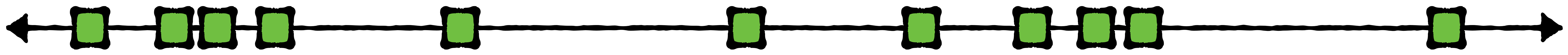
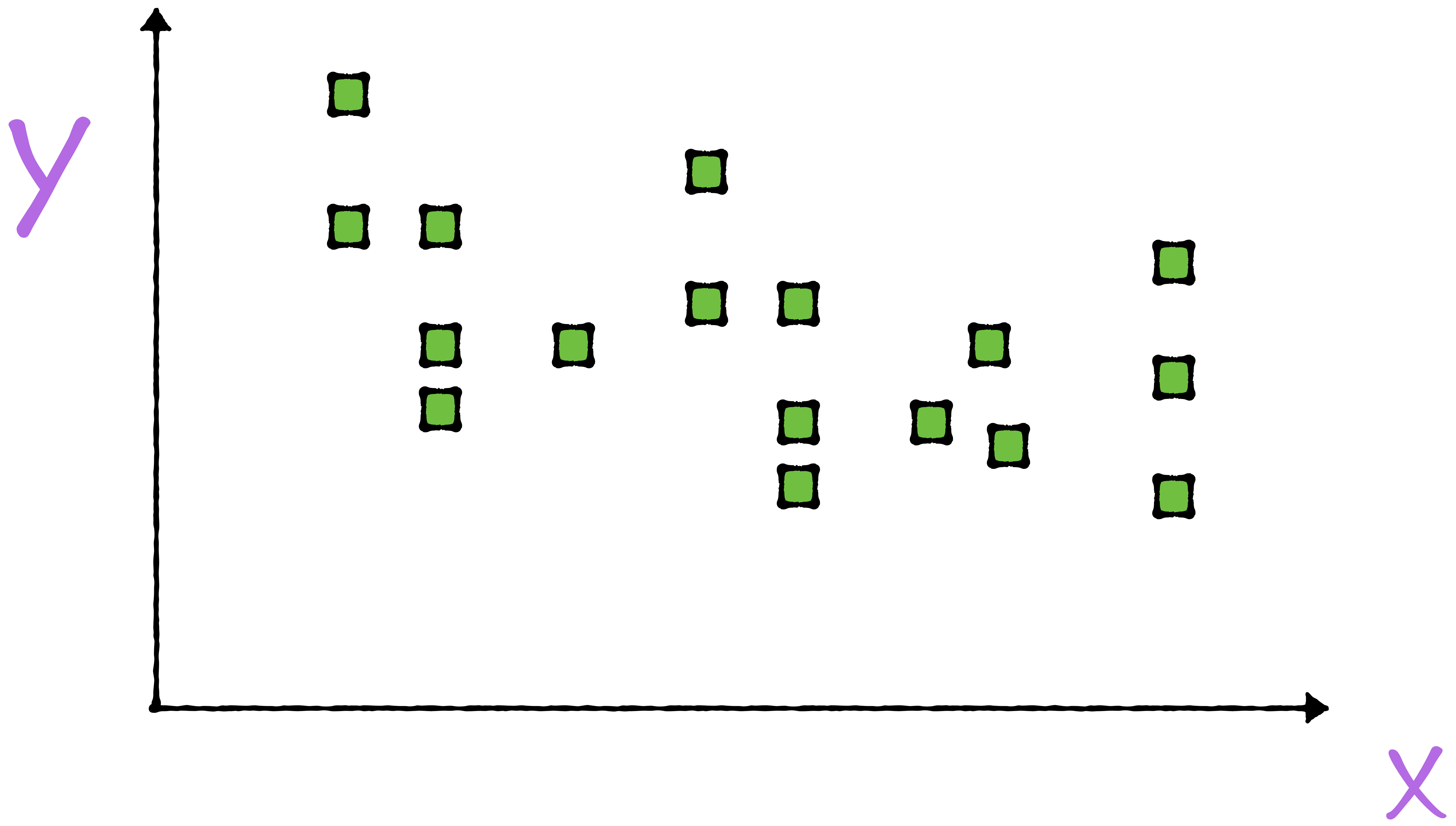
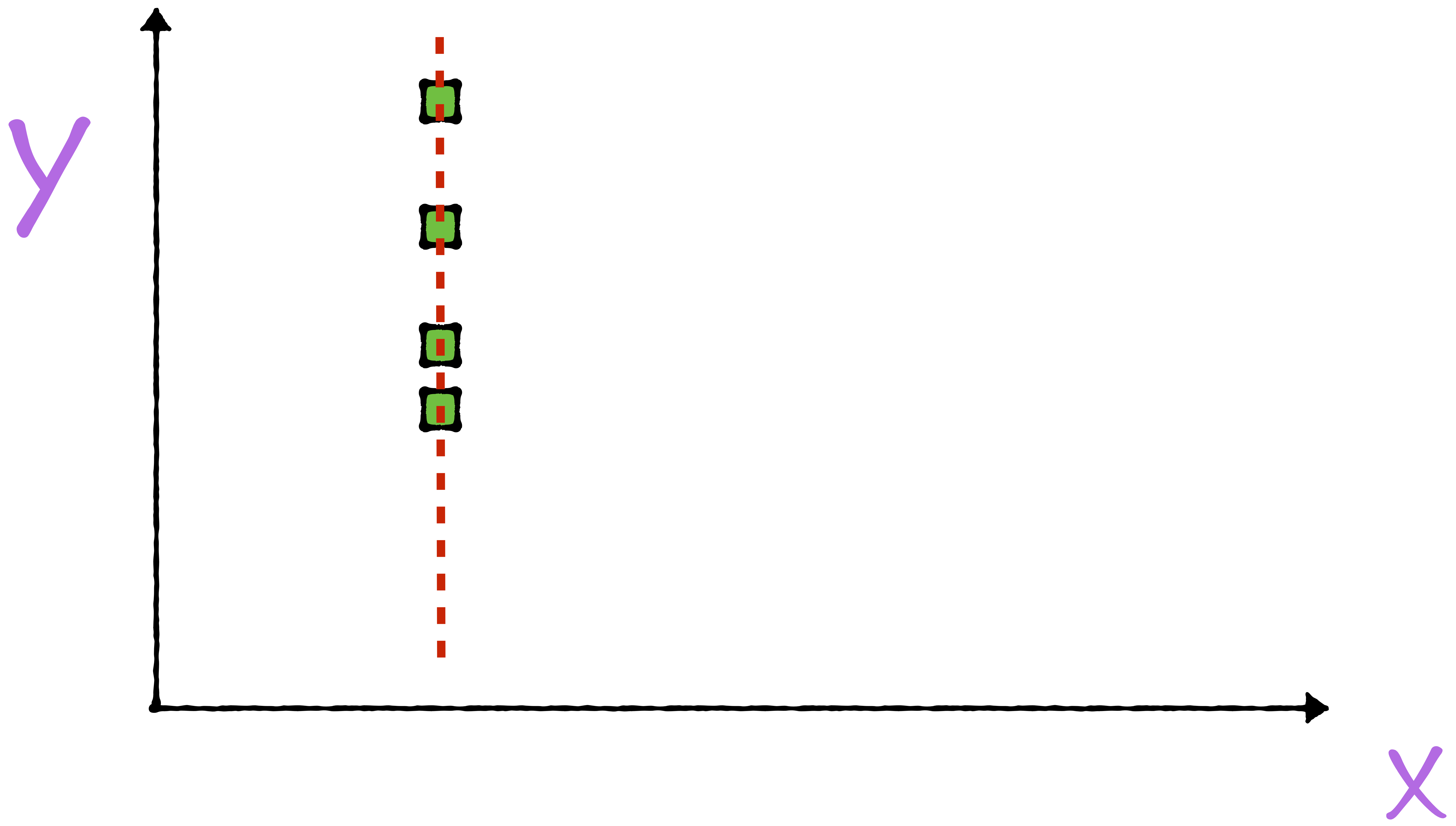
One dimensional data

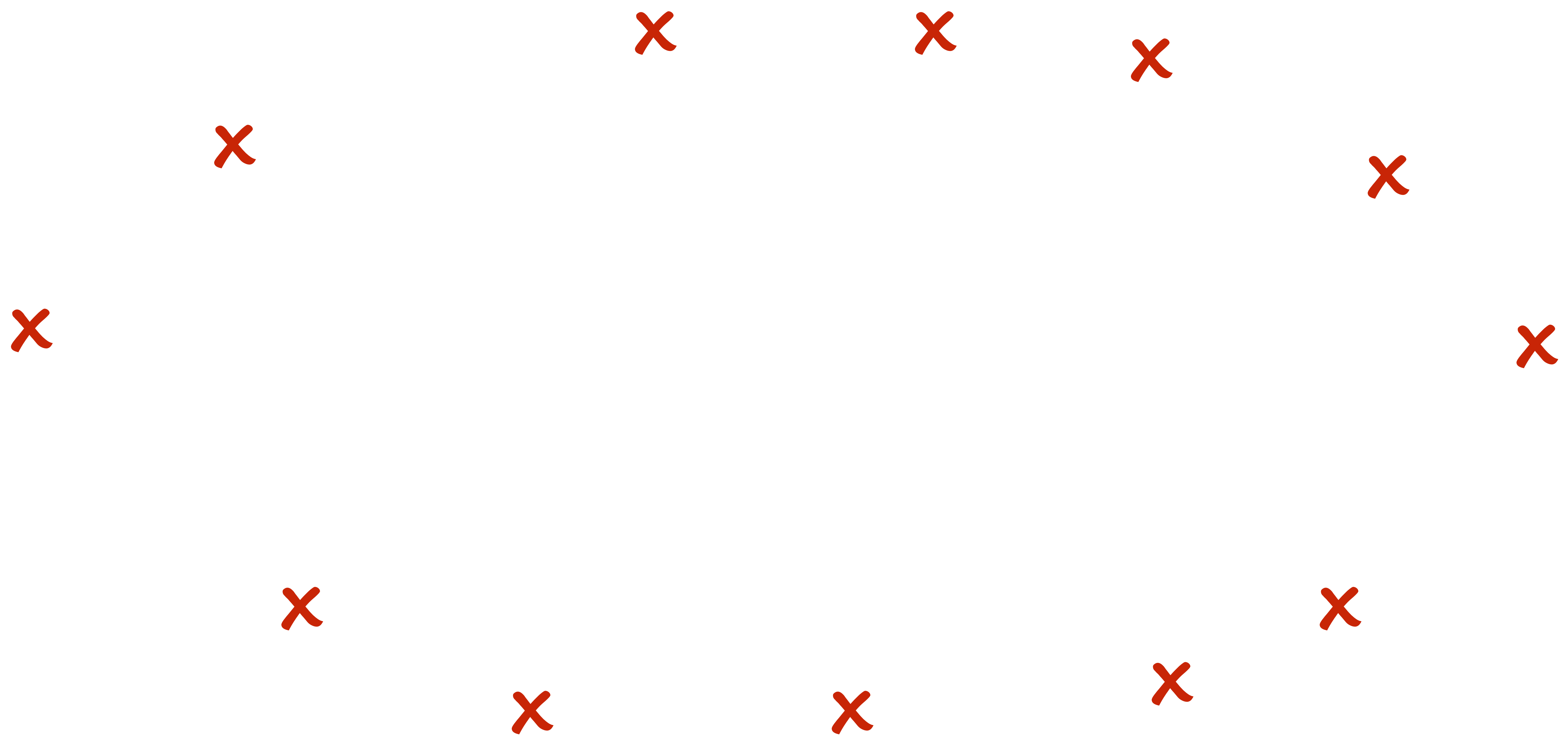


2 Dimensional data

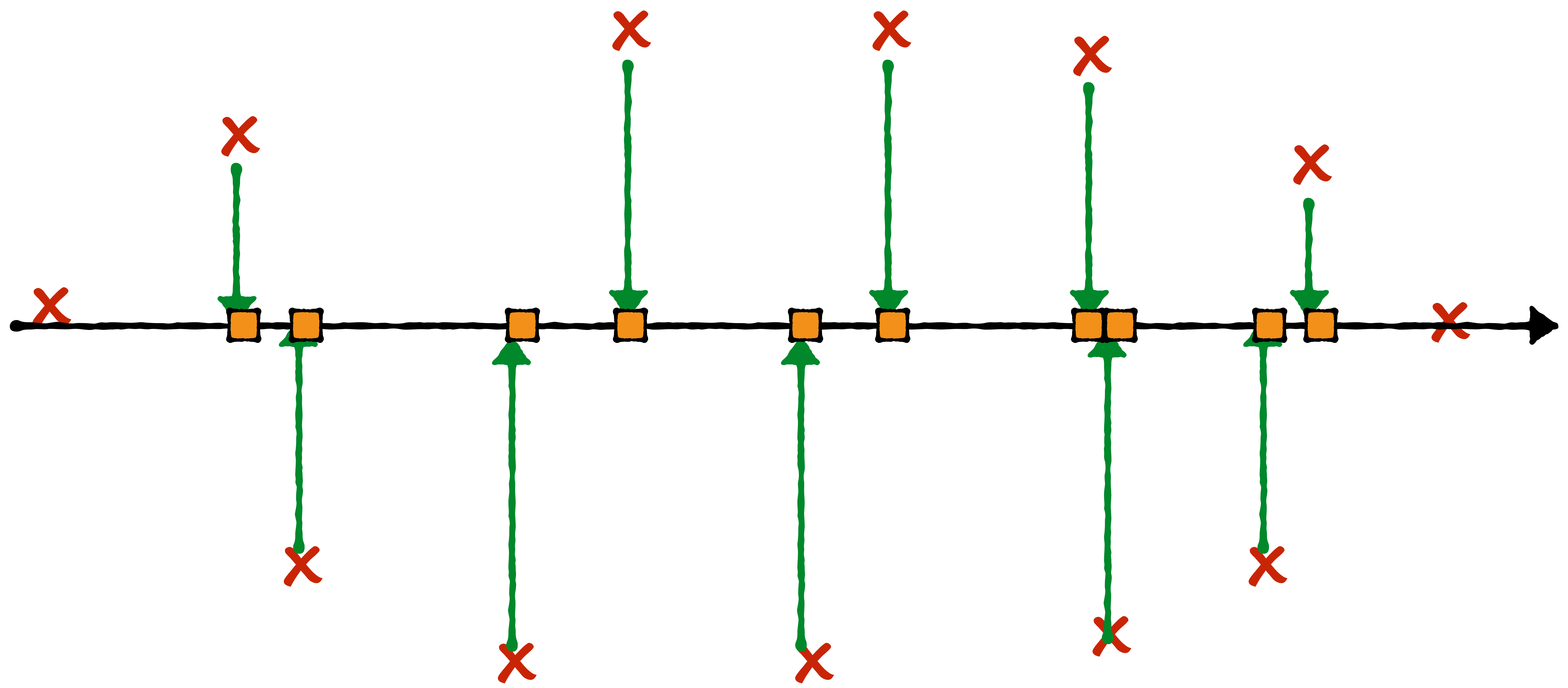


2 Dimensional data



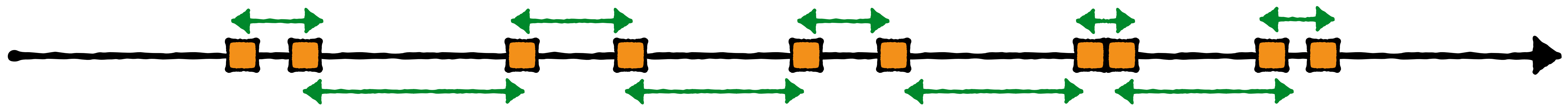


Find the “best” directions to represent this data



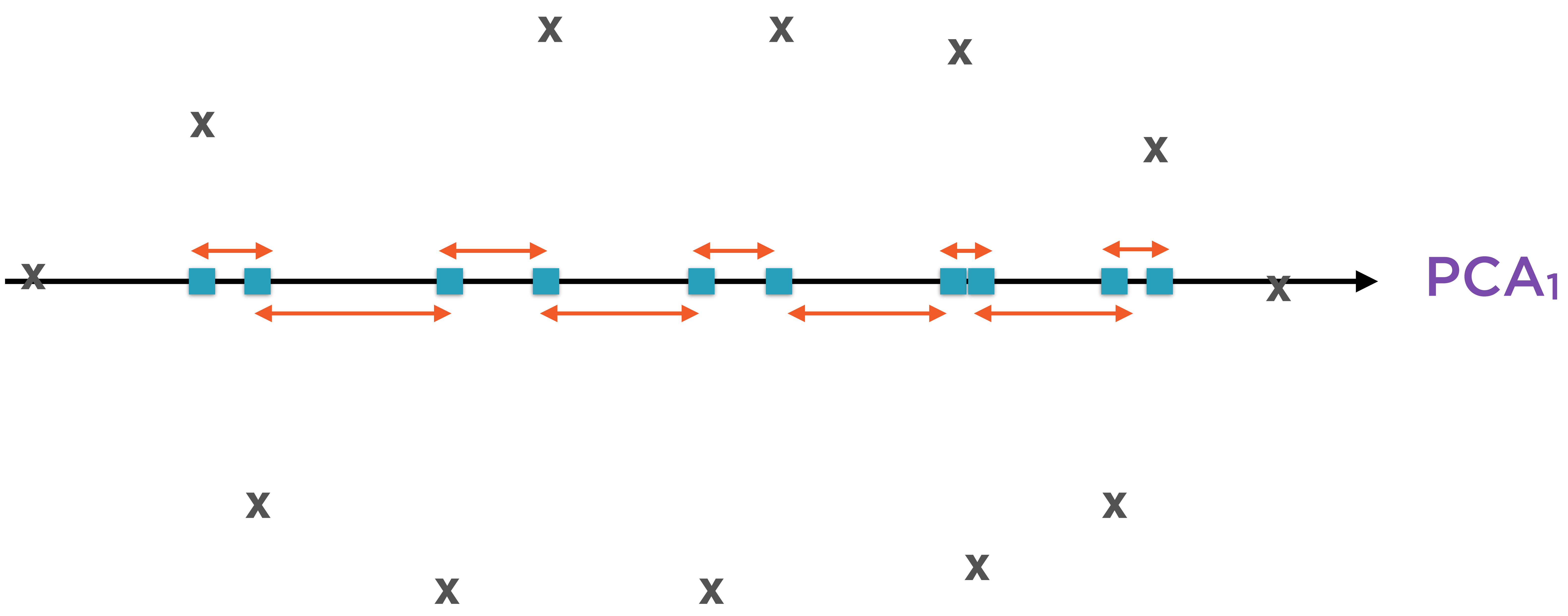
Project data onto a line

Distances between the projections carry information



Find the direction that has the largest distances between projections

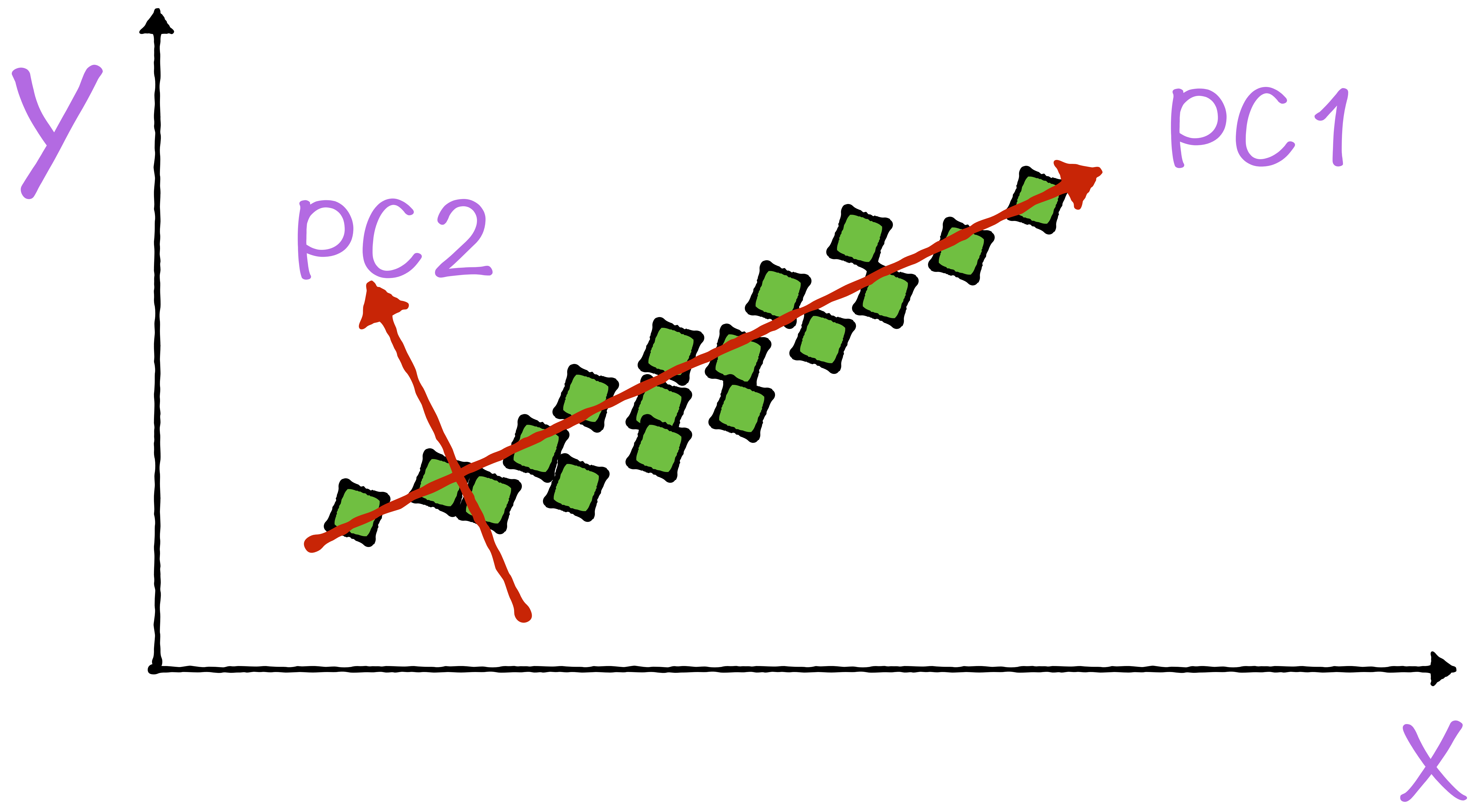
Intuition Behind PCA



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

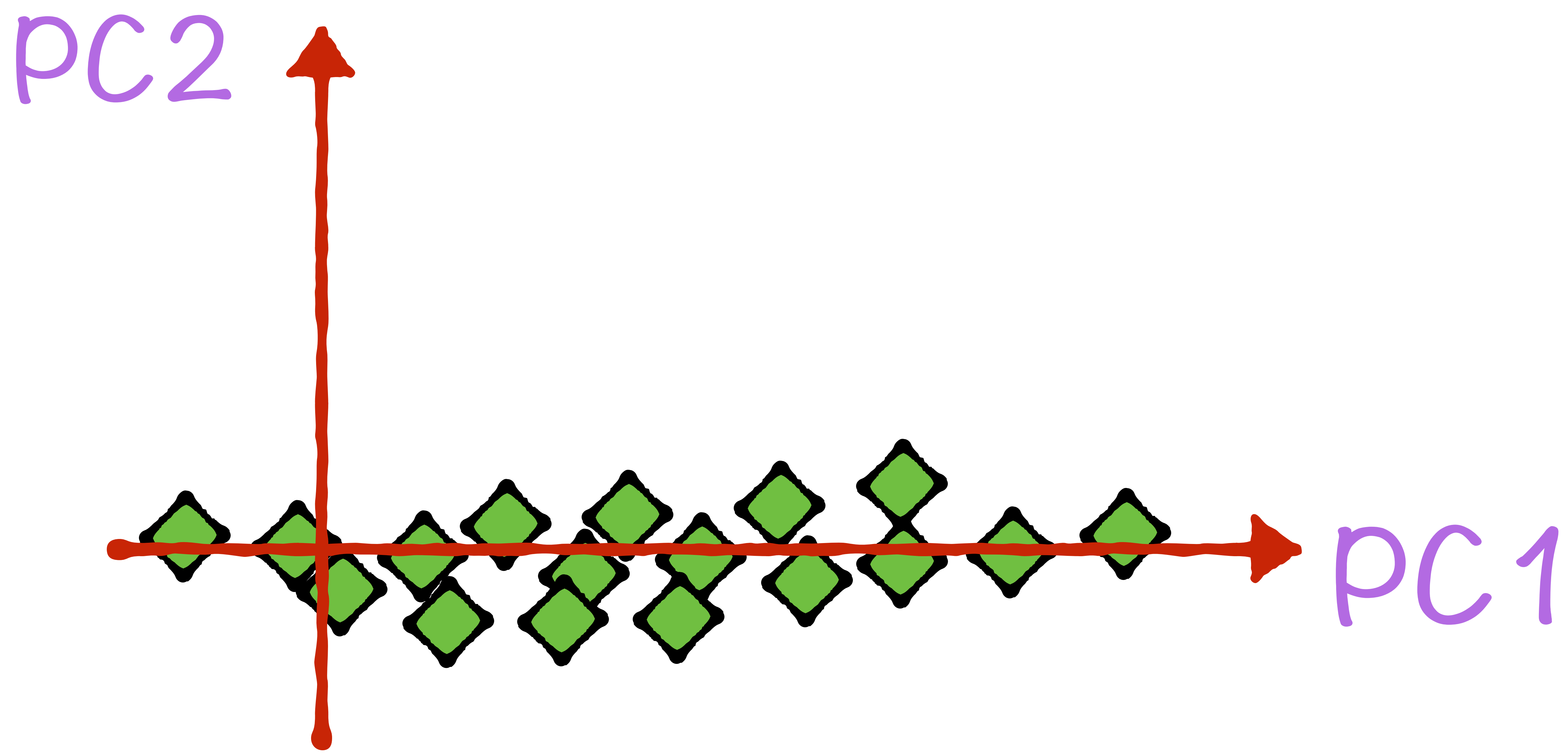
PCA

Dimensionality reduction



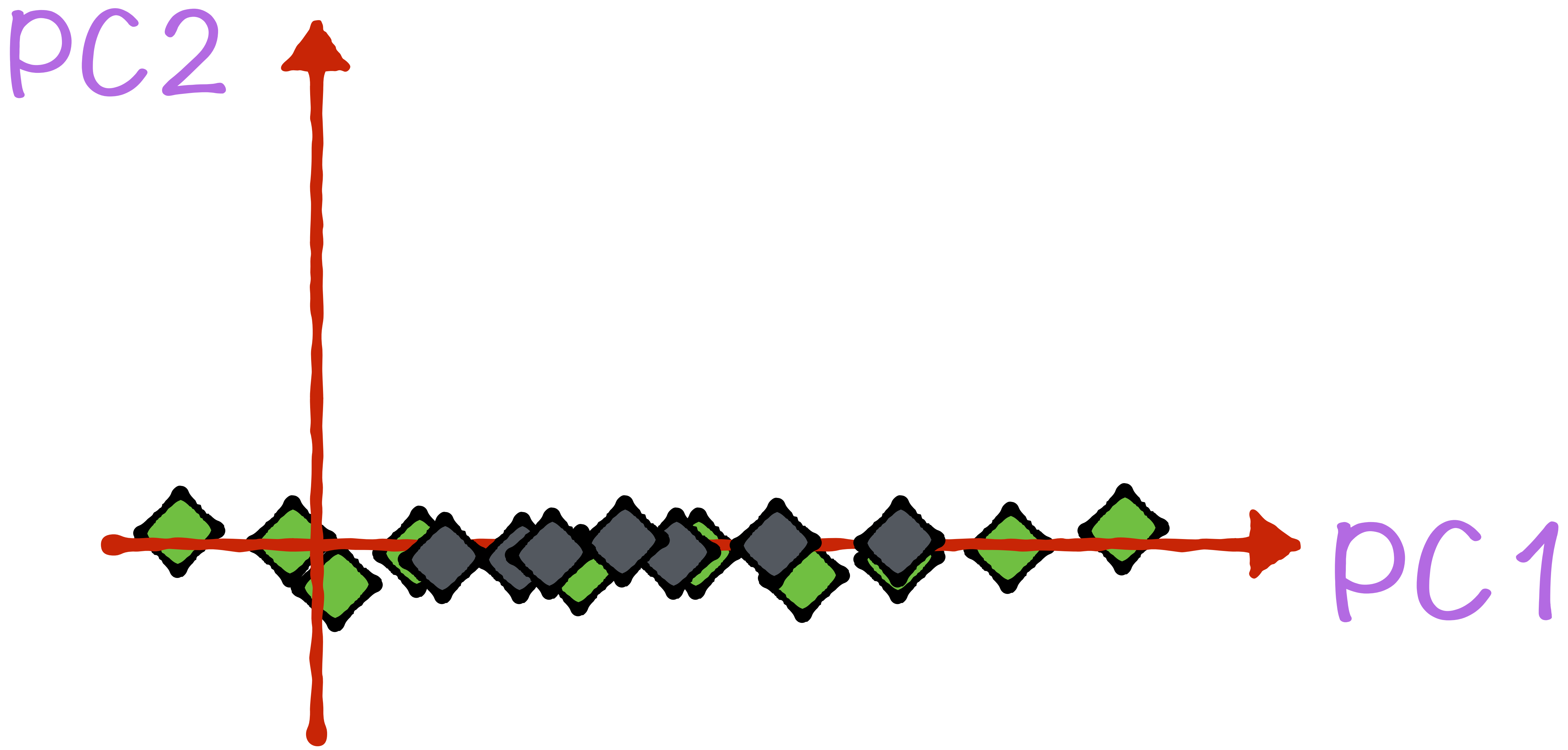
PCA

Dimensionality reduction



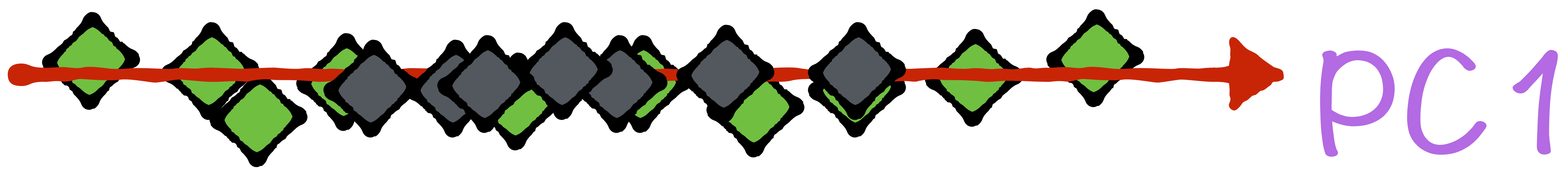
PCA

Dimensionality reduction



PCA

Dimensionality reduction



Random variables

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

Exxon Dow Jones Google Apple

Random variables

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

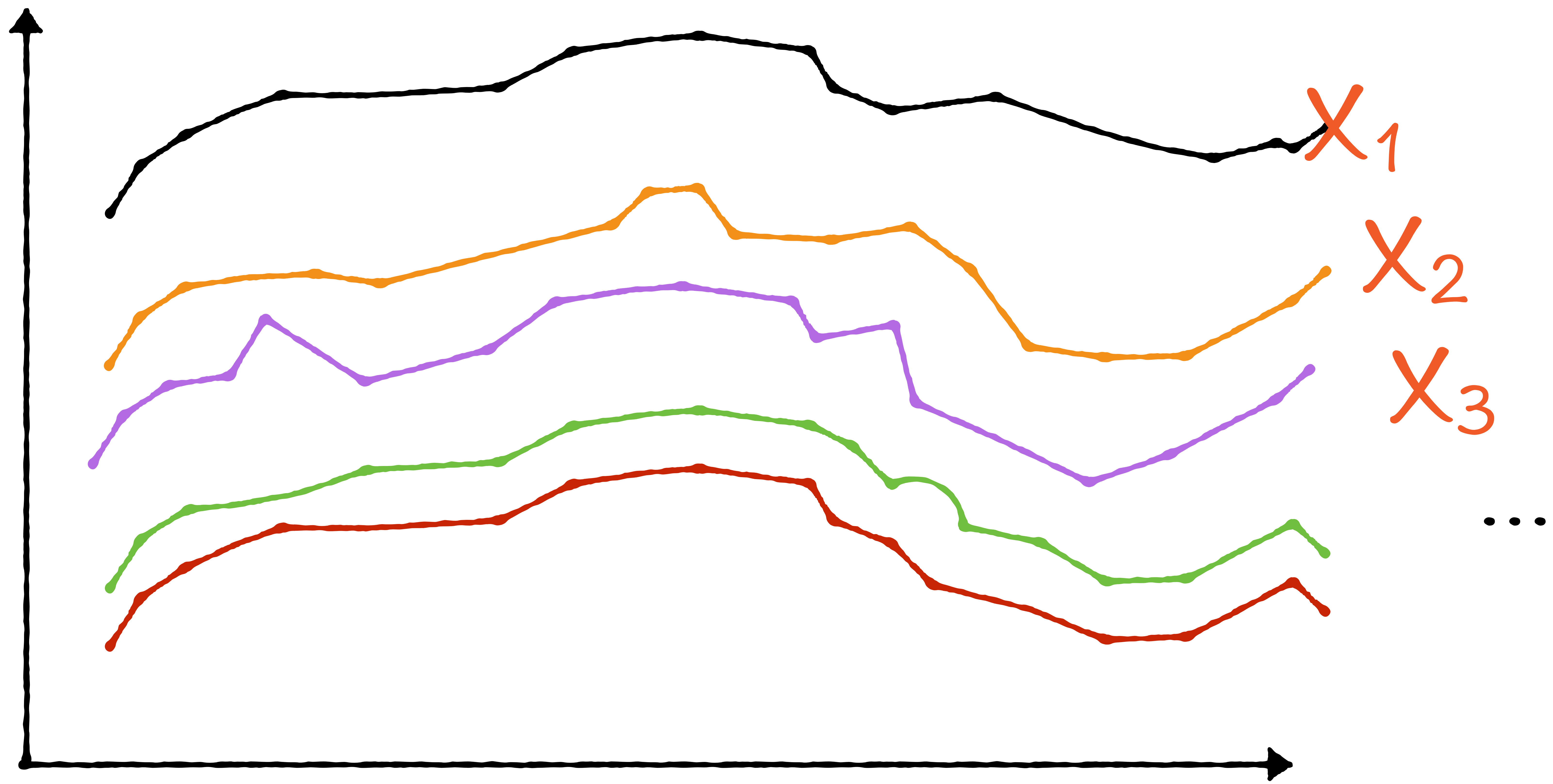
K columns

$$
\begin{array}{ccccc}
\begin{bmatrix}
x_{11} & x_{12} & x_{13} & & x_{1k} \\
x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\
x_{31} & x_{32} & x_{33} & & x_{3k} \\
\cdots & \cdots & \cdots & & \cdots \\
x_{n1} & x_{n2} & x_{n3} & & x_{nk}
\end{bmatrix} \\
\begin{matrix} X_1 & \quad X_2 & \quad X_3 & \qquad & \quad X_k \end{matrix}
\end{array}
$$

$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \end{bmatrix}$$

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

Problem -> Multicollinearity



Many of the X variables contain the same information

$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \end{bmatrix}$$

Use PCA when these random variables are highly correlated

Correlated variables

$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \end{bmatrix}$$

$$\Big\downarrow \text{PCA}$$

Uncorrelated variables

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

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

These are the principal components

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

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

Arranged in descending order of variance

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

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

$$=$$

$$var(X_1) + var(X_2) + var(x_3) + ..var(X_k)$$

$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \end{bmatrix}$$

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

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

Eigendecomposition

Solution: Finding Principal Component 1

Eigenvector

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

Principal Component

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

Eigen Value

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

Eigendecomposition

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

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

Eigendecomposition

Correlated variables

$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \end{bmatrix}$$

PCA

Uncorrelated variables

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

Results of PCA

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

Eigenvectors $\begin{bmatrix} v_1 & v_2 & v_3 & & v_k \end{bmatrix}$

Eigenvalues $\begin{bmatrix} e_1 & e_2 & e_3 & & e_k \end{bmatrix}$

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

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

Eigenvalue 1 Eigenvalue 2 Eigenvalue 3 Eigenvalue k

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

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

$$= \text{Total Variance } F$$

$$= \text{Total Variance } X$$

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

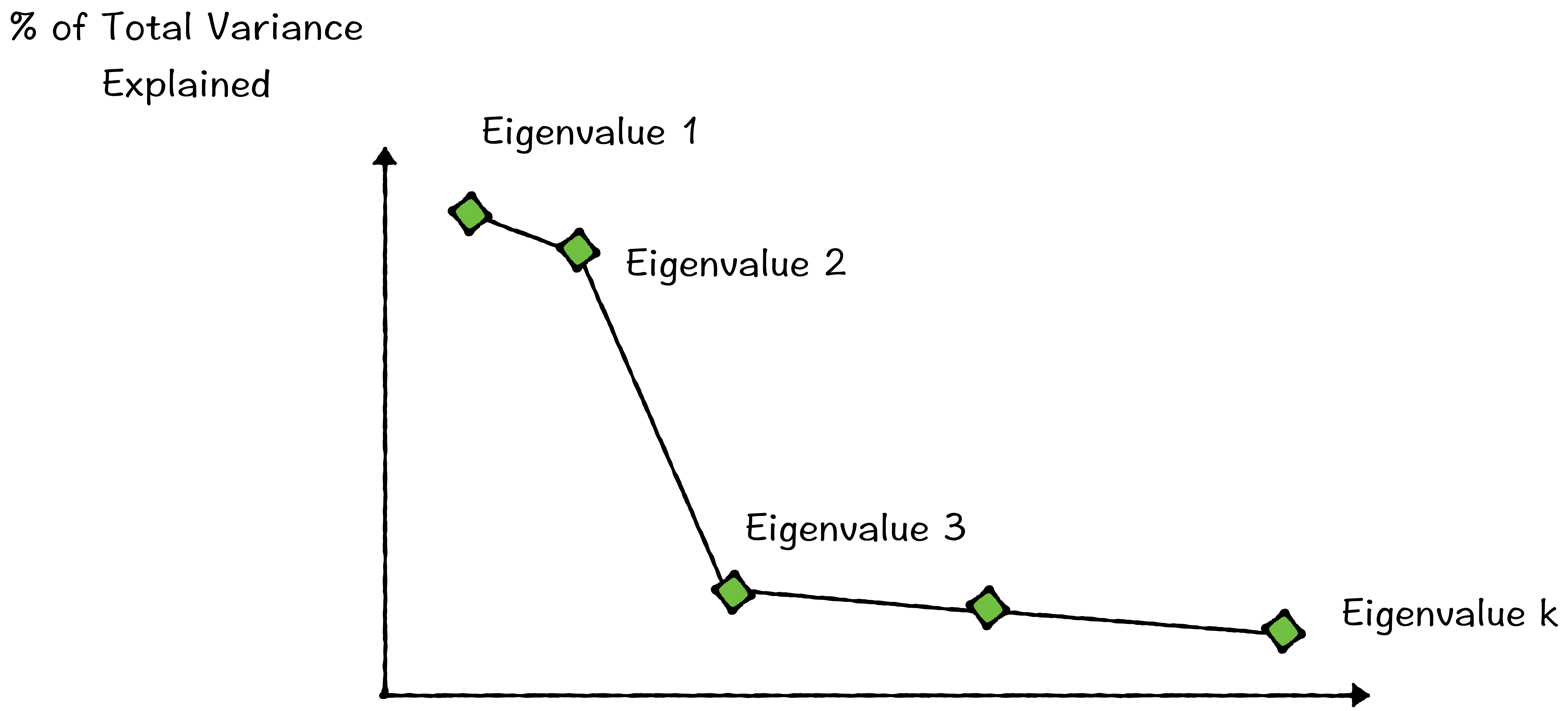
$$\frac{\text{Eigenvalue 1}}{\text{Variance(F)}}$$

$$\frac{\text{Eigenvalue 2}}{\text{Variance(F)}}$$

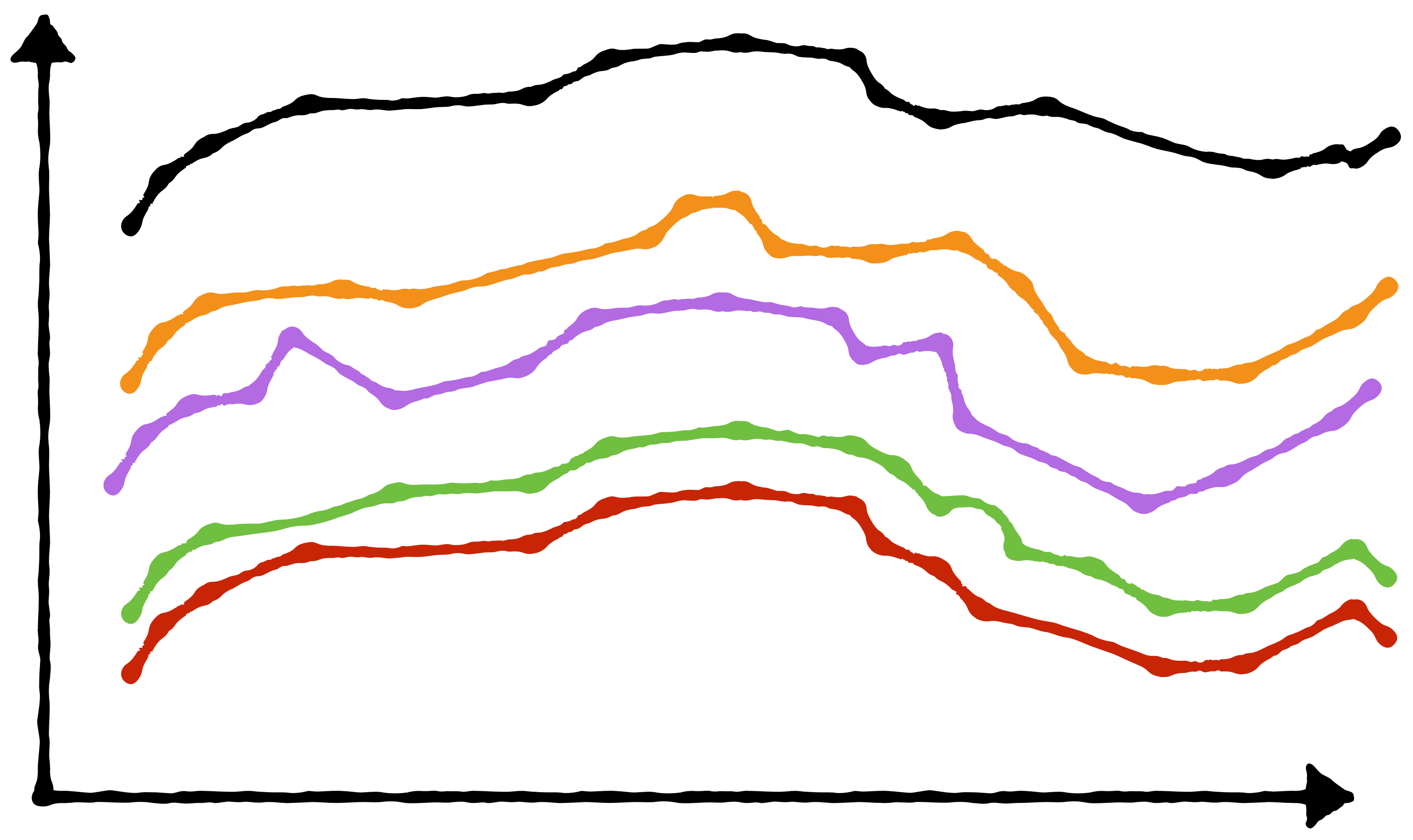
$$\frac{\text{Eigenvalue 3}}{\text{Variance(F)}}$$

$$\frac{\text{Eigenvalue 4}}{\text{Variance(F)}}$$

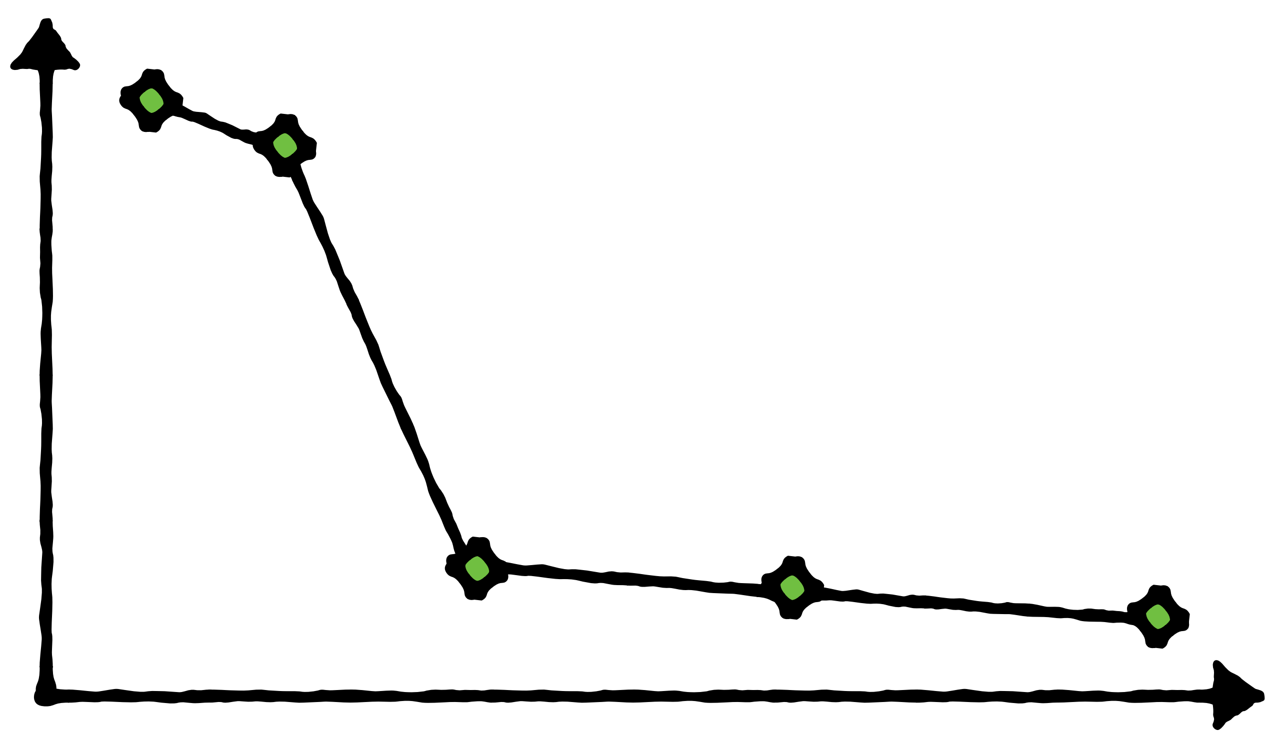
Sum= 100%


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

PCA is great when



Many, Highly Correlated
Xi



Unequal Eigenvalues

Correlation matrix

$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \end{bmatrix}$$

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

Correlation matrix

$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \end{bmatrix}$$

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

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

$$\begin{array}{c} \begin{bmatrix} x_{11} & x_{12} & x_{13} & & x_{1k} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ x_{31} & x_{32} & x_{33} & & x_{3k} \\ \cdots & \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & x_{n3} & & x_{nk} \end{bmatrix} \\ \begin{matrix} X_1 & \quad X_2 & \quad X_3 & \qquad & \quad X_k \end{matrix} \end{array}$$

$$F = X v$$

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

Matrix Multiplication

$$F = X\,v$$

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

k columns   k columns

$v_1 \quad v_2 \ldots v_k$

Matrix Multiplication

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

Matrix Multiplication

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

Matrix Multiplication

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

Matrix Multiplication

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

Matrix Multiplication

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