Principal Component Analysis (PCA)

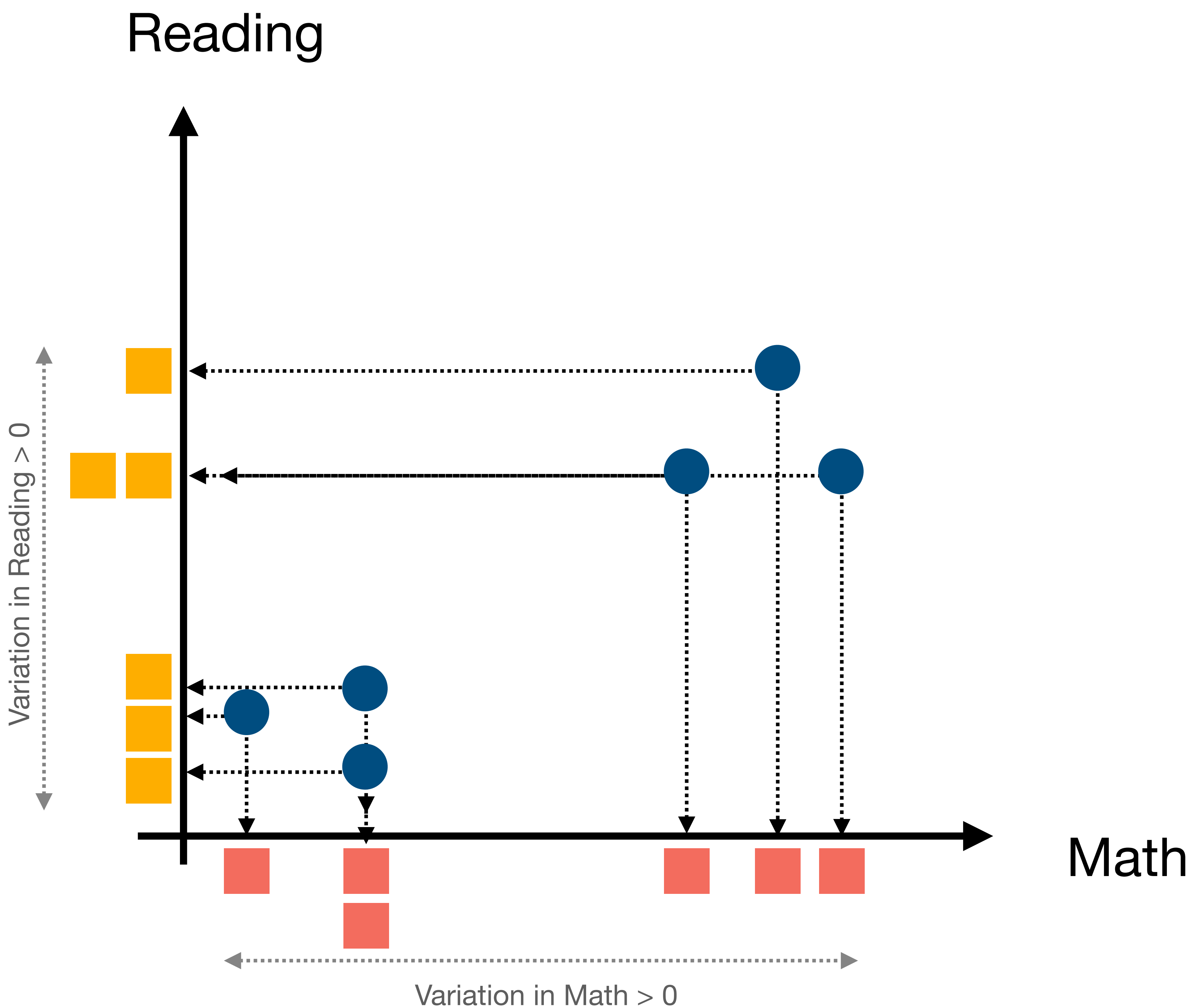
- Dimension Reduction
- Representation of Data

อ.ดร.สิวะโชติ ศรีสุทธิยากร และ อ.ดร.ประภาศิริ รัชชประภาพรกุล
ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย

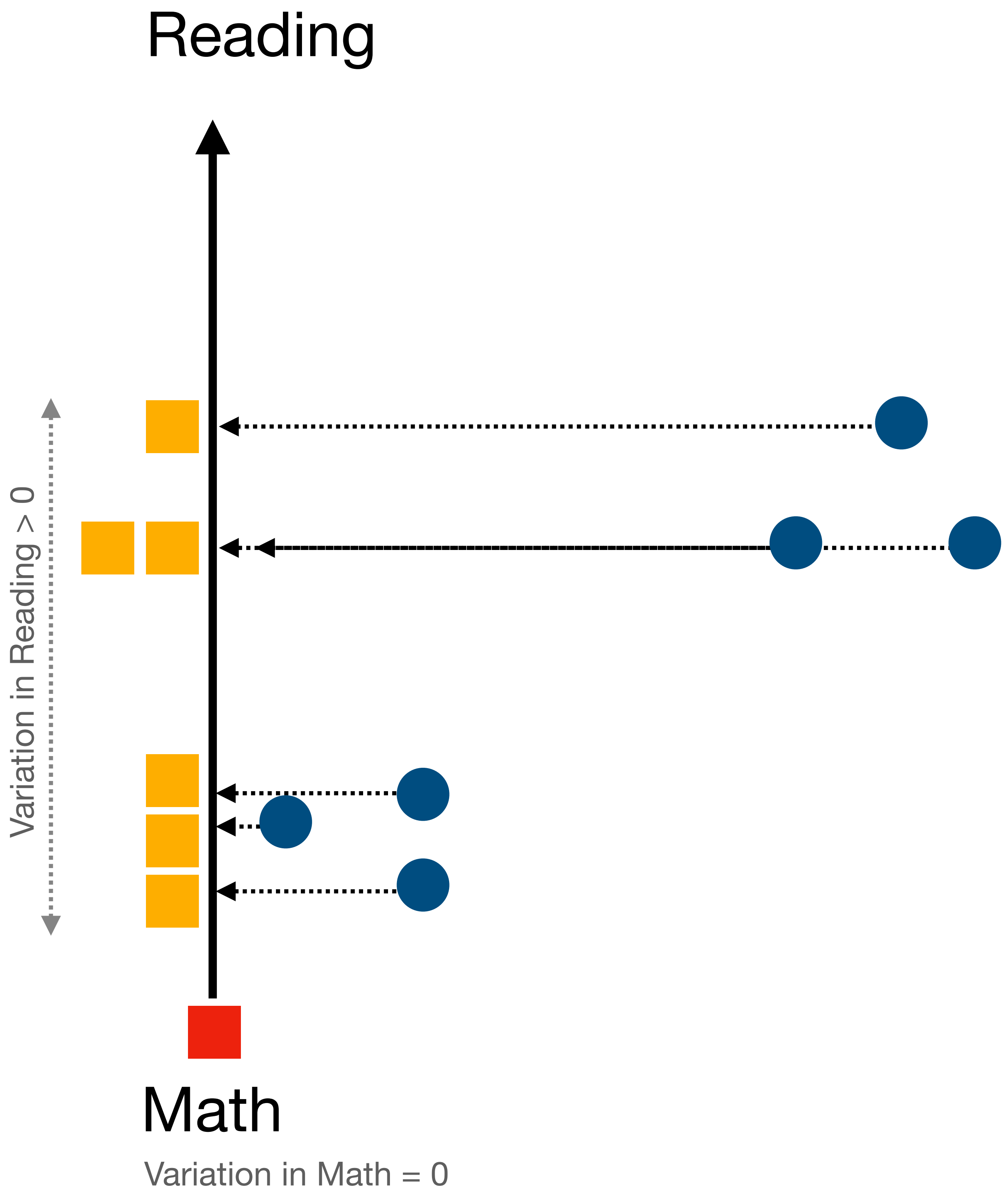
PCA - concept

Student	Reading	Math
บุญมี	8	99
บุญมา	5	43
บุญหนัก	1	16
บุญทับ	4	49
บุญถึก	7	83
บุญถึง	3	55

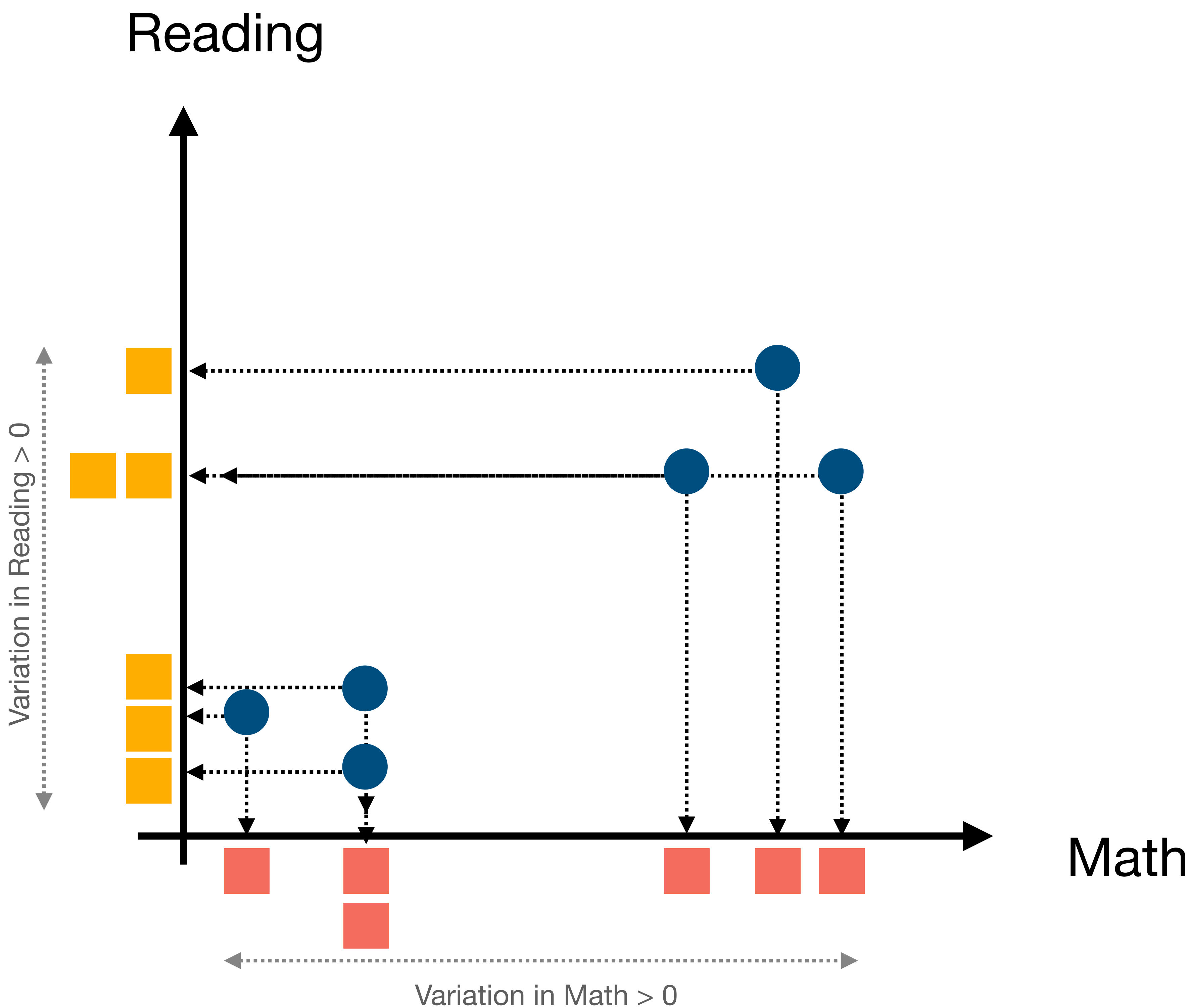
PCA - concept



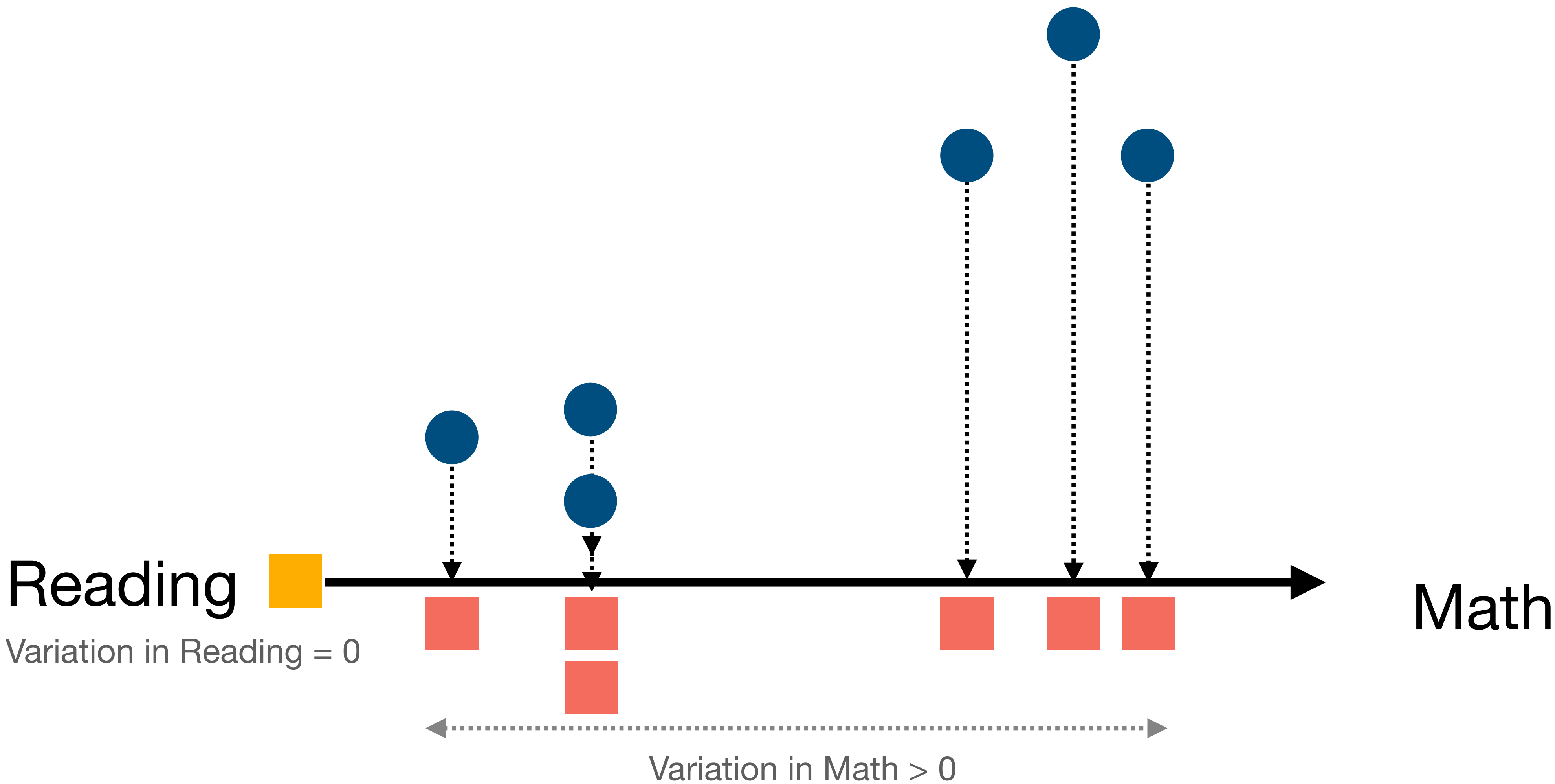
PCA - concept



PCA - concept

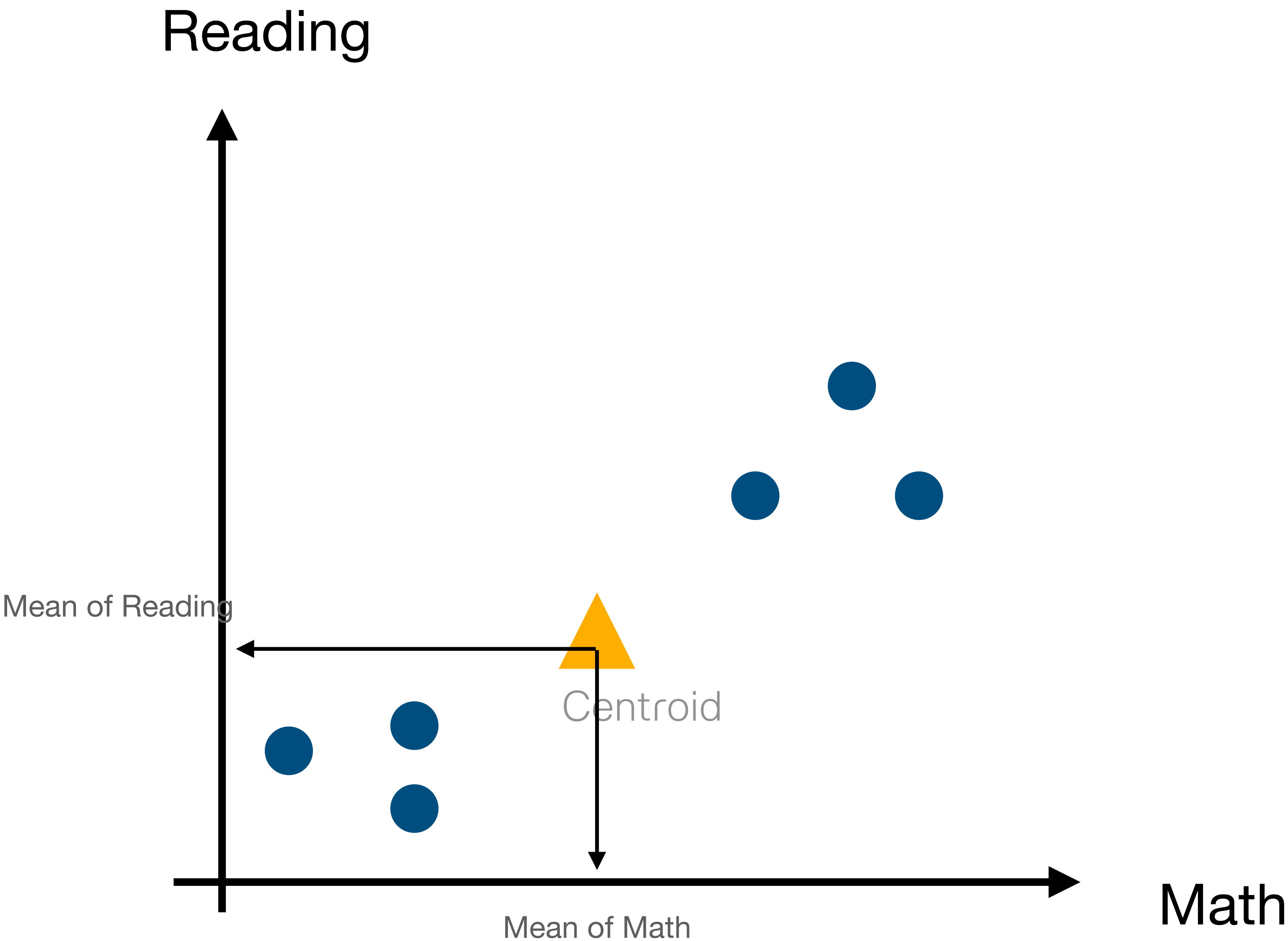


PCA - concept



PCA - concept

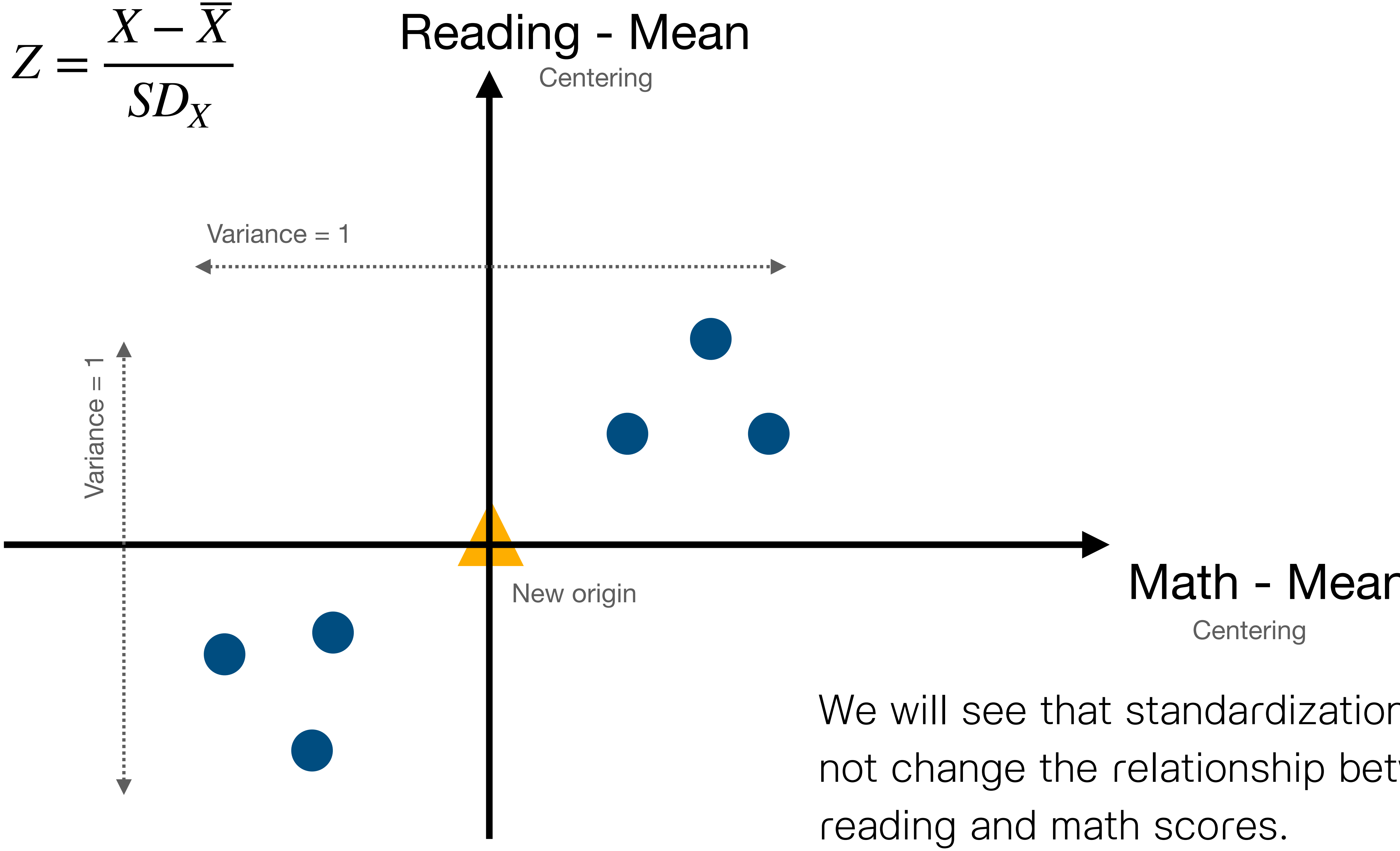
Calculate centroids and standard deviations of the data



PCA - concept

Standardized the variables

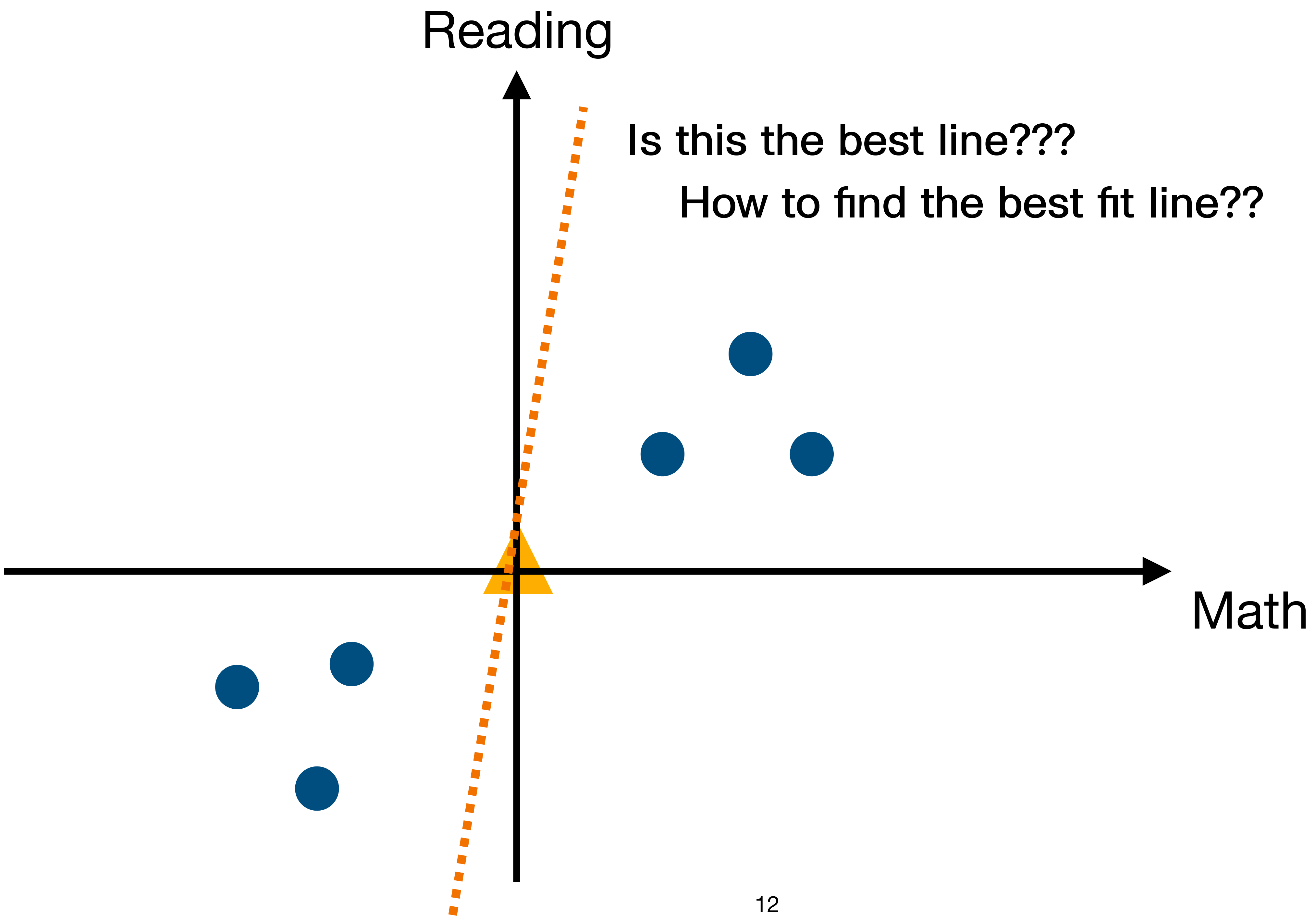
$$Z = \frac{X - \overline{X}}{SD_X}$$



We will see that standardization did not change the relationship between reading and math scores.

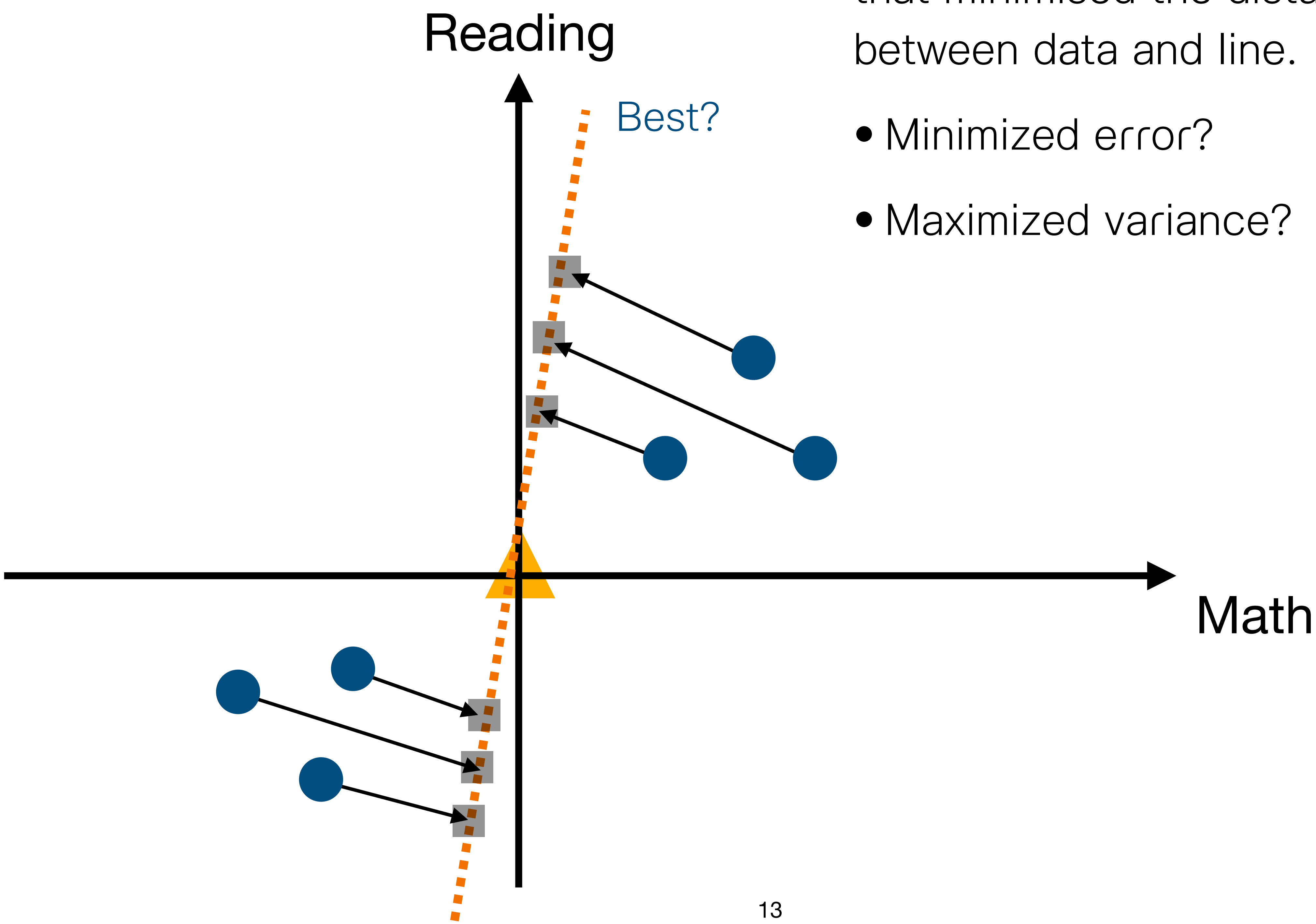
PCA - concept

Fit a line to the data…



PCA - concept

Fit a line to the data…

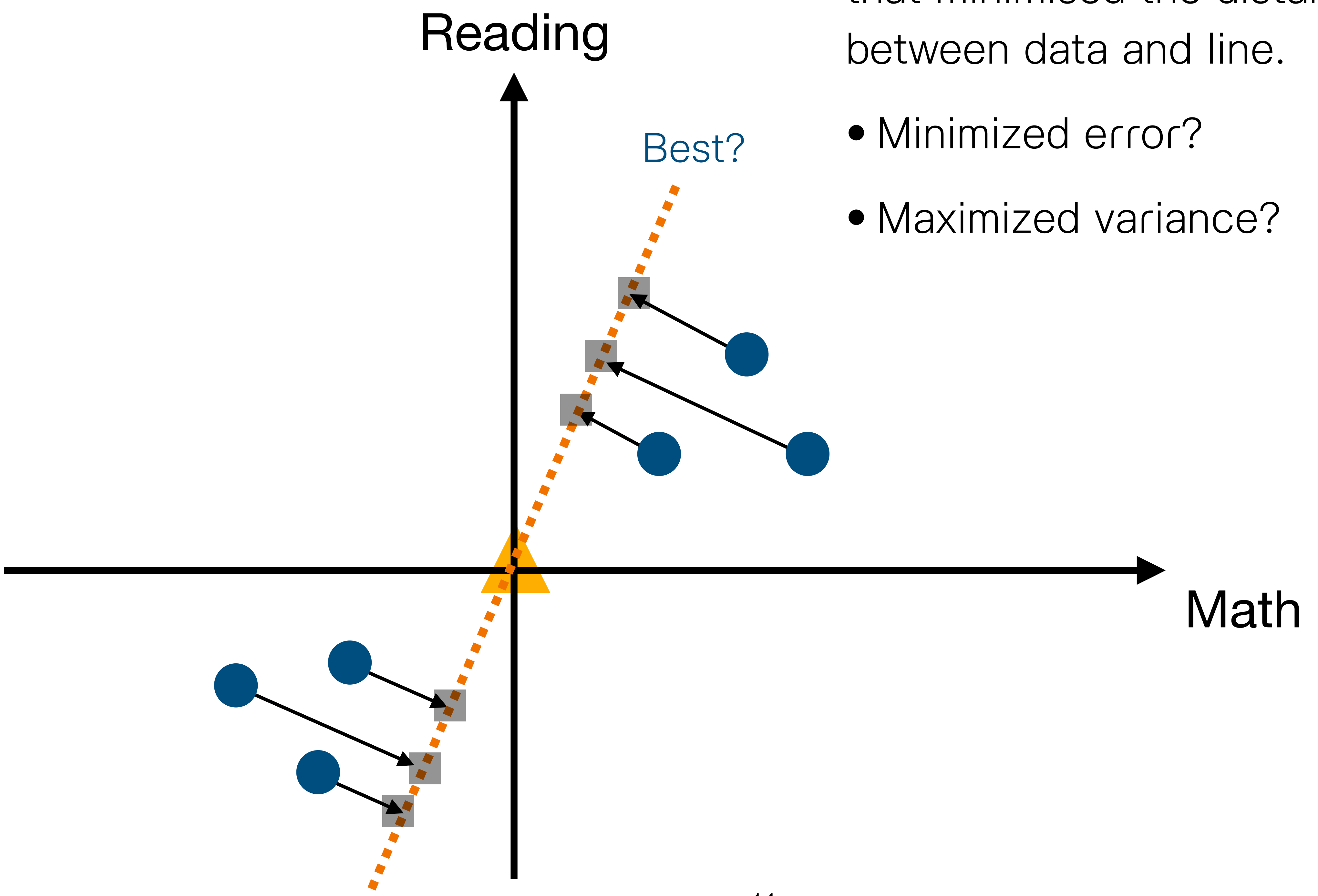


The best fit line is the line that minimised the distances between data and line.

- Minimized error?
- Maximized variance?

PCA - concept

Fit a line to the data...



The best fit line is the line that minimised the distances between data and line.

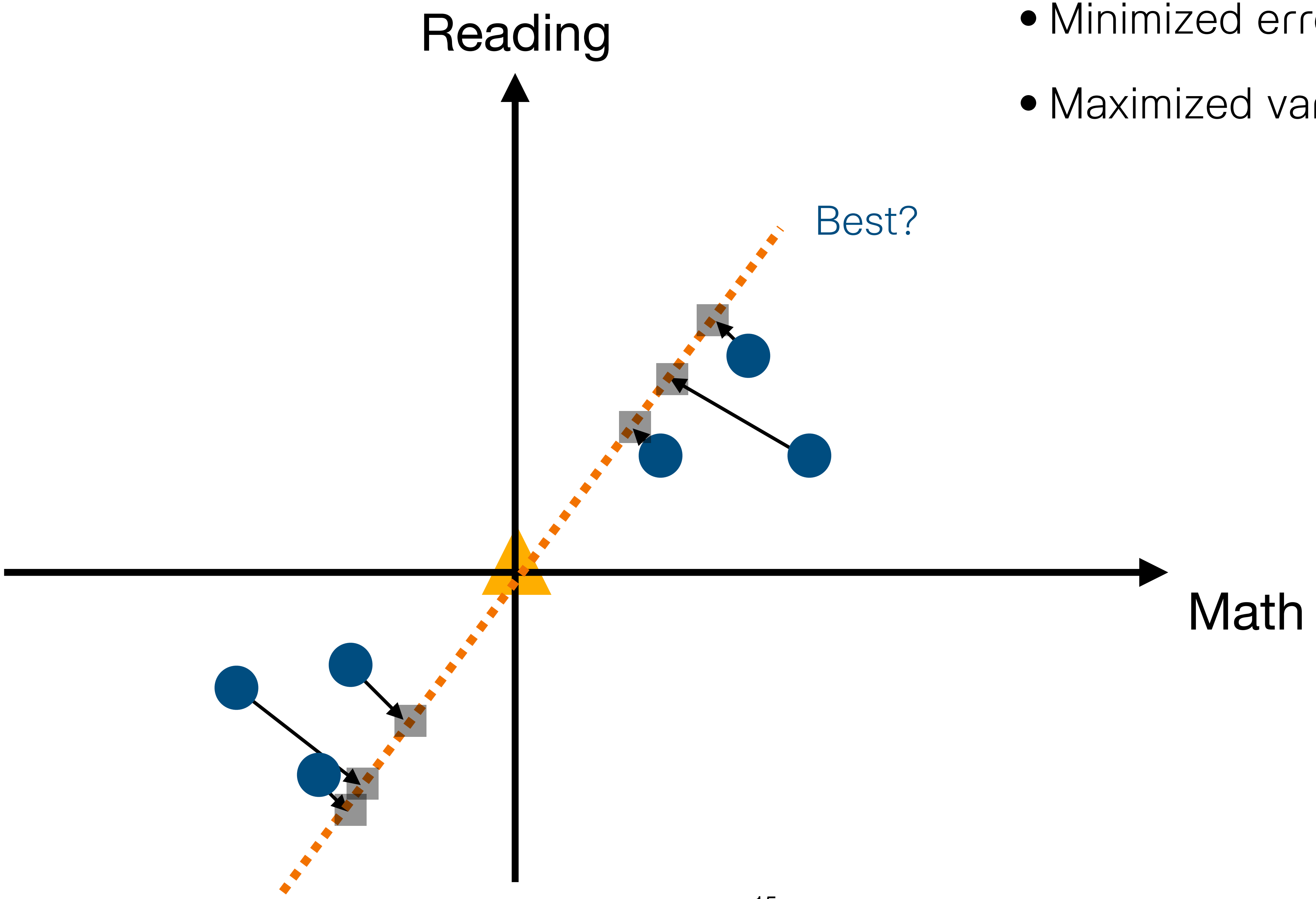
- Minimized error?
- Maximized variance?

PCA - concept

Fit a line to the data…

The best fit line is the line that minimised the distances between data and line.

- Minimized error?
- Maximized variance?

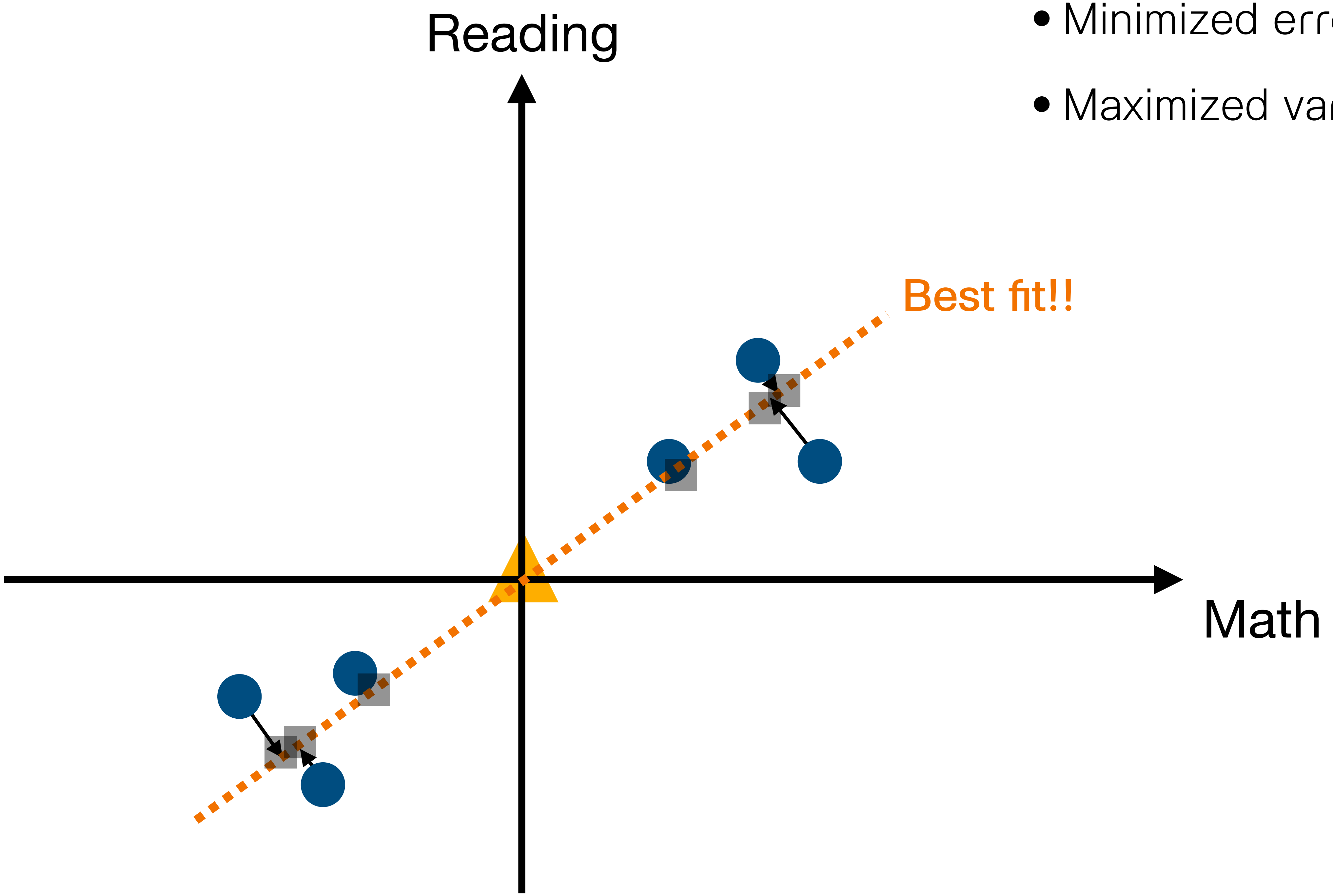


PCA - concept

Fit a line to the data...

The best fit line is the line that minimised the distances between data and line.

- Minimized error?
- Maximized variance?



PCA - concept

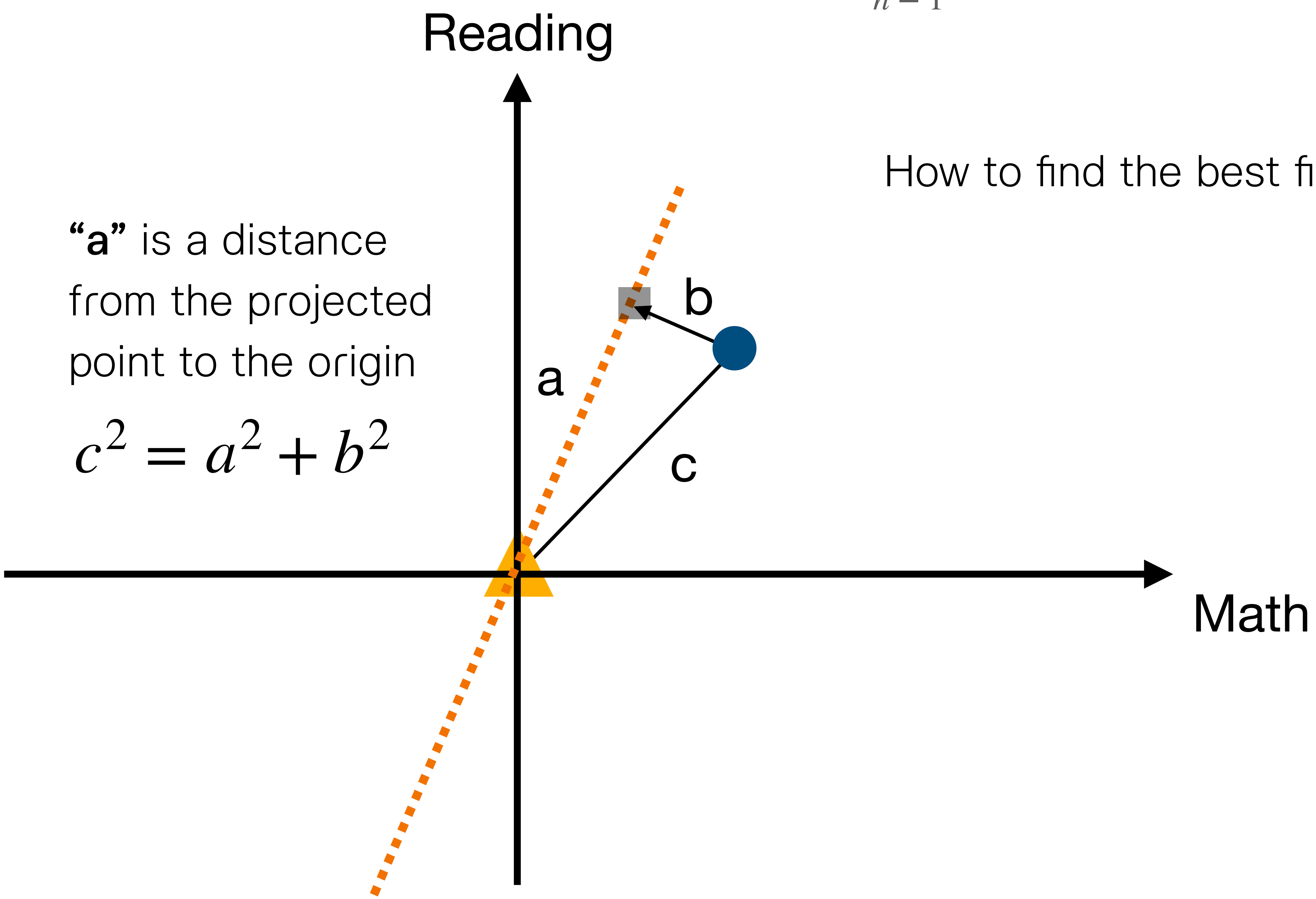
Fit a line to the data…

$$S^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

How to find the best fit line??

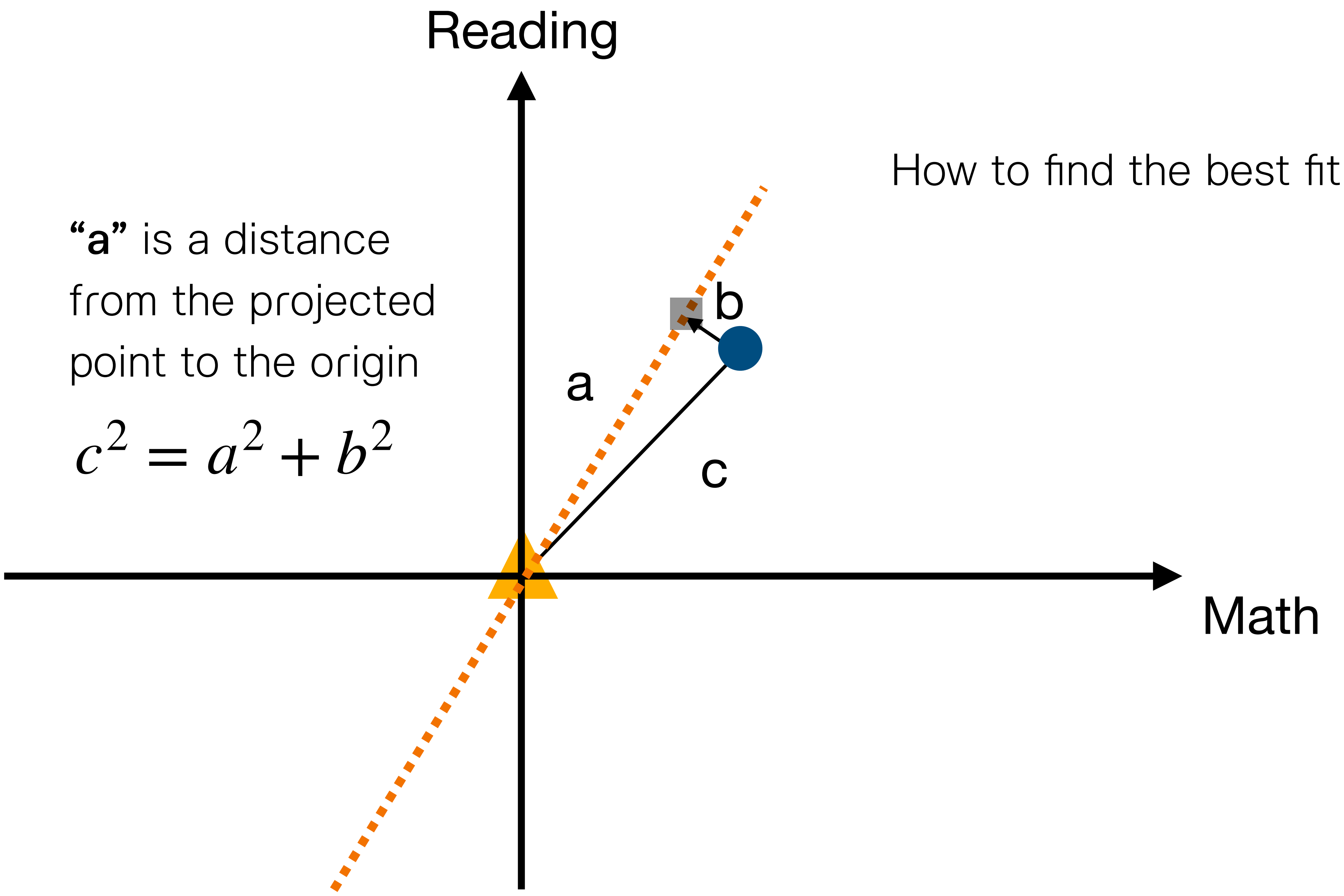
“a” is a distance from the projected point to the origin

$$c^2 = a^2 + b^2$$



PCA - concept

Fit a line to the data…



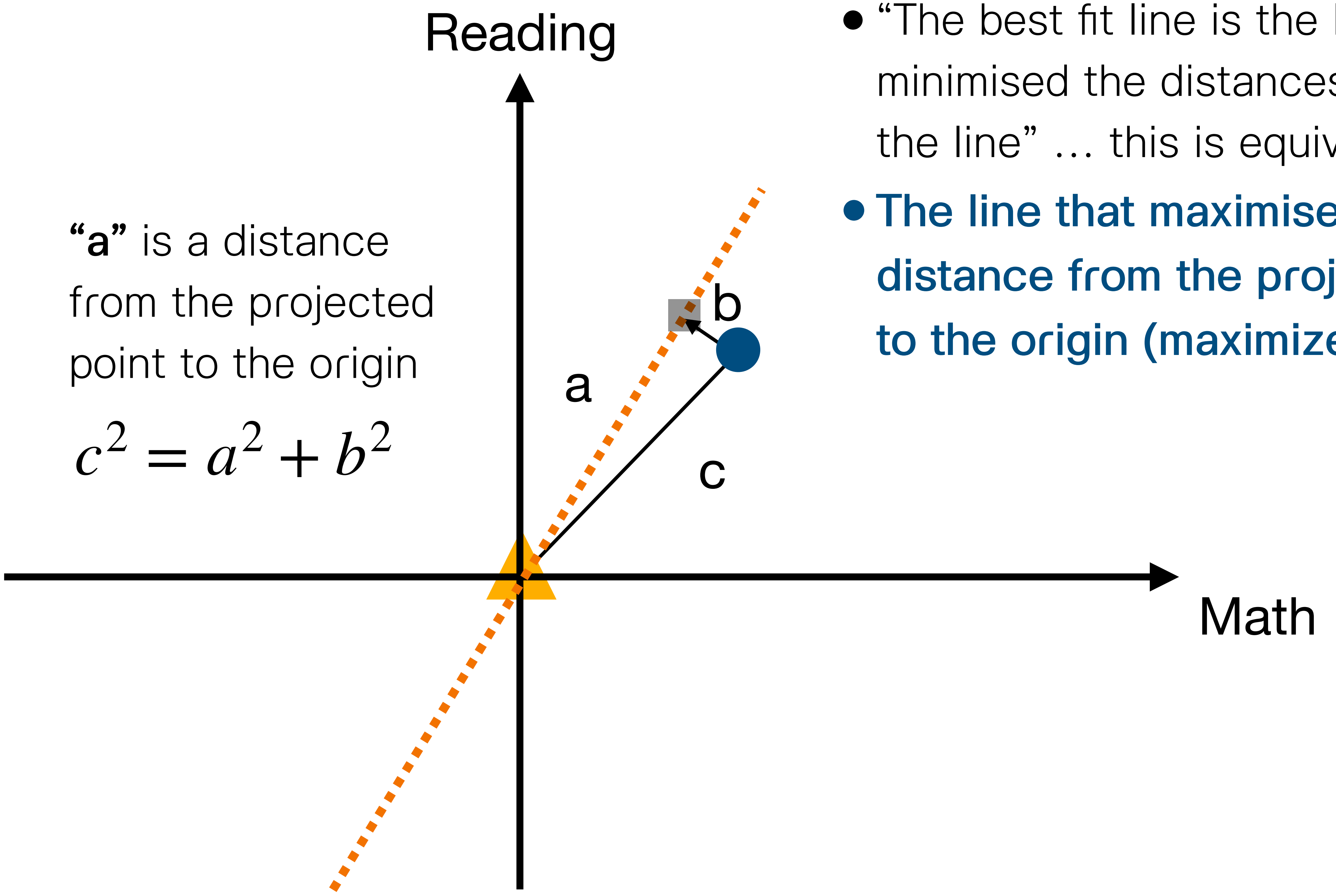
How to find the best fit line??

“a” is a distance from the projected point to the origin

$$c^2 = a^2 + b^2$$

PCA - concept

Fit a line to the data…



"a" is a distance from the projected point to the origin

$$c^2 = a^2 + b^2$$

How to find the best fit line??

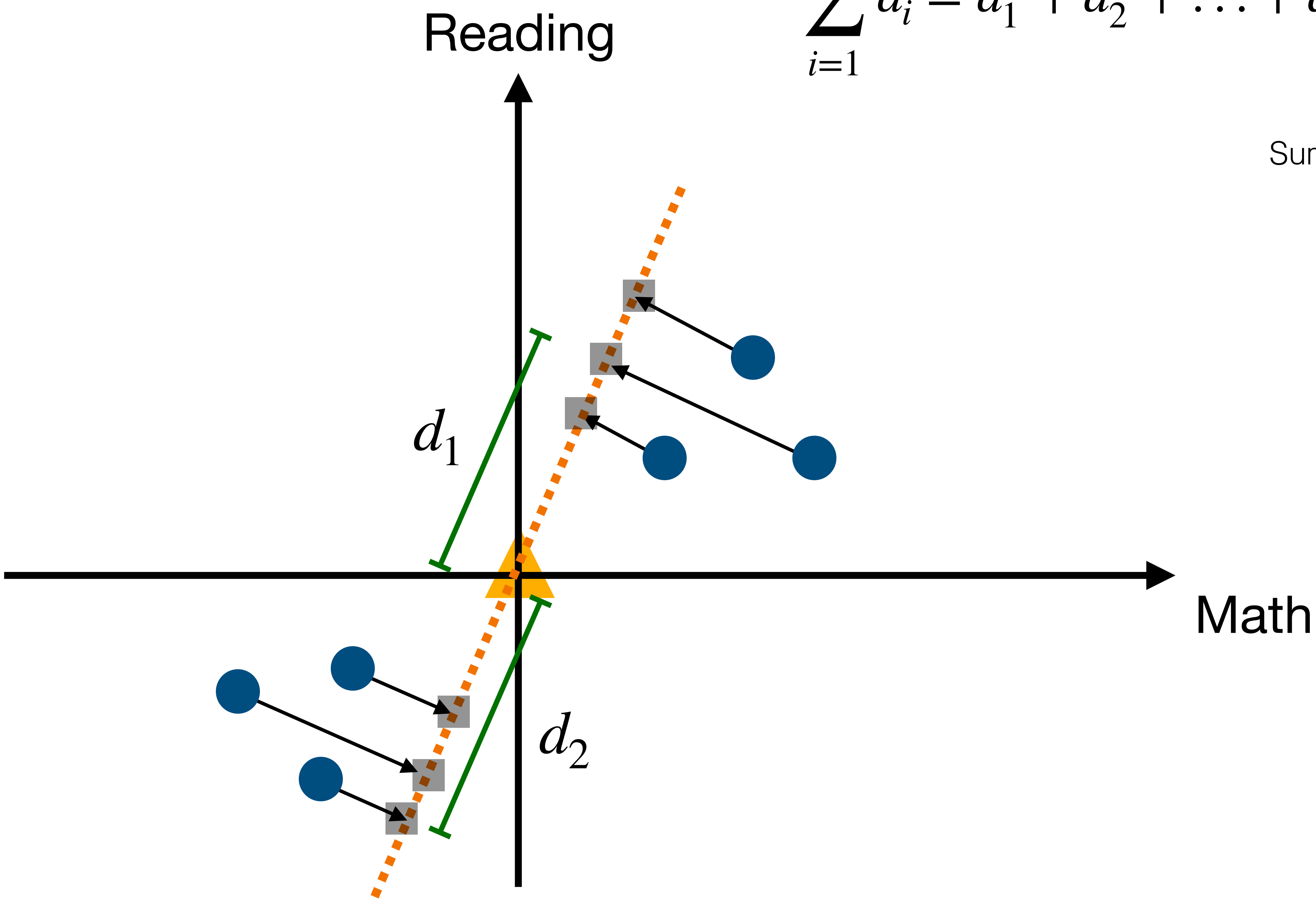
- "The best fit line is the line that minimised the distances from data to the line" … this is equivalent to….
- **The line that maximised the distance from the projected points to the origin (maximizes "a")**

PCA - concept

Let d_i be a distance from the projected point i to the origin.

$$\sum_{i=1}^{6} d_i = d_1^2 + d_2^2 + \ldots + d_6^2 = \mathrm{SS}_1$$

Sum squares distance



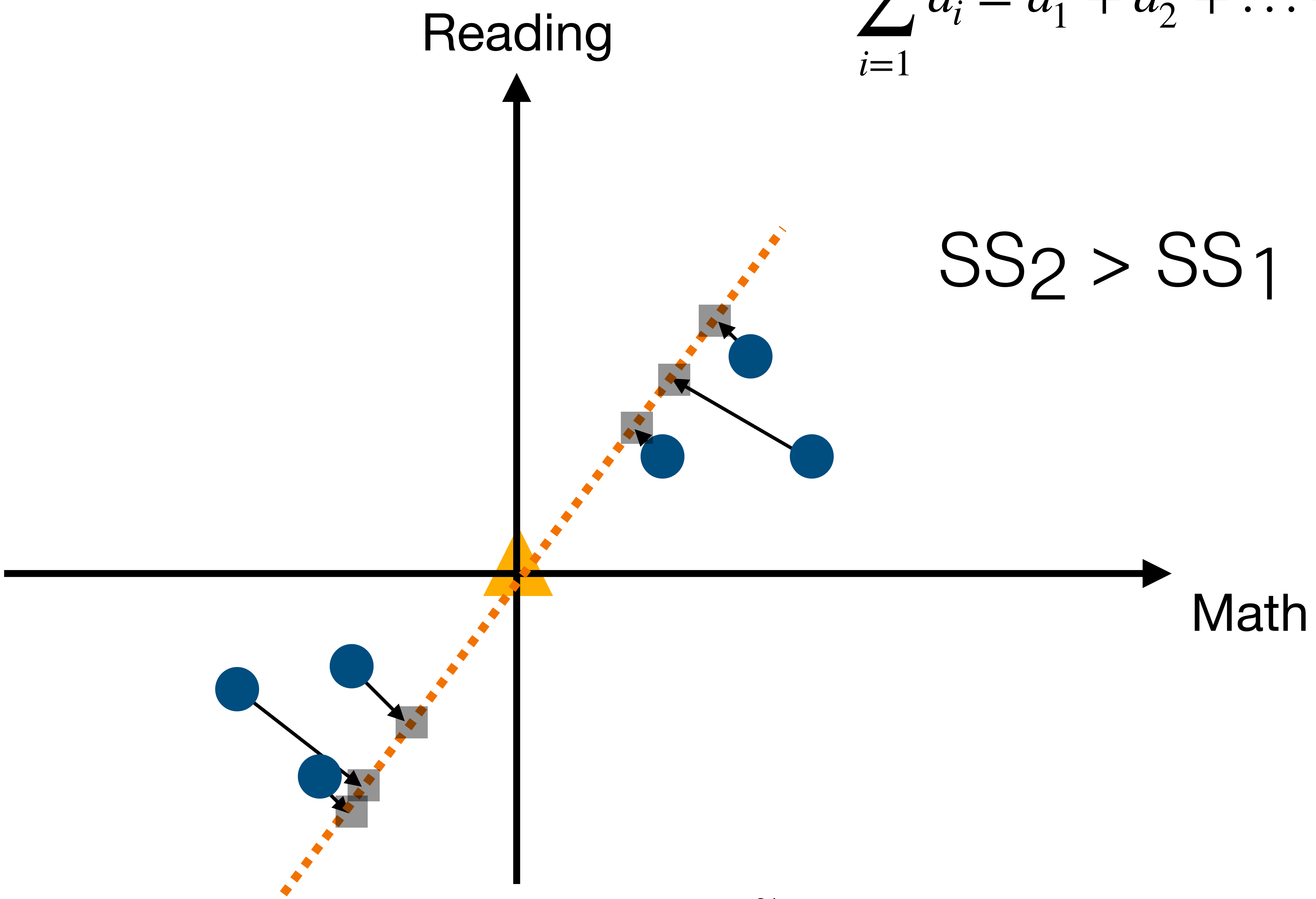
PCA - concept

Fit a line to the data...

Let d_i be a distance from the projected point i to the origin.

$$\sum_{i=1}^{6} d_i = d_1^2 + d_2^2 + \ldots + d_6^2 = \mathrm{SS}_2$$

$\mathrm{SS}_2 > \mathrm{SS}_1$

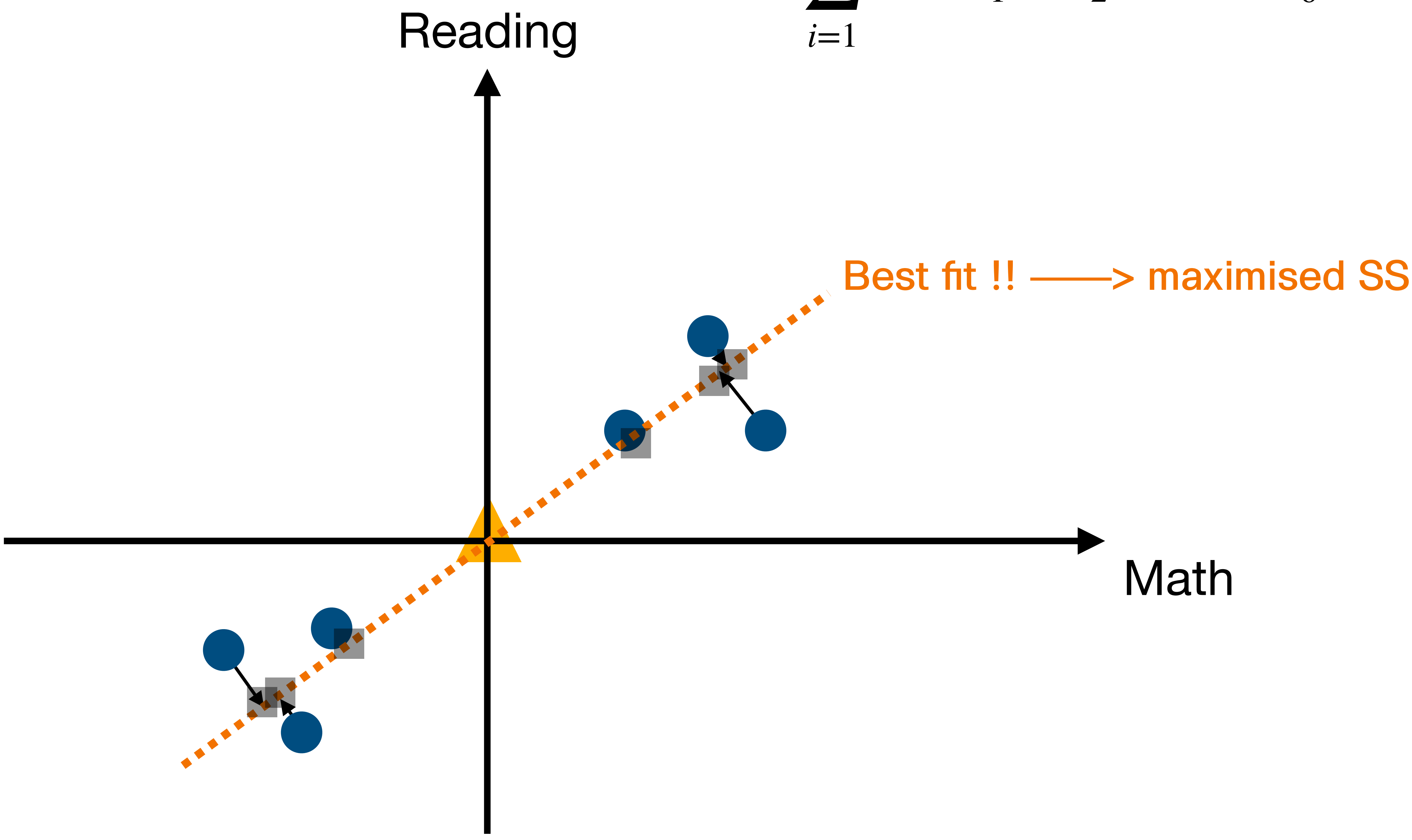


PCA - concept

Let d_i be a distance from the projected point i to the origin.

$$\sum_{i=1}^{6} d_i = d_1^2 + d_2^2 + \ldots + d_6^2 = \text{SS}$$

Fit a line to the data…

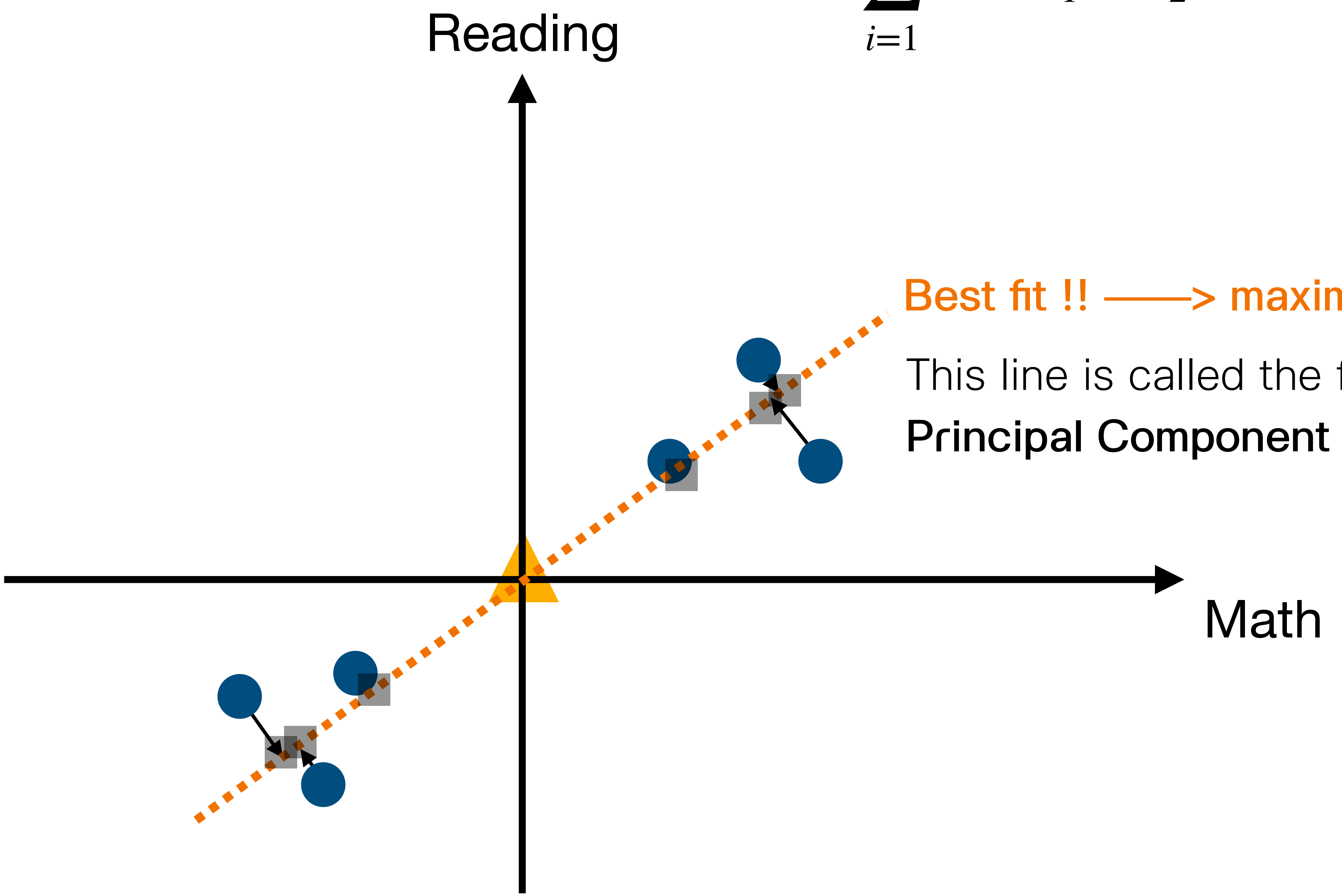


PCA - concept

Fit a line to the data...

Let d_i be a distance from the projected point i to the origin.

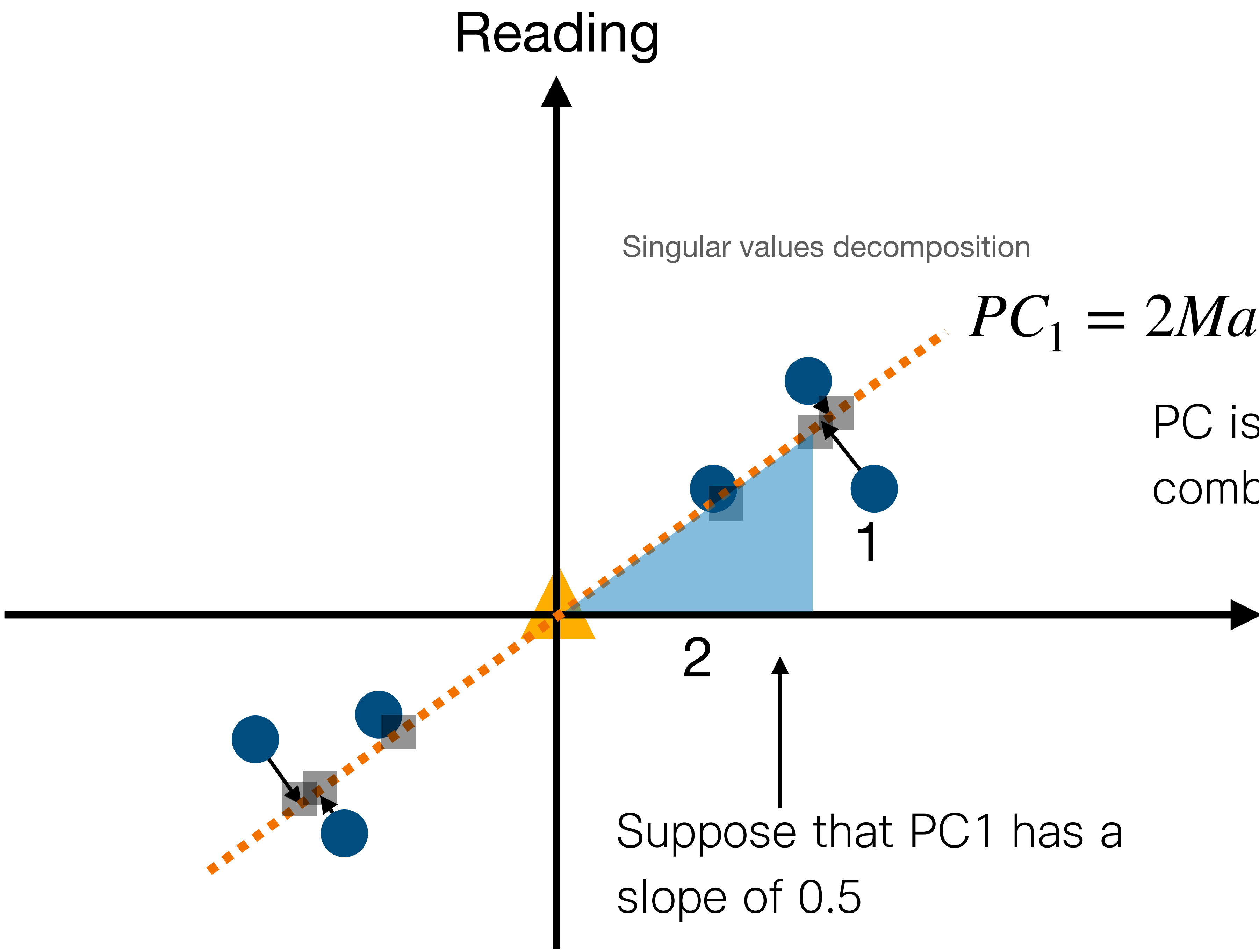
$$\sum_{i=1}^{6} d_i = d_1^2 + d_2^2 + \ldots + d_6^2 = \text{SS}$$



Best fit !! ——> maximised SS

This line is called the first **Principal Component (PC1).**

PCA - concept



Singular values decomposition

$$PC_1 = 2Math + Reading$$

PC is a linear combination of variables.

Suppose that PC1 has a slope of 0.5

Figure not to scale.

PCA - concept

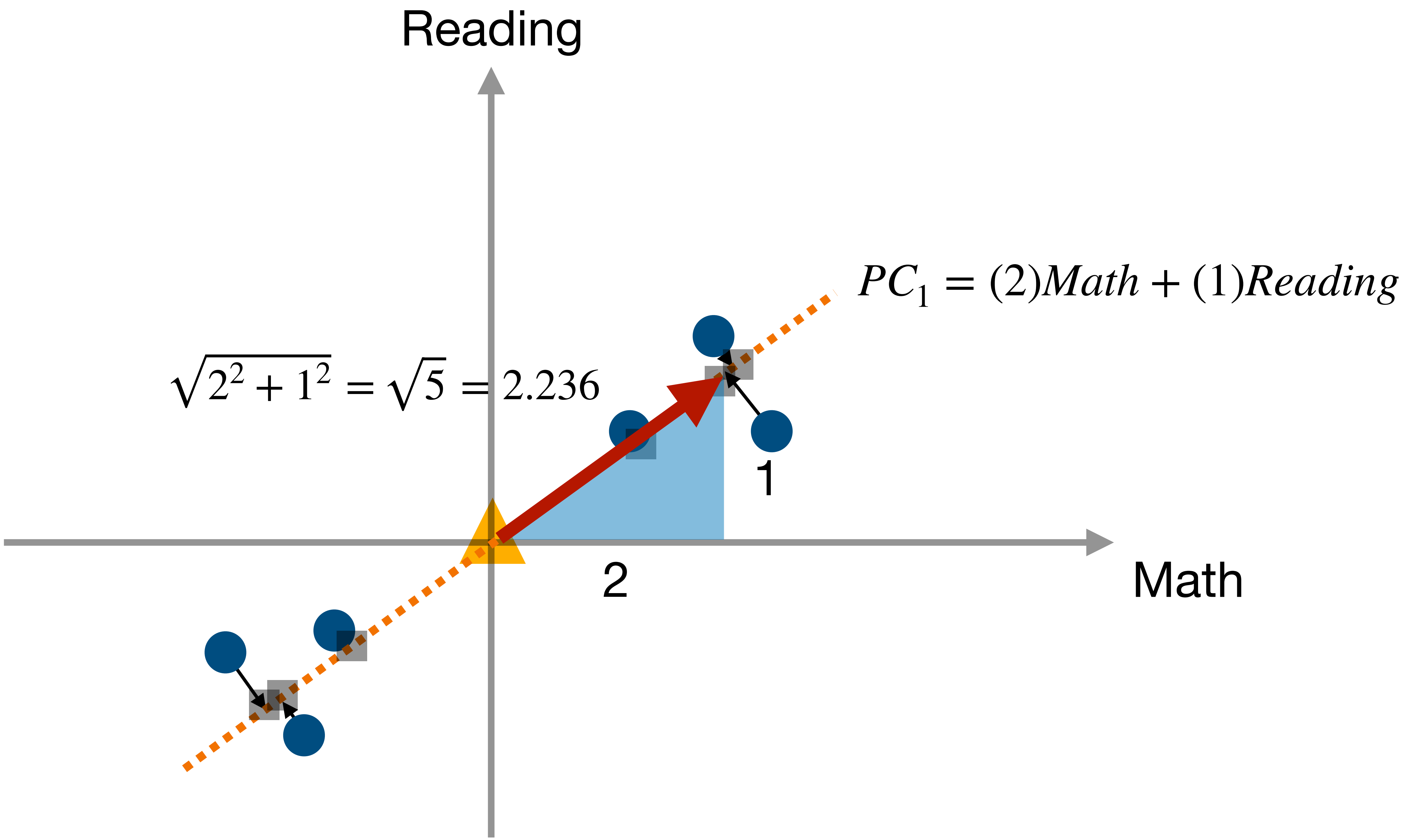


Figure not to scale.

PCA - concept

In PCA, we scaled the (red line) distance into unit value.

- 2.236/2.236 = 1
- 1/2.236 = 0.447
- 2/2.236 = 0.894

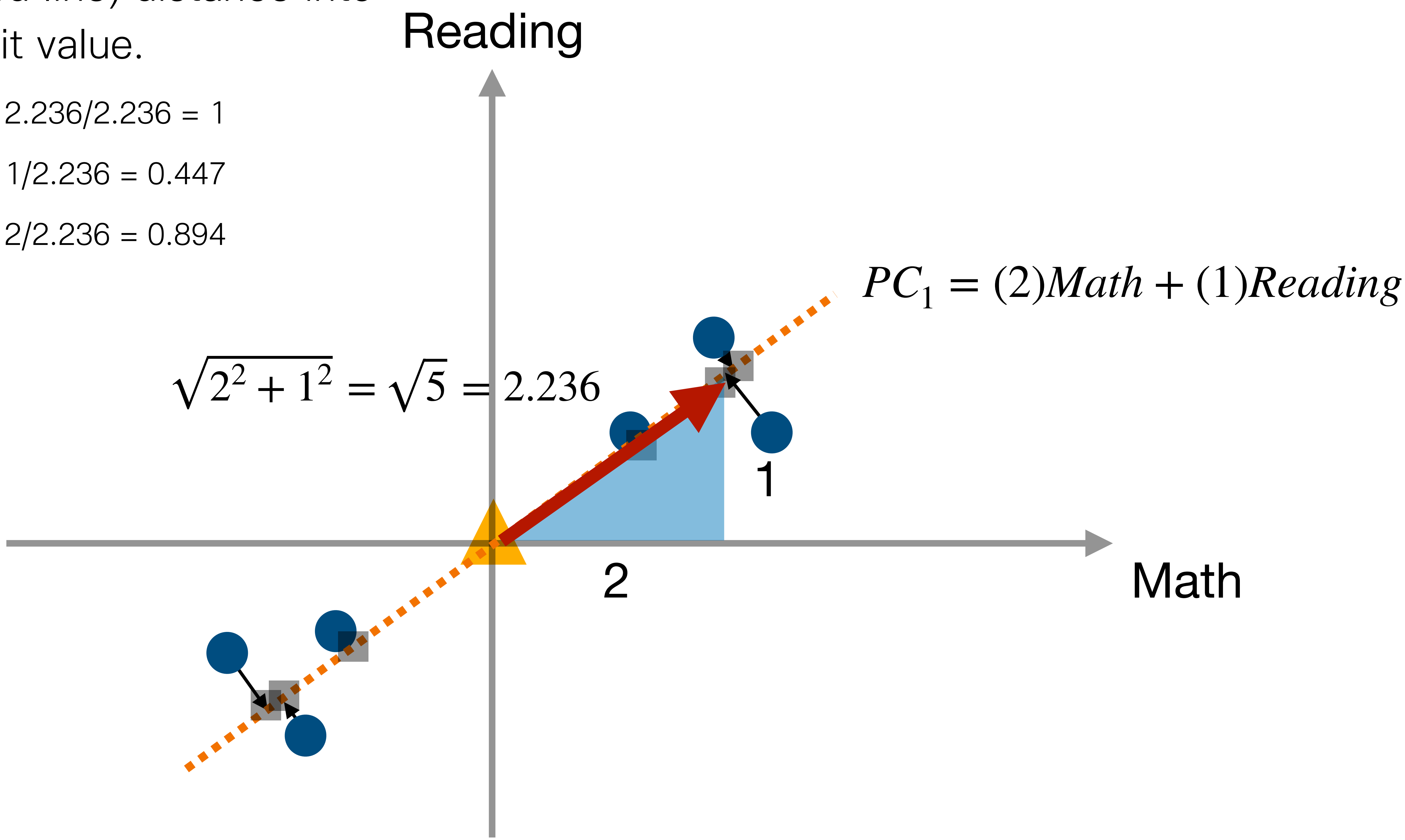


Figure not to scale.

PCA - concept

In PCA, we scaled the (red line) distance into unit value.

- 2.236/2.236 = 1
- 1/2.236 = 0.447
- 2/2.236 = 0.894

Eigenvector

(0.894,0.447) is called **"eigenvector" for PC1**

$$PC_1 = (0.894)Math + (0.447)Reading$$

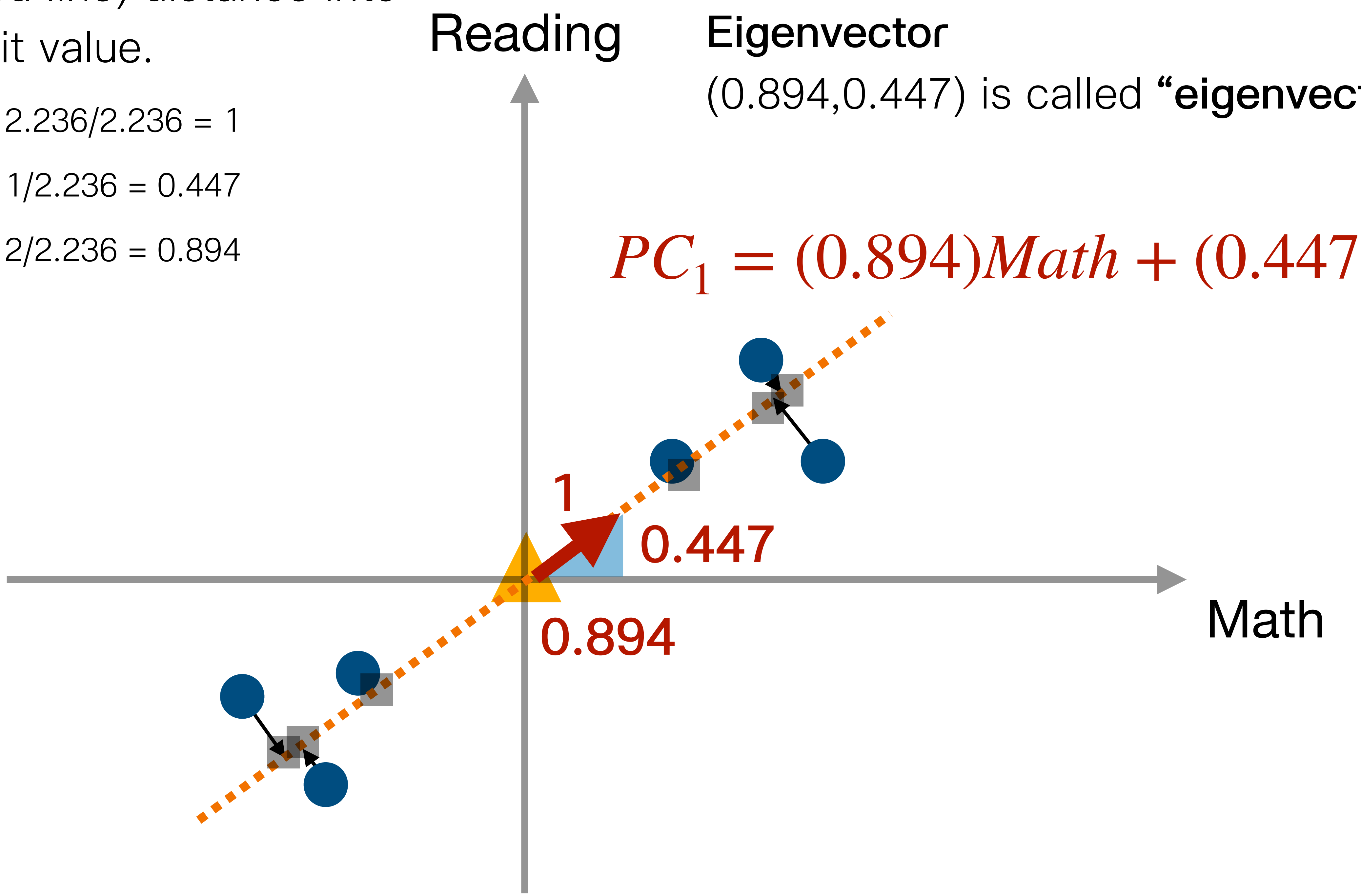
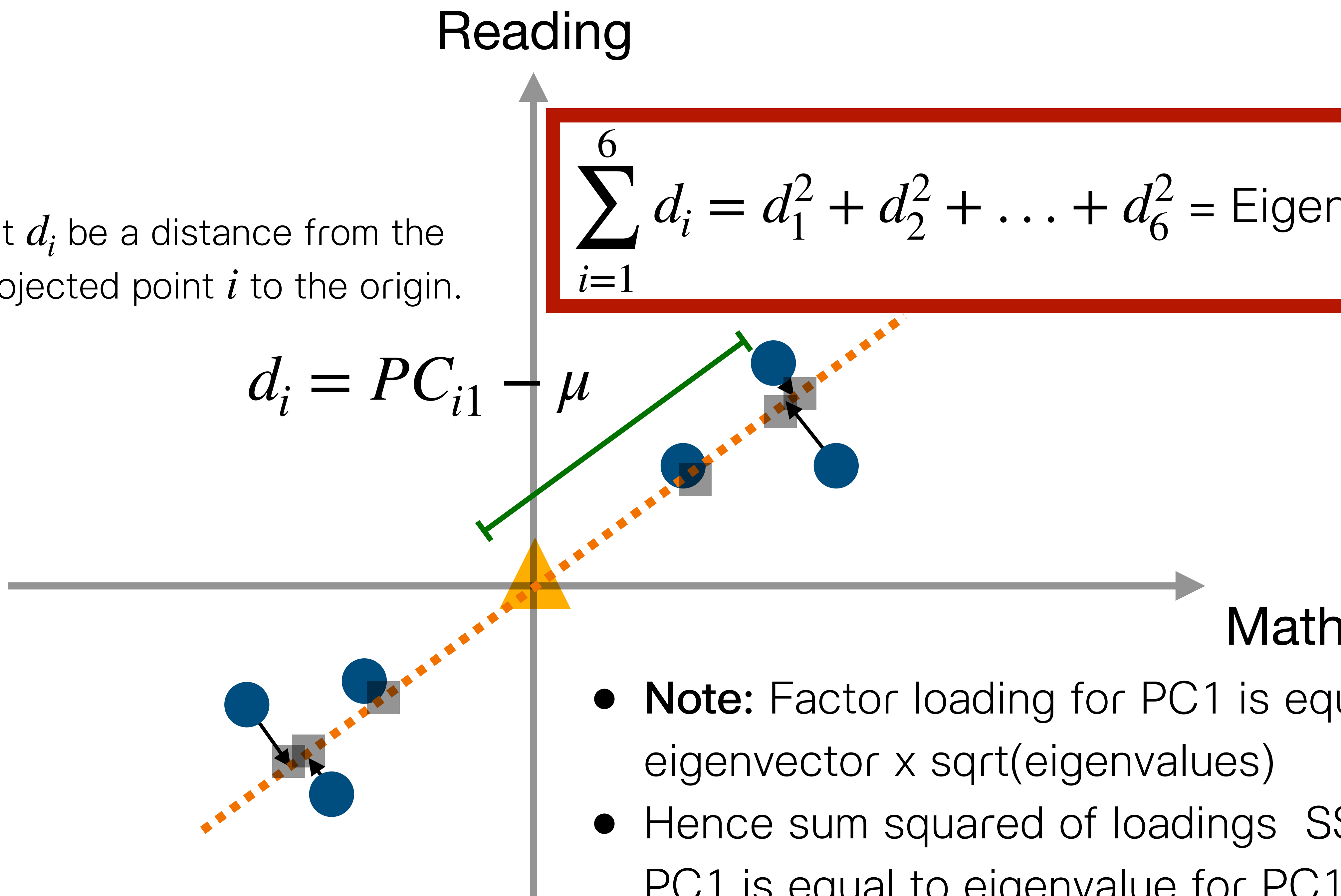


Figure not to scale.

PCA - concept



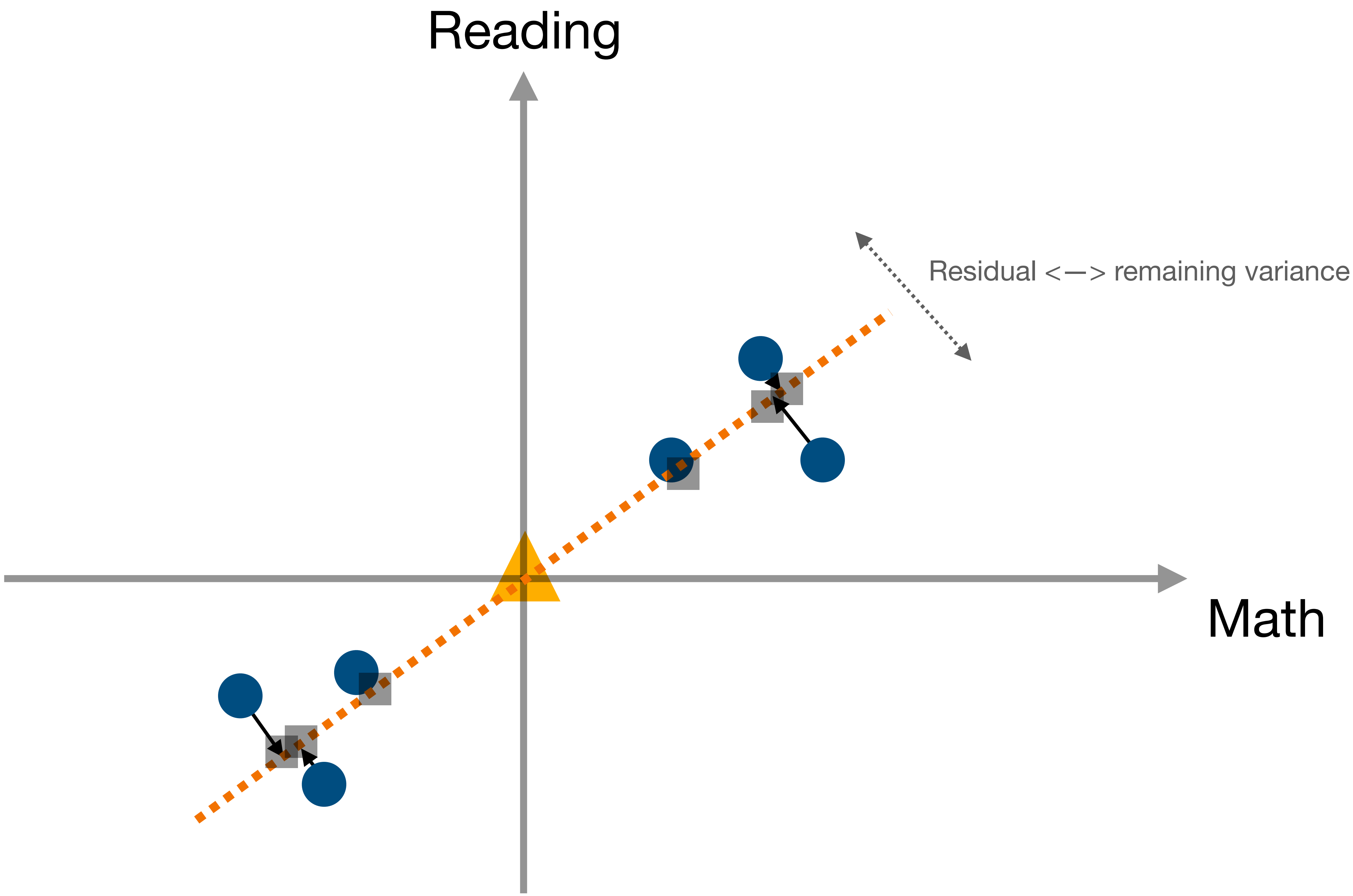
Let d_i be a distance from the projected point i to the origin.

$$d_i = PC_{i1} - \mu$$

$$\boxed{\sum_{i=1}^{6} d_i = d_1^2 + d_2^2 + \ldots + d_6^2 = \text{Eigenvalue for PC1}}$$

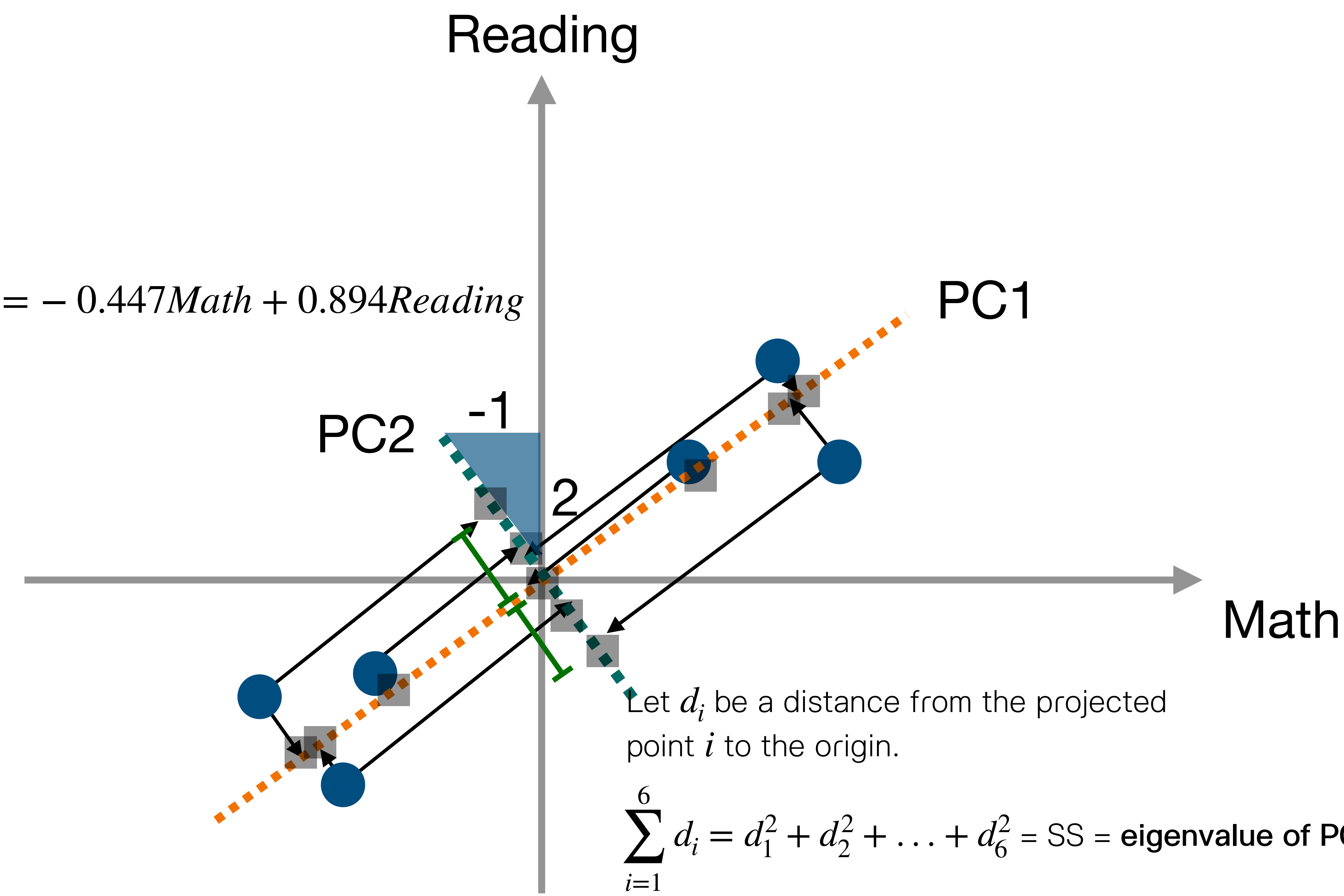
- **Note:** Factor loading for PC1 is equal to eigenvector x sqrt(eigenvalues)
- Hence sum squared of loadings SS-loading for PC1 is equal to eigenvalue for PC1

PCA - concept



PCA - concept

$$PC_2 = -0.447Math + 0.894Reading$$



Let d_i be a distance from the projected point i to the origin.

$$\sum_{i=1}^{6} d_i = d_1^2 + d_2^2 + \ldots + d_6^2 = \text{SS} = \textbf{eigenvalue of PC2}$$

Figure not to scale.

PCA - concept

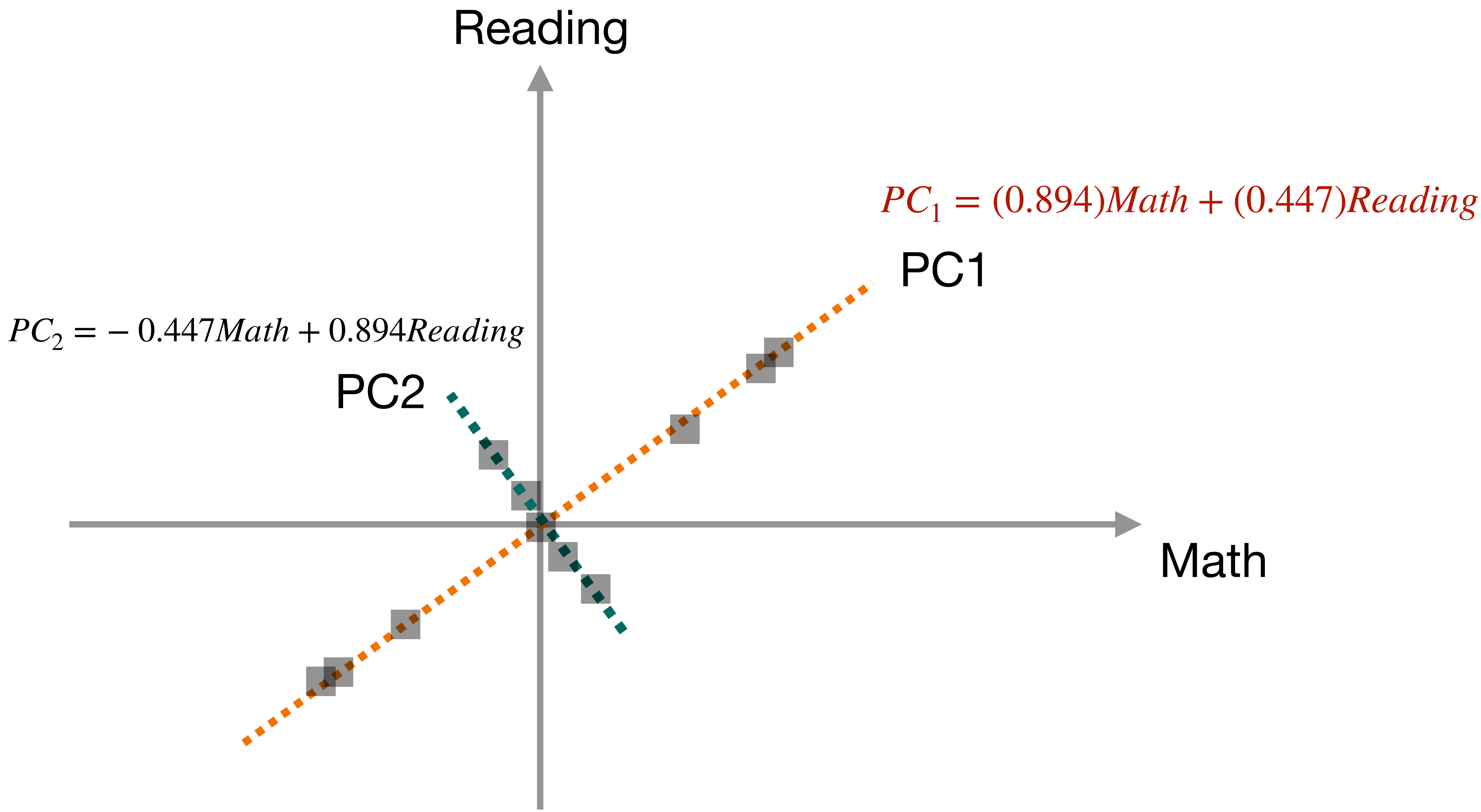


Figure not to scale.

PCA - concept

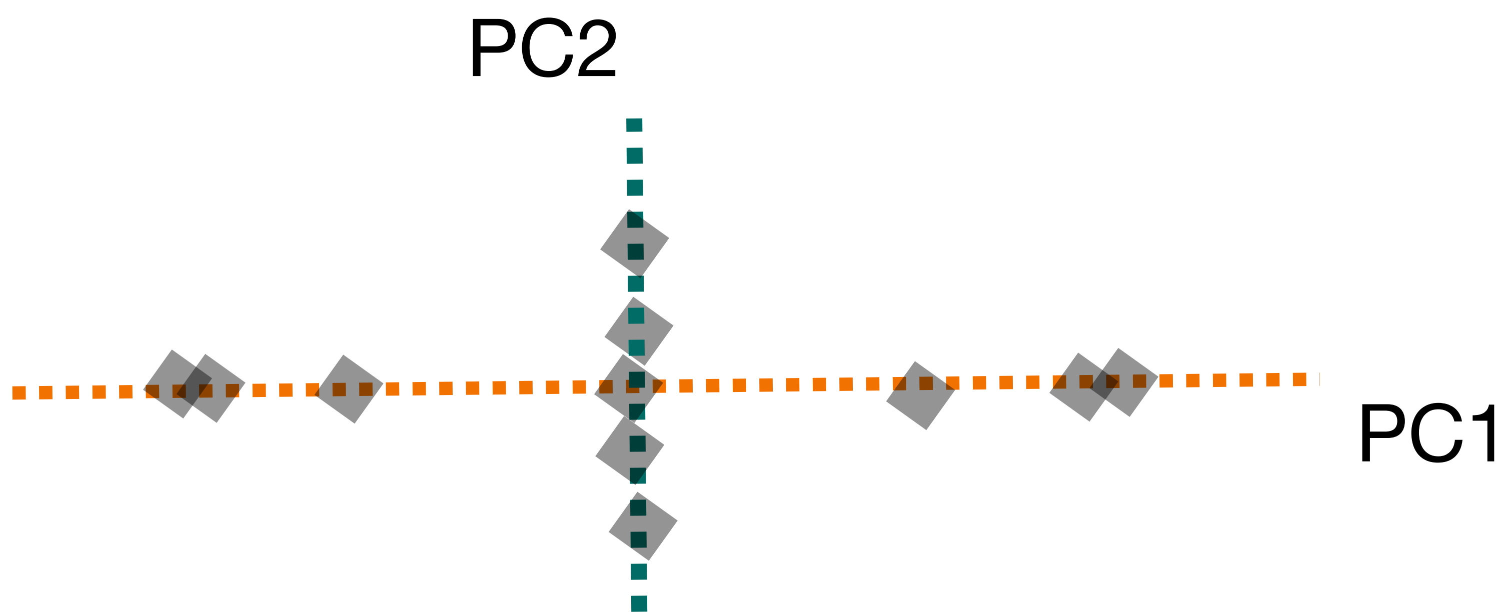
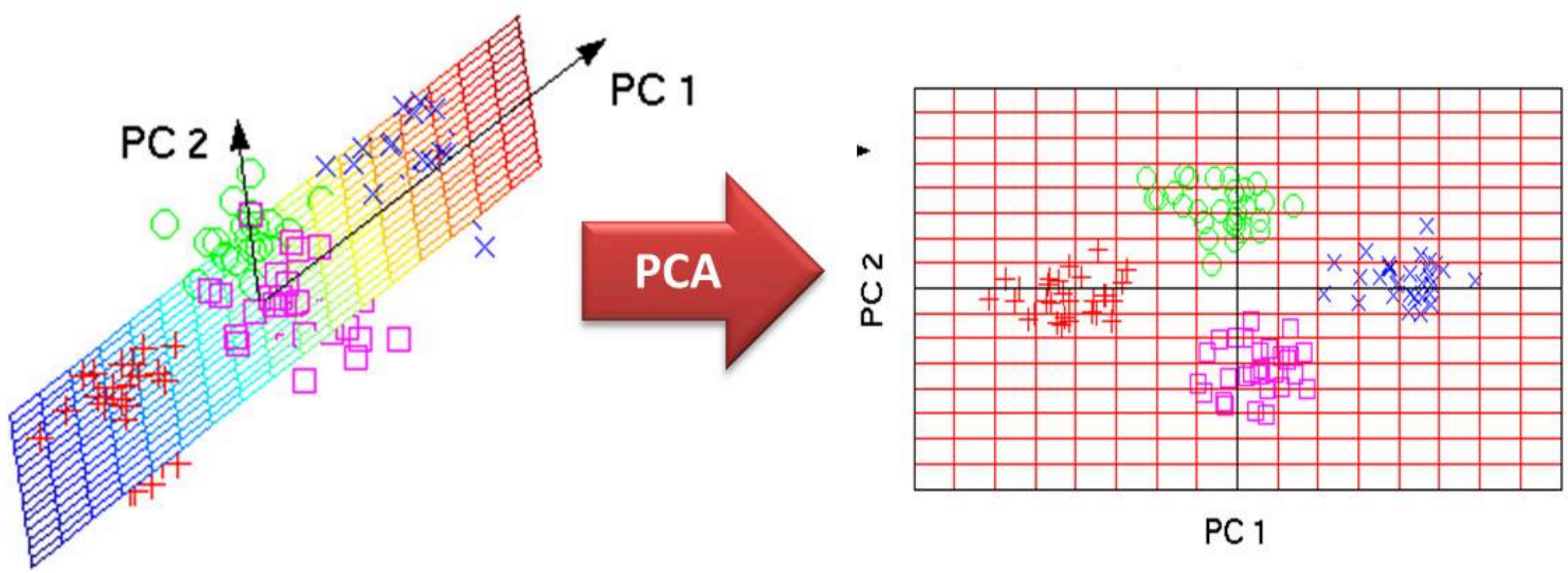


Figure not to scale.



https://www.kindsonthegenius.com/blog/2017/12/dimensionality-reduction-and-principal-component-analysis-pca.html