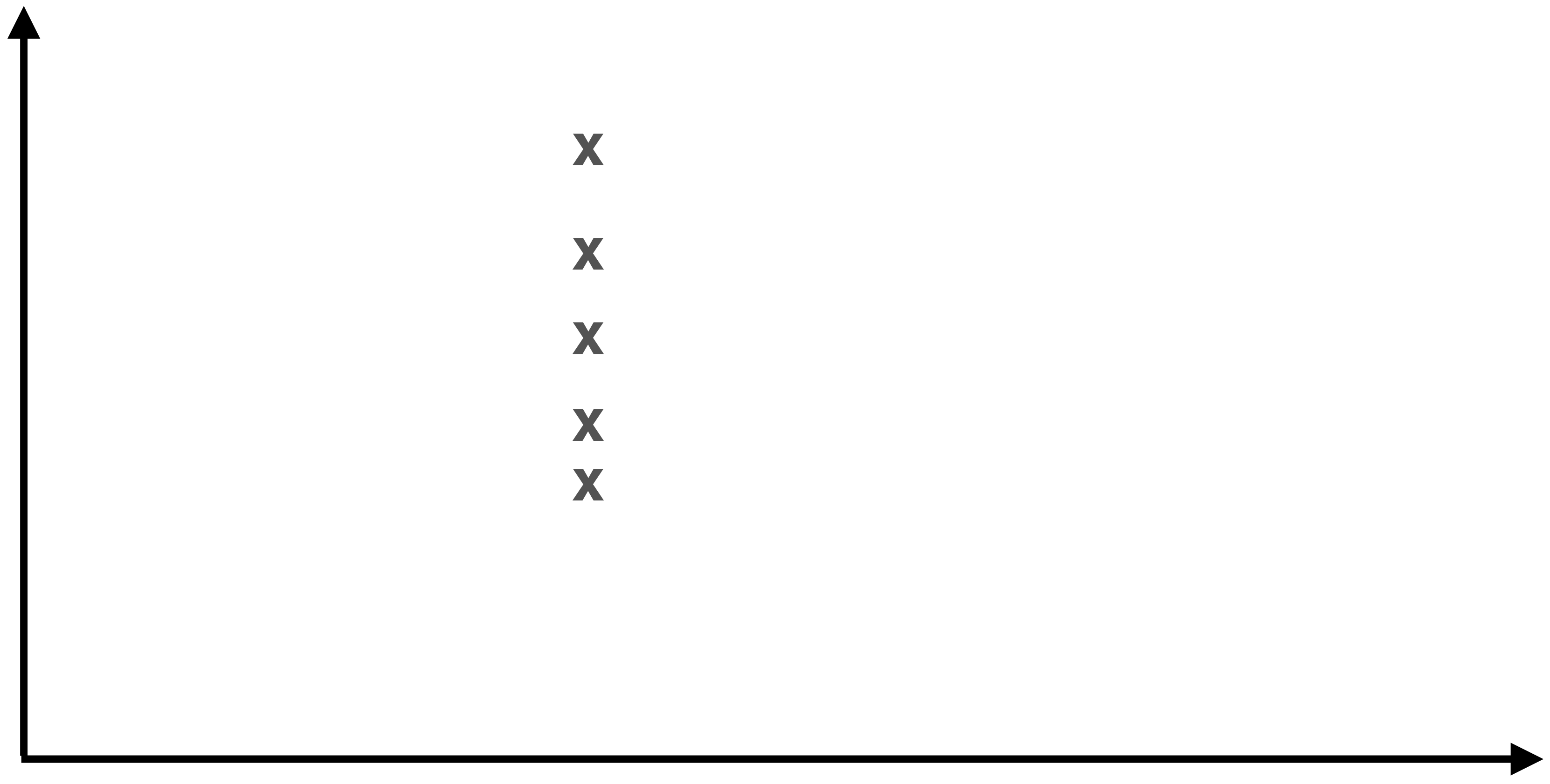
Unidimensional data points can be represented using a line, such as a number line

Data in Two Dimensions



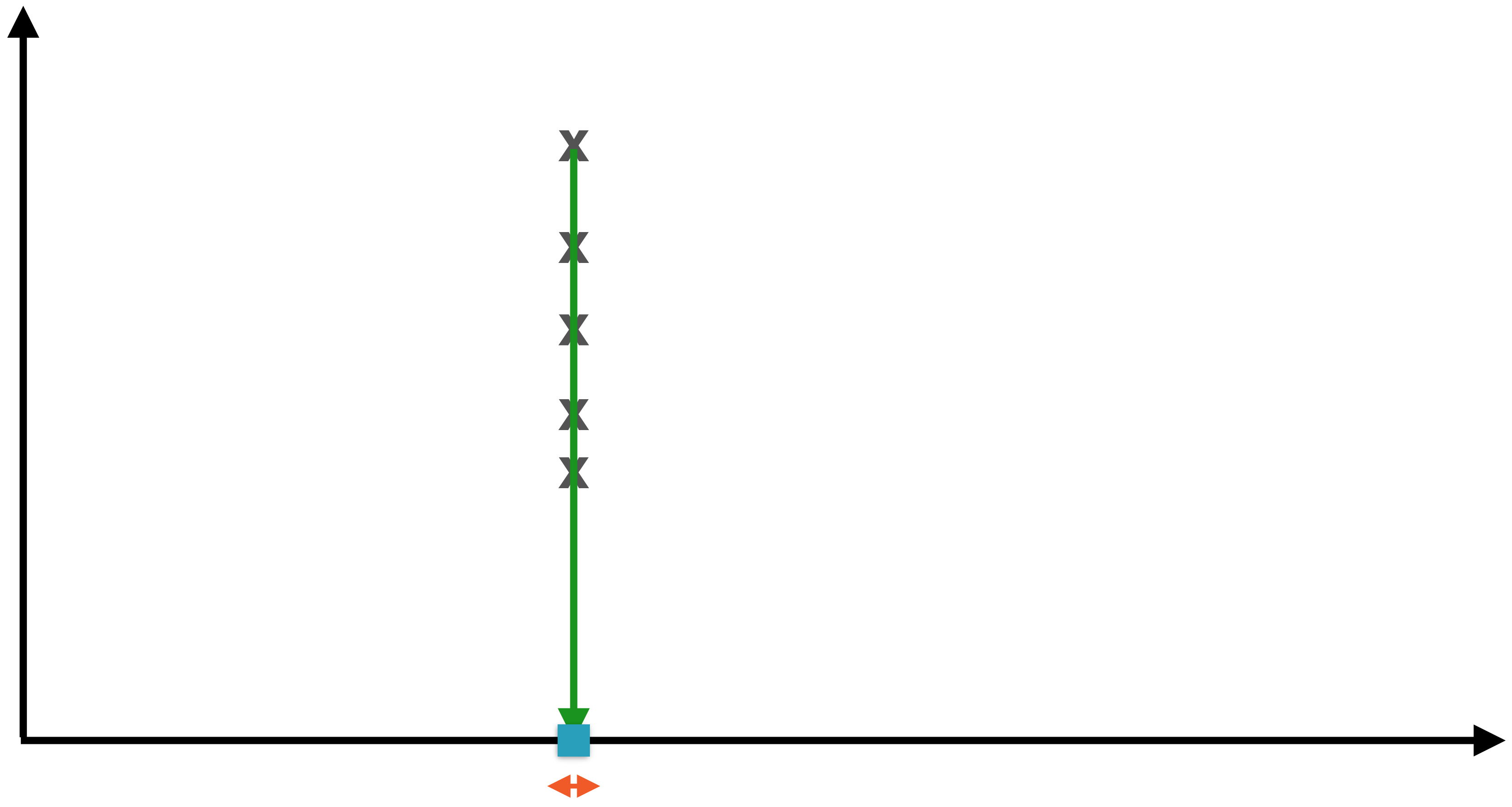
It's often more insightful to view data in relation to some other, related data

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

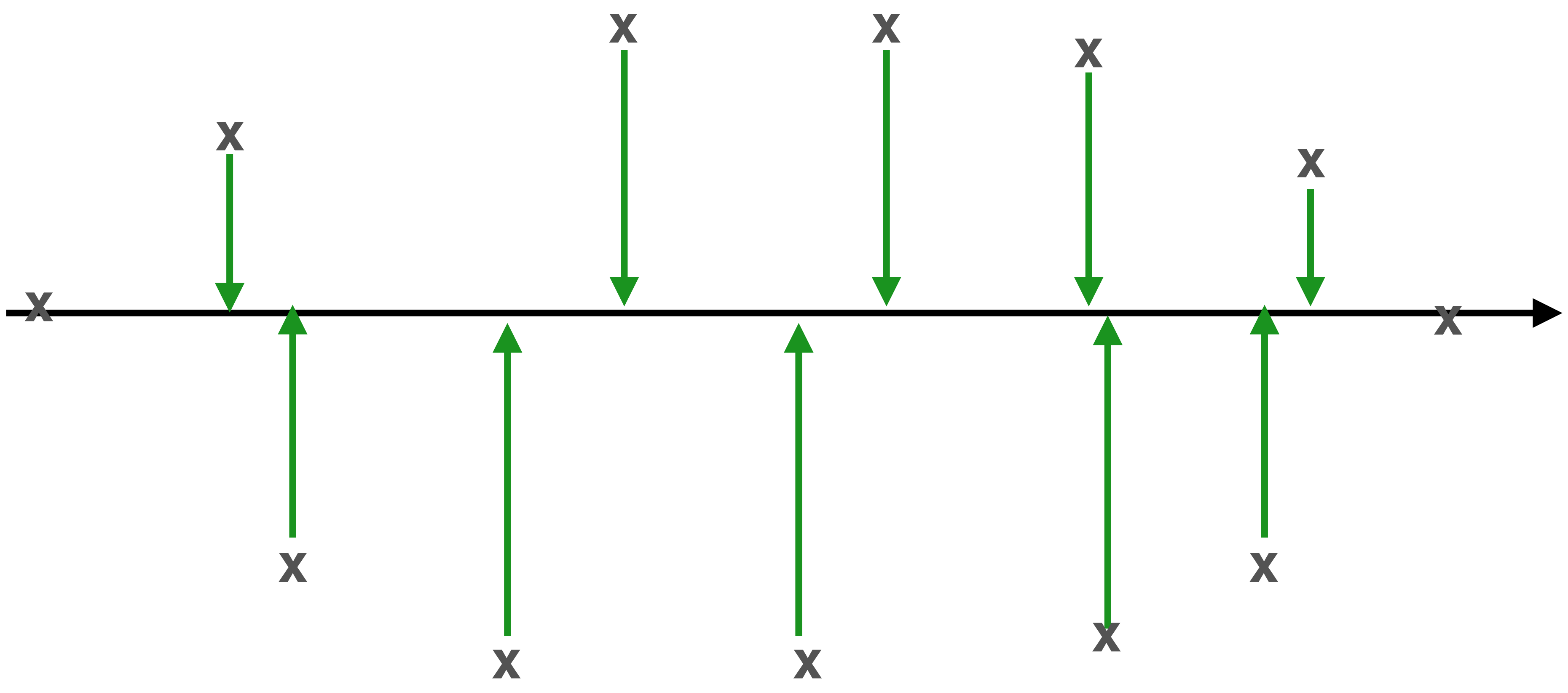
Good Choice of Dimensions

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

Intuition Behind PCA

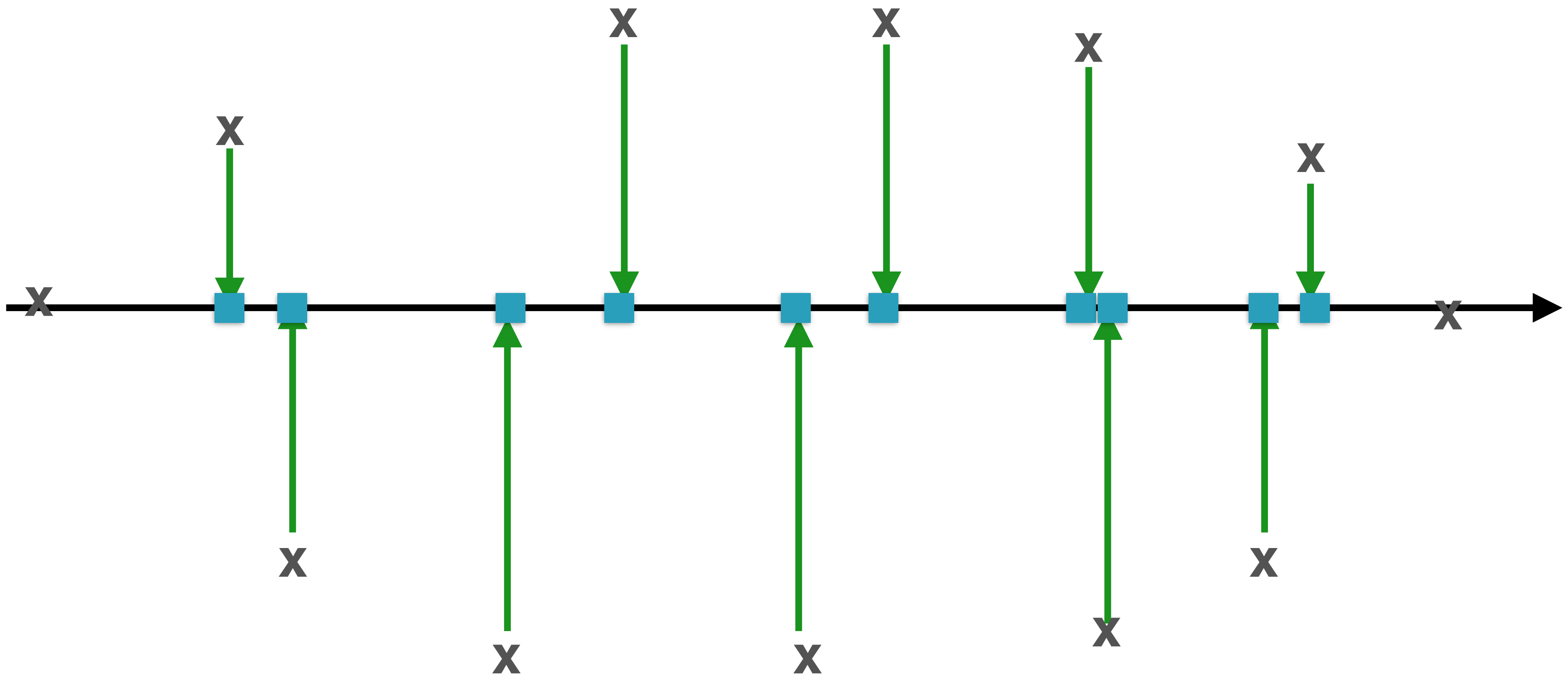
Objective: Find the "best" directions to represent this data

Intuition Behind PCA



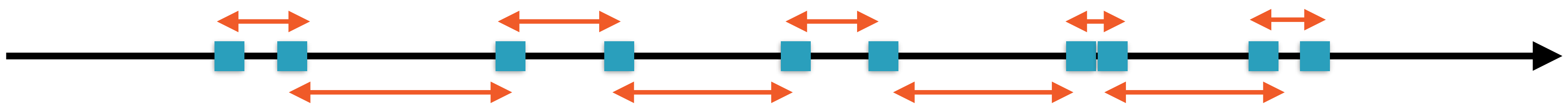
Start by "projecting" the data onto a line in some direction

Intuition Behind PCA



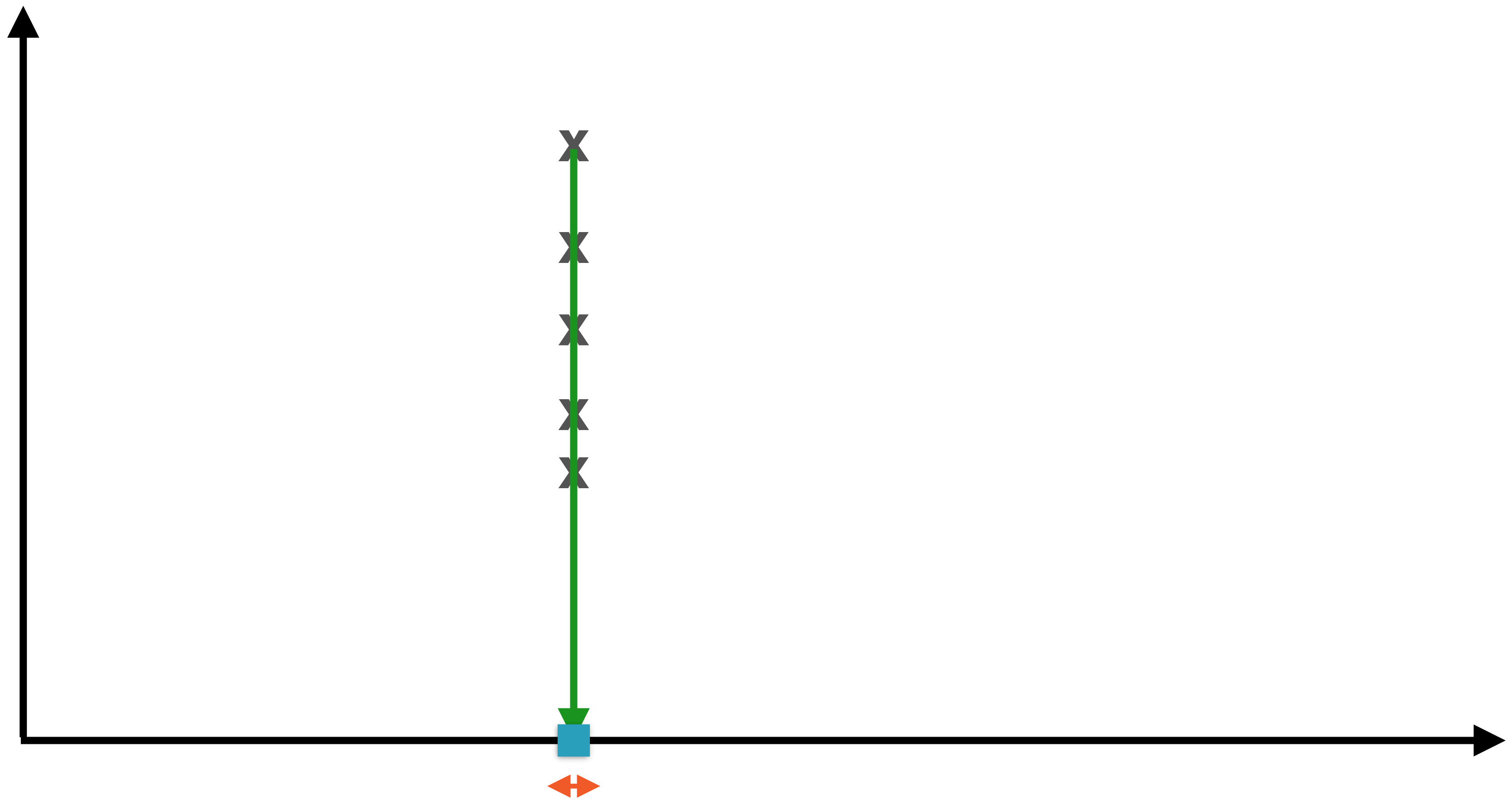
Start by "projecting" the data onto a line in some direction

Intuition Behind PCA



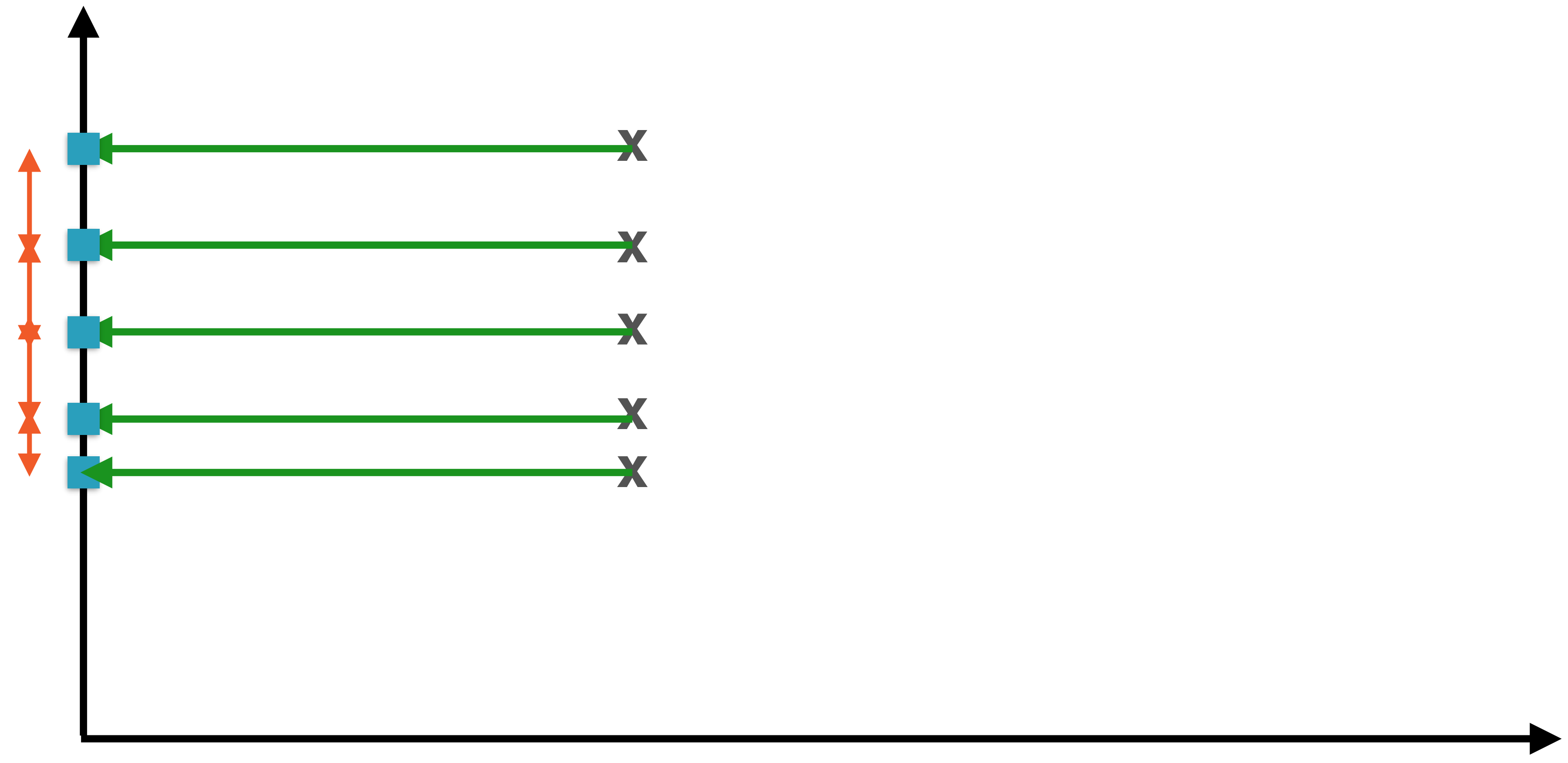
The greater the distances between these projections, the "better" the direction

Bad Projection



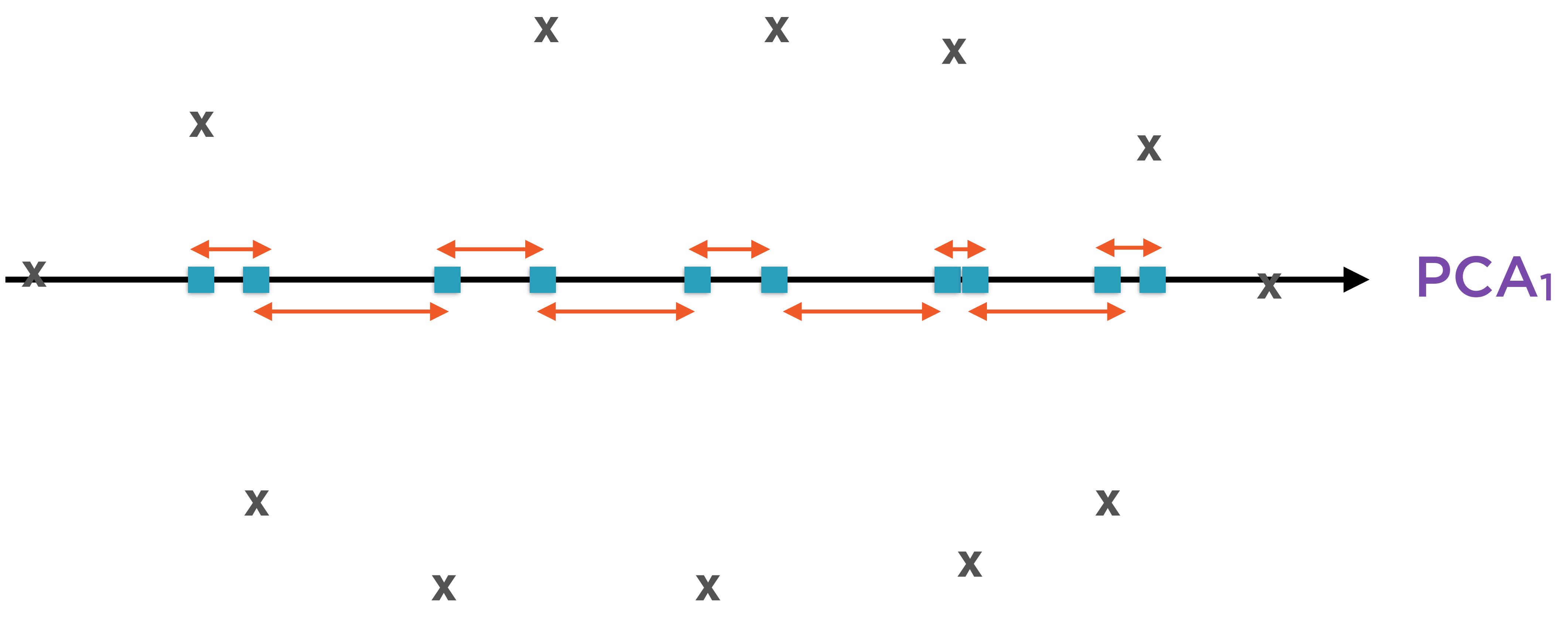
A projection where the distances are minimised is a bad one - information is lost

Good Projection



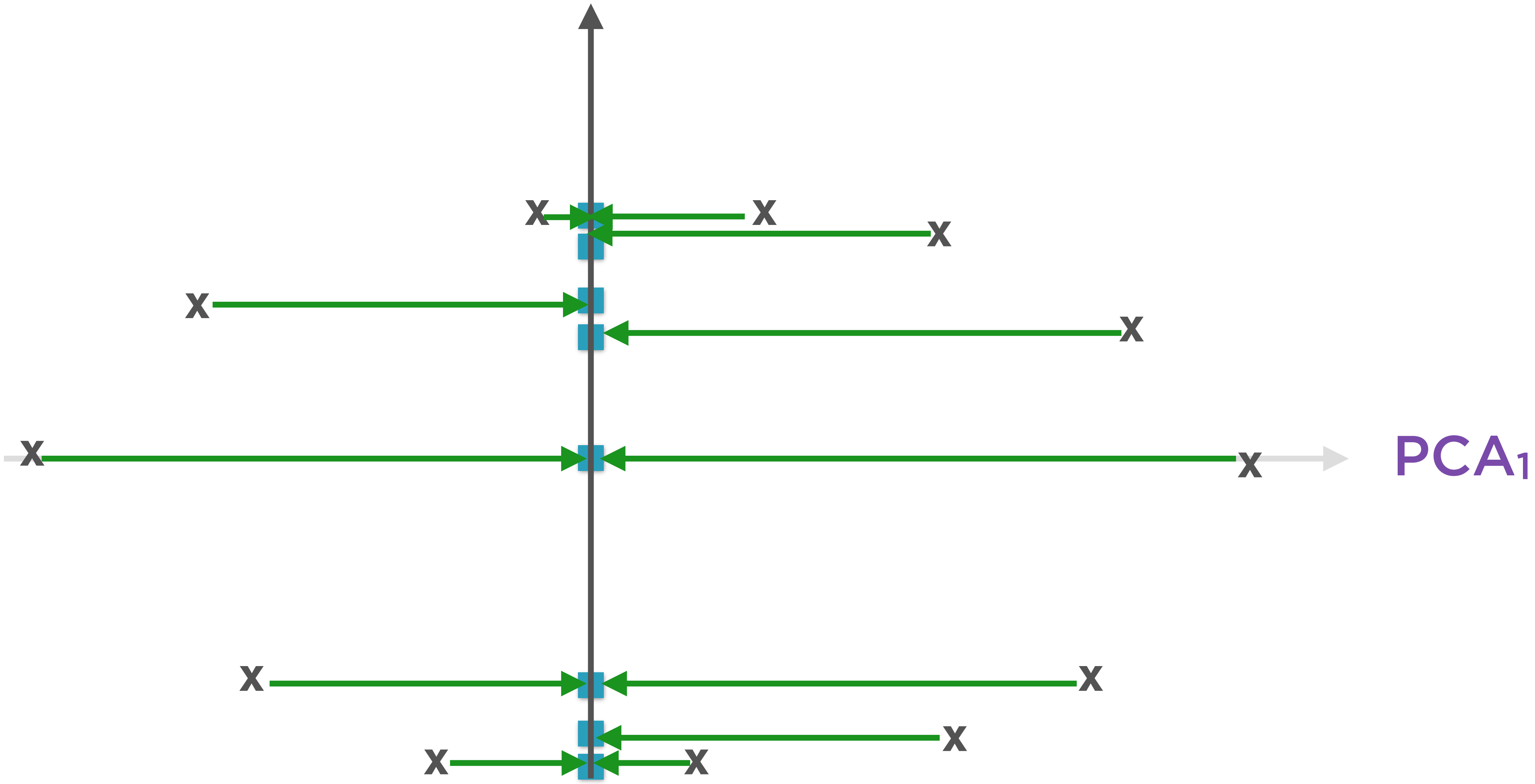
A projection where the distances are maximised is a good one - information is preserved

Intuition Behind PCA



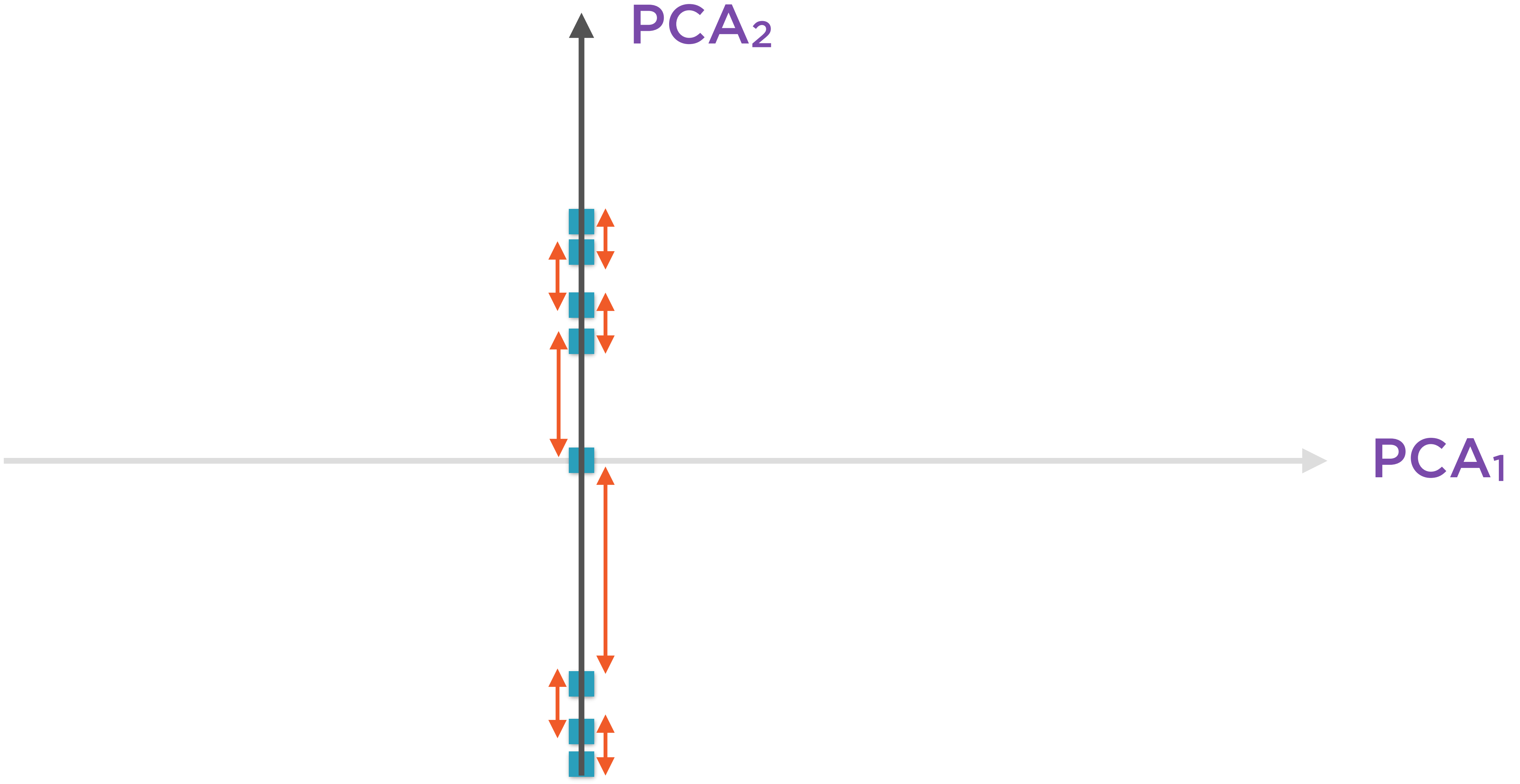
The direction along which this variance is maximised is the first principal component of the original data

Intuition Behind PCA



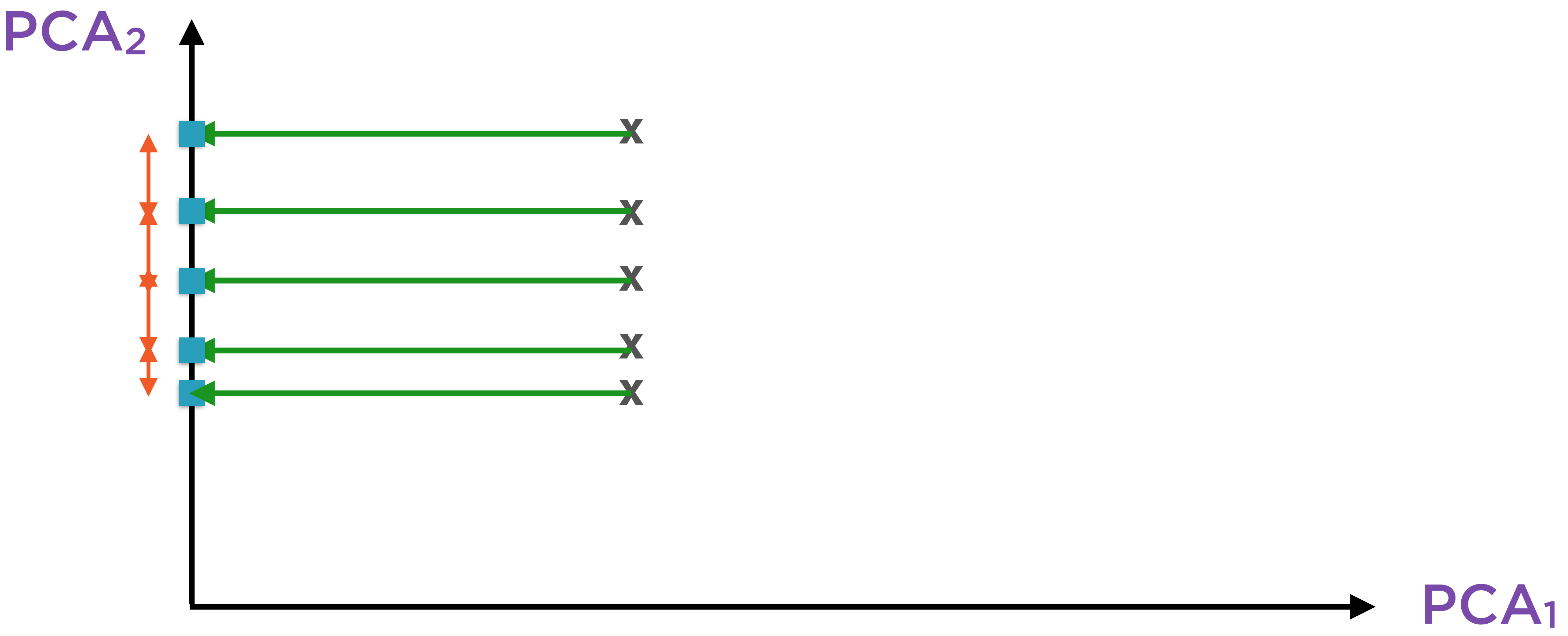
Find the next best direction, the second principal component, which must be at right angles to the first

Intuition Behind PCA



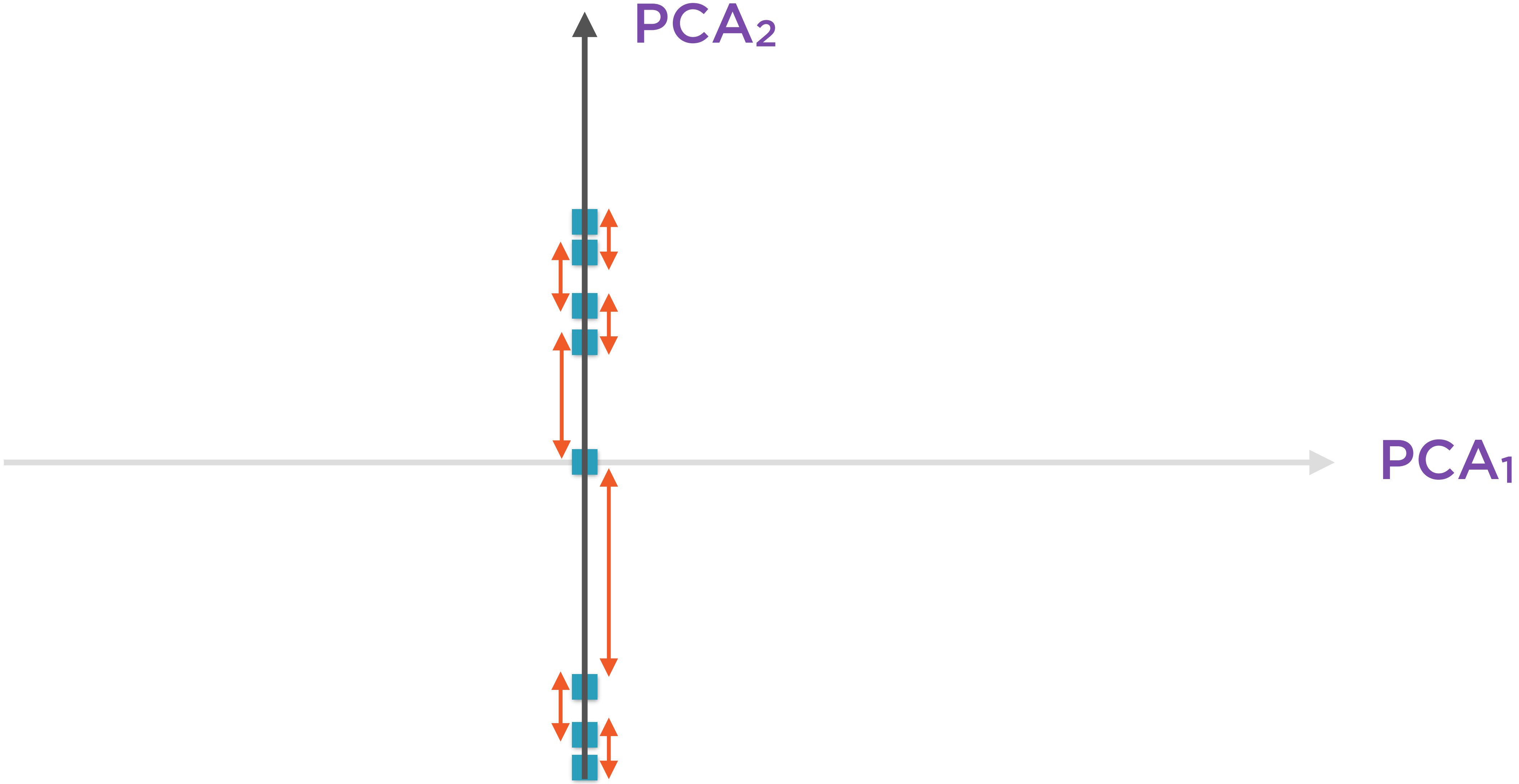
Find the next best direction, the **second principal component**, which must be at right angles to the first

Principal Components at Right Angles



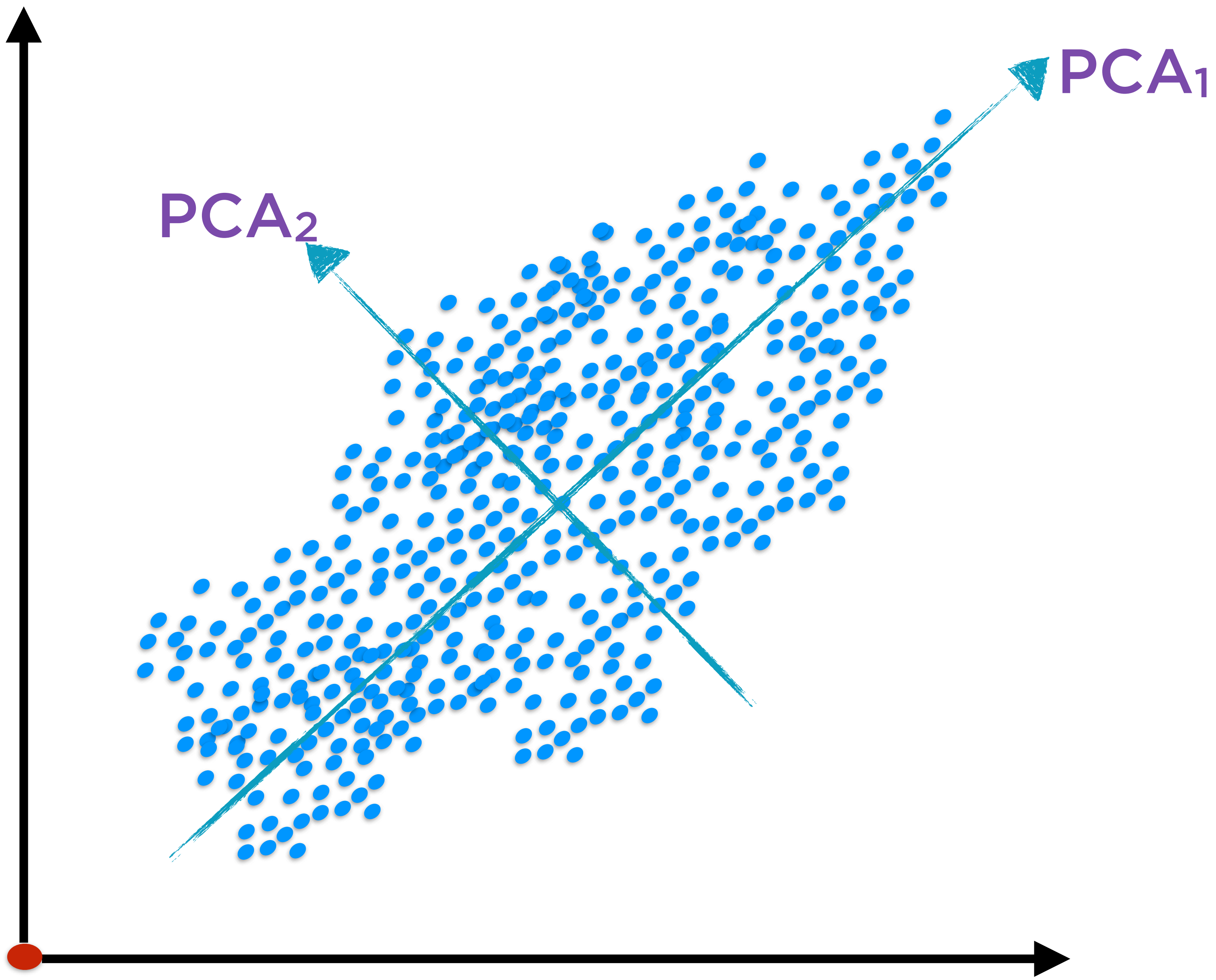
Directions at right angles help express the most variation with the smallest number of directions

Intuition Behind PCA



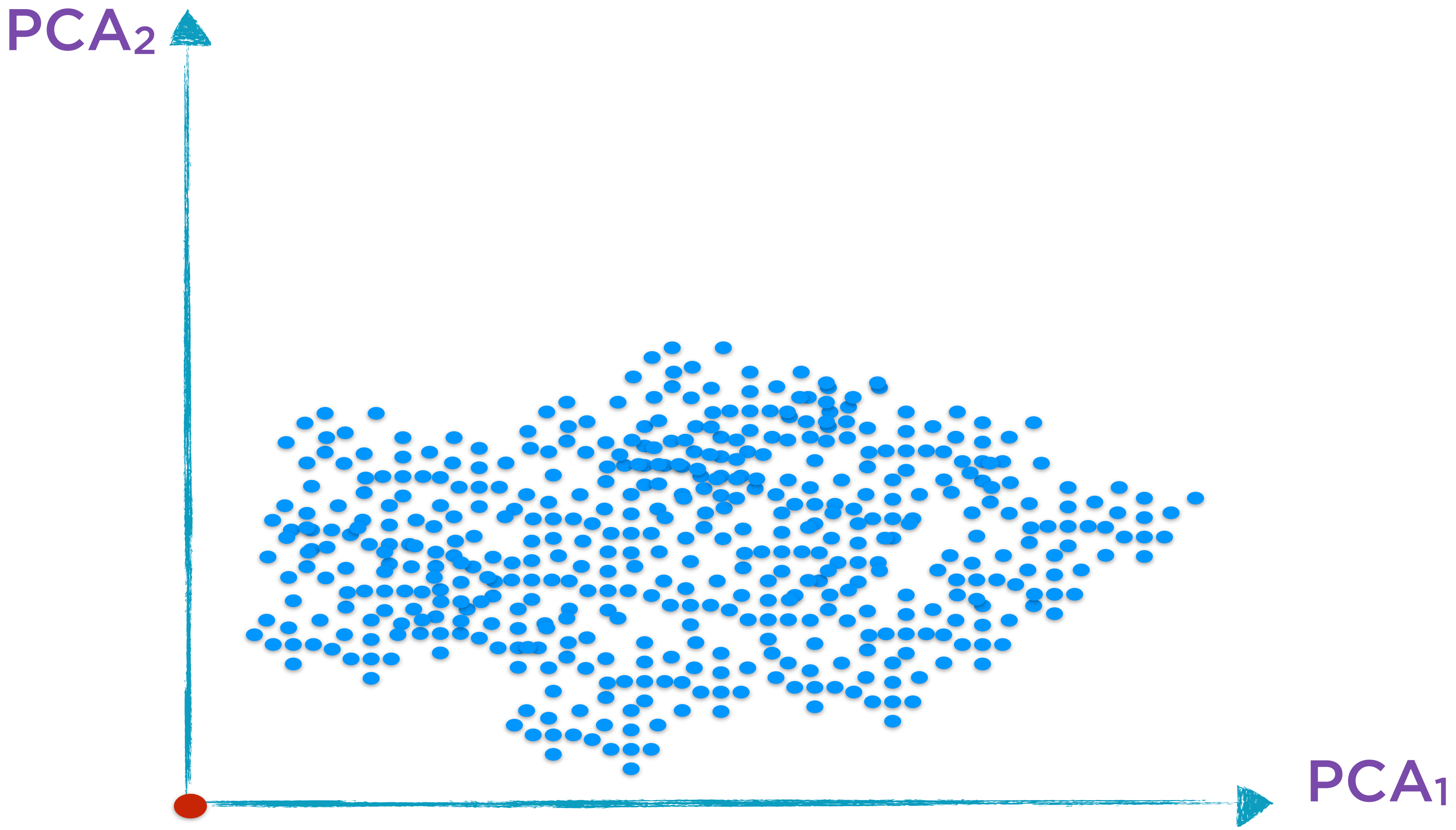
The variances are clearly smaller along this second principal component than along the first

Intuition Behind PCA



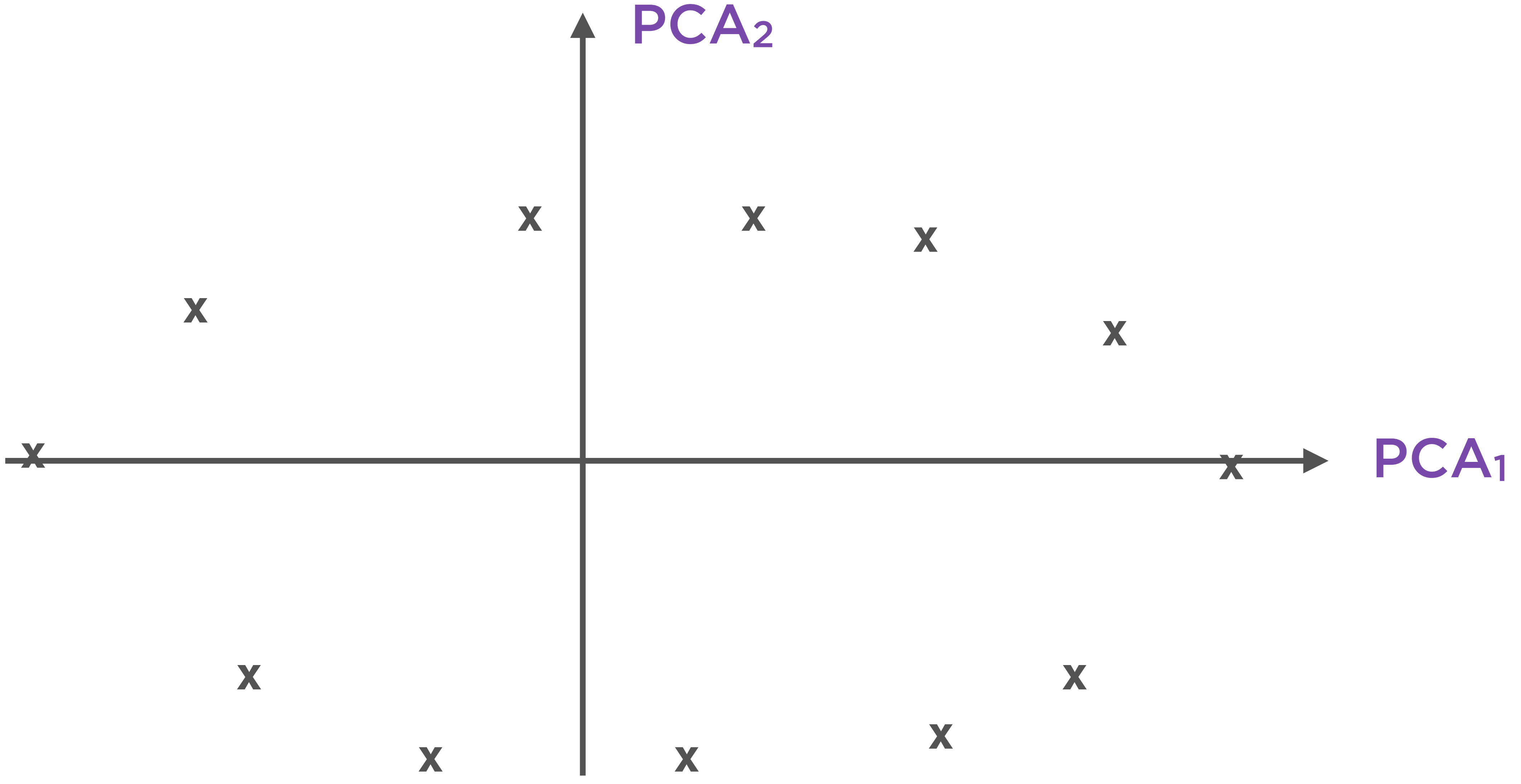
In general, there are as many principal components as there are dimensions in the original data

Intuition Behind PCA



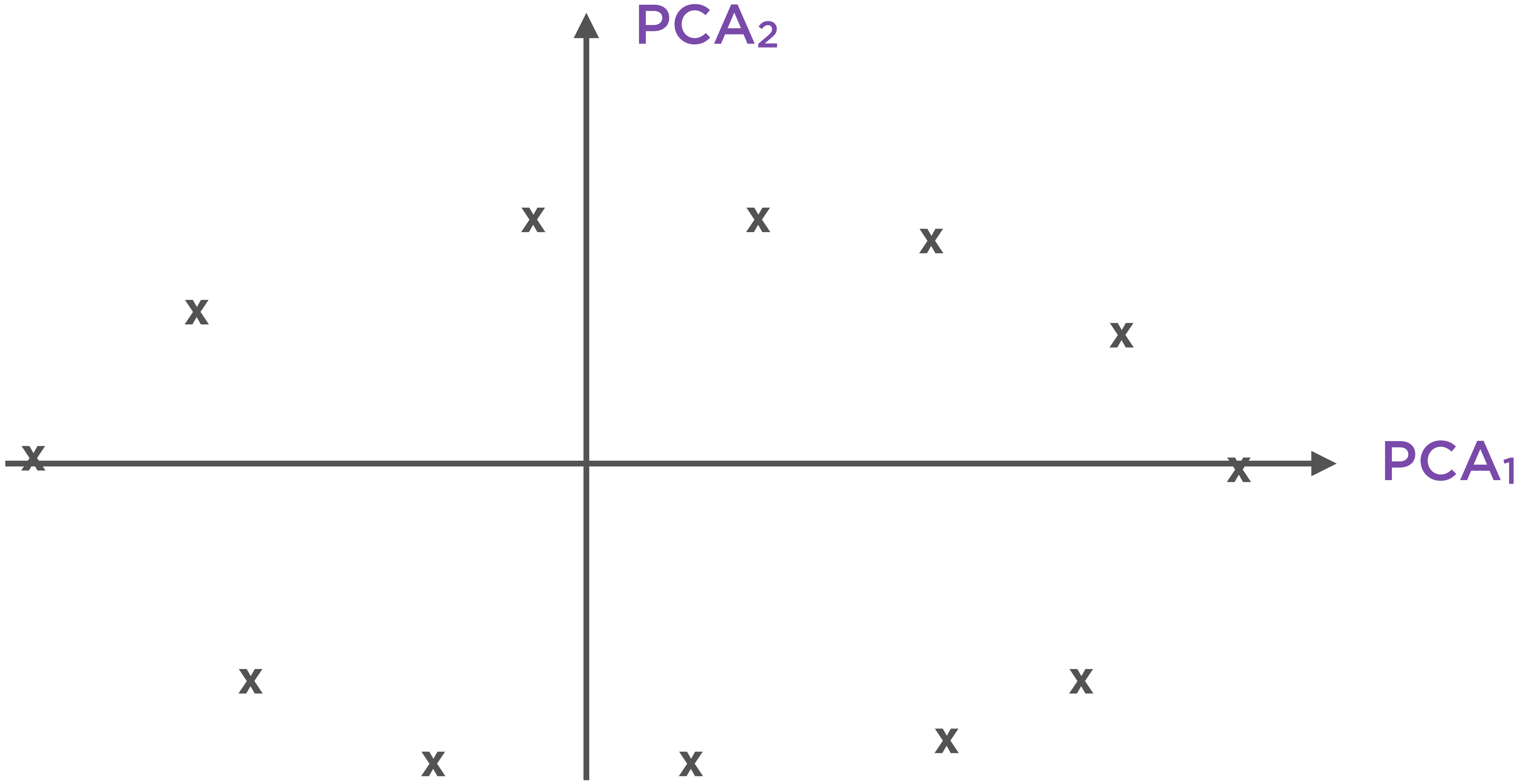
Re-orient the data along these new axes

Dimensionality Reduction



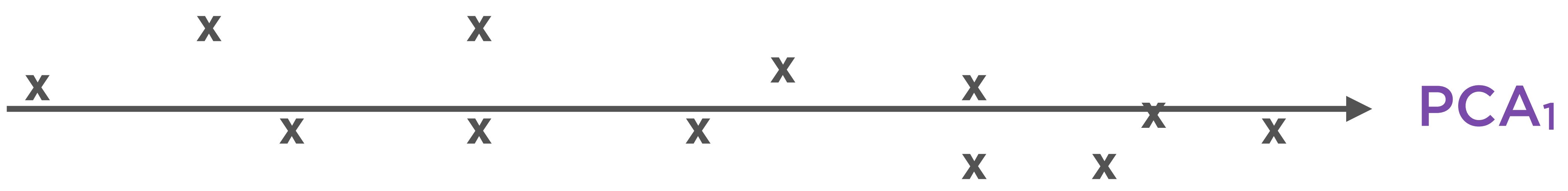
If the variance along the second principal component is small enough, we can just ignore it and use just 1 dimension to represent the data

Dimensionality Reduction



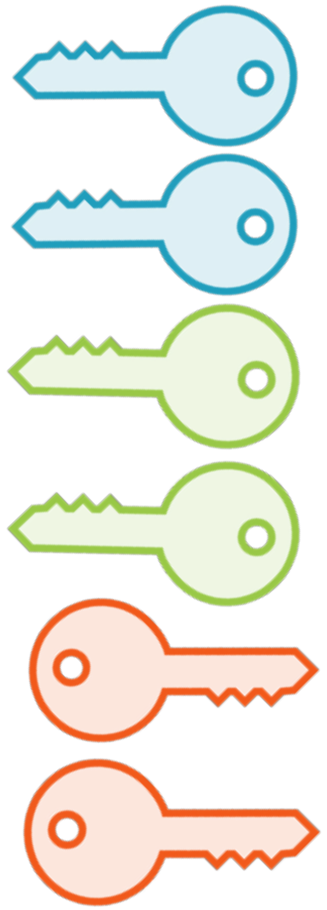
Variation along 2 dimensions: 2 principal components required

Dimensionality Reduction



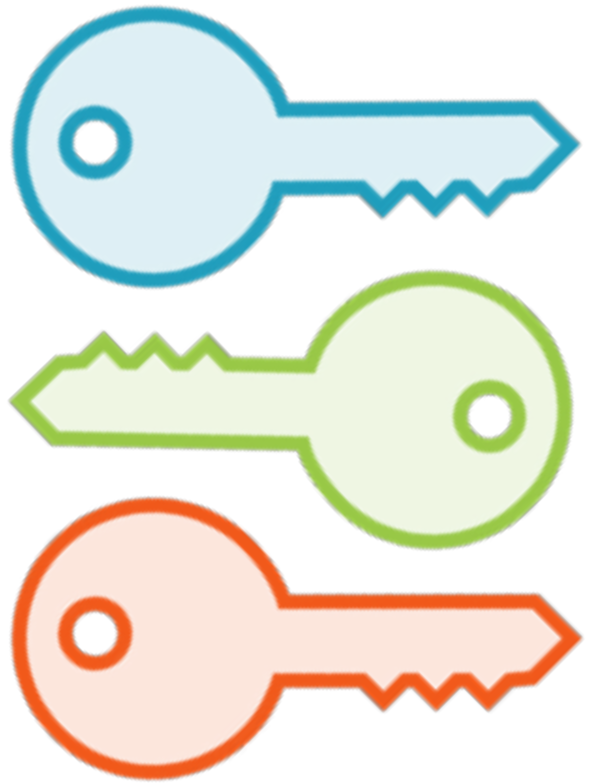
Variation along 1 dimension: 1 principal component is sufficient

Similar, yet Different



Regression

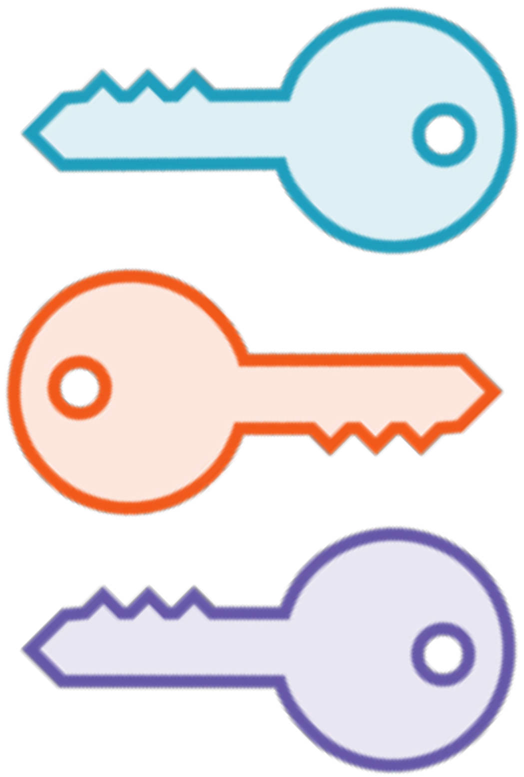
Connect the dots



Factor Analysis

Cut through the clutter

Regression



Causes

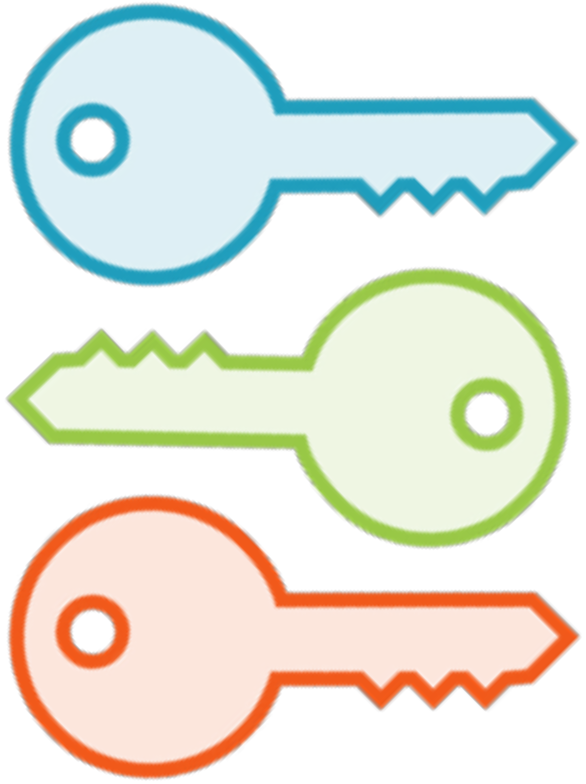
Independent variables

Effect

Dependent variable

Factor Analysis

Many Observed Causes



Few Underlying Causes

One Effect

Simplistic

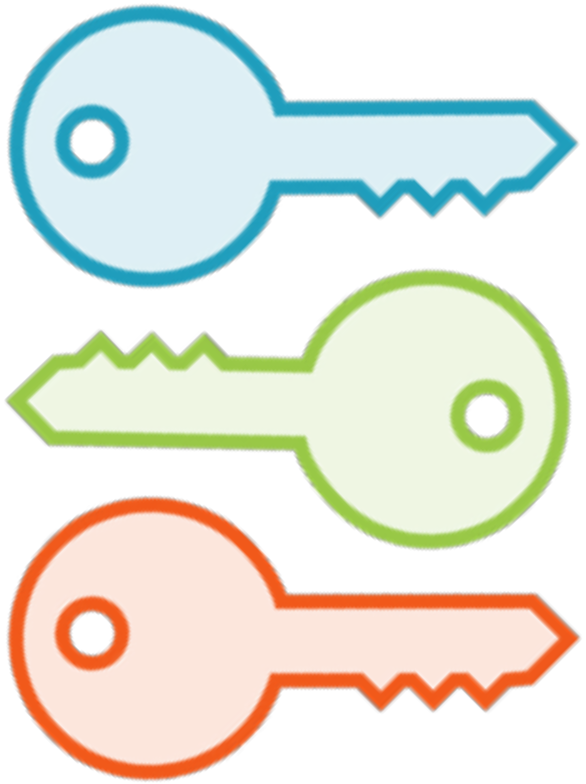
Causes

Independent variables

Effect

Dependent variable

Simple



Causes

Independent variables

Effect

Dependent variable

What and How

Cut through clutter

Extract underlying factors from a set of data

Principal components analysis (PCA)

Cookie-cutter technique that finds the ‘good’ factors from a set of data points

PCA is one solution to the factor-extraction problem - a cookie-cutter solution

What and How

Connect the dots

Fit a curve through a set of data

Regression

Cookie-cutter technique that finds the ‘best-fit’ line through a set of data points

Regression is one solution to the data-fitting problem - a cookie-cutter solution

Two Approaches to Factor Extraction

Rule-based

Human experts identify and extract factors

ML-based

Algorithm identifies and extracts factors

PCA and Factor Analysis

Principal Component Analysis is one procedure for factor analysis

It is mathematically guaranteed to result in independent factors

However, those factors may not actually correspond to intuition

Correlated Random Variables

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \cdots \\ E_n \end{bmatrix} \quad \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \cdots \\ D_n \end{bmatrix} \quad \begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ \cdots \\ G_n \end{bmatrix} \quad \cdots \quad \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \cdots \\ A_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

G_i = % return of Google stock on day i

A_i = % return of Apple stock on day i

Correlated Random Variables

$$
\begin{array}{cc}
\begin{bmatrix}
E_1 & D_1 & G_1 & & A_1 \\
E_2 & D_2 & G_2 & & A_2 \\
E_3 & D_3 & G_3 & \cdots & A_3 \\
\cdots & \cdots & \cdots & & \cdots \\
E_n & D_n & G_n & & A_n
\end{bmatrix} & \text{n rows} \\
\text{k columns} &
\end{array}
$$

Summarise the returns of k stocks, each over n days, into an nxk matrix

Correlated Random Variables

$$\begin{array}{cc} \begin{bmatrix} x_{11} & x_{12} & x_{13} & & x_{1k} \\ x_{21} & x_{22} & x_{23} & & x_{2k} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \cdots & \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & x_{n3} & & x_{nk} \end{bmatrix} & \text{n rows} \\ \text{k columns} & \end{array}$$

Summarise the returns of k stocks, each over n days, into an nxk matrix

Correlated Random Variables

$$\begin{array}{cc} \begin{bmatrix} x_{11} & x_{12} & x_{13} & & x_{1k} \\ x_{21} & x_{22} & x_{23} & & x_{2k} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \cdots & \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & x_{n3} & & x_{nk} \end{bmatrix} & \text{n rows} \end{array}$$

X_1 (n rows, 1 column)

k columns

Correlated Random Variables

$$\begin{bmatrix} x_{11} & \boxed{x_{12}} & x_{13} & & x_{1k} \\ x_{21} & x_{22} & x_{23} & & x_{2k} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \cdots & \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & x_{n3} & & x_{nk} \end{bmatrix} \text{ n rows}$$

X_2 (n rows, 1 column)

k columns

Correlated Random Variables

$$\begin{array}{cc}
\begin{bmatrix}
x_{11} & x_{12} & x_{13} & & x_{1k} \\
x_{21} & x_{22} & x_{23} & & x_{2k} \\
x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\
\cdots & \cdots & \cdots & & \cdots \\
x_{n1} & x_{n2} & x_{n3} & & x_{nk}
\end{bmatrix} & \text{n rows} \\
\text{k columns} &
\end{array}$$

X_k (n rows, 1 column)

Correlated Random Variables

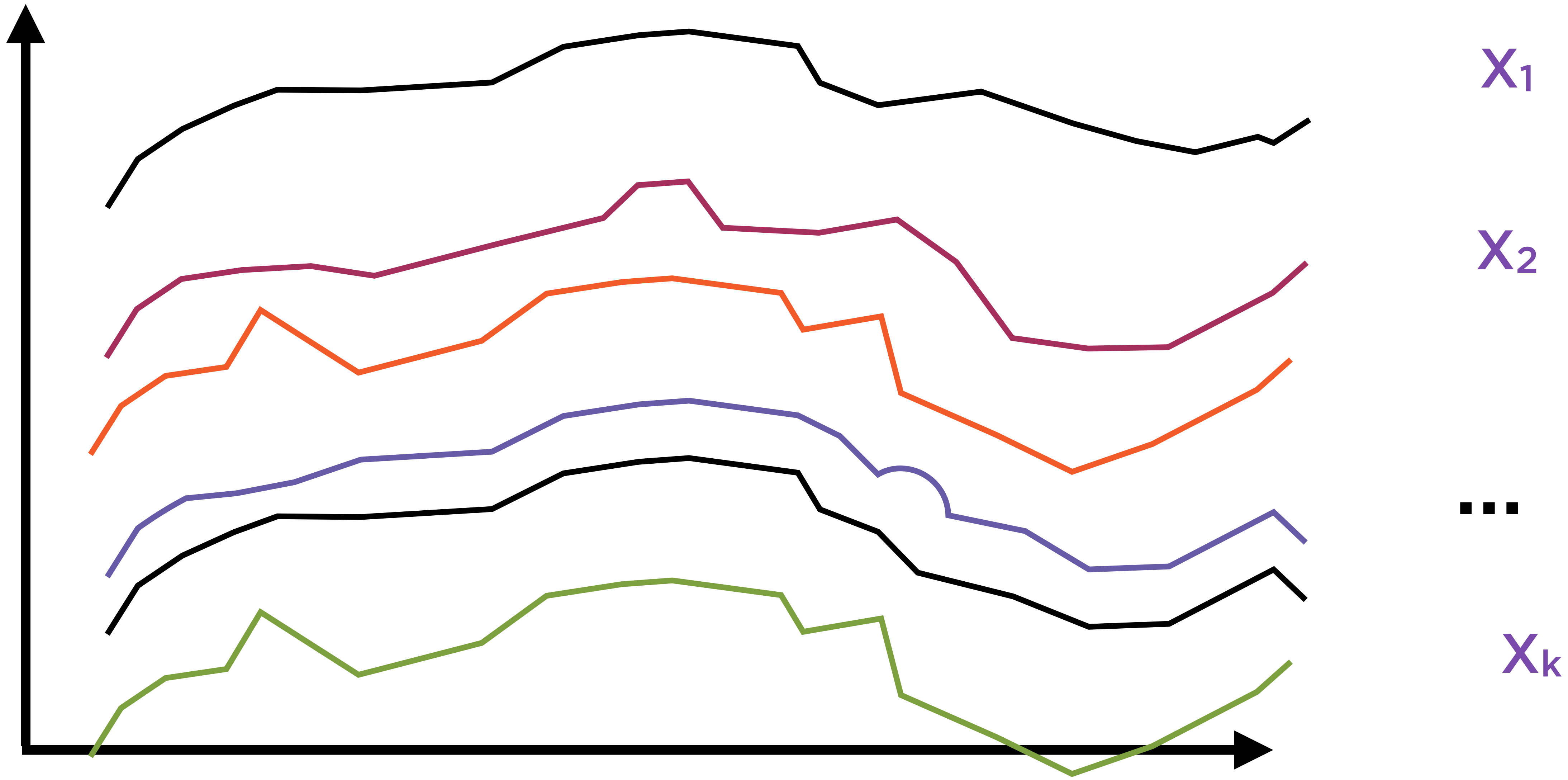
$$[\ X_1 \quad X_2 \quad X_3 \quad \ldots \quad X_k\ ]$$

n rows

k columns

Each element X_i of this matrix is a vector with 1 column and n rows

Correlated Random Variables



Highly correlated variables are not suitable for use in regression

Correlated Random Variables

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

n rows

k columns

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

Principal Components 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

PCA

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

Principal Components Analysis

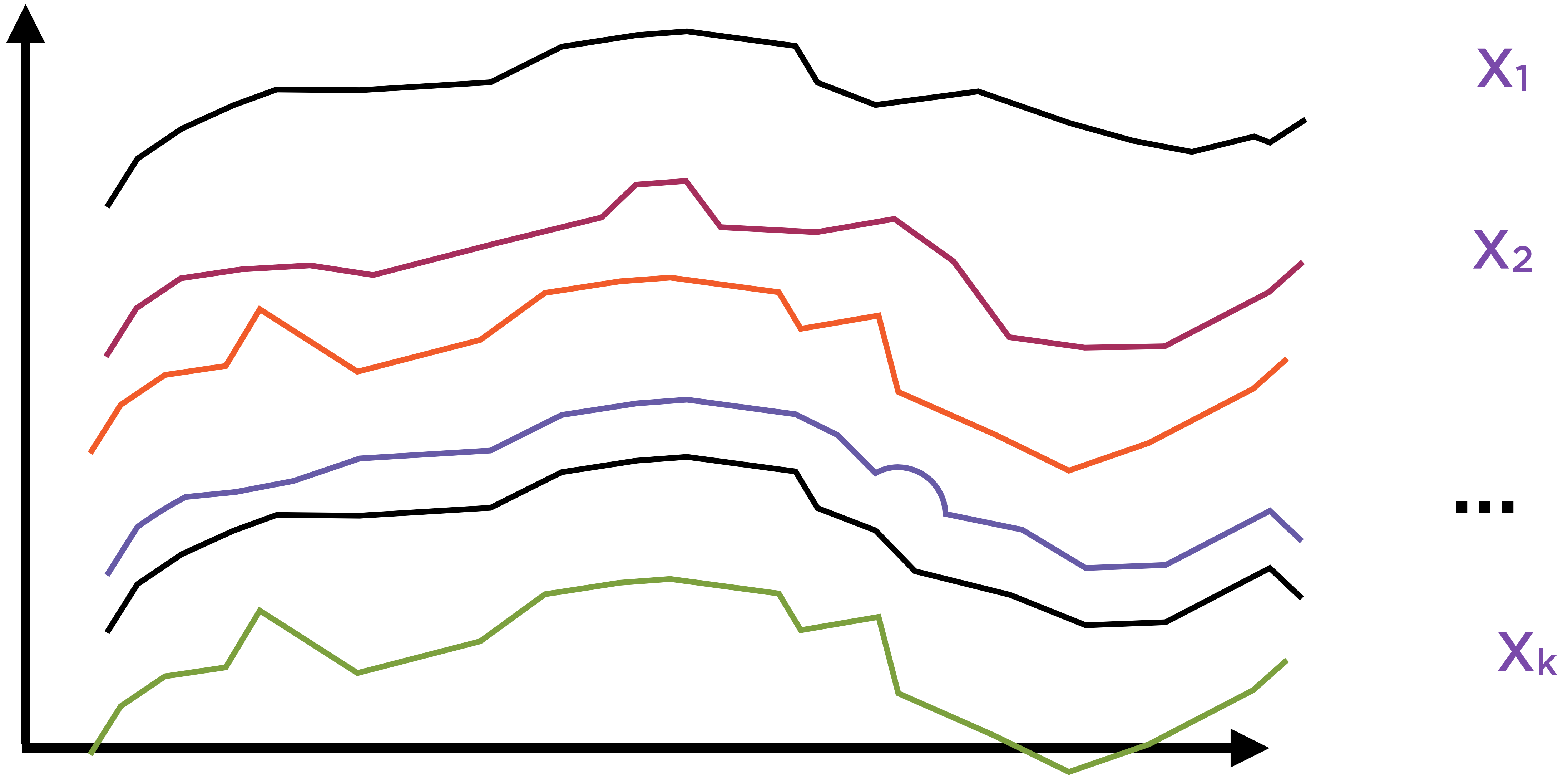
$$[\ F_1 \quad F_2 \quad F_3 \quad \ldots \quad F_k\ ]$$

n rows

k columns

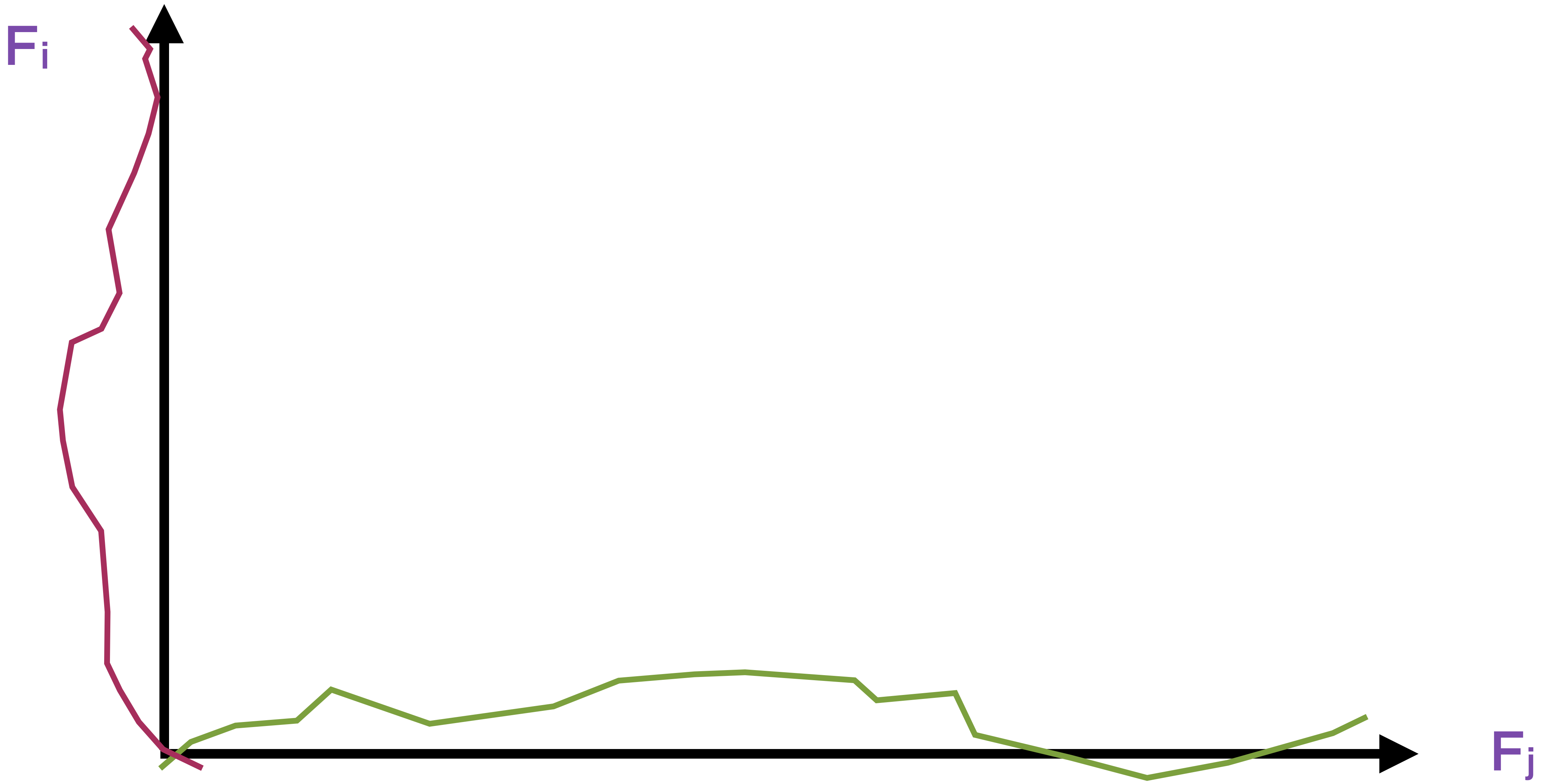
These vectors F_i are the principal components of the original vectors X_i

Correlated X_i



Highly correlated variables are not suitable for use in regression

Uncorrelated F_i



Any of the principal components is perfectly uncorrelated with all others

Principal Components Analysis

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

$$\mathrm{var}(F_1) > \mathrm{var}(F_2) > \mathrm{var}(F_3) \quad > \mathrm{var}(F_k)$$

These vectors F_i are arranged in order of decreasing variance

The greater the variance of a principal component, the more important it is

The greater the variance of a principal component, the more important it is

Principal Components Analysis

$$[\; F_1 \quad F_2 \quad F_3 \quad \dots \quad F_k \;]$$

$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

$$var(F_1) + var(F_2) + var(F_3) + var(F_k)$$

$$=$$

$$var(X_1) + var(X_2) + var(X_3) + var(X_k)$$

$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

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

Principal Components Analysis

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

$$\mathrm{var}(F_1) + \mathrm{var}(F_2) + \mathrm{var}(F_3) \qquad + \mathrm{var}(F_k)$$

$$=$$

$$\mathrm{var}(X_1) + \mathrm{var}(X_2) + \mathrm{var}(X_3) \qquad + \mathrm{var}(X_k)$$

The sum of the variances of vectors F_i is equal to sum of variances of original X_i

Principal Components

How

are such principal components found?

Why

are they more useful than the original data?

What

do we do with the PCs once we have them?

How Principal Components Are Found

Principal Components 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

PCA

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

Problem: Finding Principal Component 1

Find F_1

$$F_1 = a_1X_1 + a_2X_2 + a_3X_3 \dots + a_kX_k$$

such that

Variance(F_1) is maximised

subject to constraint

$$a_1^2 + a_2^2 + \dots + a_k^2 = 1$$

This problem has a cookie-cutter solution in linear algebra - eigen decomposition

Solution: Finding Principal Component 1

Eigenvector:

$$v_1 = [\ a_1, a_2, a_3 \dots a_k\ ]$$

Principal Component:

$$F_1 = a_1X_1 + a_2X_2 + a_3X_3 \dots + a_kX_k$$

Eigenvalue:

$$e = \text{Variance}(F_1)$$

Eigen decomposition gives us the answer

Problem: Finding Principal Component 2

Given F_1, find F_2

$$F_2 = a_1(X_1 - F_1) + a_2(X_2 - F_1) + a_3(X_3 - F_1) \ldots + a_k(X_k - F_1)$$

such that

Variance(F_2) is maximised

subject to constraint

$$a_1^2 + a_2^2 + \ldots + a_k^2 = 1$$

Eigen decomposition finds all of these solutions in one go

Principal Components 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

PCA

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

Principal Components Analysis

[X_1 X_2 X_3 ... X_k] →

Eigenvalue Decomposition →

Principal Components:

[F_1 F_2 F_3 ... F_k]

n rows

k columns

Eigenvectors:

[v_1 v_2 v_3 ... v_k]

k rows

k columns

Eigenvalues:

[e_1 e_2 e_3 ... e_k]

1 row

k columns

Results of PCA

Eigenvalues

tell importance of each principal component

Principal Components

for the largest eigenvalues can be used in regression

Eigenvectors

are needed to calculate the principal components

Interpreting Eigenvalues

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

PCA

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

Interpreting Eigenvalues

$$\begin{bmatrix} F_1 & F_2 & F_3 & \cdots & F_k \end{bmatrix}$$

n rows

k columns

These vectors F_i are the principal components of the original vectors X_i

Interpreting Eigenvalues

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

$$\mathrm{var}(F_1) > \mathrm{var}(F_2) > \mathrm{var}(F_3) \quad > \mathrm{var}(F_k)$$

These vectors F_i are arranged in order of decreasing variance

The greater the variance of a principal component, the more important it is

Interpreting Eigenvalues

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

$$\mathrm{var}(F_1) > \mathrm{var}(F_2) > \mathrm{var}(F_3) \quad > \mathrm{var}(F_k)$$

Eigenvalue 1

Eigenvalue 2

Eigenvalue 3

Eigenvalue k

The greater the eigenvalue of a principal component, the more important it is

Principal Components Analysis

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

$$\mathrm{var}(F_1) + \mathrm{var}(F_2) + \mathrm{var}(F_3) \quad + \mathrm{var}(F_k)$$
$$=$$
$$\mathrm{var}(X_1) + \mathrm{var}(X_2) + \mathrm{var}(X_3) \quad + \mathrm{var}(X_k)$$

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

Principal Components Analysis

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

$$\mathrm{var}(F_1) + \mathrm{var}(F_2) + \mathrm{var}(F_3) \quad + \mathrm{var}(F_k)$$

$$=$$

$$\mathrm{var}(X_1) + \mathrm{var}(X_2) + \mathrm{var}(X_3) \quad + \mathrm{var}(X_k)$$

The sum of the variances of vectors F_i is equal to sum of variances of original X_i

Principal Components Analysis

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

$$\downarrow \qquad \downarrow \qquad \downarrow \qquad \qquad \downarrow$$

$$var(F_1) + var(F_2) + var(F_3) \qquad + var(F_k)$$

$$=$$

$$var(X_1) + var(X_2) + var(X_3) \qquad + var(X_k)$$

$$= \quad \text{Total Variance}(X)$$

$$= \quad \text{Total Variance}(F)$$

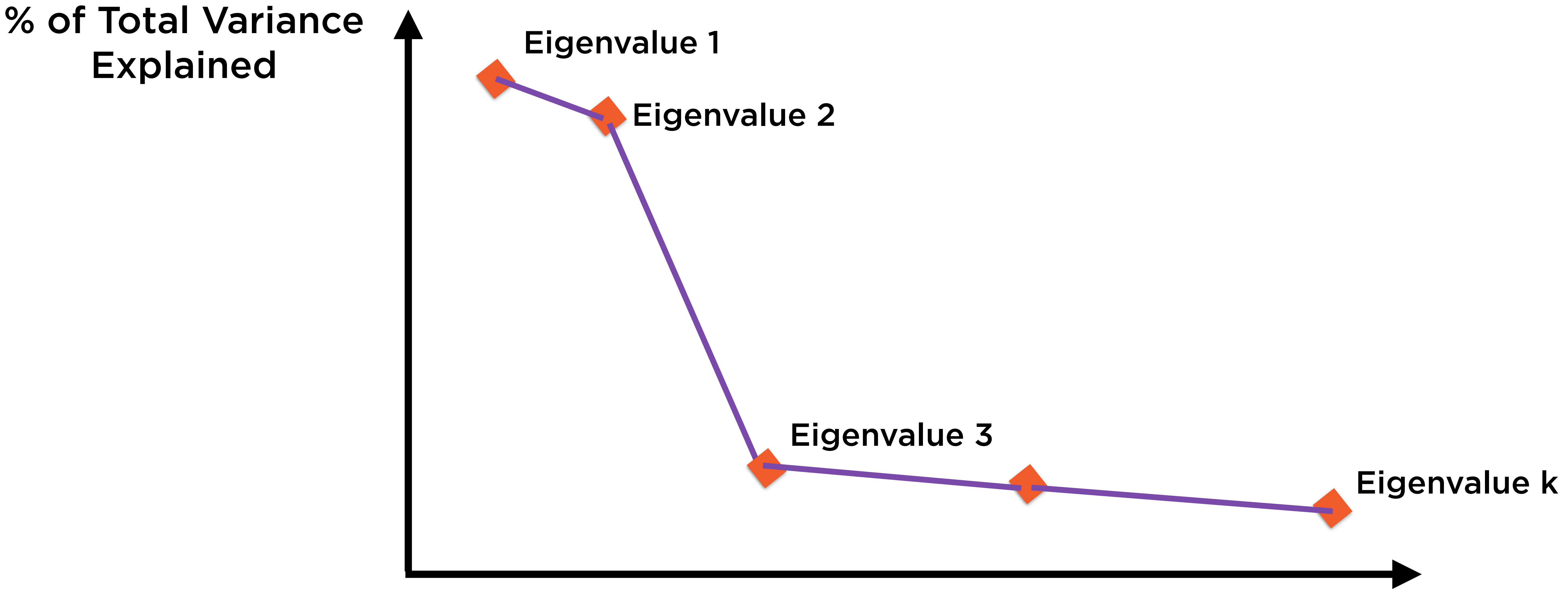
Interpreting Eigenvalues

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

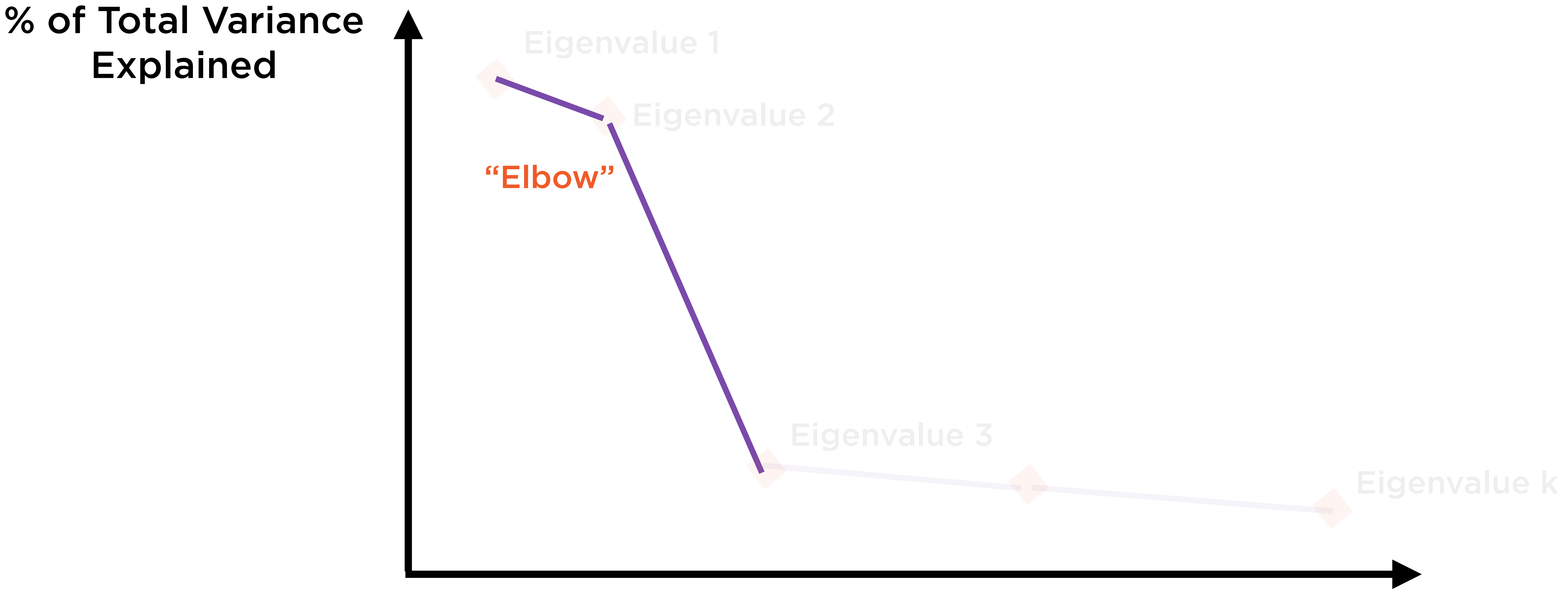
$$\frac{\text{Eigenvalue 1}}{\text{Variance(F)}} + \frac{\text{Eigenvalue 2}}{\text{Variance(F)}} + \cdots + \frac{\text{Eigenvalue k}}{\text{Variance(F)}}$$

$$= 100\%$$

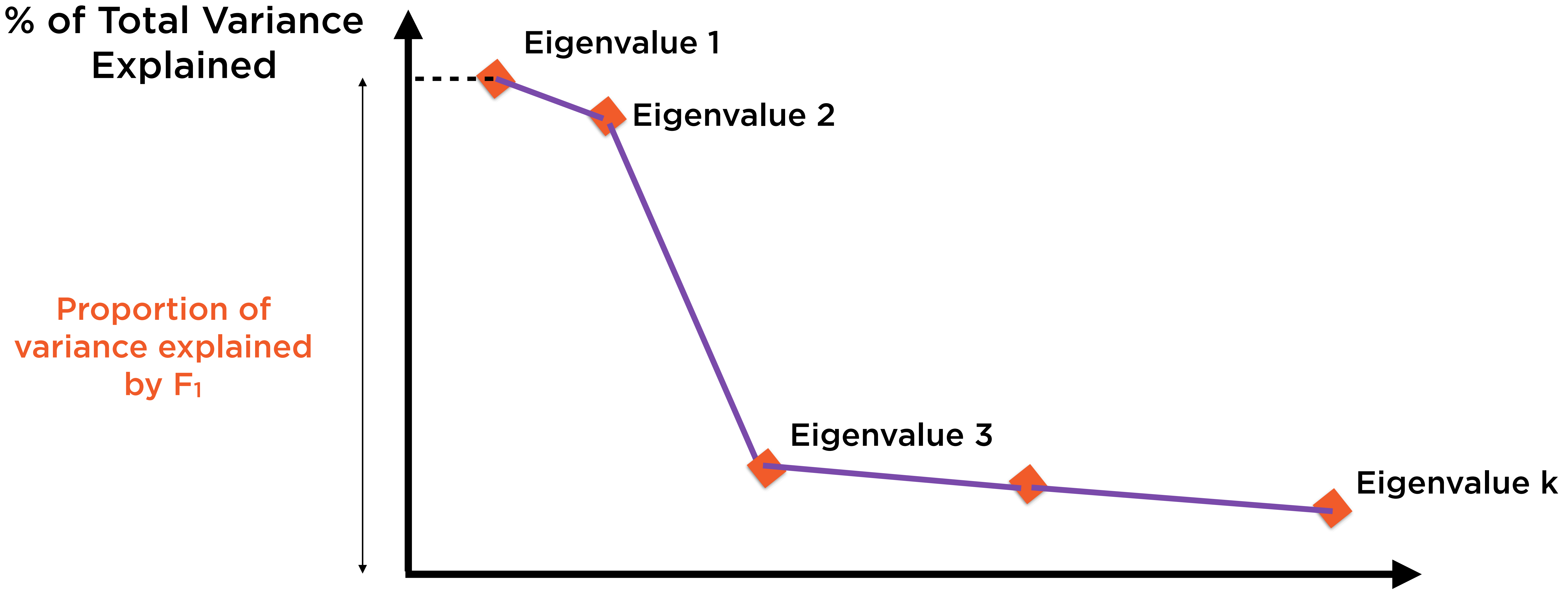
Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k

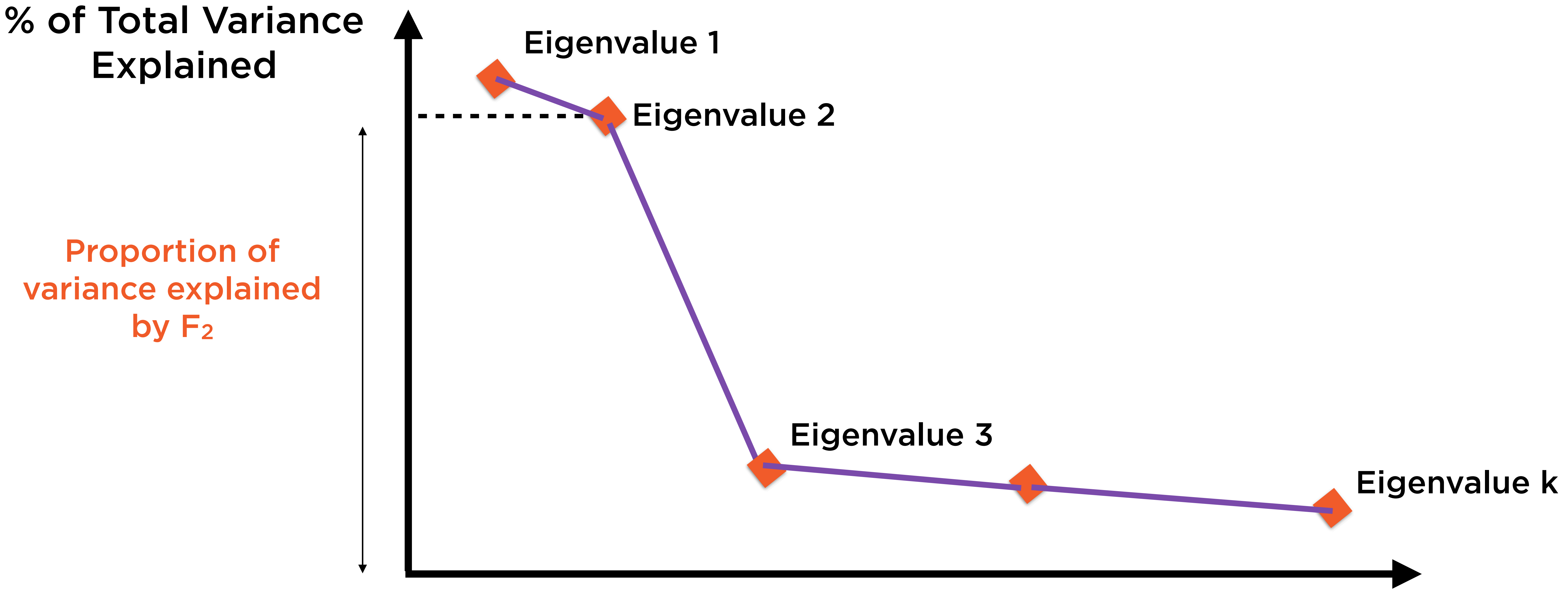
Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
"Elbow"
Eigenvalue 3
Eigenvalue k

Scree Plots

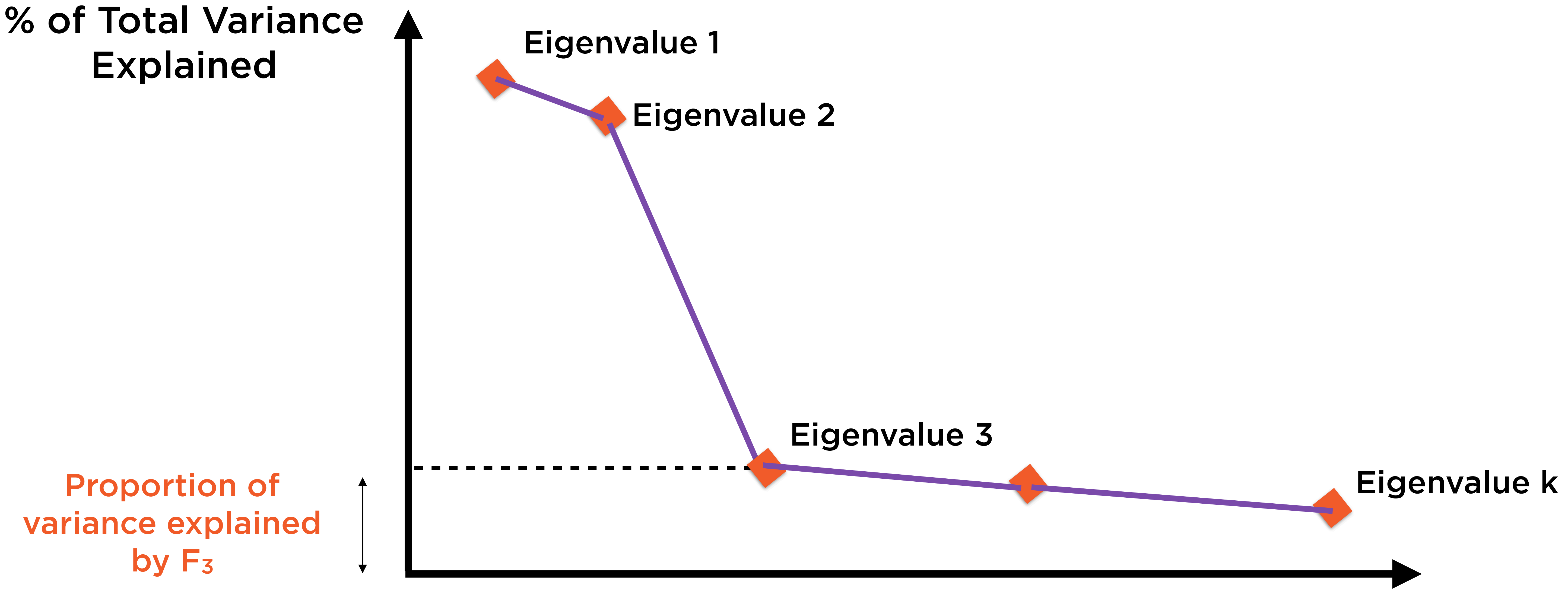

% of Total Variance Explained
Proportion of variance explained by F1
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k

Scree Plots

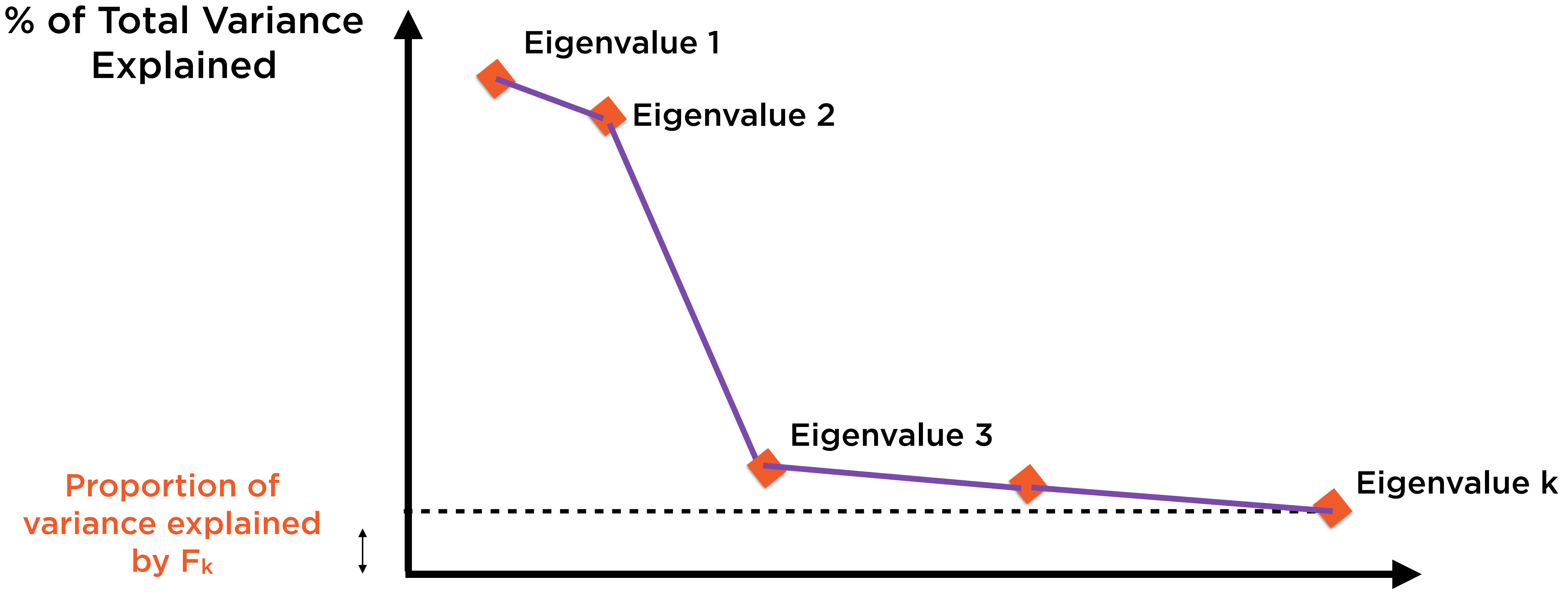


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Proportion of variance explained by F2
Eigenvalue 3
Eigenvalue k

Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k
Proportion of variance explained by F3

Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k
Proportion of variance explained by F_k

Use the Scree plot to determine how many principal components to discard

Results of PCA

Eigenvalues

tell importance of each principal component

Principal Components

for the largest eigenvalues can be used in regression

Eigenvectors

are needed to calculate the principal components

Correlated Random Variables

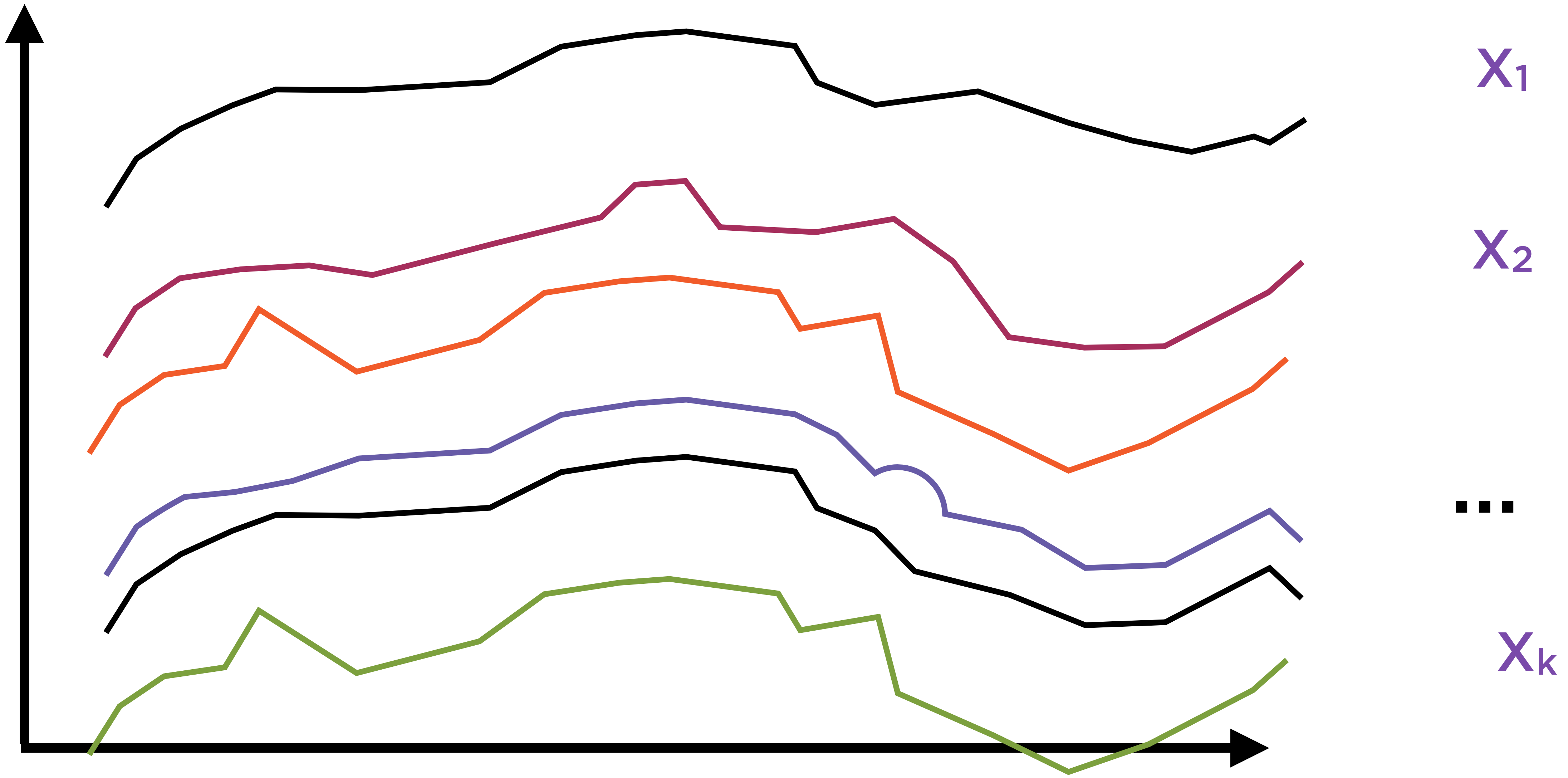
$$[\ X_1 \quad X_2 \quad X_3 \quad \dots \quad X_k\ ]$$

n rows

k columns

Each element X_i of this matrix is a vector with 1 column and n rows

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

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

Principal Components 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

PCA

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

Principal Components Analysis

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

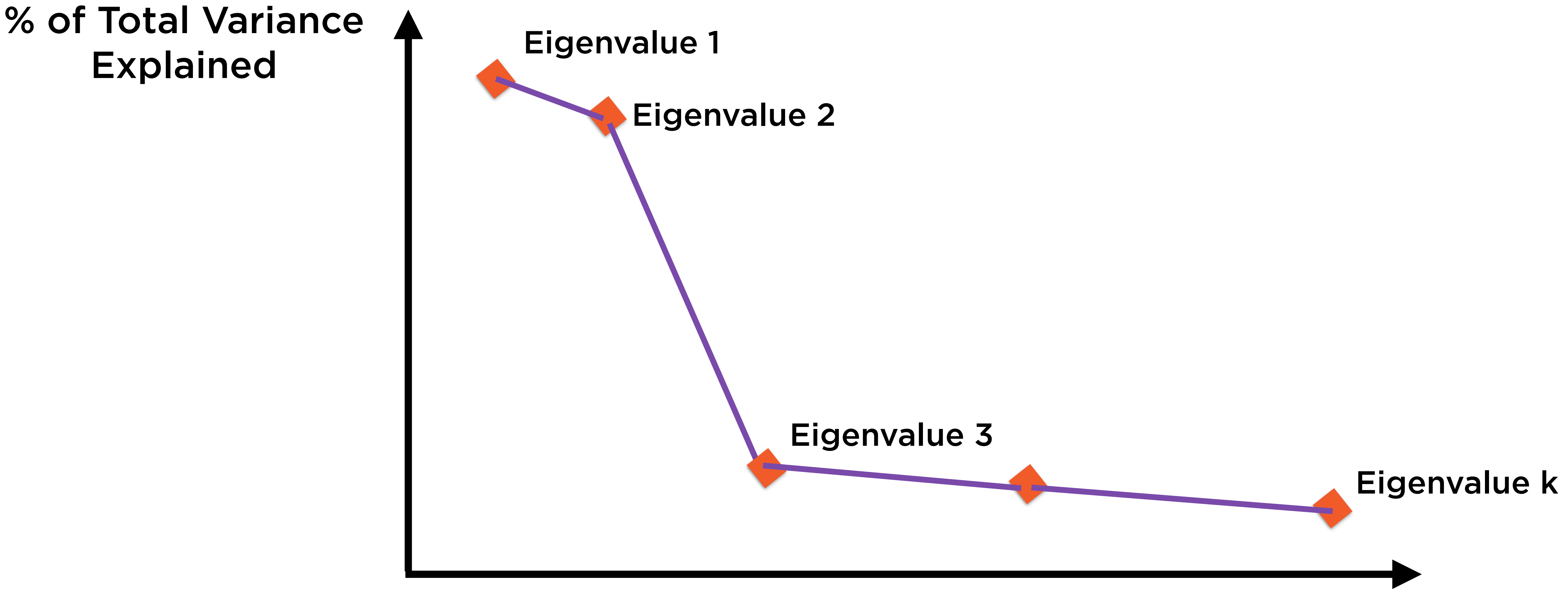
n rows

k columns

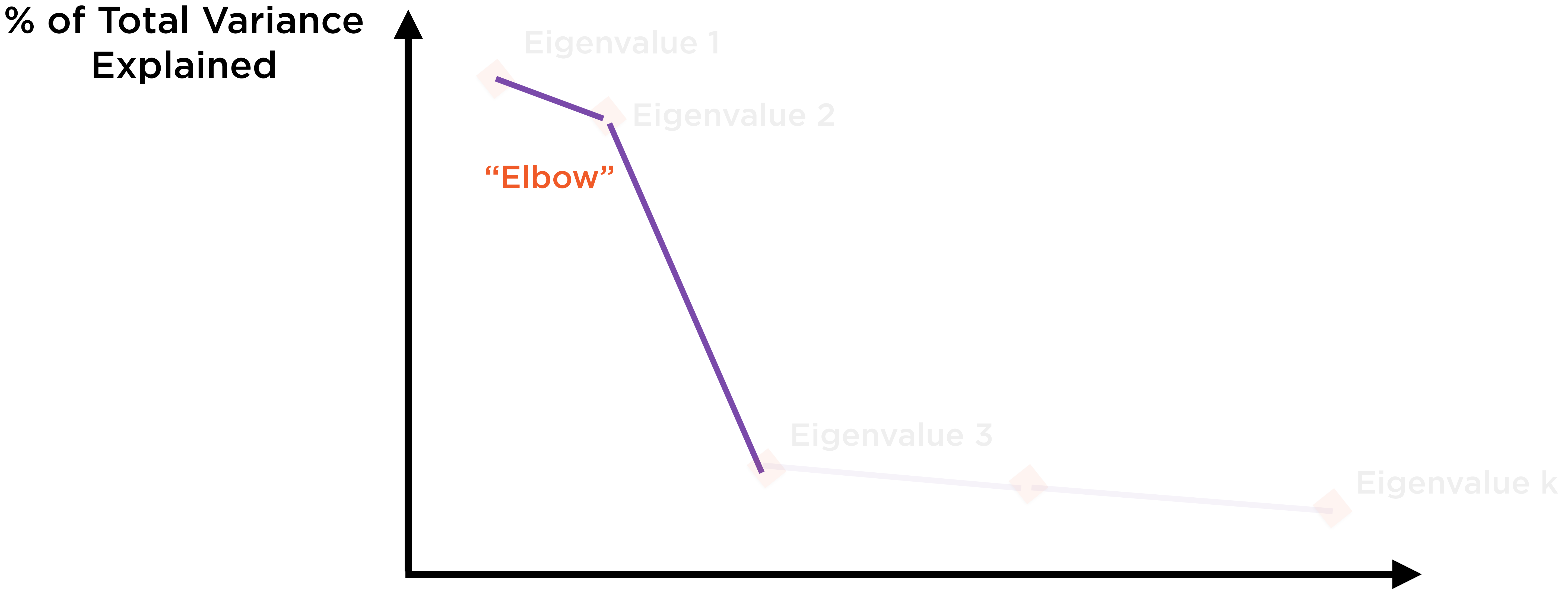
These vectors F_i are the principal components of the original vectors X_i

Discard “low-value” principal components using the eigenvalues e_i

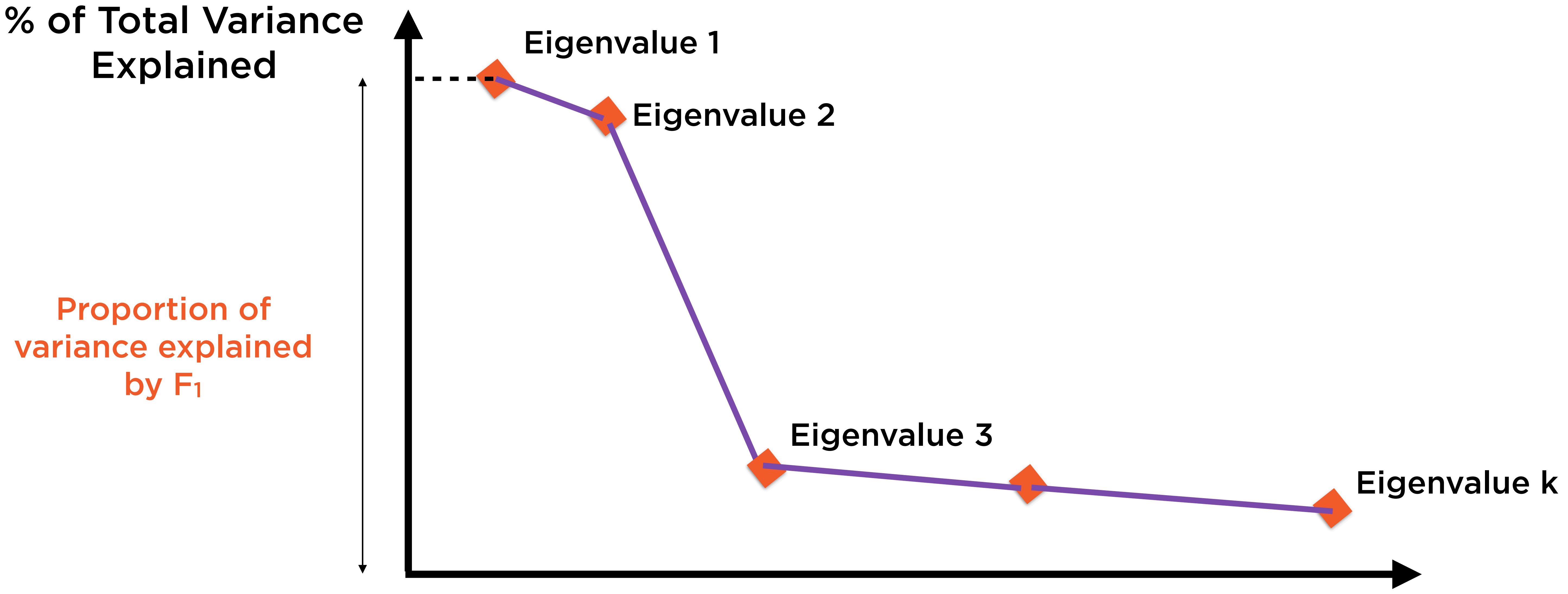
Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k

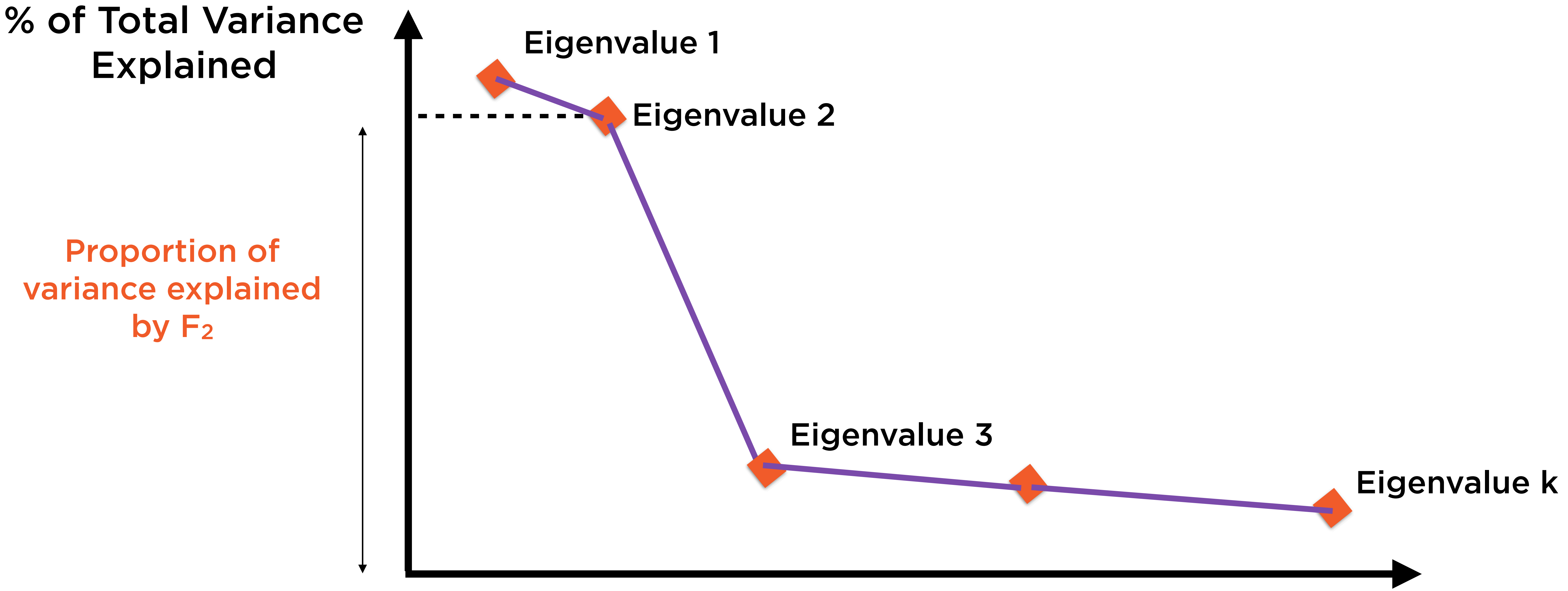
Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
"Elbow"
Eigenvalue 3
Eigenvalue k

Scree Plots

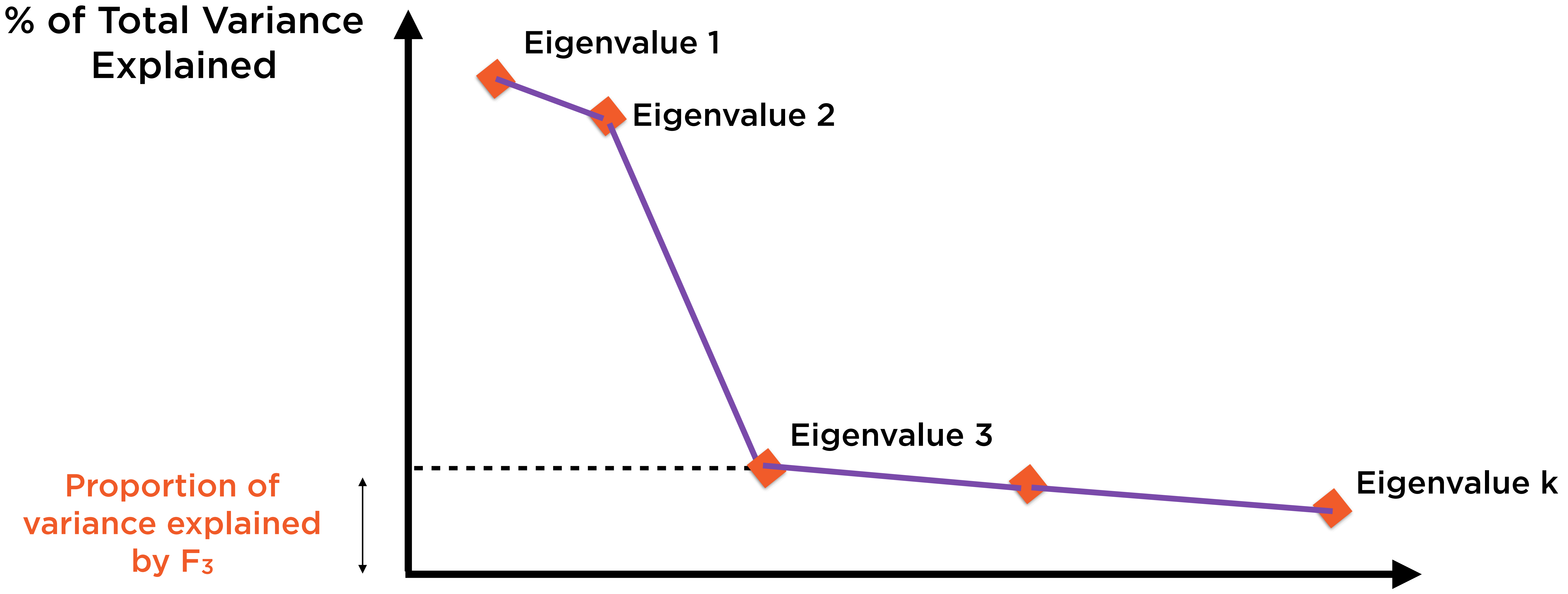

% of Total Variance Explained
Proportion of variance explained by F_1
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k

Scree Plots



% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Proportion of variance explained by F_2
Eigenvalue 3
Eigenvalue k

Scree Plots



% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k
Proportion of variance explained by F3

Principal Components Analysis

$$\left[\; F_1 \quad F_2 \quad \cancel{F_3} \quad \cancel{\cdots} \quad \cancel{F_k} \;\right]$$

n rows

k columns

Keep F_1 and F_2, discard the rest

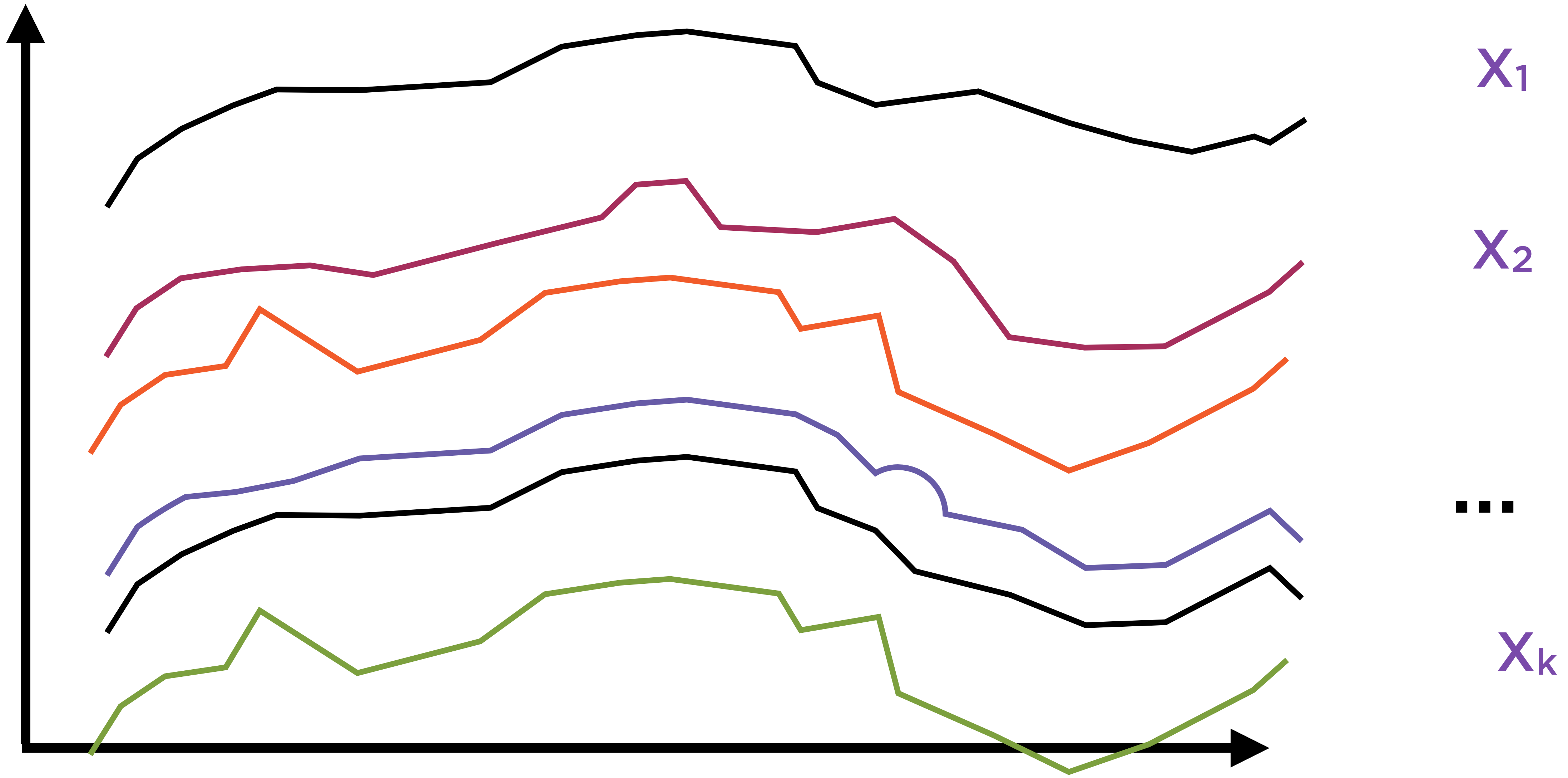
These 2 principal components explain the vast majority of the total variance in the original data

$$\left[\; F_1 \quad F_2 \;\right]$$

n rows

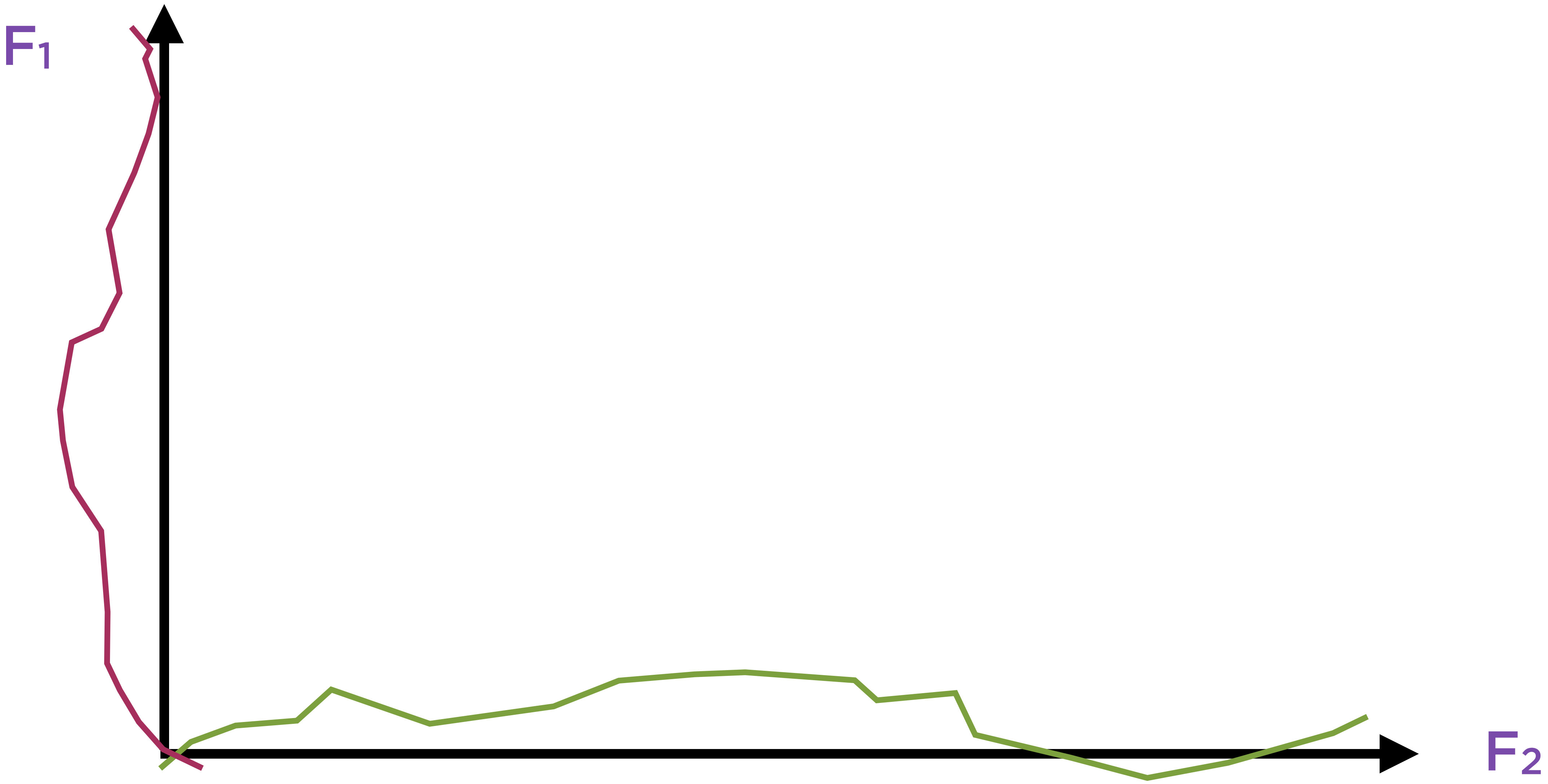
2 columns

Correlated X_i



Highly correlated variables are not suitable for use in regression

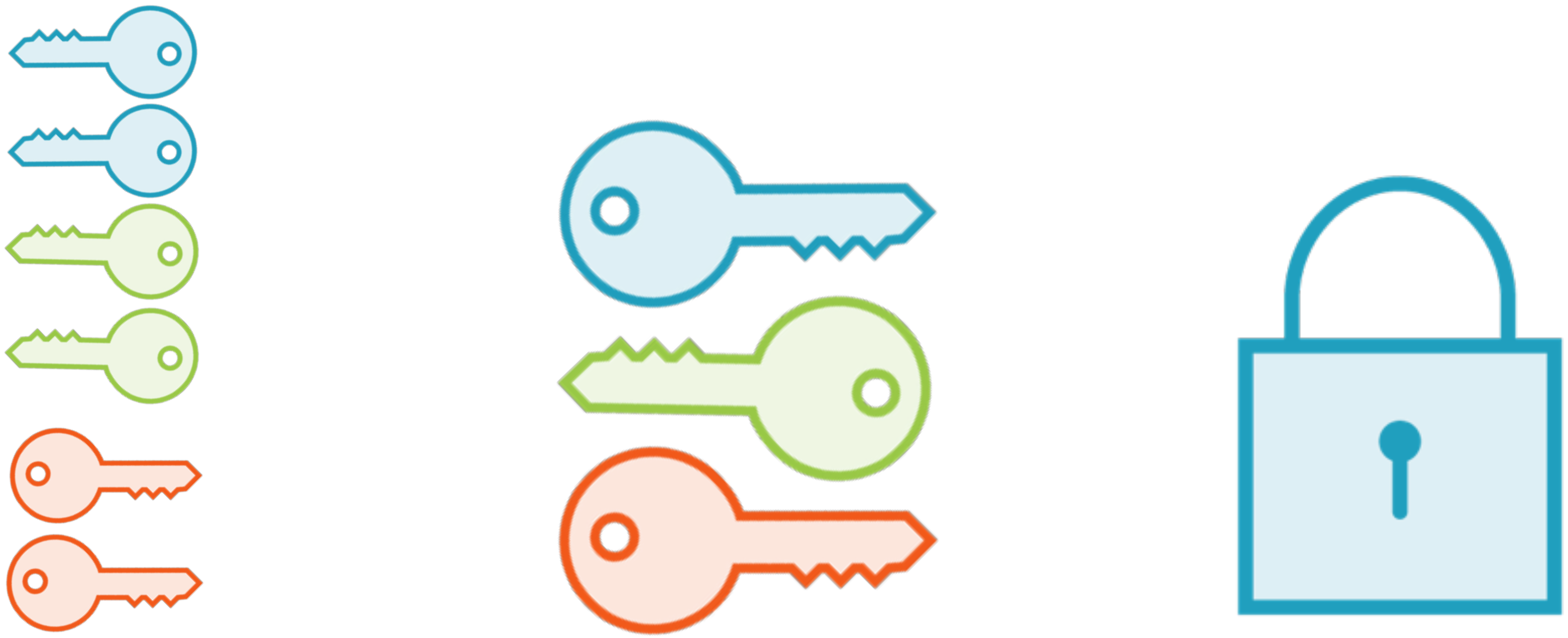
Uncorrelated F_i



Any of the principal components is perfectly uncorrelated with all others

Factor analysis: eliminating low-value principal components

Factor Analysis



Many Observed Causes

Few Underlying Causes

One Effect

Dimensionality Reduction

$$\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \cdots & x_{3k} \\ \cdots & & \cdots \\ x_{n1} & & x_{nk} \end{bmatrix}$$

k columns

PCA Factor Reduction

$$\begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ f_{31} & f_{32} \\ \cdots & \cdots \\ f_{n1} & f_{n2} \end{bmatrix}$$

2 columns

Results of PCA

Eigenvalues

tell importance of each principal component

Principal Components

for the largest eigenvalues can be used in regression

Eigenvectors

are needed to calculate the principal components

Principal Components Analysis

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

Eigenvalue Decomposition

Principal Components:

$[\ F_1\ F_2\ F_3 \ldots F_k\ ]$ n rows

k columns

Eigenvectors:

$[\ v_1\ v_2\ v_3 \ldots v_k\ ]$ k rows

k columns

Eigenvalues:

$[\ e_1\ e_2\ e_3 \ldots e_k\ ]$ 1 row

k columns

Problem: Finding Principal Component 1

Find F_1

$$F_1 = a_1X_1 + a_2X_2 + a_3X_3 \ldots + a_kX_k$$

such that

Variance(F_1) is maximised

subject to constraint

$$a_1^2 + a_2^2 + \ldots + a_k^2 = 1$$

This problem has a cookie-cutter solution in linear algebra - eigen decomposition

Solution: Finding Principal Component 1

Eigenvector:

$$v_1 = [\ a_1, a_2, a_3 \dots a_k\ ]$$

Principal Component:

$$F_1 = a_1X_1 + a_2X_2 + a_3X_3 \dots + a_kX_k$$

Each principal component is simply the matrix product of the original data matrix and the corresponding eigenvector

$$F = X \quad V$$

F	X	V
n rows, k columns	n rows, k columns	k rows, k columns

Matrix Multiplication

$$F = X v$$

$$= \begin{array}{c} \begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \dots & x_{3k} \\ \dots & & \dots \\ x_{n1} & & x_{nk} \end{bmatrix} \text{ n rows} \\ \text{k columns} \end{array} \quad \begin{array}{c} \begin{bmatrix} v_1 & v_2 & \dots & v_k \end{bmatrix} \text{ k rows} \\ \text{k columns} \end{array}$$

Matrix Multiplication

$$F = X v$$

$$= \begin{array}{c} \begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \dots & x_{3k} \\ \dots & & \dots \\ x_{n1} & & x_{nk} \end{bmatrix} \\ \text{k columns} \end{array} \text{n rows} \quad \begin{array}{c} \begin{bmatrix} a_1 & b_1 & k_1 \\ a_2 & b_2 & k_2 \\ a_3 & b_3 & k_3 \\ \dots & \dots & \dots \\ a_k & b_k & k_k \end{bmatrix} \\ \text{k columns} \\ \\ v_1 \quad v_2 \dots v_k \end{array} \text{k rows}$$

Matrix Multiplication

$$\underbrace{\begin{bmatrix} F_{11} & & F_{1k} \\ F_{21} & & F_{2k} \\ F_{31} & \dots & F_{3k} \\ \dots & & \dots \\ F_{n1} & & F_{nk} \end{bmatrix}}_{\text{k columns}} \text{n rows} = \underbrace{\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \dots & x_{3k} \\ \dots & & \dots \\ x_{n1} & & x_{nk} \end{bmatrix}}_{\text{k columns}} \text{n rows} \underbrace{\begin{bmatrix} a_1 & b_1 & k_1 \\ a_2 & b_2 & k_2 \\ a_3 & b_3 & k_3 \\ \dots & \dots & \dots \\ a_k & b_k & k_k \end{bmatrix}}_{\text{k columns}} \text{k rows}$$

$v_1 \quad v_2 \;\dots\; v_k$

Matrix Multiplication

$$
\begin{bmatrix} F_{11} & & F_{1k} \\ F_{21} & & F_{2k} \\ F_{31} & \ldots & F_{3k} \\ \ldots & & \ldots \\ F_{n1} & & F_{nk} \end{bmatrix}
=
\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \ldots & x_{3k} \\ \ldots & & \ldots \\ x_{n1} & & x_{nk} \end{bmatrix}
\begin{bmatrix} a_1 & b_1 & k_1 \\ a_2 & b_2 & k_2 \\ a_3 & b_3 & k_3 \\ \ldots & \ldots & \ldots \\ a_k & b_k & k_k \end{bmatrix}
$$

Matrix Multiplication

$$\begin{bmatrix} F_{11} & & F_{1k} \\ F_{21} & & F_{2k} \\ F_{31} & \cdots & F_{3k} \\ \cdots & & \cdots \\ F_{n1} & & F_{nk} \end{bmatrix} = \begin{bmatrix} x_{11} & & x_{1k} \\ \boxed{x_{21}} & & \boxed{x_{2k}} \\ x_{31} & \cdots & x_{3k} \\ \cdots & & \cdots \\ x_{n1} & & x_{nk} \end{bmatrix} \begin{bmatrix} a_1 & b_1 & k_1 \\ a_2 & b_2 & k_2 \\ a_3 & b_3 & k_3 \\ \cdots & \cdots & \cdots \\ a_k & b_k & k_k \end{bmatrix}$$

Matrix Multiplication

$$\begin{bmatrix} F_{11} & & F_{1k} \\ F_{21} & & F_{2k} \\ F_{31} & \dots & F_{3k} \\ \dots & & \dots \\ F_{n1} & & F_{nk} \end{bmatrix} = \begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \dots & x_{3k} \\ \dots & & \dots \\ x_{n1} & & x_{nk} \end{bmatrix} \begin{bmatrix} a_1 & b_1 & k_1 \\ a_2 & b_2 & k_2 \\ a_3 & b_3 & k_3 \\ \dots & \dots & \dots \\ a_k & b_k & k_k \end{bmatrix}$$

Matrix Multiplication

$$\underbrace{\begin{bmatrix} F_{11} \\ F_{21} \\ F_{31} \\ \dots \\ F_{n1} \end{bmatrix}}_{\text{1 column}} \ \text{n rows} \quad = \quad \underbrace{\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \dots & x_{3k} \\ \dots & & \dots \\ x_{n1} & & x_{nk} \end{bmatrix}}_{\text{k columns}} \ \text{n rows} \quad \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_k \end{bmatrix}}_{\text{1 column}} \ \text{k rows}$$

Matrix Multiplication

F_i	$=$	X	v_i
n rows, 1 column		n rows, k columns	k rows, 1 column

Each principal component is the matrix product of the original data and the corresponding eigenvector

Why Principal Components Are Useful

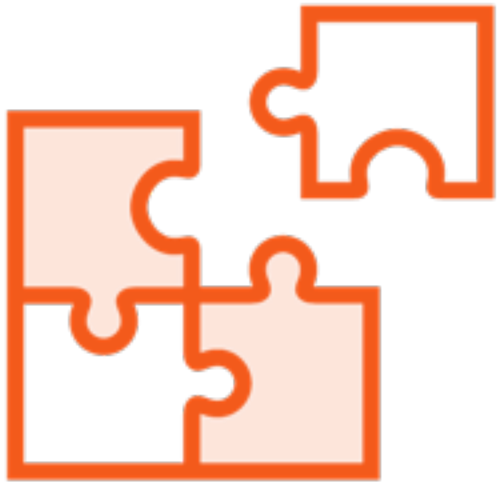
Benefits of Principal Components

Dimensionality Reduction

Cut through the clutter

Latent Factor Identification

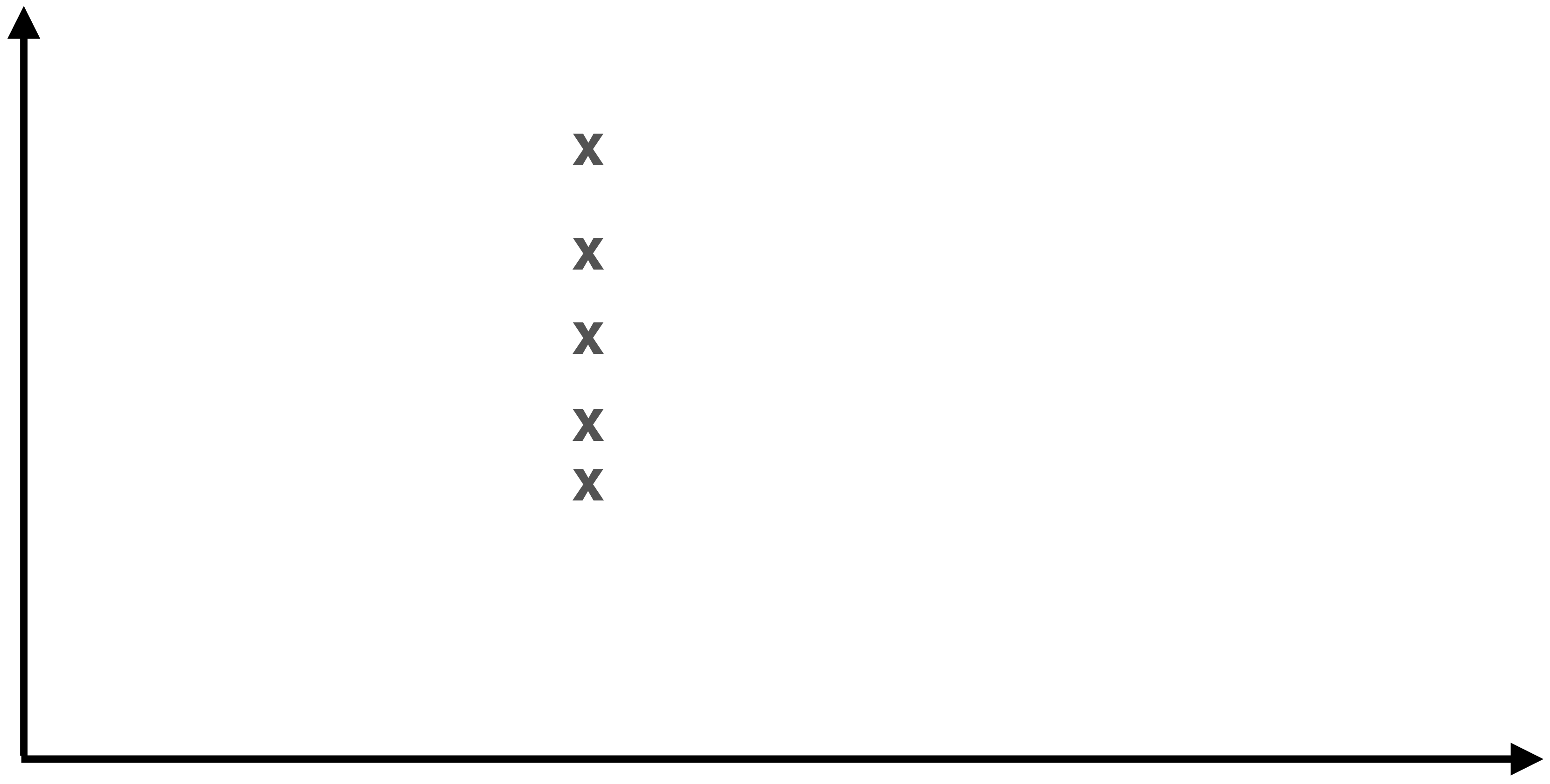
Find underlying causes



Missing Data & Scenario Generation

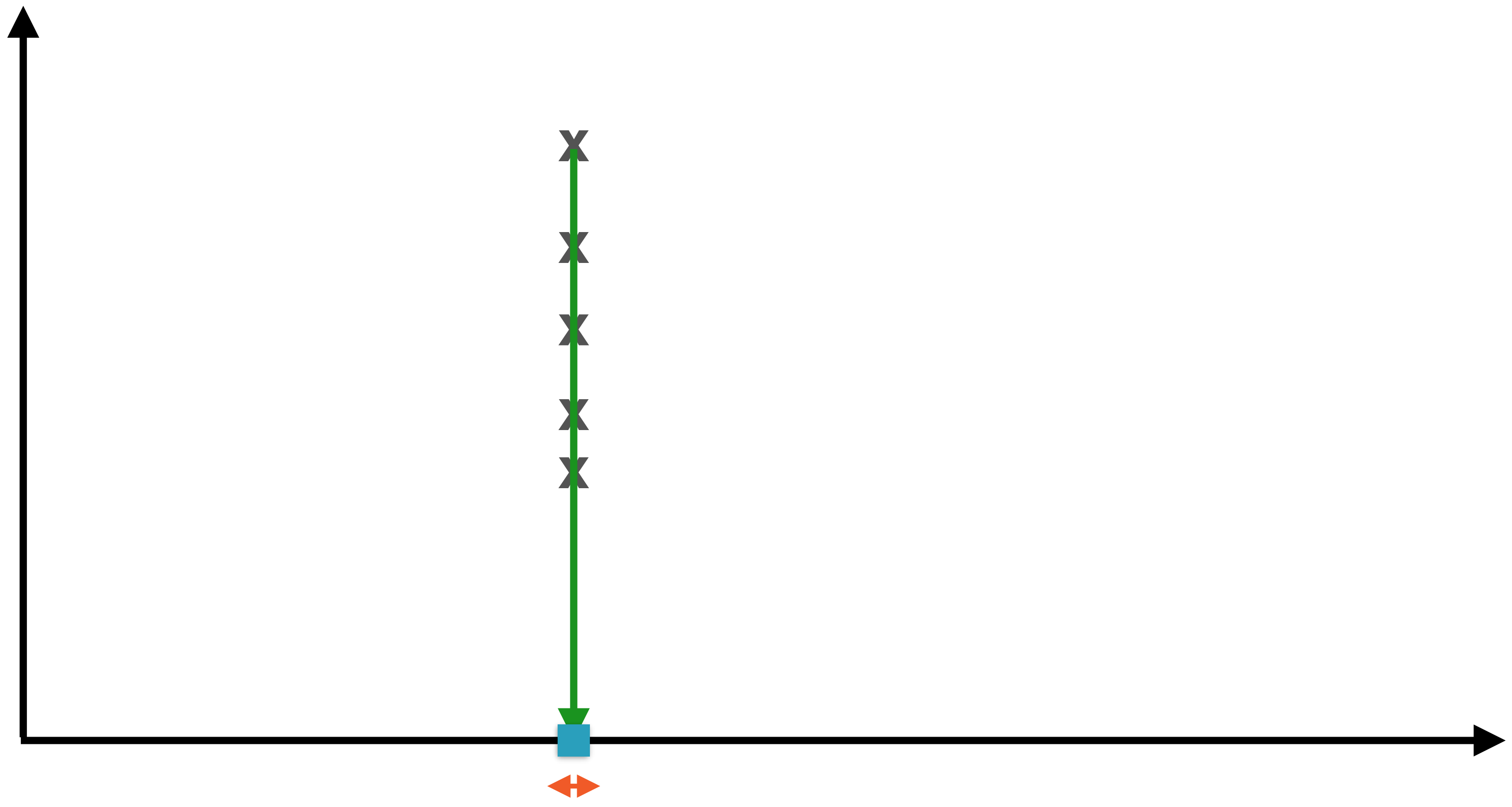
Extrapolate or interpolate data

A Question of Dimensionality



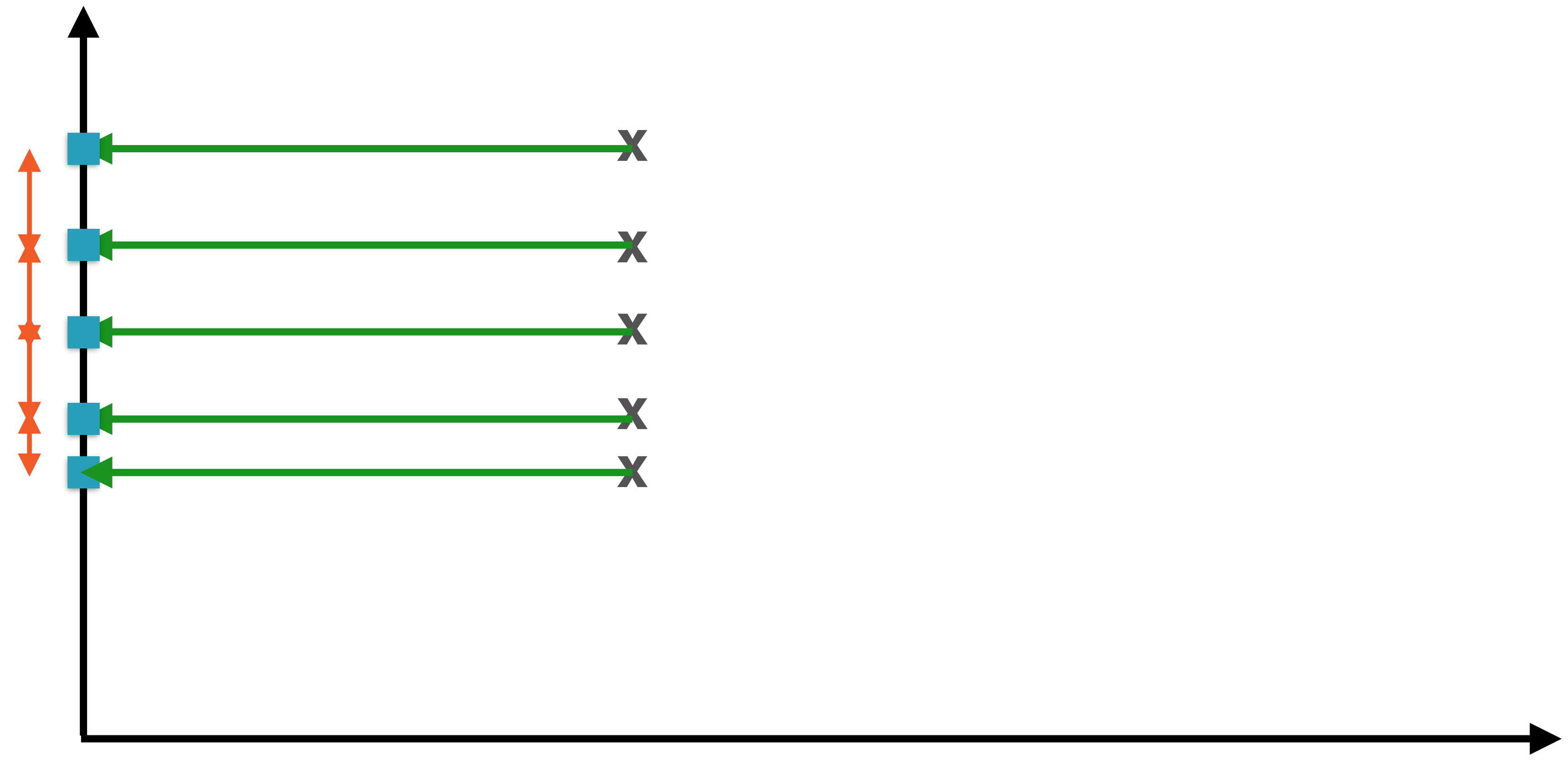
Pop quiz: Do we really need two dimensions to represent this data?

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

Principal Components Analysis

[X_1 X_2 X_3 ... X_k] →

Eigenvalue Decomposition

→

Principal Components:

[F_1 F_2 F_3 ... F_k] n rows

k columns

Eigenvectors:

[v_1 v_2 v_3 ... v_k] k rows

k columns

Eigenvalues:

[e_1 e_2 e_3 ... e_k] 1 row

k columns

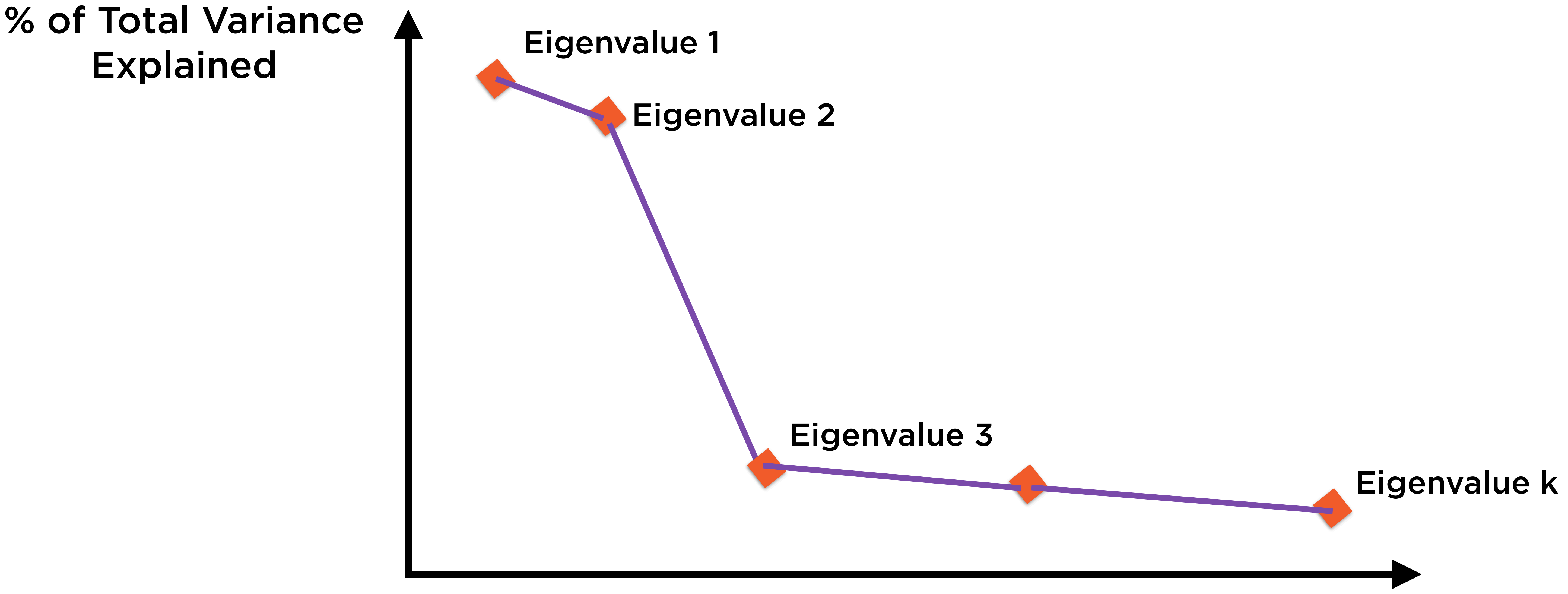
PCA for Dimensionality Reduction

$$\underbrace{[\; F_1 \quad F_2 \quad F_3 \quad \dots \quad F_k \;]}_{\text{k columns}} \quad \updownarrow \text{ n rows}$$

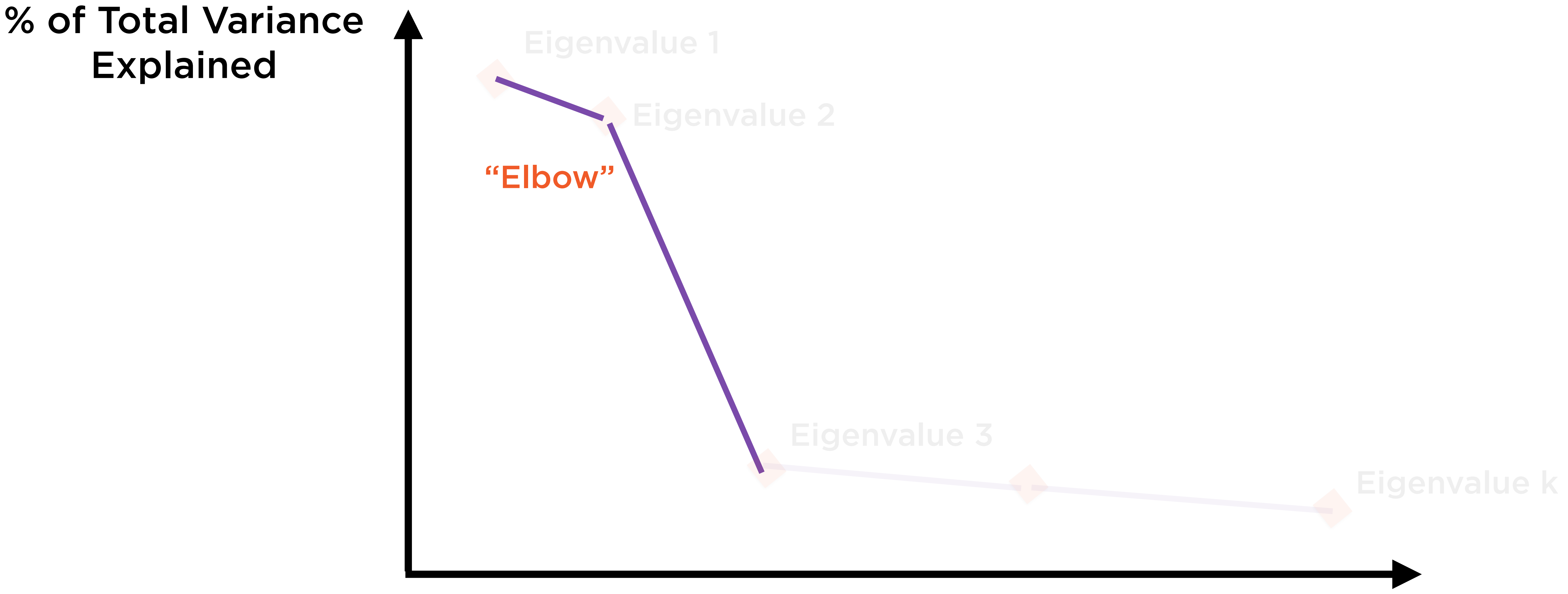
These vectors F_i are the principal components of the original vectors X_i

Discard “low-value” principal components using the eigenvalues e_i

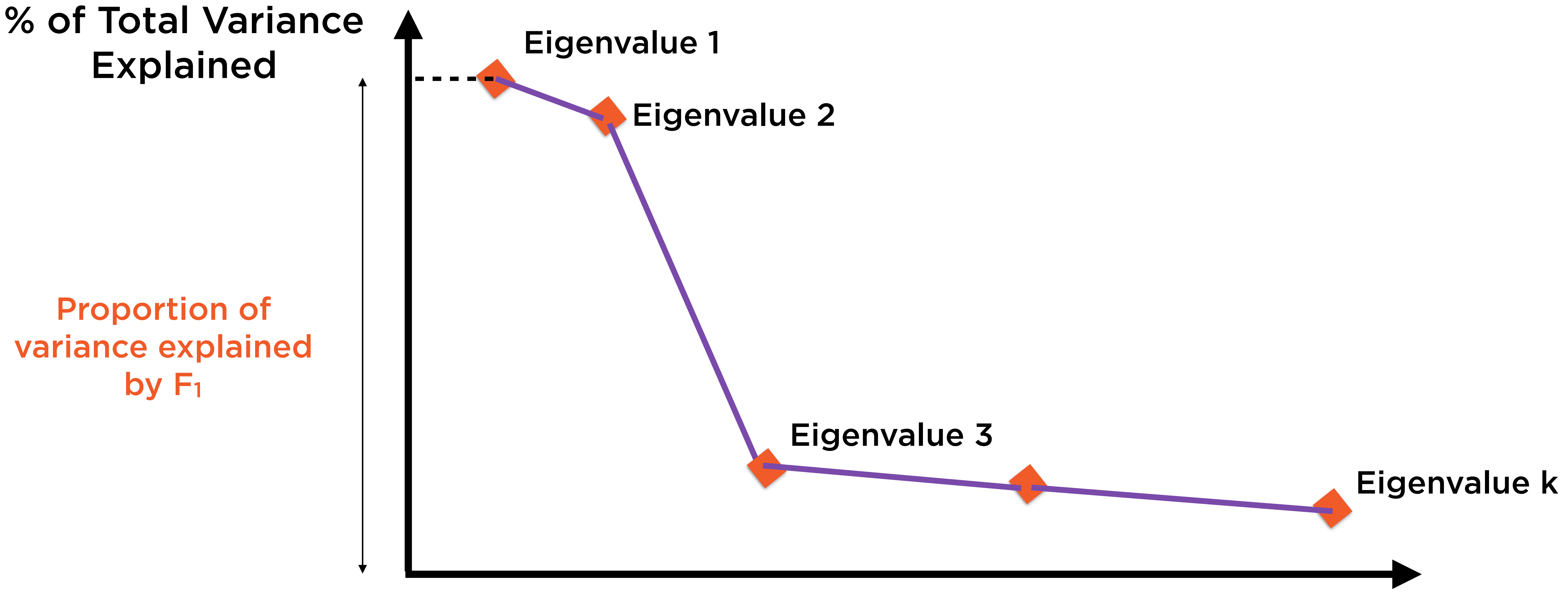
PCA for Dimensionality Reduction



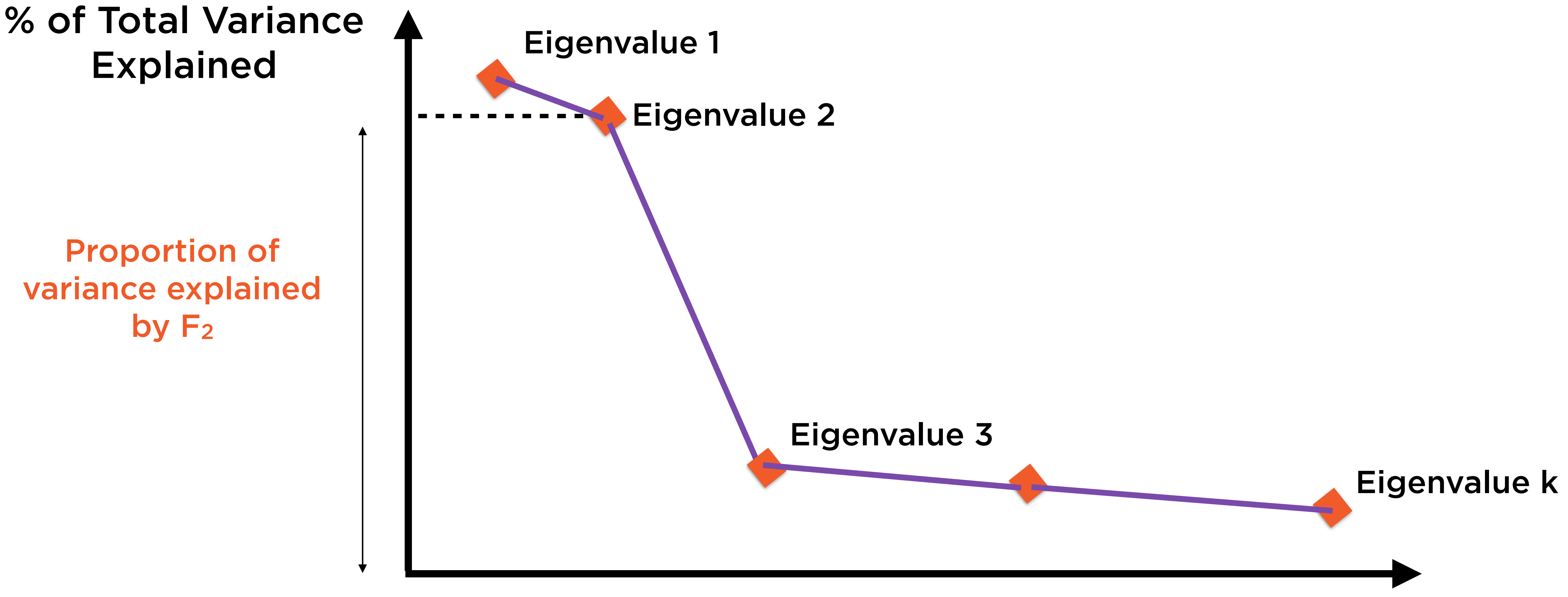
PCA for Dimensionality Reduction



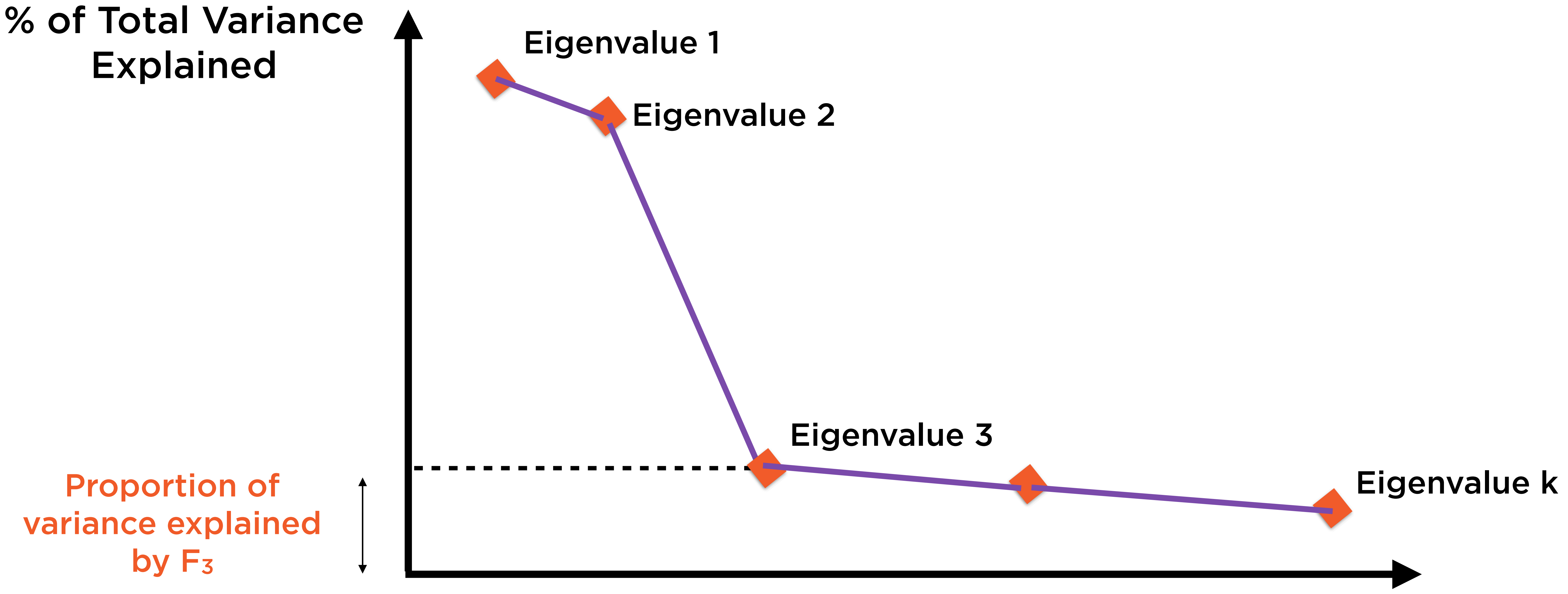
PCA for Dimensionality Reduction



PCA for Dimensionality Reduction



PCA for Dimensionality Reduction



PCA for Dimensionality Reduction

$$\left[\; F_1 \quad F_2 \quad \cancel{F_3} \quad \cancel{\cdots} \quad \cancel{F_k} \;\right]$$

n rows

k columns

Keep F_1 and F_2, discard the rest

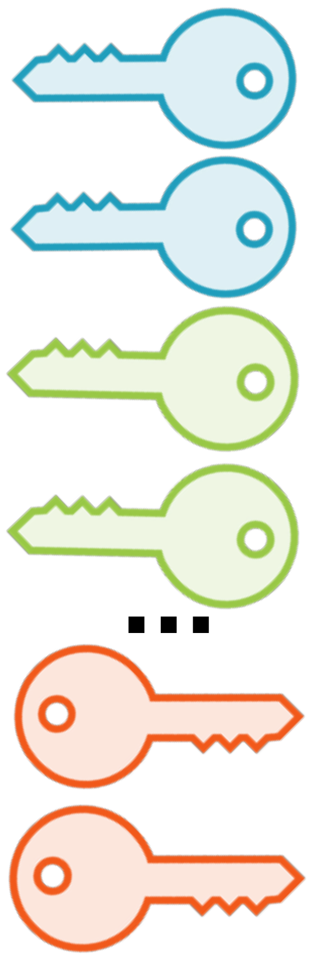
These 2 principal components explain the vast majority of the total variance in the original data

$$\left[\; F_1 \quad F_2 \;\right]$$

n rows

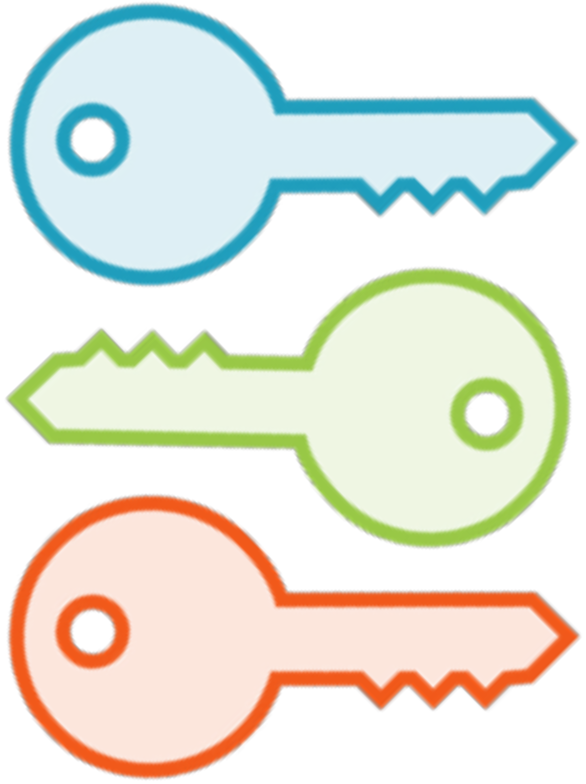
2 columns

Success as a Salesperson



Many Observed Causes

Cold calls, experience, social media followers, perceived honesty, billing punctuality...



Few Underlying Causes

Personality traits

One Effect

Success as a salesperson

Kitchen Sink Regression

Proposed Regression Equation:

BONUS = A + B COLDCALLS + C EXPERIENCE + D NUMFOLLOWERS + E HONESTY + F PUNCTUALITY + ...

PCA Regression

Proposed Regression Equation:

BONUS = A + B COLDCALLS + C EXPERIENCE + D NUMFOLLOWERS + E HONESTY + F PUNCTUALITY + ...

PCA

Modified Regression Equation:

BONUS = A + B F_1 + C F_2

Adding Random Variables

$$P = w_1E + w_2D + w_3G \ldots + w_kA$$

P_i = % return of stock portfolio on day i

Portfolio P consists of w_1 stocks of Exxon, w_2 of the Dow, w_3 of Google and w_k of Apple

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

Analysing the sum of random variables is an extremely common use-case

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

n rows

k columns

PCA

$$y = F_1 + F_2$$

n rows

2 columns

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

Mean(y)

Simple - mean of sum is sum of means

Variance(y)

Tricky - requires use of covariance matrix

Adding related variables is difficult,
adding independent variables is easy

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \dots + X_k$$

$$\text{Variance}(y) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Covariance}(X_i, X_j)$$

k^2 terms

If the X variables are independent, we can easily find the variance of the sum

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var(X_k) \end{bmatrix}$$

k rows

k columns

Diagonal elements are the variances

Adding Independent Random Variables

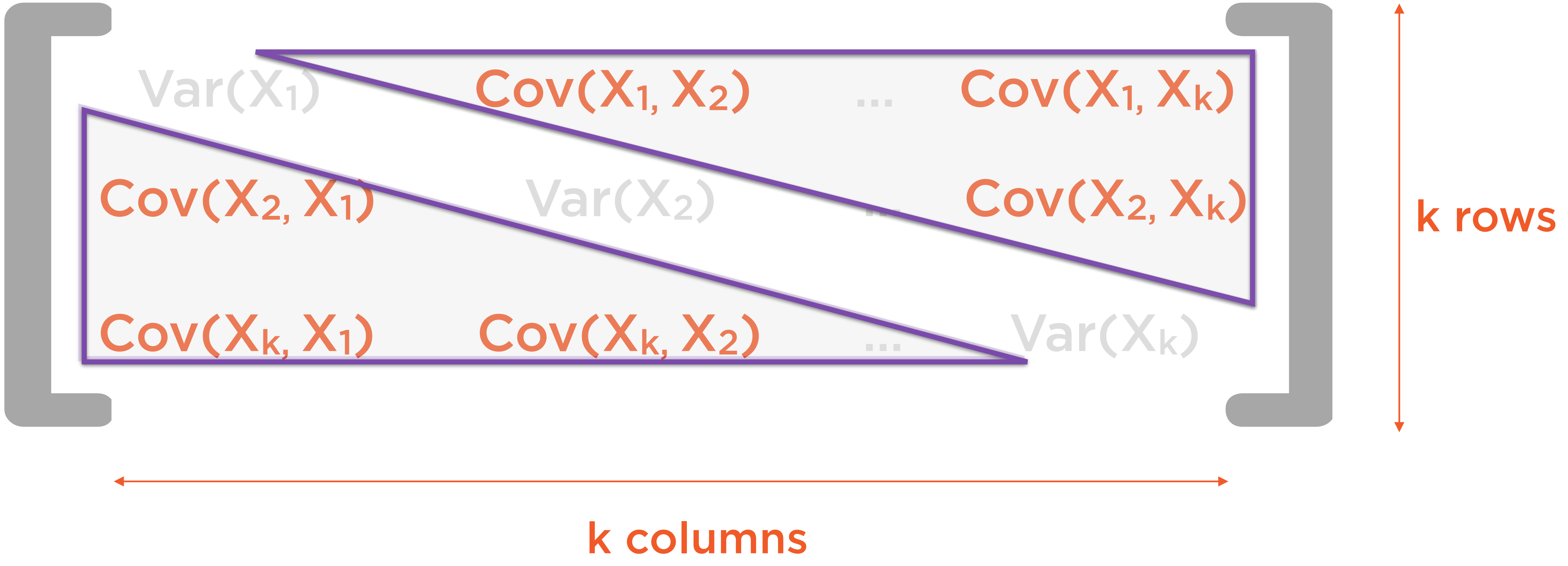
$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{array}{cl}
\begin{bmatrix}
Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\
Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_k) \\
Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var(X_k)
\end{bmatrix} & \text{k rows} \\
\text{k columns} &
\end{array}$$

Add all the diagonal elements...

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$



...and half the sum of the off-diagonal entries

Adding Independent Random Variables

$$y = F_1 + F_2$$

$$\begin{array}{cl} \begin{bmatrix} Var(F_1) & 0 \\ 0 & Var(F_2) \end{bmatrix} & \updownarrow \text{2 rows} \\ \xleftrightarrow{\text{2 columns}} & \end{array}$$

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\text{Variance}(y) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Covariance}(X_i, X_j)$$

k^2 terms

Calculating kxk full covariance matrix is difficult

Adding Independent Random Variables

$$y = F_1 + F_2$$

$$\text{Variance}(y) = \text{Variance}(F_1) + \text{Variance}(F_2)$$

2 terms

Calculating 2x2 diagonal covariance matrix after PCA is very simple

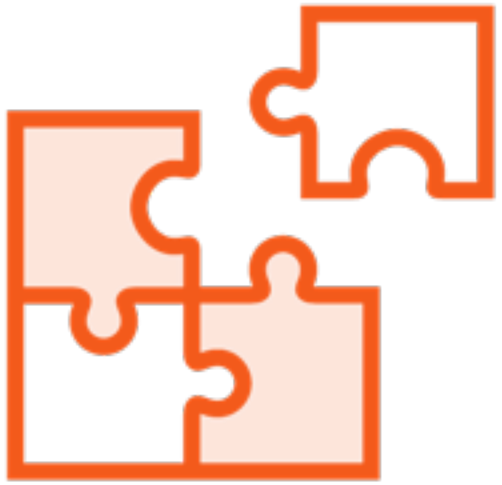
Benefits of Principal Components

Dimensionality Reduction

Cut through the clutter

Latent Factor Identification

Find underlying causes



Missing Data & Scenario Generation

Extrapolate or interpolate data

PCA as ML-based Factor Extraction

Rule-based

Human experts identify and extract factors

ML-based

Algorithm identifies and extracts factors

PCA for Latent Factor Identification

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

PCA

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

PCA for Latent Factor Identification

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

$$[\ L_1 \quad L_2 \quad L_3 \quad \ldots \quad L_k\ ]$$

Exploratory Factor Analysis: Experts trace back principal components to observable factors

5 Latent Factors in Psychology

Openness

Conscientiousness

Extraversion

Agreeableness

Neuroticism

3 Latent Factors in Stock Returns

Market Movements

Interest Rates

Industry Sectors

3 Latent Factors in Bond Returns

Trend

Tilt

Convexity

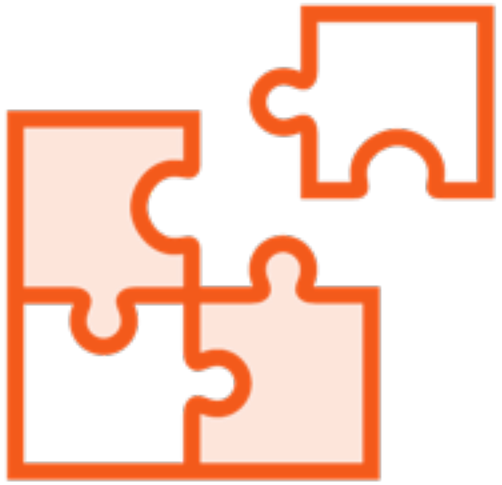
Benefits of Principal Components

Dimensionality Reduction

Cut through the clutter

Latent Factor Identification

Find underlying causes



Missing Data & Scenario Generation

Extrapolate or interpolate data

Missing Data Generation

$$FB = w_1GOOG + w_2AAPL + w_3SP500 + \dots + w_kMSFT$$

5 years

Facebook's IPO was in 2012, several years after other major tech companies

Missing Data Generation

$$FB = w_1 GOOG + w_2 AAPL + w_3 SP500 + \ldots + w_k MSFT$$

5 years

PCA

$$FB = F_1 + F_2$$

5 years

Missing Data Generation

$$FB = F_1 + F_2$$

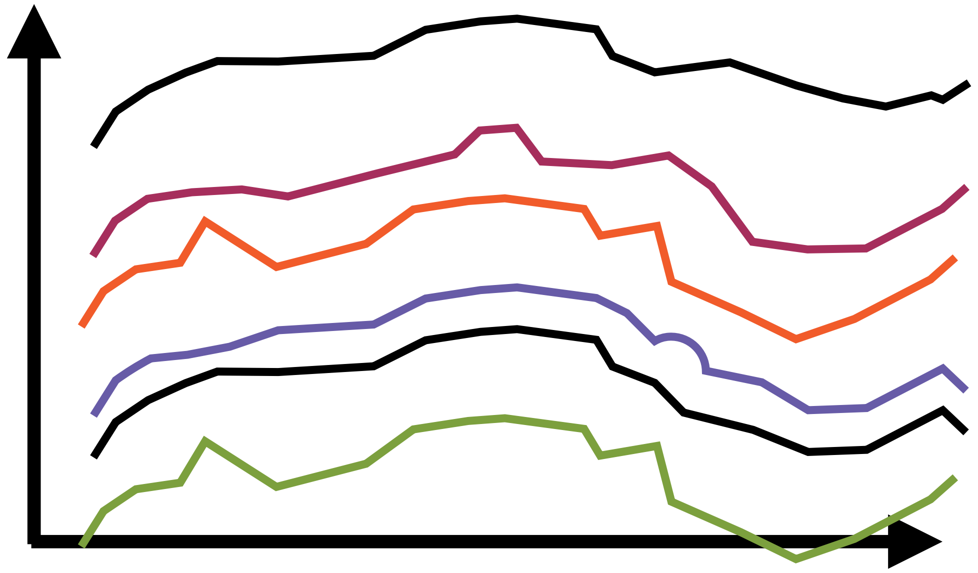
5 years

$$FB_{extrapolated} = F_1 + F_2$$

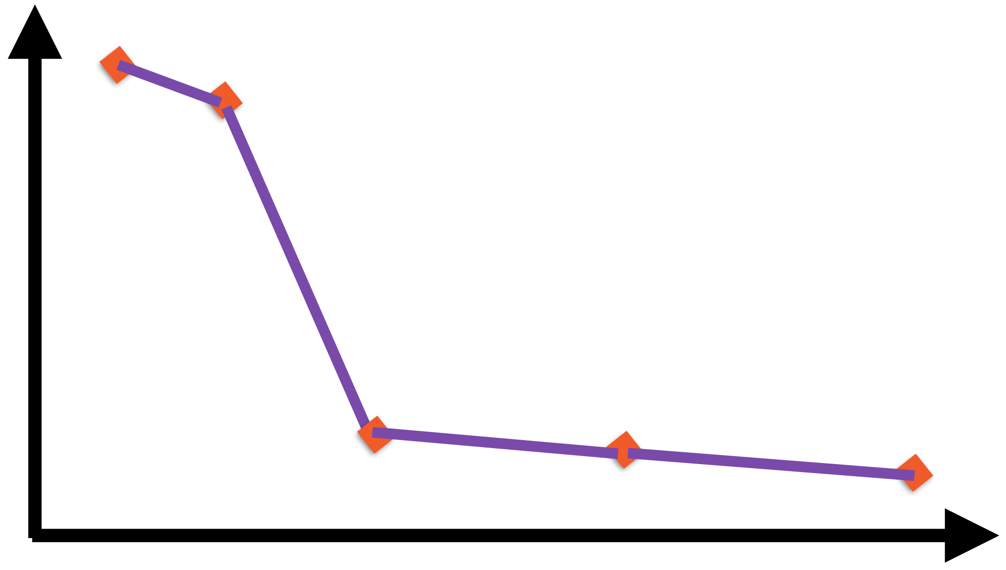
10 years

When Not to Use PCA

PCA’s Forte

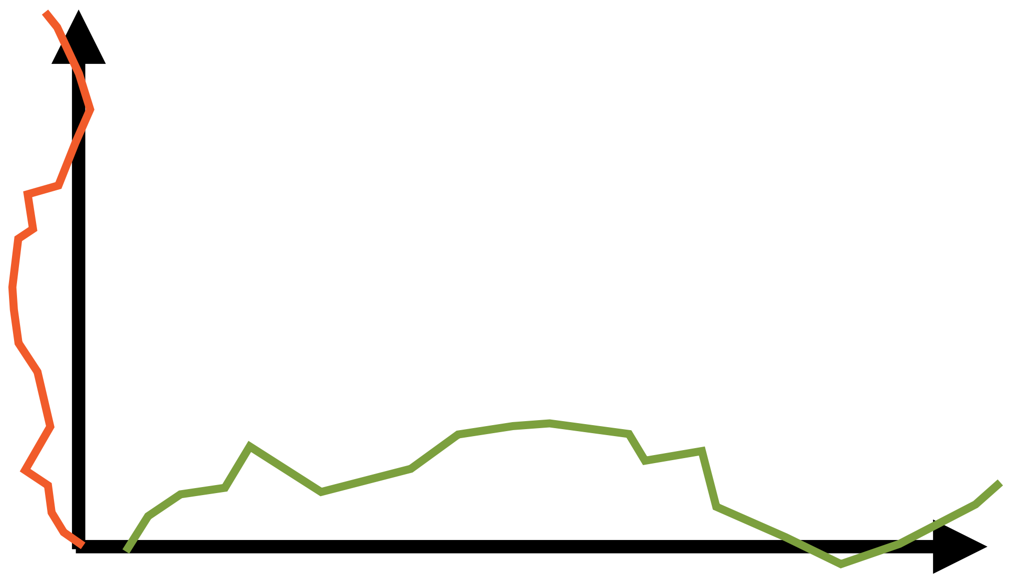


Many, Highly Correlated Xi

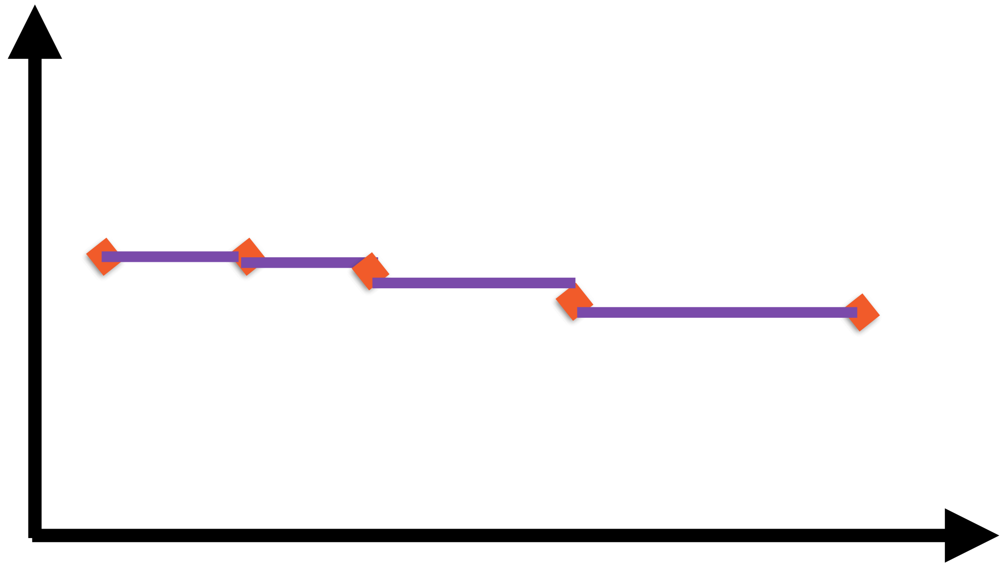


Unequal Eigenvalues

PCA’s Weak Spots

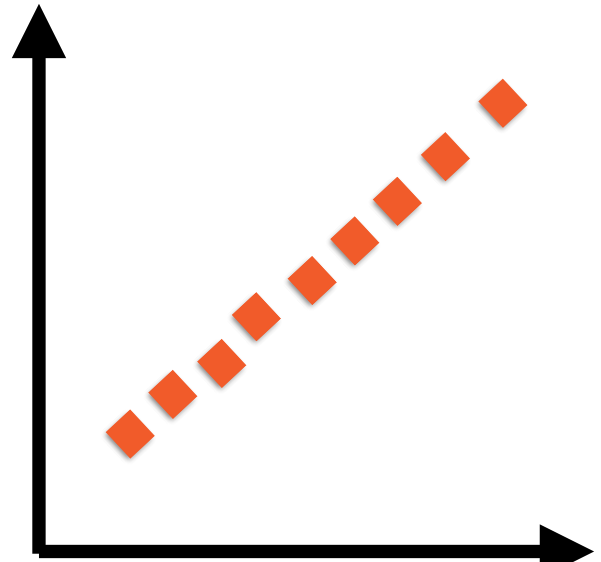


Few, Uncorrelated Xi



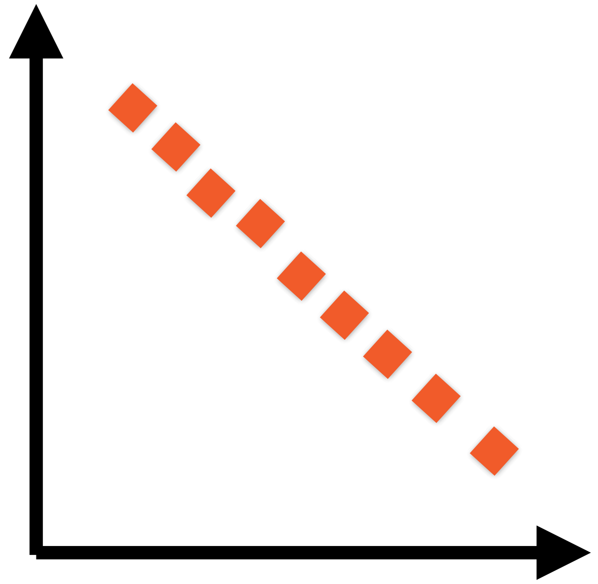
Almost Equal Eigenvalues

PCA for Highly Correlated Data



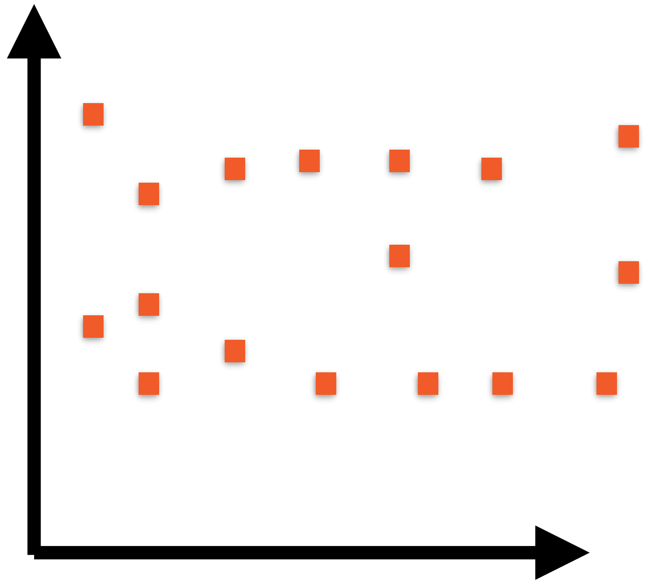
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



Correlation = 0

Changes in X independent* of changes in Y

Correlation and Covariance

$$\text{Correlation }(x,y) = \frac{\text{Covariance }(x,y)}{\sqrt{\text{Variance }(x)}\sqrt{\text{Variance }(y)}}$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Covariance Matrix

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

$$\begin{bmatrix} \sigma^2_{x_1} & \sigma_{x_1 x_2} & \cdots & \sigma_{x_1 x_k} \\ \sigma_{x_2 x_1} & \sigma^2_{x_2} & \cdots & \sigma_{x_2 x_k} \\ \sigma_{x_k x_1} & \sigma_{x_k x_2} & \cdots & \sigma^2_{x_k} \end{bmatrix}$$

k rows

k columns

Each element is the **covariance** of two random variables

Correlation Matrix

$$
\begin{array}{c}
[\ X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\ ] \\
\begin{bmatrix}
\rho_{x_1} & \rho_{x_1 x_2} & \cdots & \rho_{x_1 x_k} \\
\rho_{x_2 x_1} & \rho_{x_2} & \cdots & \rho_{x_2 x_k} \\
\rho_{x_k x_1} & \rho_{x_k x_2} & \cdots & \rho_{x_k}
\end{bmatrix}
\end{array}
\quad \text{k rows}
$$

k columns

Each element is the correlation of two random variables

Correlation Matrix

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

$$\begin{bmatrix} 1 & \rho_{x_1x_2} & \cdots & \rho_{x_1x_k} \\ \rho_{x_2x_1} & 1 & \cdots & \rho_{x_2x_k} \\ \rho_{x_kx_1} & \rho_{x_kx_2} & \cdots & 1 \end{bmatrix}$$

k rows

k columns

Diagonal elements are always 1

PCA for Highly Correlated Data

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

$$\begin{bmatrix} 1 & \rho_{x_1 x_2} & \cdots & \rho_{x_1 x_k} \\ \rho_{x_2 x_1} & 1 & \cdots & \rho_{x_2 x_k} \\ \rho_{x_k x_1} & \rho_{x_k x_2} & \cdots & 1 \end{bmatrix}$$

Rule-of-thumb: If average absolute values of off-diagonal entries is less than 0.3, PCA not a great idea

Factor Analysis: Excel, R or Python?

Excel

Need to implement using VBA

R

In-built functionality

Python

In-built functionality

Summary

Principal components contain within them all of the information in a dataset

PCA relies on a common mathematical technique called eigen decomposition

Eigenvalues help us decide which components to keep and discard

PCA helps with dimensionality reduction as well as exploratory factor analysis