BMI 206

Linear Regression and Principal Component Analysis

Vasilis Ntranos 2024

The geometry of linear regression

- A collection of **n features** across **m individuals** + one observation per **individual**
- **datapoints:** $x_1, x_2, x_3, \ldots, x_m \in R^n$, **observation:** $y \in R^m$
  - **Model**: for some reason, we think that y_i depends on x_{ij}'s
    Specifically, the ith measurement of y could be given by a **linear combination** of x_{ij}'s
    (approximated) (plus a constant / intercept)

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$y_1 \sim \alpha_1 x_{11} + \alpha_2 x_{12} + \alpha_3 x_{13} + \alpha_4 x_{14} + \beta$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$y_1 \sim \alpha_1 x_{11} + \alpha_2 x_{12} + \alpha_3 x_{13} + \alpha_4 x_{14} + \beta$$

$$y_2 \sim \alpha_1 x_{21} + \alpha_2 x_{22} + \alpha_3 x_{23} + \alpha_4 x_{24} + \beta$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$y_1 \sim \alpha_1 x_{11} + \alpha_2 x_{12} + \alpha_3 x_{13} + \alpha_4 x_{14} + \beta$$

$$y_2 \sim \alpha_1 x_{21} + \alpha_2 x_{22} + \alpha_3 x_{23} + \alpha_4 x_{24} + \beta$$

$$y_3 \sim \alpha_1 x_{31} + \alpha_2 x_{32} + \alpha_3 x_{33} + \alpha_4 x_{34} + \beta$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$y_1 \sim \alpha_1 x_{11} + \alpha_2 x_{12} + \alpha_3 x_{13} + \alpha_4 x_{14} + \beta$$

$$y_2 \sim \alpha_1 x_{21} + \alpha_2 x_{22} + \alpha_3 x_{23} + \alpha_4 x_{24} + \beta$$

$$y_3 \sim \alpha_1 x_{31} + \alpha_2 x_{32} + \alpha_3 x_{33} + \alpha_4 x_{34} + \beta$$

$$y_4 \sim \alpha_1 x_{41} + \alpha_2 x_{42} + \alpha_3 x_{43} + \alpha_4 x_{44} + \beta$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$y_1 \sim \alpha_1 x_{11} + \alpha_2 x_{12} + \alpha_3 x_{13} + \alpha_4 x_{14} + \beta$$

$$y_2 \sim \alpha_1 x_{21} + \alpha_2 x_{22} + \alpha_3 x_{23} + \alpha_4 x_{24} + \beta$$

$$y_3 \sim \alpha_1 x_{31} + \alpha_2 x_{32} + \alpha_3 x_{33} + \alpha_4 x_{34} + \beta$$

$$y_4 \sim \alpha_1 x_{41} + \alpha_2 x_{42} + \alpha_3 x_{43} + \alpha_4 x_{44} + \beta$$

$$y_5 \sim \alpha_1 x_{51} + \alpha_2 x_{52} + \alpha_3 x_{53} + \alpha_4 x_{54} + \beta$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$y_1 \sim \alpha_1 x_{11} + \alpha_2 x_{12} + \alpha_3 x_{13} + \alpha_4 x_{14} + \beta$$
$$y_2 \sim \alpha_1 x_{21} + \alpha_2 x_{22} + \alpha_3 x_{23} + \alpha_4 x_{24} + \beta$$
$$y_3 \sim \alpha_1 x_{31} + \alpha_2 x_{32} + \alpha_3 x_{33} + \alpha_4 x_{34} + \beta$$
$$y_4 \sim \alpha_1 x_{41} + \alpha_2 x_{42} + \alpha_3 x_{43} + \alpha_4 x_{44} + \beta$$
$$y_5 \sim \alpha_1 x_{51} + \alpha_2 x_{52} + \alpha_3 x_{53} + \alpha_4 x_{54} + \beta$$
$$y_6 \sim \alpha_1 x_{61} + \alpha_2 x_{62} + \alpha_3 x_{63} + \alpha_4 x_{64} + \beta$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$\begin{aligned}
y_1 &\sim \alpha_1 x_{11} + \alpha_2 x_{12} + \alpha_3 x_{13} + \alpha_4 x_{14} + \beta \\
y_2 &\sim \alpha_1 x_{21} + \alpha_2 x_{22} + \alpha_3 x_{23} + \alpha_4 x_{24} + \beta \\
y_3 &\sim \alpha_1 x_{31} + \alpha_2 x_{32} + \alpha_3 x_{33} + \alpha_4 x_{34} + \beta \\
y_4 &\sim \alpha_1 x_{41} + \alpha_2 x_{42} + \alpha_3 x_{43} + \alpha_4 x_{44} + \beta \\
y_5 &\sim \alpha_1 x_{51} + \alpha_2 x_{52} + \alpha_3 x_{53} + \alpha_4 x_{54} + \beta \\
y_6 &\sim \alpha_1 x_{61} + \alpha_2 x_{62} + \alpha_3 x_{63} + \alpha_4 x_{64} + \beta
\end{aligned}$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} \sim \alpha_1 \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \\ x_{51} \\ x_{61} \end{bmatrix} + \alpha_2 \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ x_{52} \\ x_{62} \end{bmatrix} + \alpha_3 \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \\ x_{53} \\ x_{63} \end{bmatrix} + \alpha_4 \begin{bmatrix} x_{14} \\ x_{24} \\ x_{34} \\ x_{44} \\ x_{54} \\ x_{64} \end{bmatrix} + \beta \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij} + \beta$$

(Unknown) Parameters

(m=6, n=4)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} \sim \alpha_1 \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \\ x_{51} \\ x_{61} \end{bmatrix} + \alpha_2 \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ x_{52} \\ x_{62} \end{bmatrix} + \alpha_3 \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \\ x_{53} \\ x_{63} \end{bmatrix} + \alpha_4 \begin{bmatrix} x_{14} \\ x_{24} \\ x_{34} \\ x_{44} \\ x_{54} \\ x_{64} \end{bmatrix} + \beta \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Let's rename the intercept

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n+1} \alpha_j x_{ij} \qquad x_{i,n+1} = 1, \forall i$$

(Unknown) Parameters

(m=6, n=4)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} \sim \alpha_1 \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \\ x_{51} \\ x_{61} \end{pmatrix} + \alpha_2 \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ x_{52} \\ x_{62} \end{pmatrix} + \alpha_3 \begin{pmatrix} x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \\ x_{53} \\ x_{63} \end{pmatrix} + \alpha_4 \begin{pmatrix} x_{14} \\ x_{24} \\ x_{34} \\ x_{44} \\ x_{54} \\ x_{64} \end{pmatrix} + \alpha_5 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Let’s rename the intercept

The geometry of linear regression

Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n+1} \alpha_j x_{ij} \qquad x_{i,n+1} = 1, \forall i$$

(Unknown) Parameters

(m=6, n=4)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} \sim \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & 1 \\ x_{21} & x_{22} & x_{23} & x_{24} & 1 \\ x_{31} & x_{32} & x_{33} & x_{34} & 1 \\ x_{41} & x_{42} & x_{43} & x_{44} & 1 \\ x_{51} & x_{52} & x_{53} & x_{54} & 1 \\ x_{61} & x_{62} & x_{63} & x_{64} & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix}$$

The geometry of linear regression

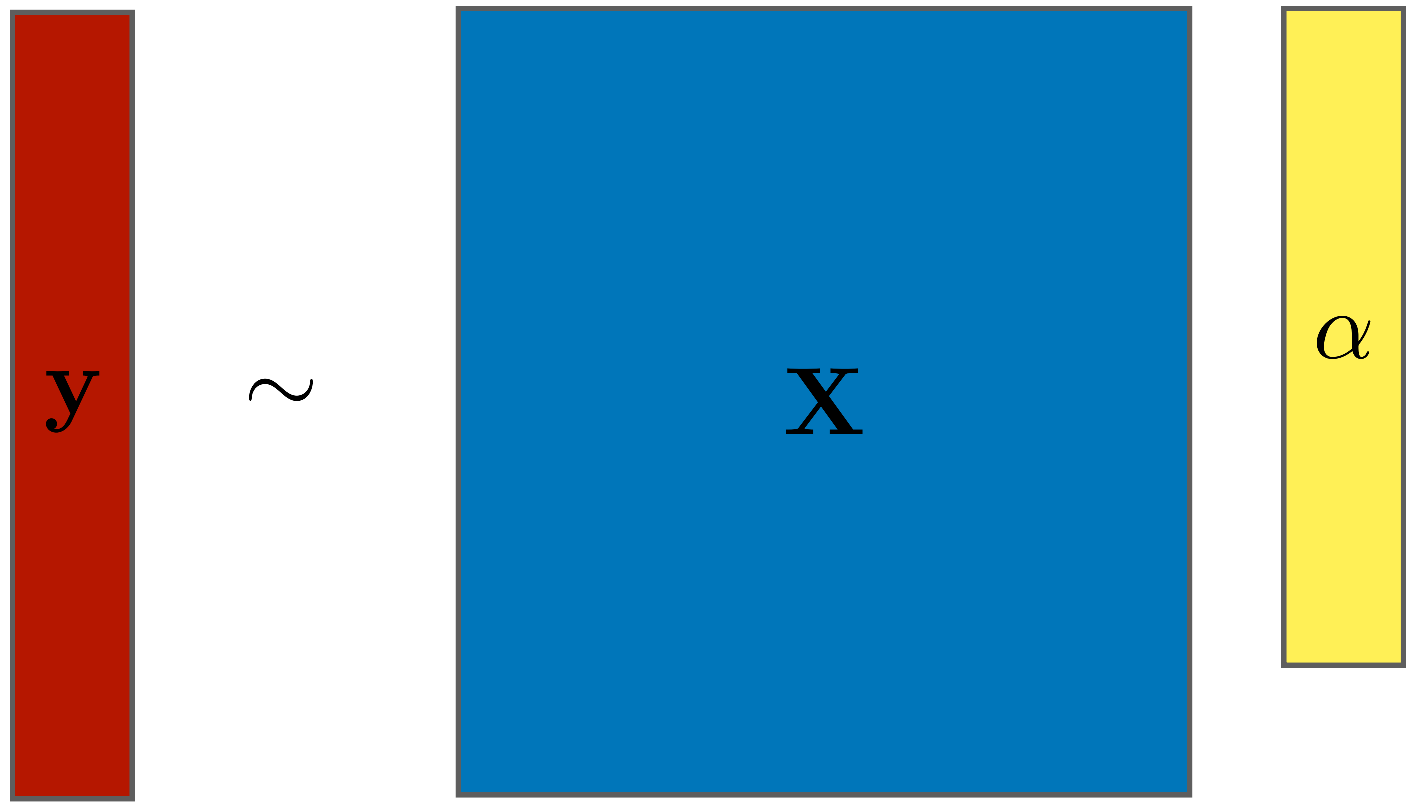
Value of the jth feature of the ith individual

observation/measurement of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij}$$

(Unknown) Parameters

(In general)



"A system of **m** equations with **n** unknowns"

$$\mathbf{y} \sim \mathbf{X}\alpha$$

$$\mathbf{y} \in \mathbb{R}^m$$

$$\mathbf{X} \in \mathbb{R}^{m \times n}$$

$$\mathbf{a} \in \mathbb{R}^n$$

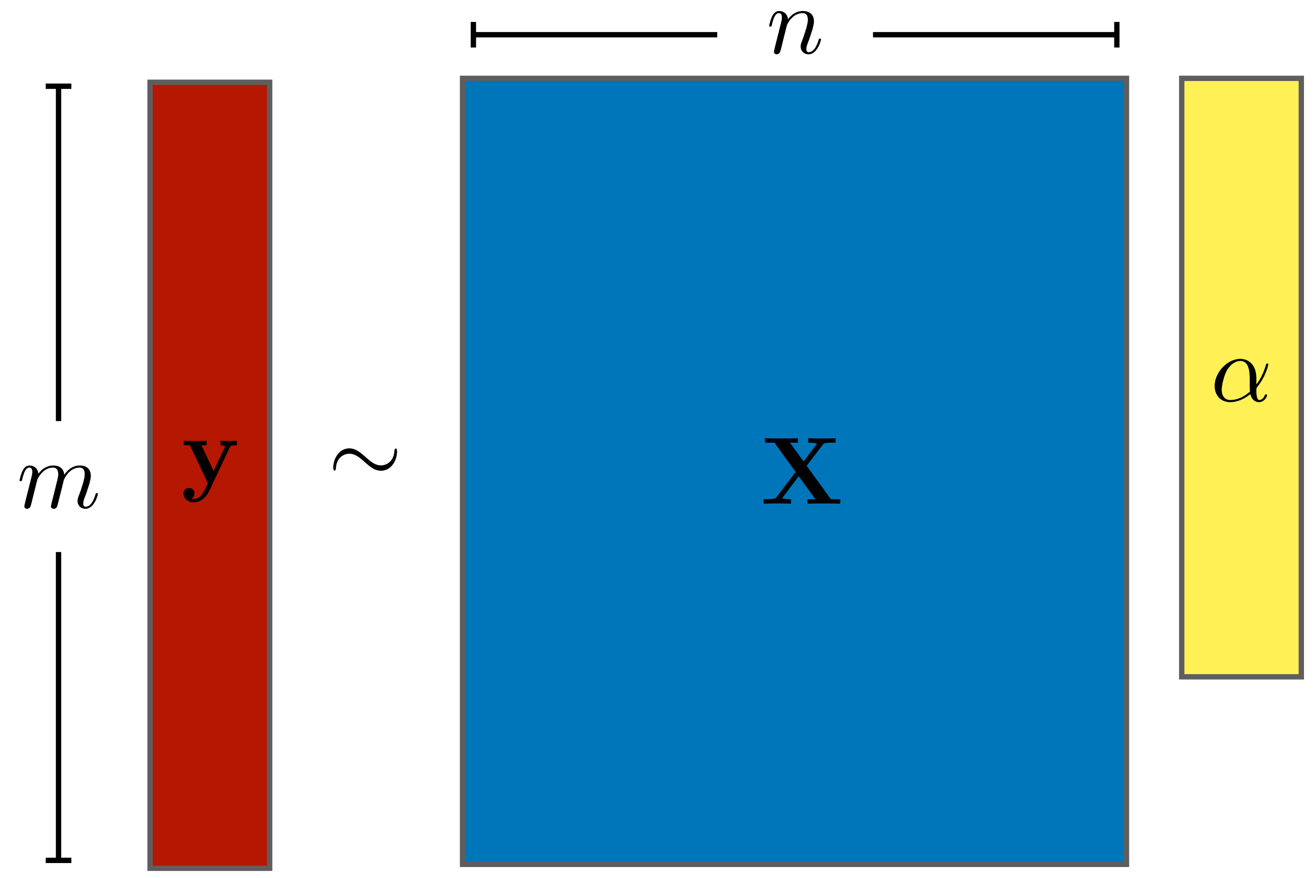
The geometry of linear regression

observation/measurement of the ith individual

Value of the jth feature of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij}$$

(Unknown) Parameters



“A system of **m** equations with **n** unknowns”

$$\mathbf{y} \sim \mathbf{X}\alpha$$

$$\mathbf{y} \in \mathbb{R}^m$$

$$\mathbf{X} \in \mathbb{R}^{m \times n}$$

$$\mathbf{a} \in \mathbb{R}^n$$

- No solution when m>n and rank(X)=n, i.e., y≠Xa, for any a.

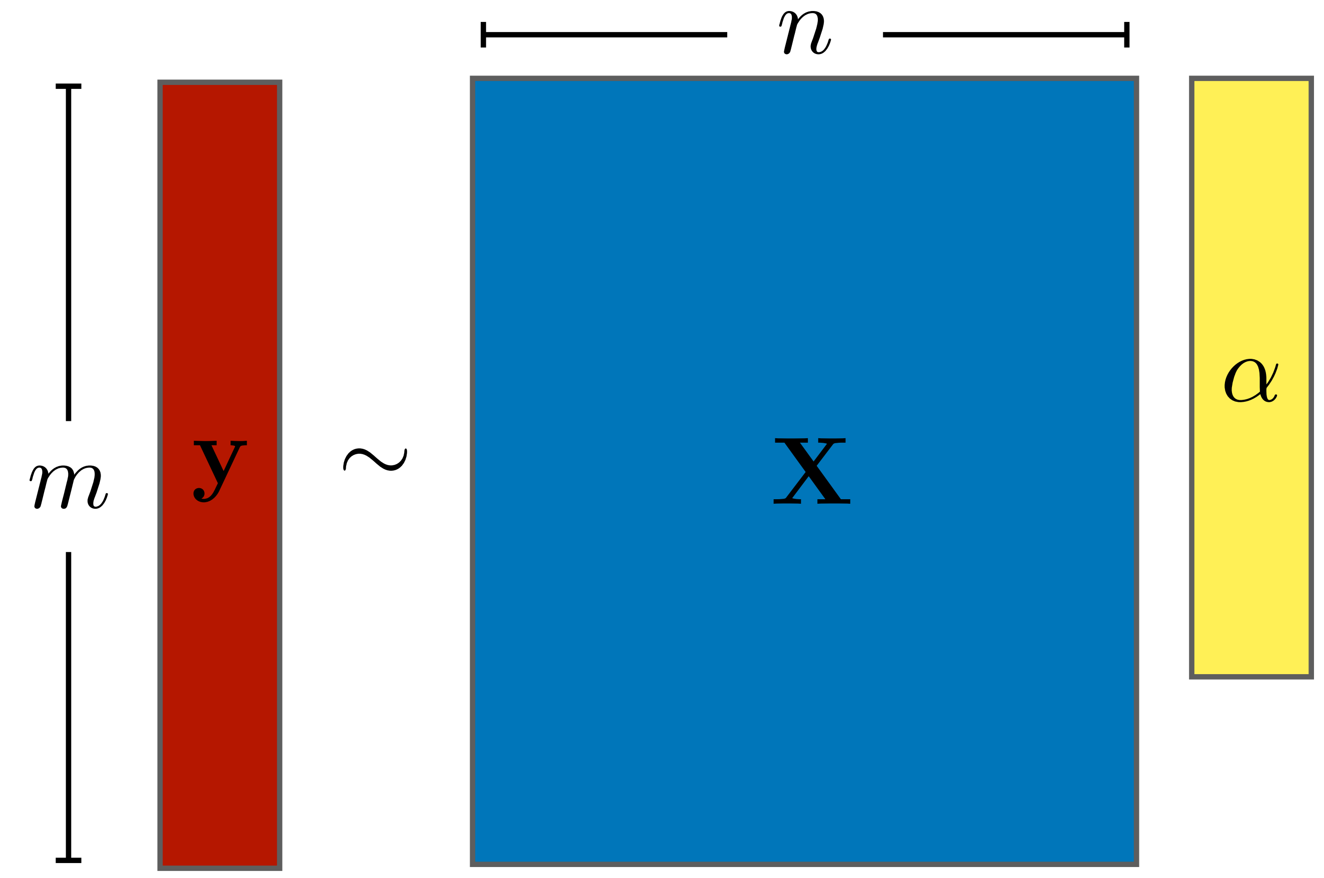
The geometry of linear regression

observation/measurement of the ith individual

Value of the jth feature of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij}$$

(Unknown) Parameters



"A system of **m** equations with **n** unknowns"

$$\mathbf{y} \sim \mathbf{X}\alpha$$

$$\mathbf{y} \in \mathbb{R}^m$$

$$\mathbf{X} \in \mathbb{R}^{m \times n}$$

$$\mathbf{a} \in \mathbb{R}^n$$

- No solution when m>n and rank(X)=n, i.e., y≠Xa, for any a.
- Approach: **minimize error** / find parameters that bring **Xa** *as close as possible* to **y**:

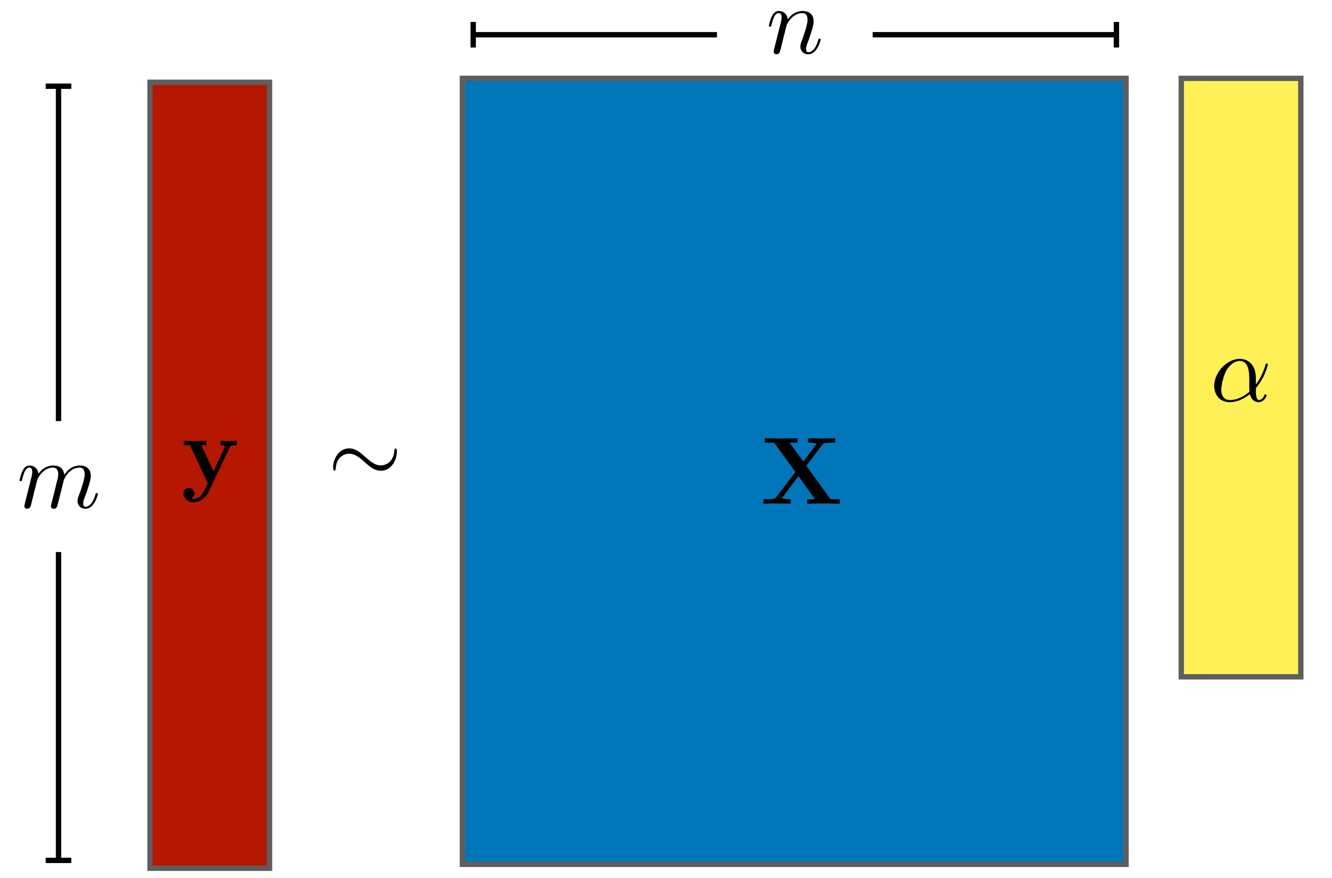
The geometry of linear regression

observation/measurement of the ith individual

Value of the jth feature of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij}$$

(Unknown) Parameters



"A system of **m** equations with **n** unknowns"

$$\mathbf{y} \sim \mathbf{X}\alpha$$

$$\mathbf{y} \in \mathbb{R}^m$$

$$\mathbf{X} \in \mathbb{R}^{m \times n}$$

$$\mathbf{a} \in \mathbb{R}^n$$

- No solution when m>n and rank(X)=n, i.e., y≠Xa, for any a.
- Approach: **minimize error** / find parameters that bring **Xa** *as close as possible* to **y**:

$$\min_{\alpha \in \mathbb{R}^n} \underbrace{\frac{1}{m} \sum_i \underbrace{\left(\underbrace{y_i - \sum_j \alpha_j x_{ij}}_{\text{Error}} \right)^2}_{\text{Square}}}_{\text{Mean}} \Leftrightarrow \min_{\alpha \in \mathbb{R}^n} ||\mathbf{y} - \mathbf{X}\alpha||^2$$

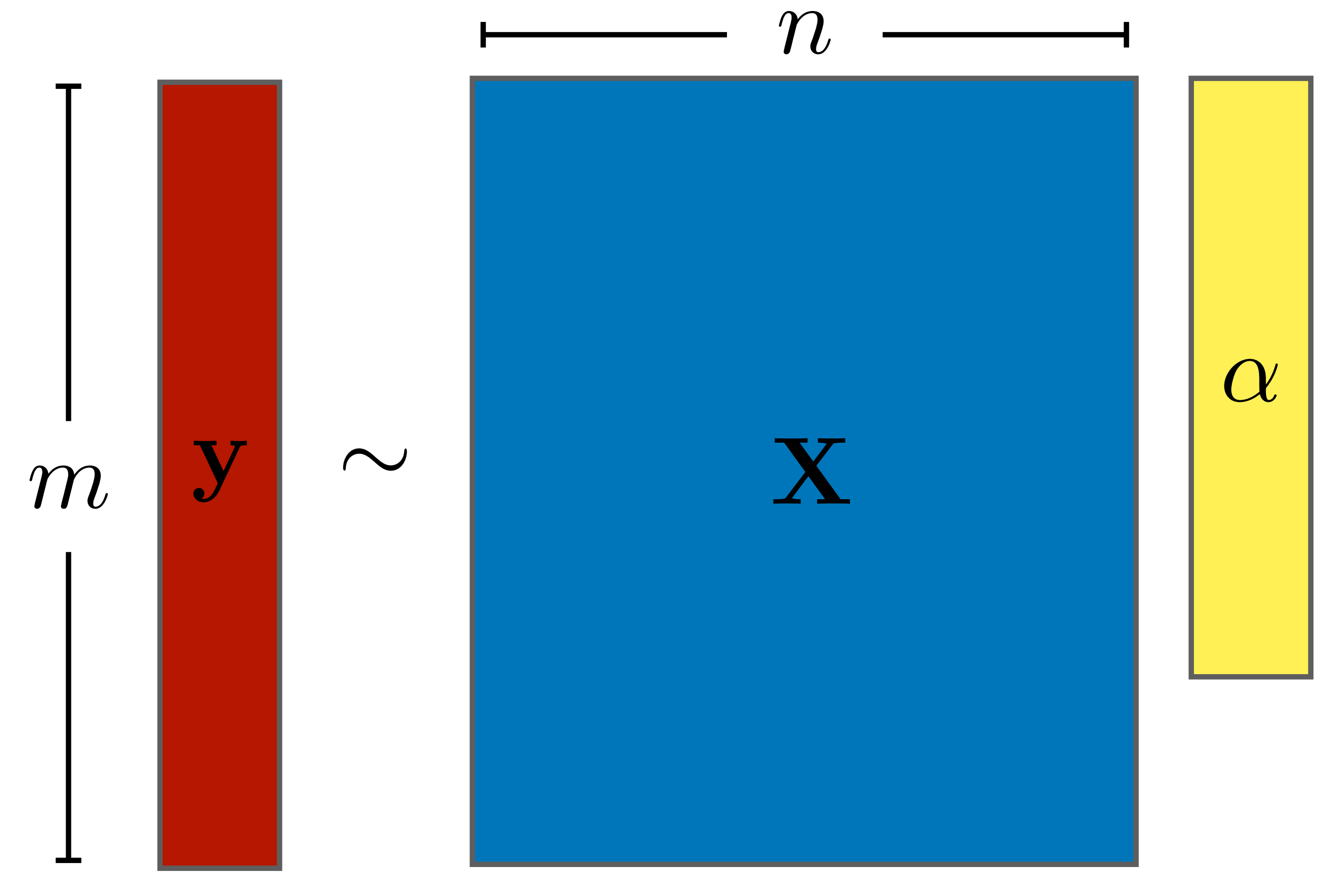
The geometry of linear regression

observation/measurement of the ith individual

Value of the jth feature of the ith individual

$$\mathbf{y}_i \sim \sum_{j=1}^{n} \alpha_j x_{ij}$$

(Unknown) Parameters



"A system of **m** equations with **n** unknowns"

$$\mathbf{y} \sim \mathbf{X}\alpha$$

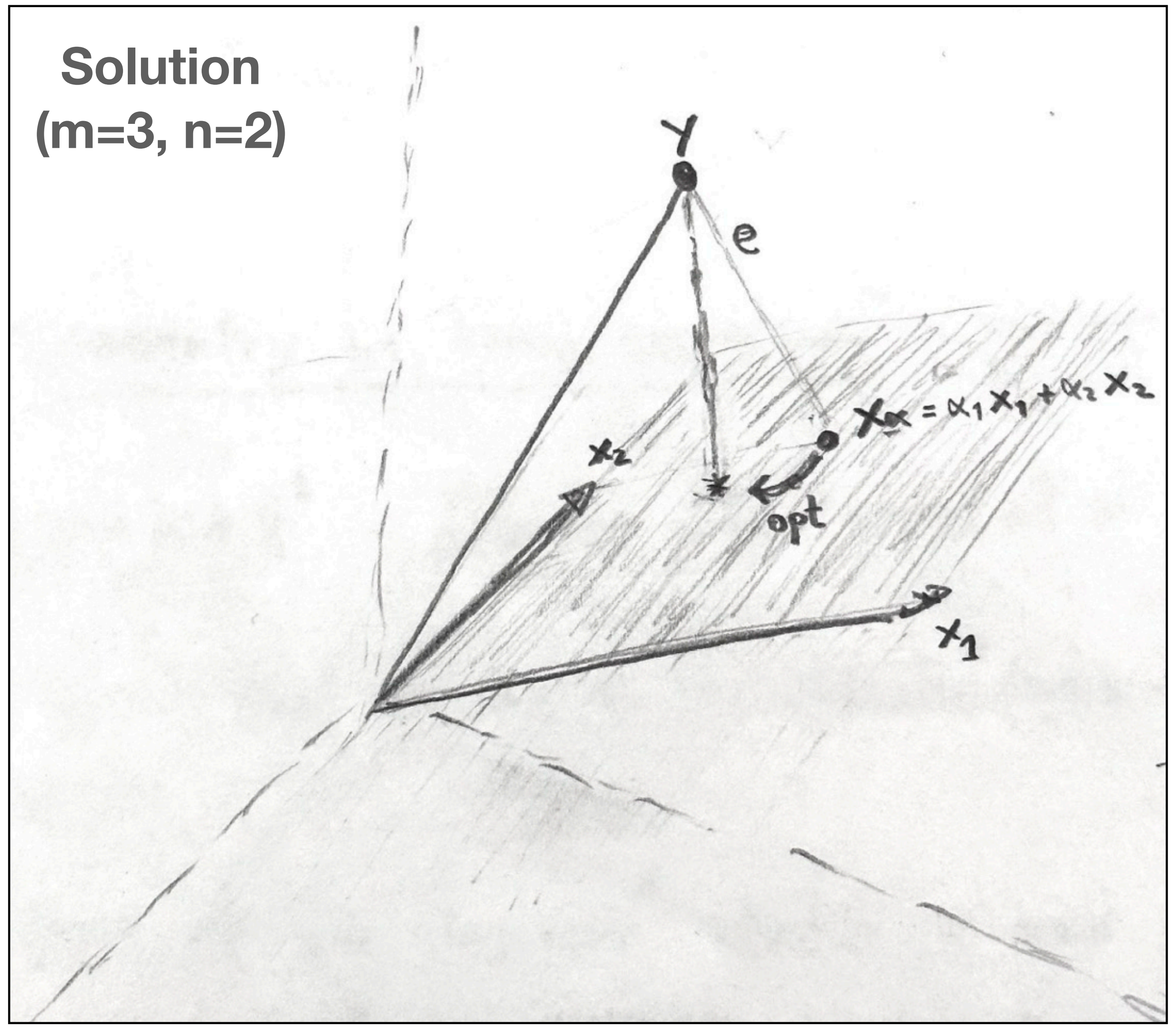
$$\mathbf{y} \in \mathbb{R}^m$$

$$\mathbf{X} \in \mathbb{R}^{m \times n}$$

$$\mathbf{a} \in \mathbb{R}^n$$

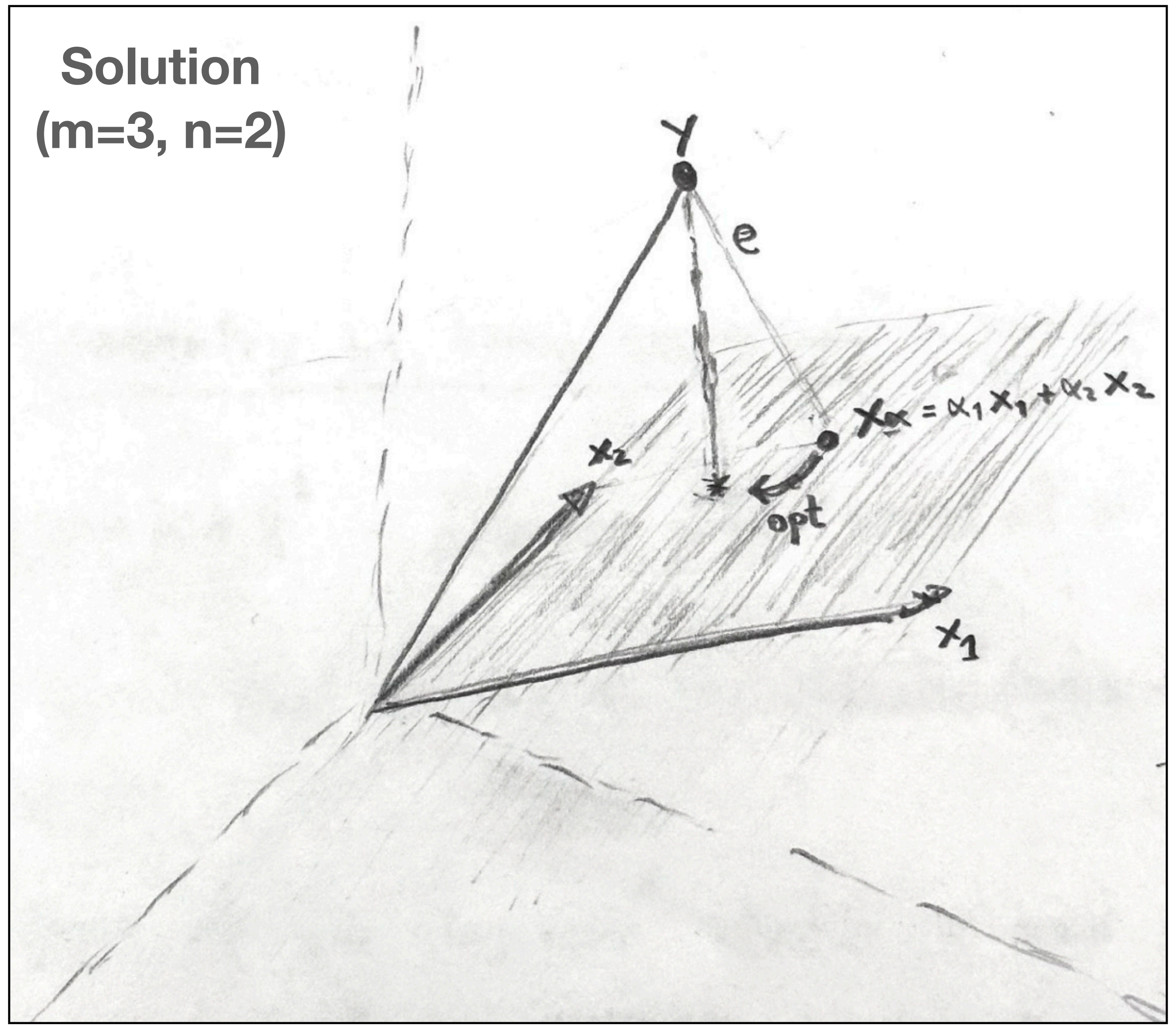
- No solution when m>n and rank(X)=n, i.e., y≠Xa, for any a.
- Approach: **minimize error** /
  find parameters that bring **Xa** *as close as possible* to **y**:

$$\underbrace{\min_{\alpha \in \mathbb{R}^n} \frac{1}{m} \sum_i \underbrace{\Big(\underbrace{y_i - \sum_j \alpha_j x_{ij}}_{\text{Error}} \Big)^2}_{\text{Square}}}_{\text{Mean}} \Leftrightarrow \min_{\alpha \in \mathbb{R}^n} ||\mathbf{y} - \mathbf{X}\alpha||^2$$



The geometry of linear regression

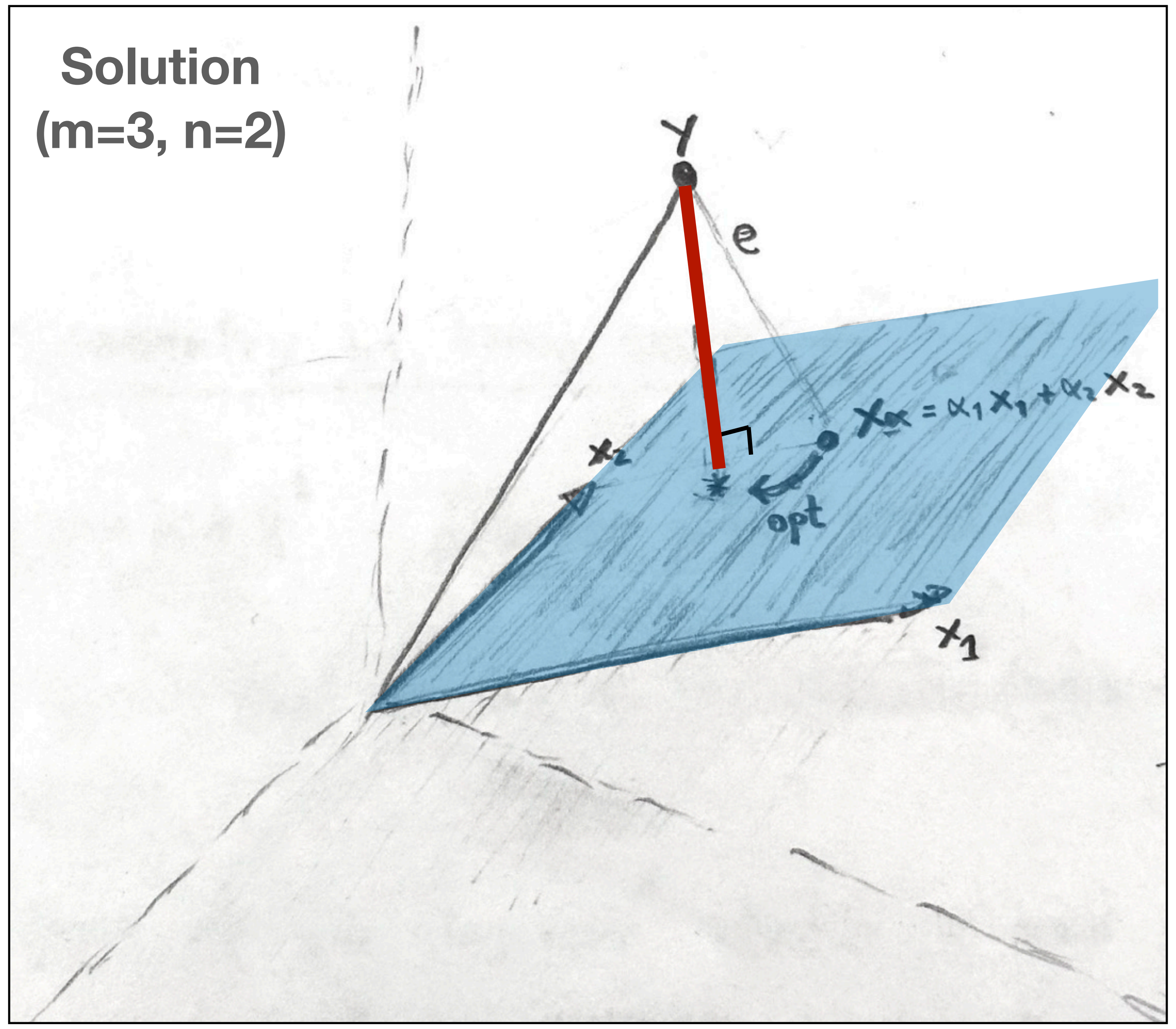
Calculating $\alpha^* = \arg\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$



The geometry of linear regression

Calculating $\alpha^* = \arg\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$

- **Orthogonality Principle:** "the **minimum** is achieved when the error is perpendicular to all columns of X"



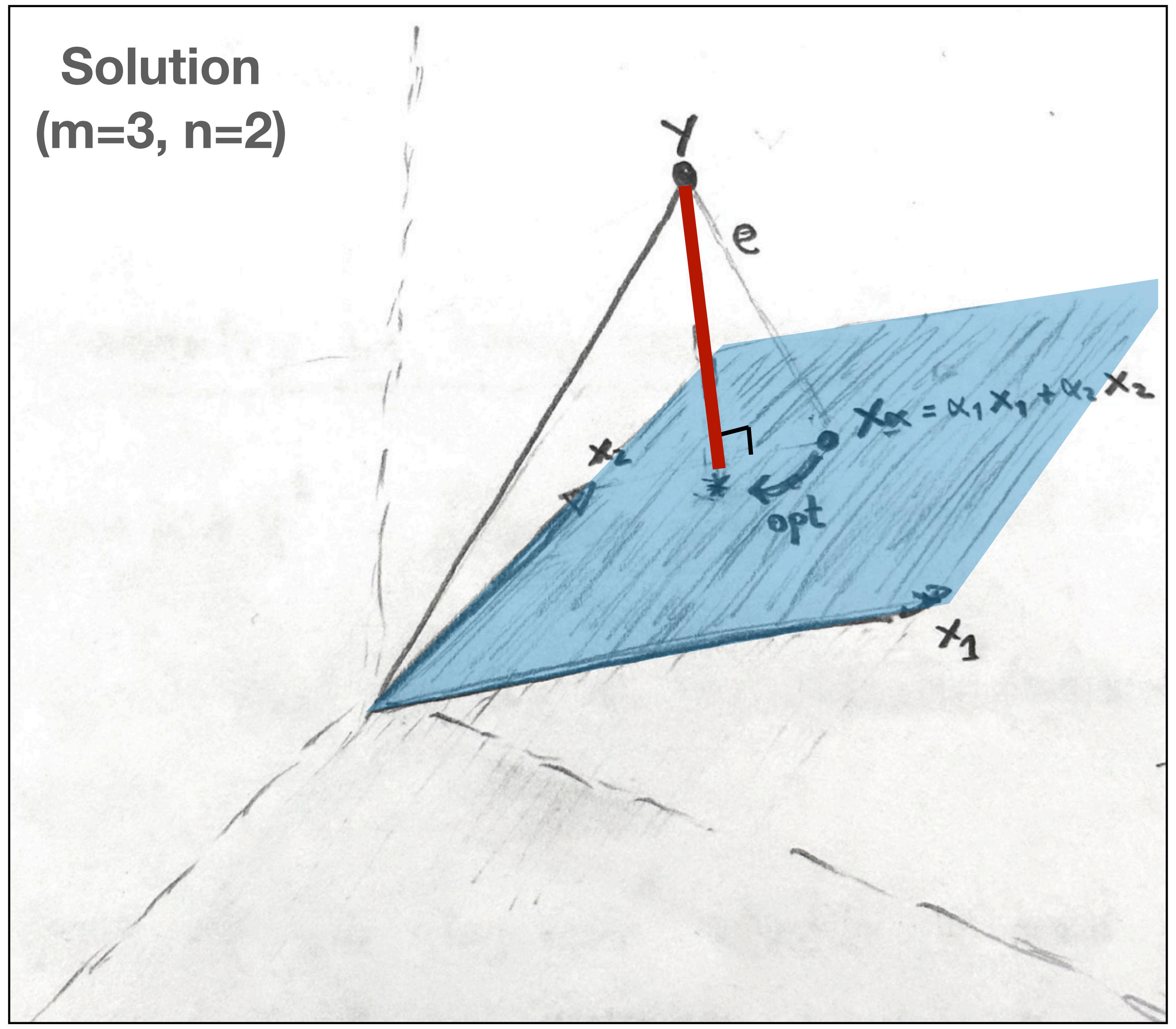
The geometry of linear regression

Calculating $\alpha^* = \arg\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$

- **Orthogonality Principle:** "the **minimum** is achieved when the error is perpendicular to all columns of X"

$$\mathbf{x}_i^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\alpha^*) = 0, \forall i$$

Show this by setting the partial derivatives of the MSE equal to zero



Solution (m=3, n=2)

The geometry of linear regression

Calculating $\alpha^* = \arg\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$

- **Orthogonality Principle:** "the **minimum** is achieved when the error is perpendicular to all columns of X"

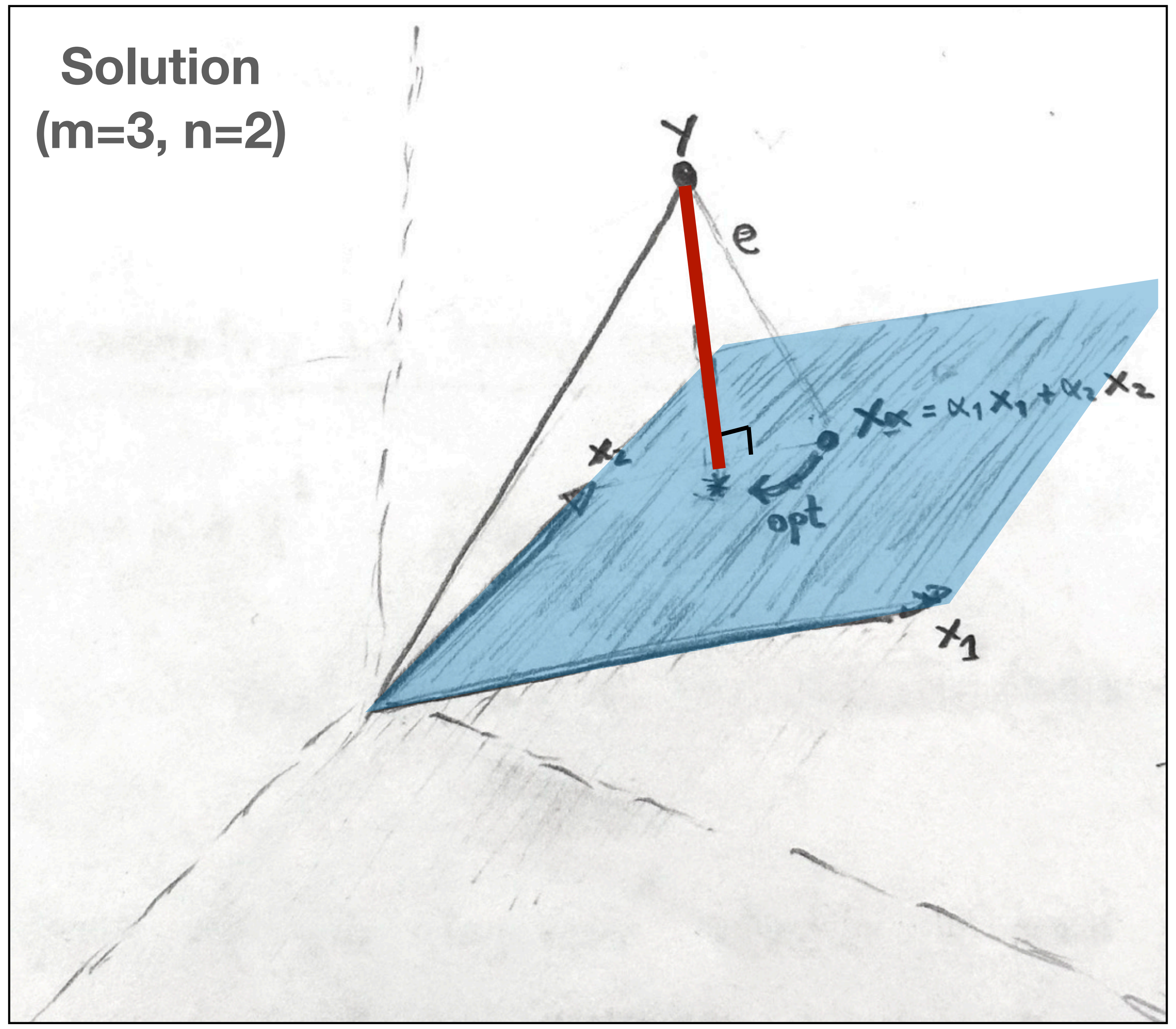
$$\mathbf{x}_i^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\alpha^*) = 0, \forall i$$

- **The solution α* satisfying the above equations is given by:**

$$\boxed{\alpha^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}$$

Proof:

$$\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\alpha^*) = 0$$
$$\Leftrightarrow \mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\alpha^* = 0$$
$$\Leftrightarrow \mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\alpha^*$$
$$\Leftrightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}}_{=\text{ Identity}}\,\alpha^*$$
$$\Leftrightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \alpha^*$$



The geometry of linear regression

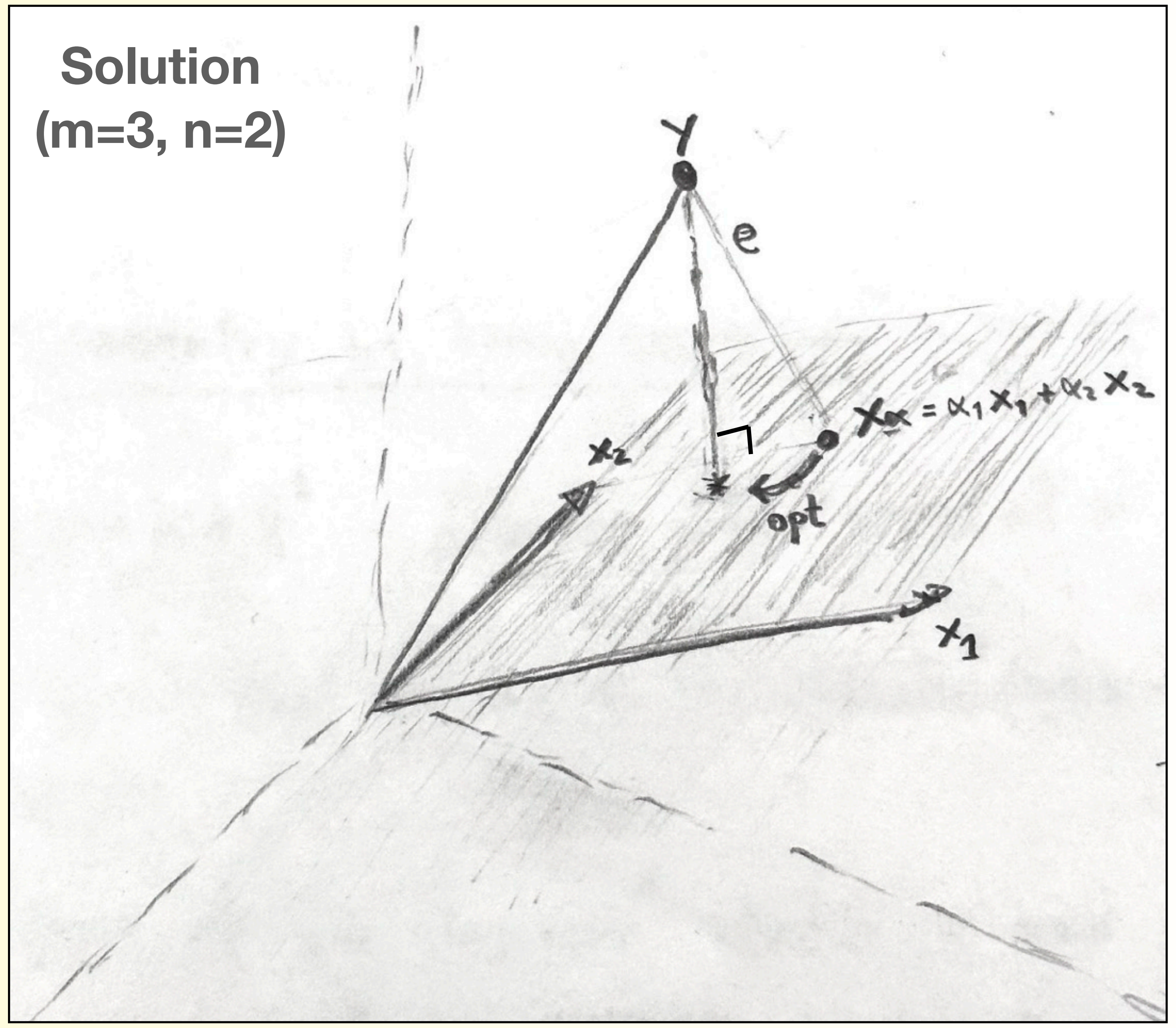
Summary

Linear regression: $\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$

Solution is given by **projecting** y onto the data

$$\alpha^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

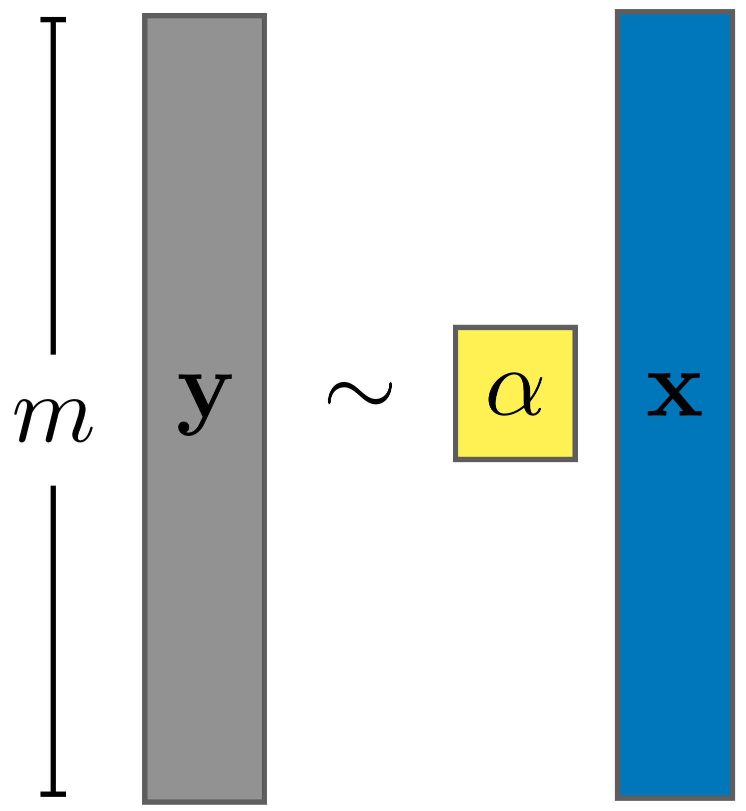
Solution (m=3, n=2)



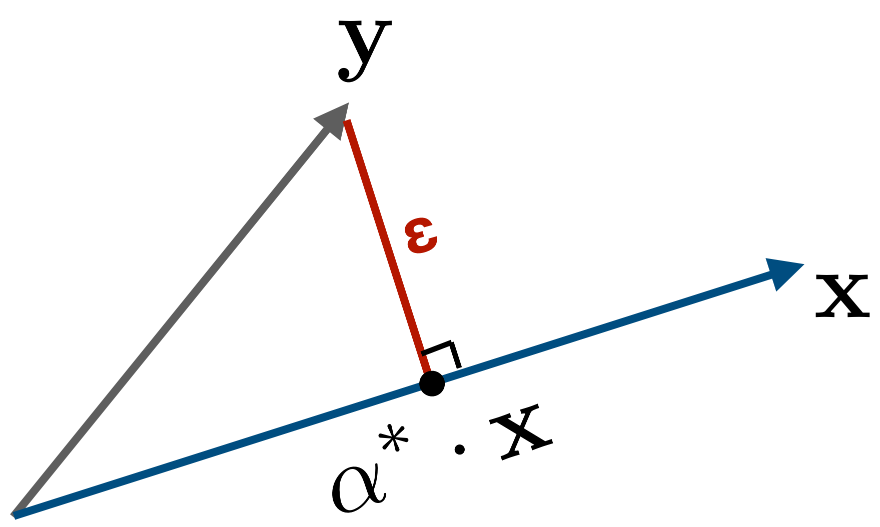
The geometry of linear regression

Where is the "regression line" ?

- n=1, general m



"2 vectors in **m dimensions**"



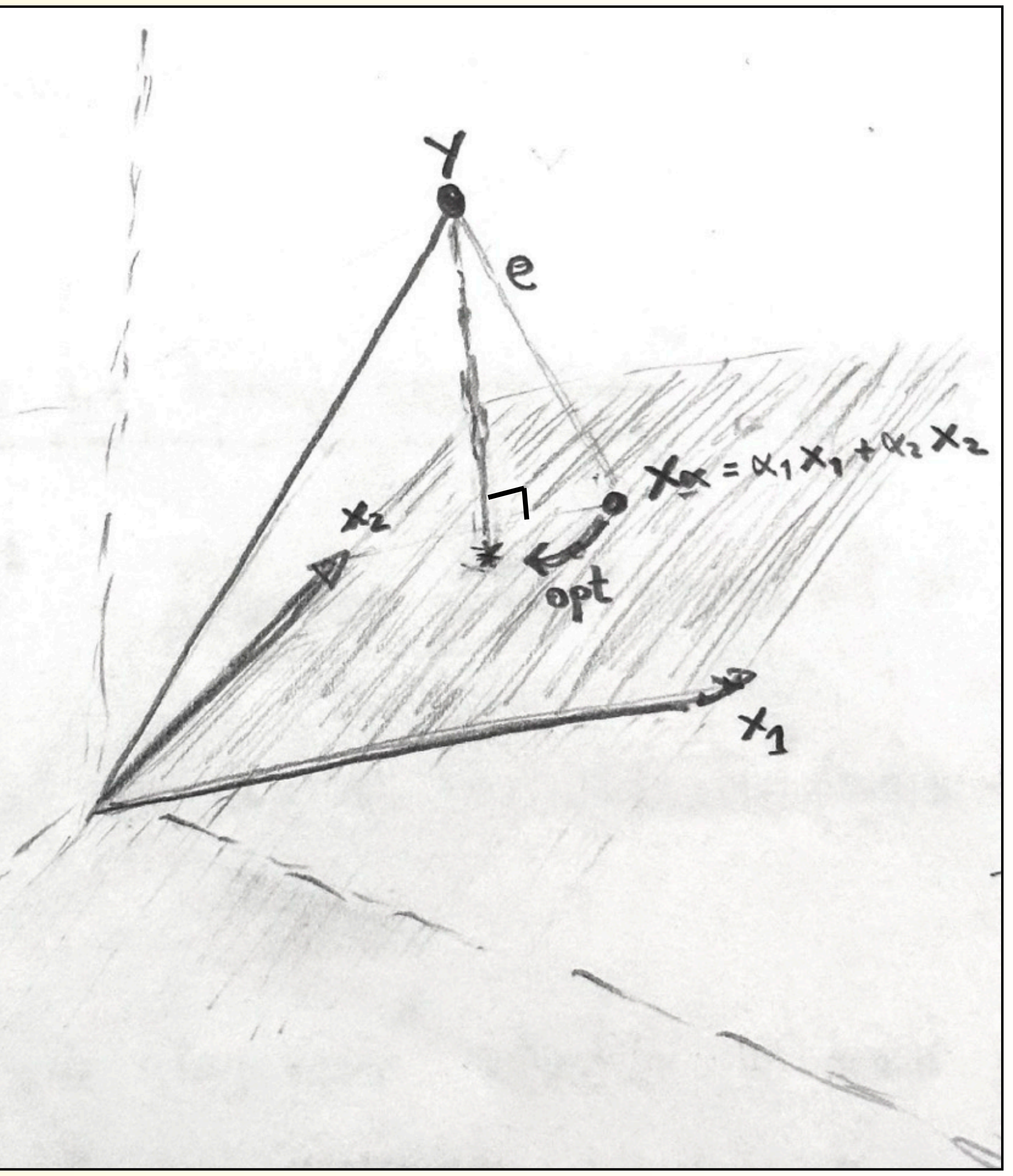
Summary

Linear regression: $\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$

Solution is given by **projecting** y onto the data

$$\alpha^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

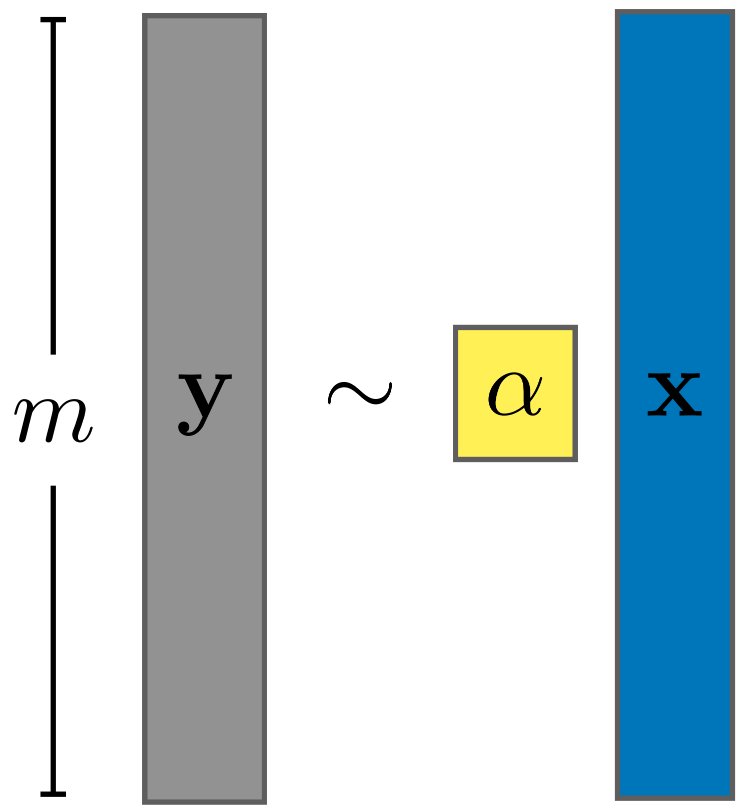
Solution (m=3, n=2)



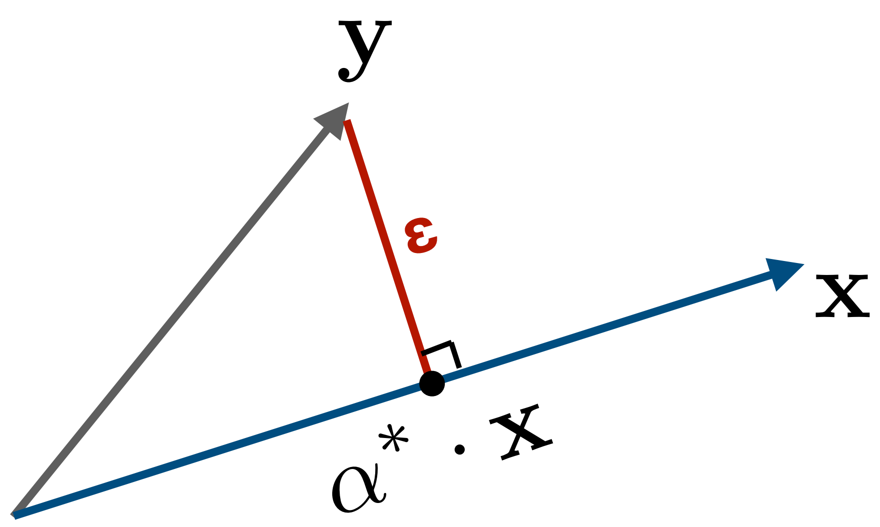
The geometry of linear regression

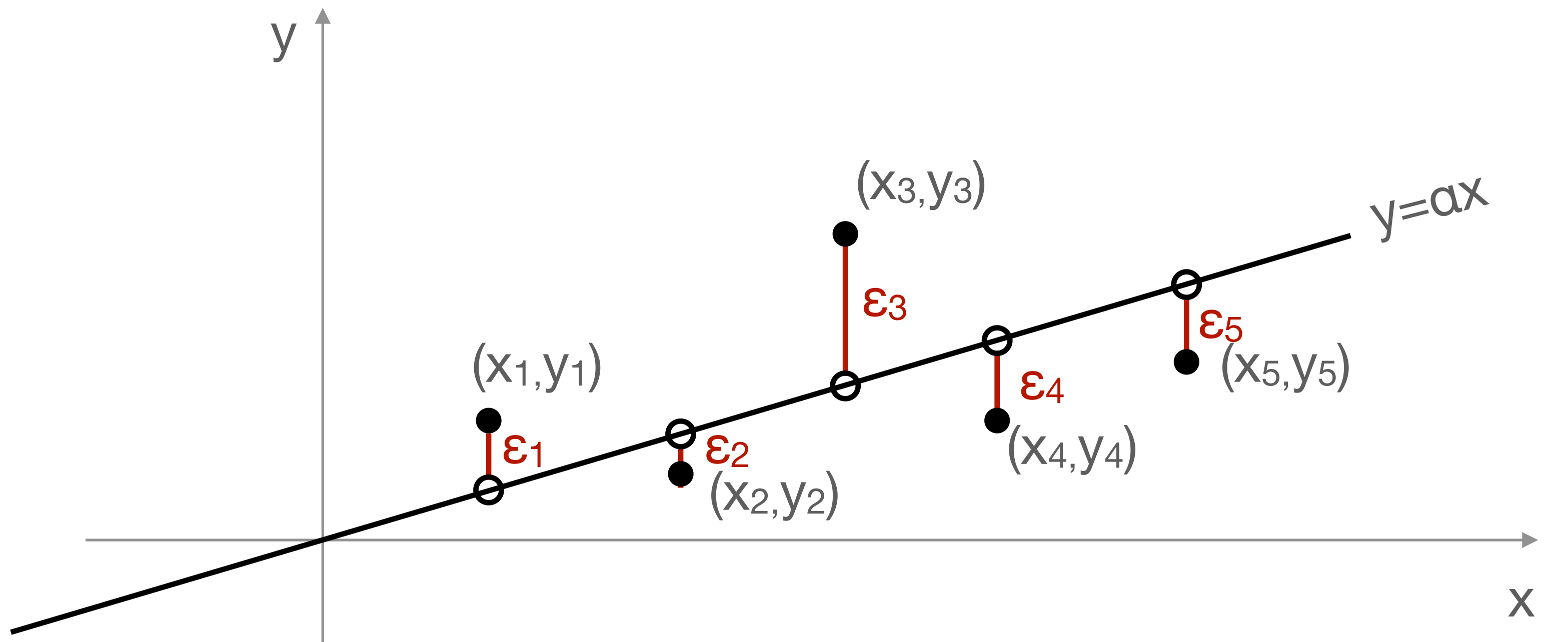
Where is the "regression line" ?

- n=1, general m



"2 vectors in **m dimensions**"





$\boldsymbol{\varepsilon}^2 = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_m^2$

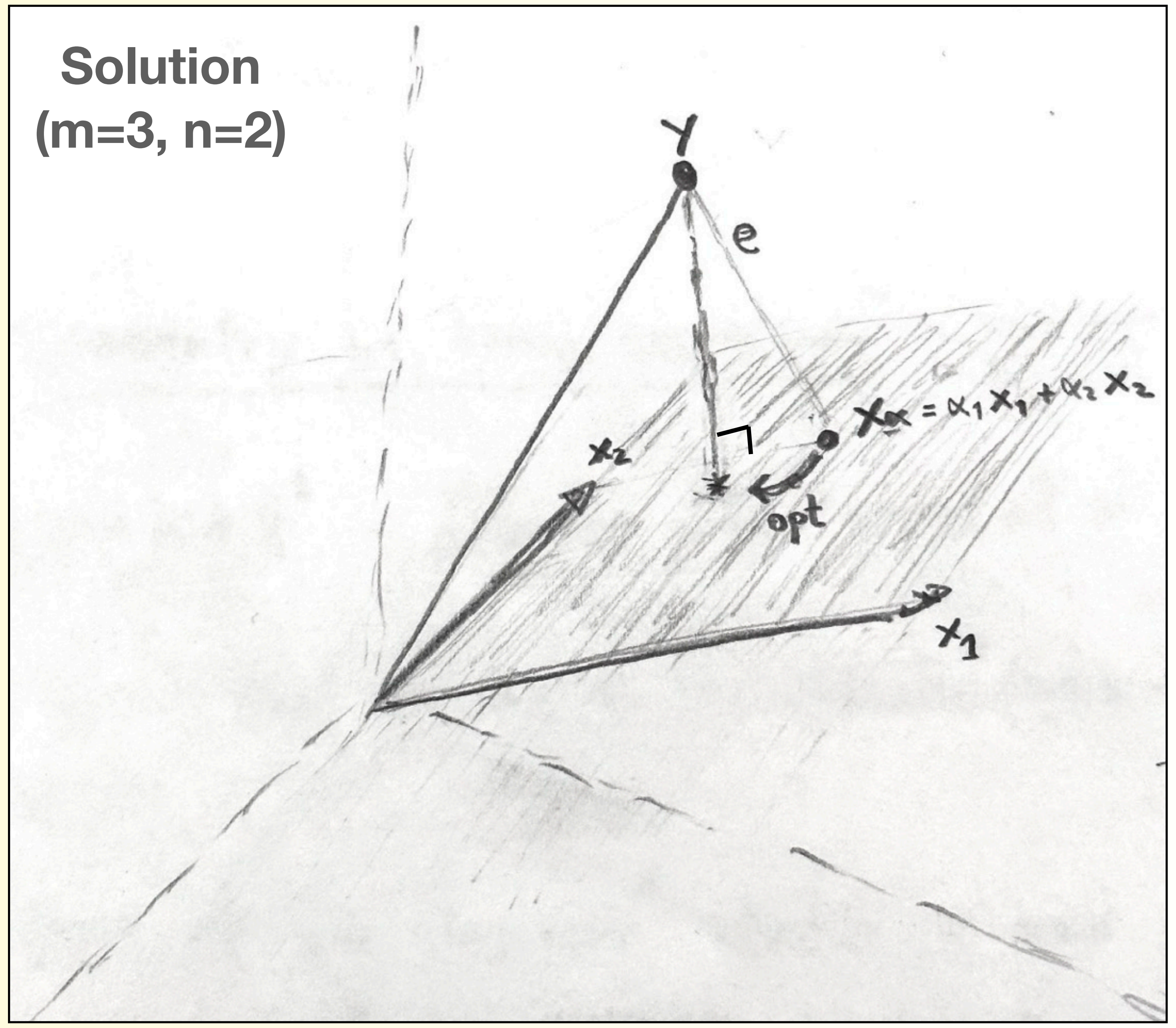
"**m vectors** in 2 dimensions"

Summary

Linear regression: $\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$

Solution is given by **projecting** y onto the data

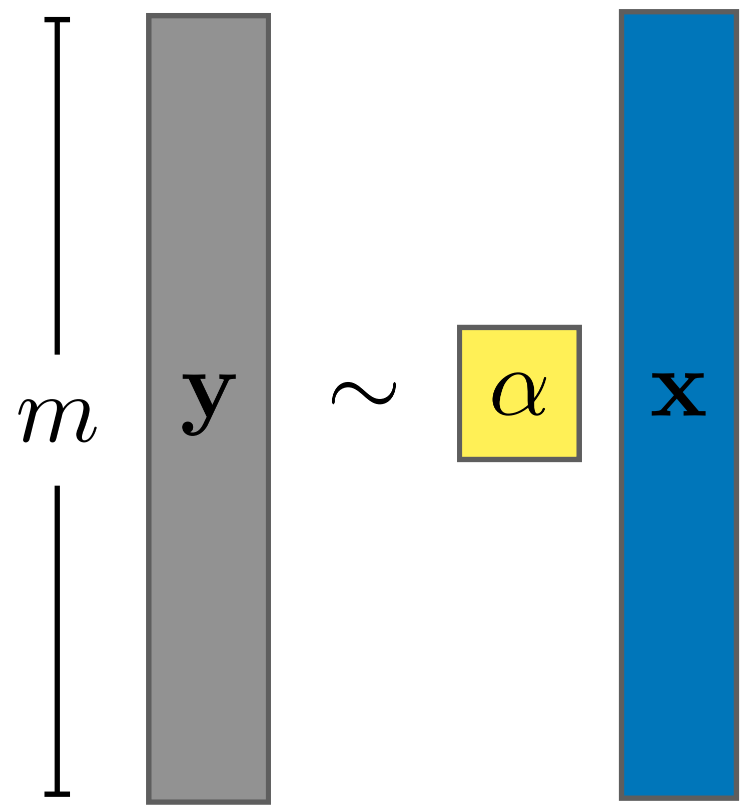
$$\alpha^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$



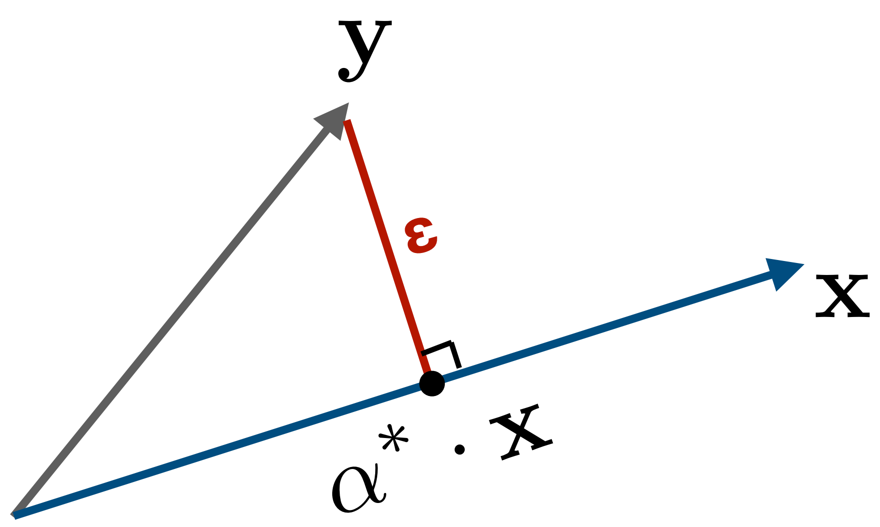
The geometry of linear regression

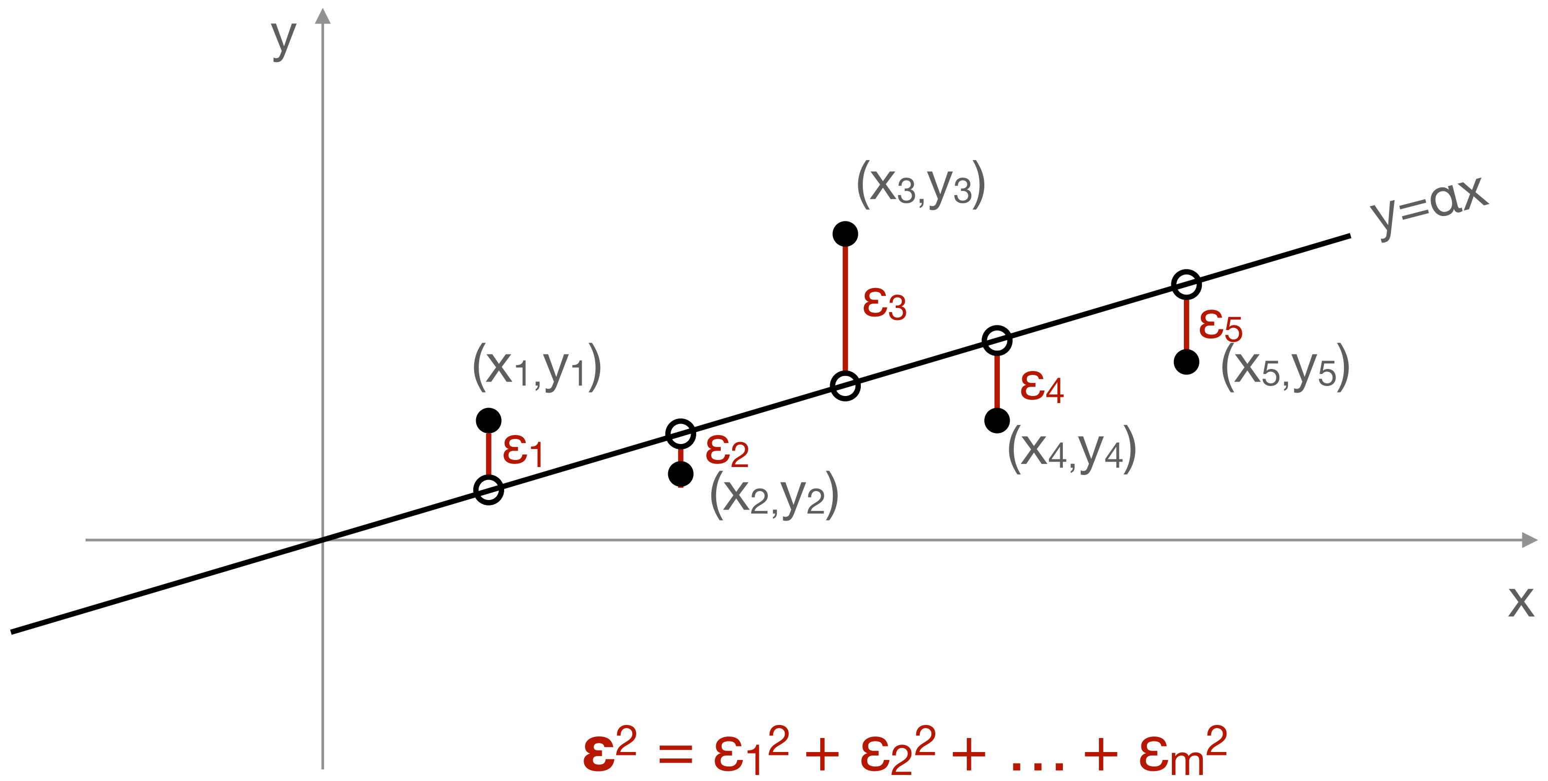
Where is the “regression line” ?

- n=1, general m



“2 vectors in **m dimensions**”





$\varepsilon^2 = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_m^2$

“**m vectors** in 2 dimensions”

- Note that solution minimizes the **vertical** (squared) distances from points to the fitted line.
— the points on the line are of the form $(x_i, \alpha x_i)$ and $\varepsilon_i = |x_i - \alpha x_i|$

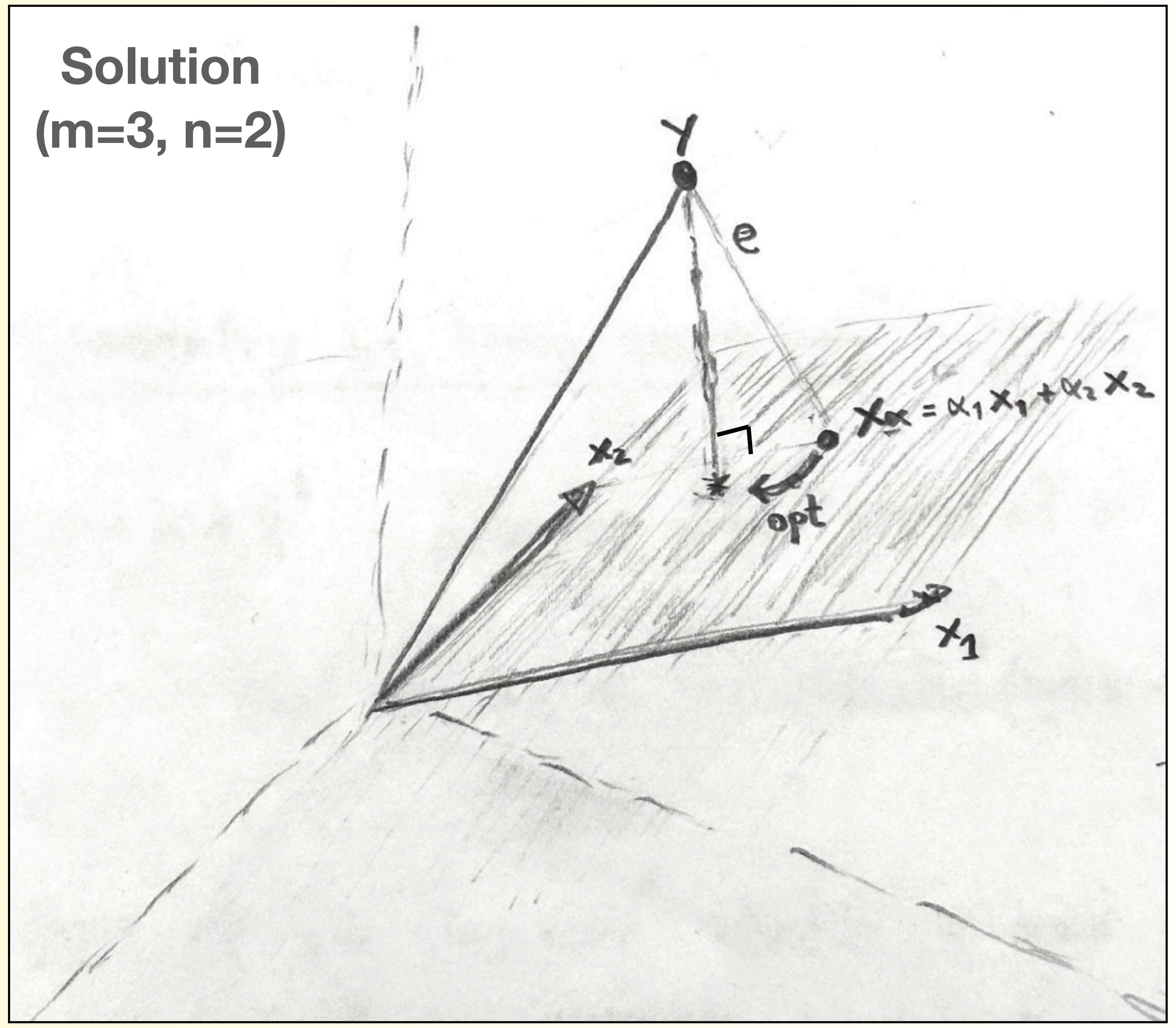
“Complete trust in x-coordinates”

Summary

Linear regression: $\min_{\alpha} ||\mathbf{y} - \mathbf{X}\alpha||^2$

Solution is given by **projecting** y onto the data

$$\alpha^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$



The geometry of linear regression

“Asymmetry in linear regression”

- $\min_{\alpha} ||\mathbf{y} - \alpha\mathbf{x}||^2$ –> Describe y in terms of x (y~x)
- Note that we can do the opposite $\min_{\alpha} ||\mathbf{x} - \alpha\mathbf{y}||^2$ to describe x~y

The geometry of linear regression

"Asymmetry in linear regression"

- $\min_{\alpha} ||\mathbf{y} - \alpha\mathbf{x}||^2$ —> Describe y in terms of x (y~x)
- Note that we can do the opposite $\min_{\alpha} ||\mathbf{x} - \alpha\mathbf{y}||^2$ to describe x~y

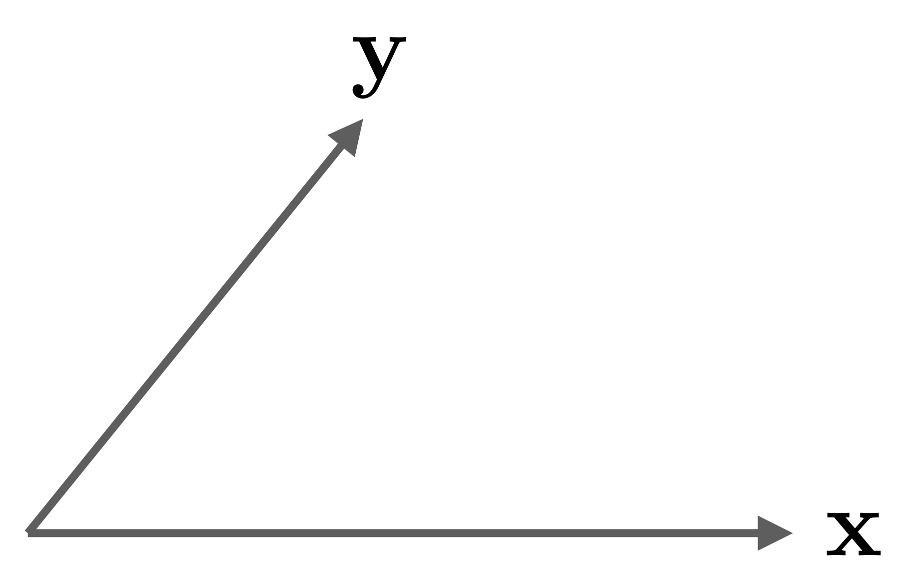
Ideas for a "symmetric" version?

The geometry of linear regression

“Asymmetry in linear regression”

- $\min_{\alpha} ||\mathbf{y} - \alpha\mathbf{x}||^2$ —> Describe y in terms of x (y~x)
- Note that we can do the opposite $\min_{\alpha} ||\mathbf{x} - \alpha\mathbf{y}||^2$ to describe x~y

What if we “describe” **both** x and y in terms of a **common** z?

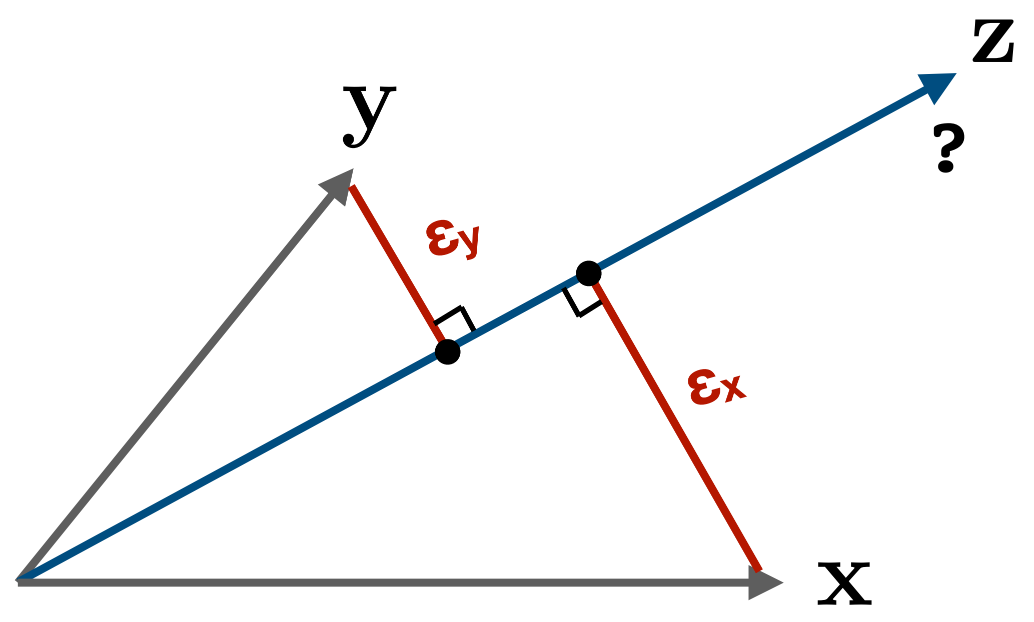


The geometry of linear regression

“Asymmetry in linear regression”

- $\min_{\alpha} ||\mathbf{y} - \alpha\mathbf{x}||^2$ —> Describe y in terms of x (y~x)
- Note that we can do the opposite $\min_{\alpha} ||\mathbf{x} - \alpha\mathbf{y}||^2$ to describe x~y

What if we “describe” **both** x and y in terms of a **common** z?

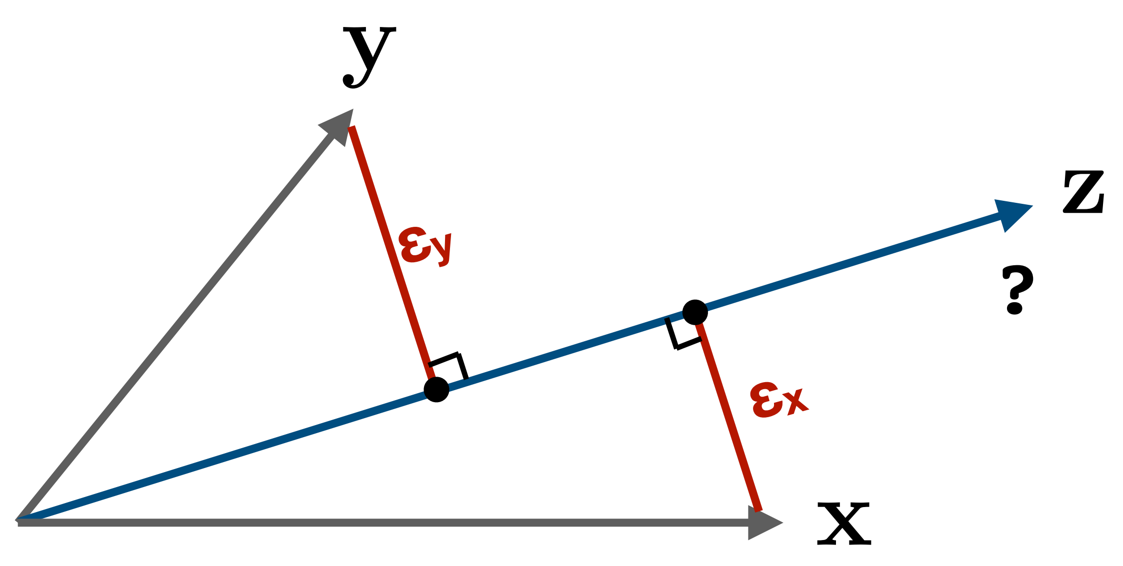


The geometry of linear regression

“Asymmetry in linear regression”

- $\min_{\alpha} ||\mathbf{y} - \alpha \mathbf{x}||^2$ $-->$ Describe y in terms of x (y~x)
- Note that we can do the opposite $\min_{\alpha} ||\mathbf{x} - \alpha \mathbf{y}||^2$ to describe x~y

What if we “describe” **both** x and y in terms of a **common** z?

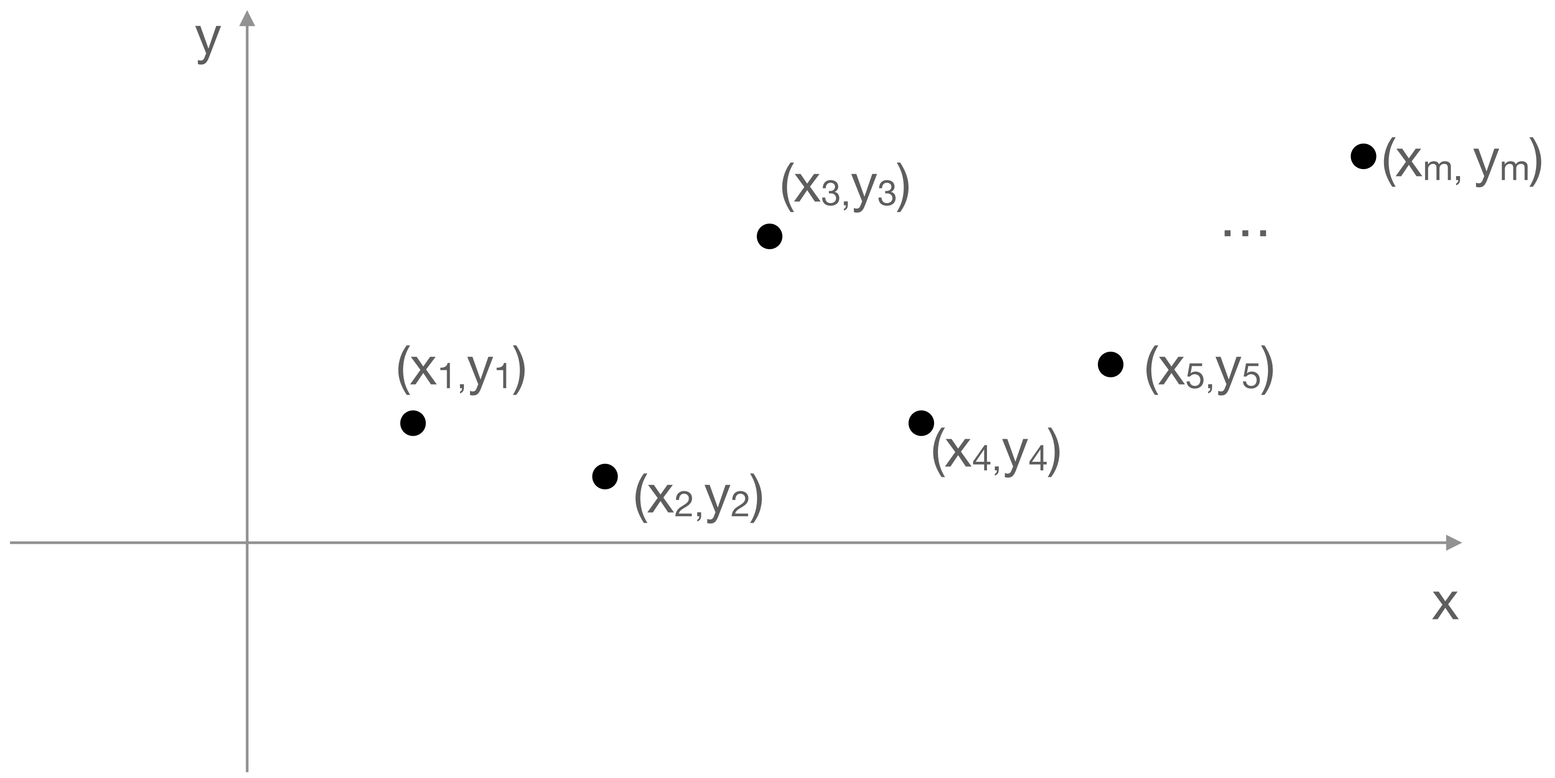


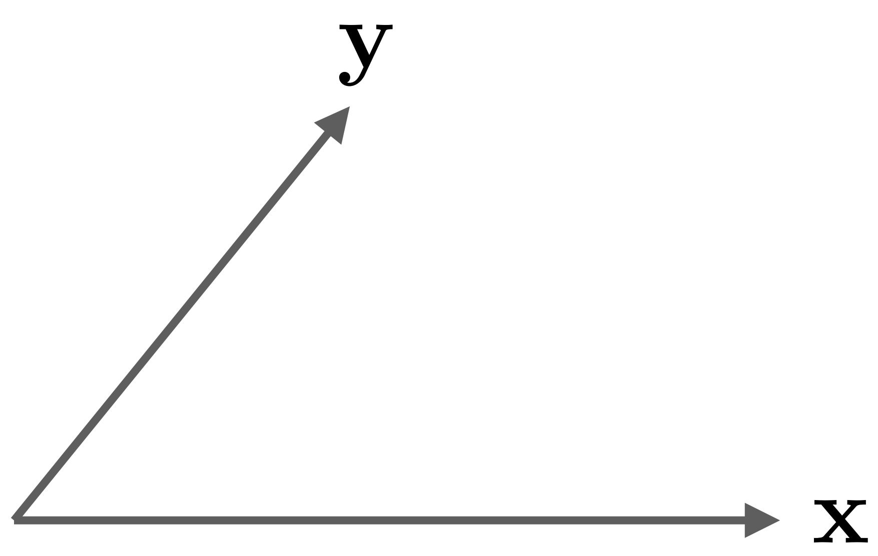
Towards Principal Component Analysis

Find z that best describes x and y

- Points in two dimensions as in previous example

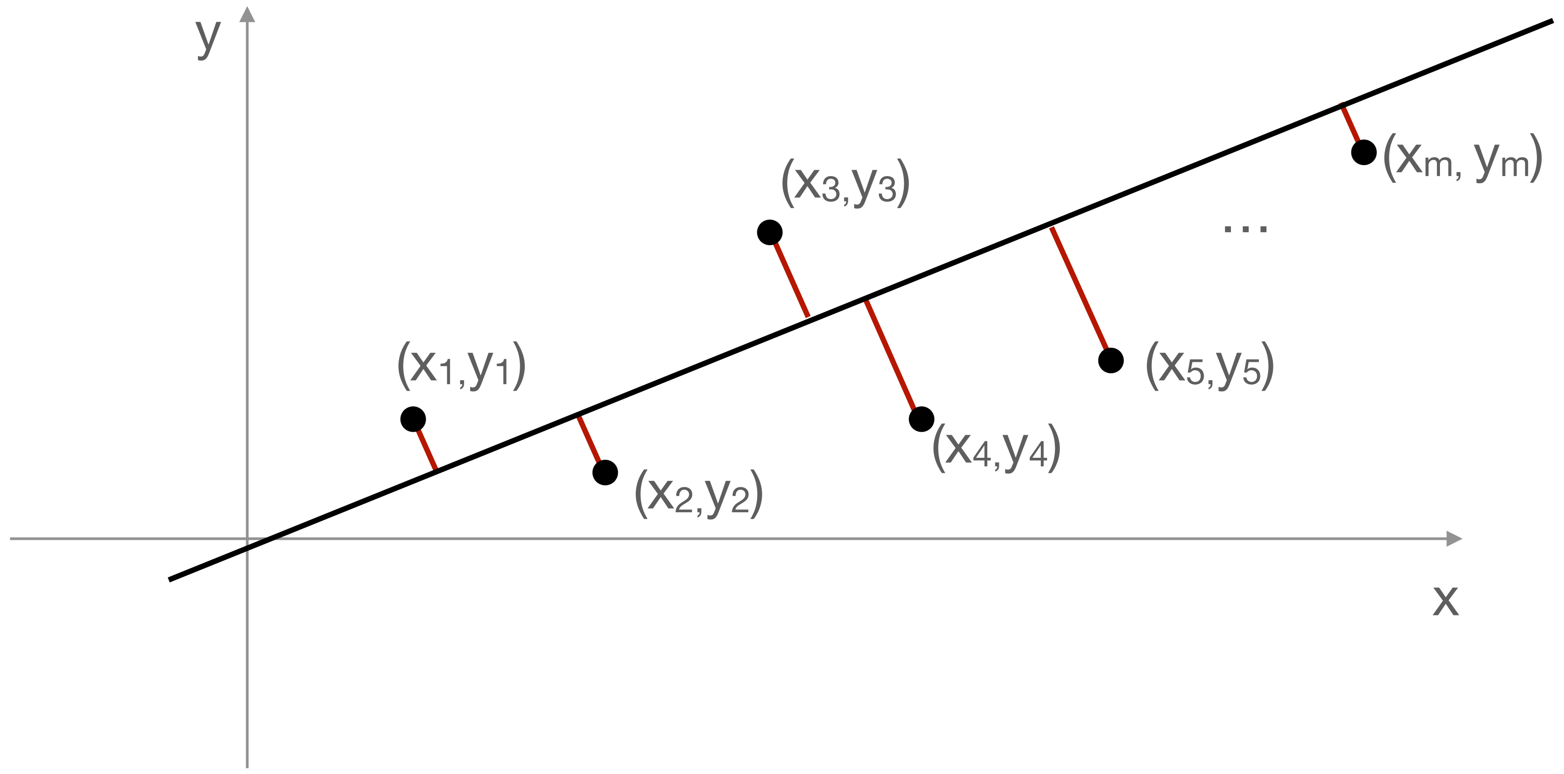
$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$

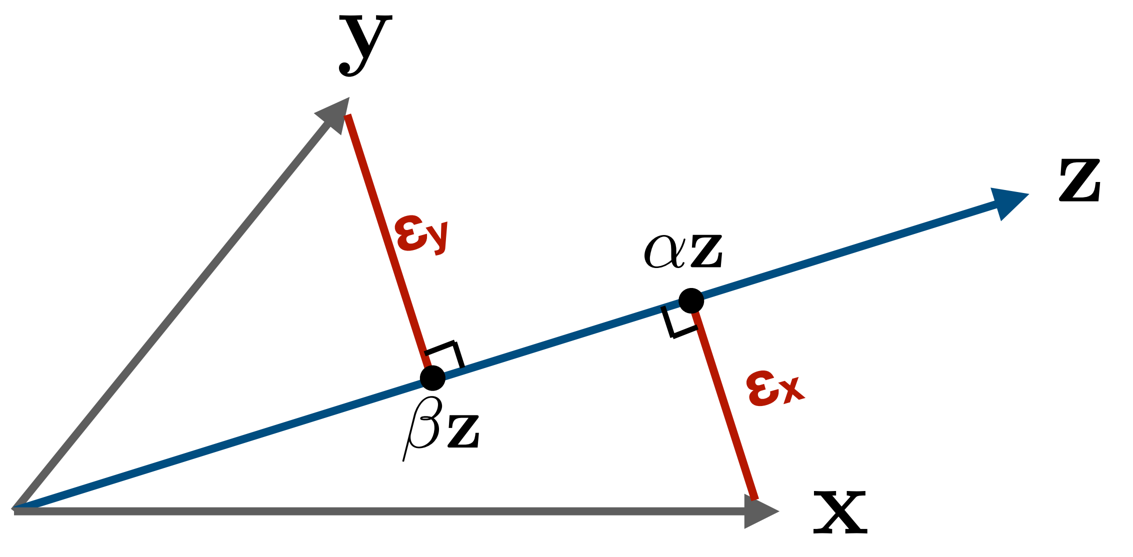




Towards Principal Component Analysis

Find z that best describes x and y





- Points in two dimensions as in previous example

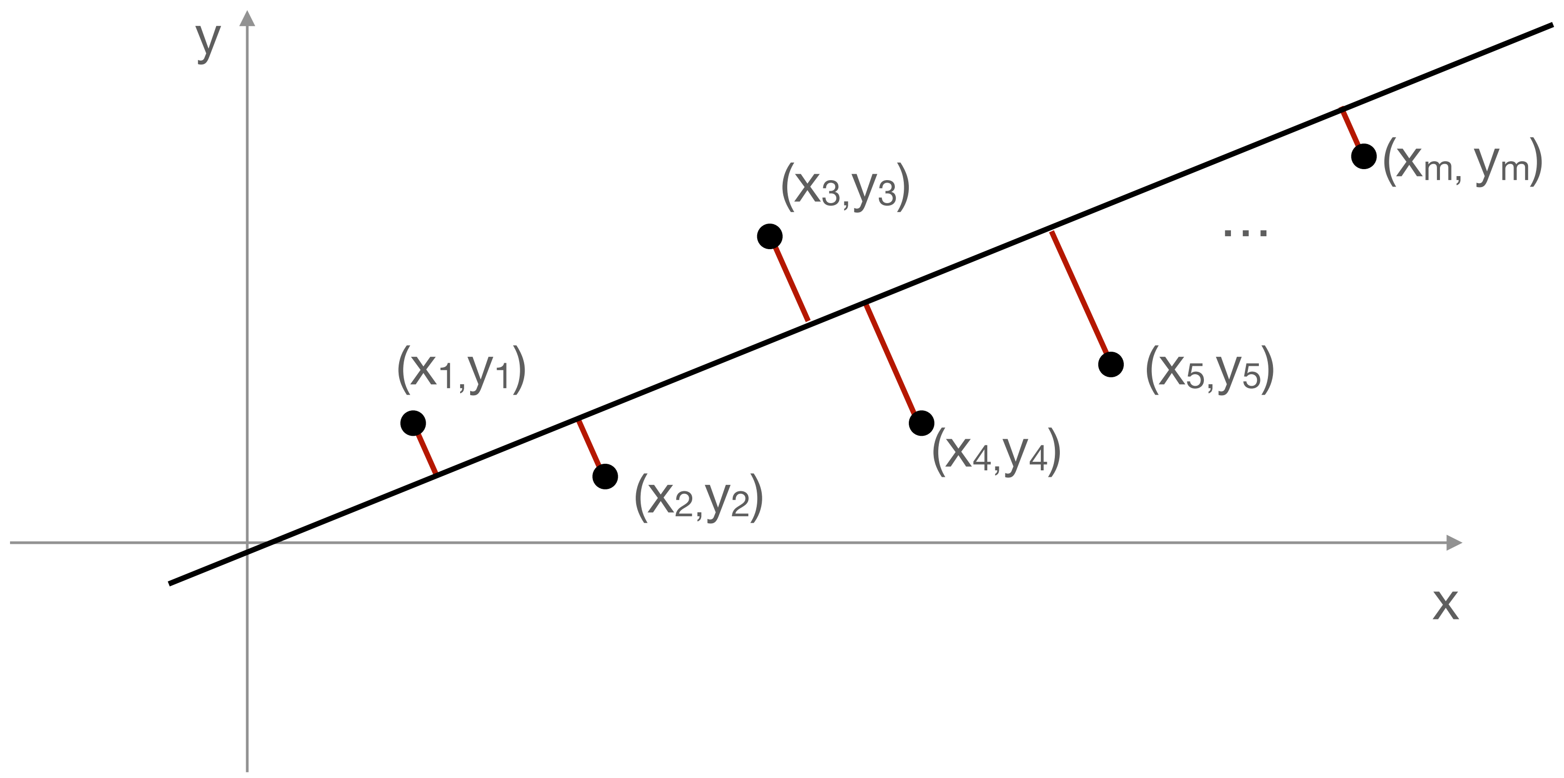
$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$

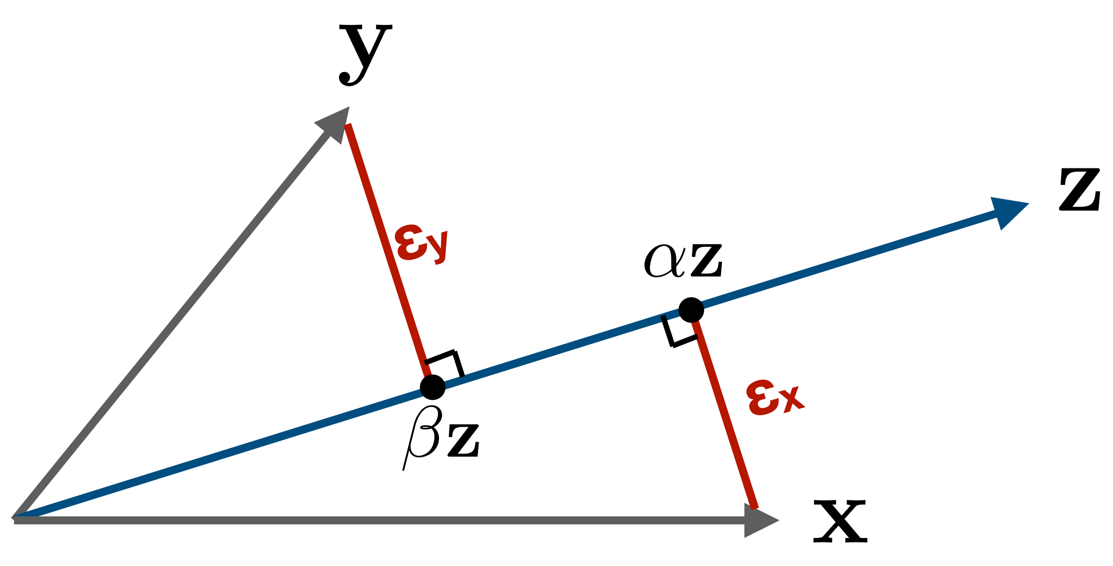
- Goal: find z that minimizes **both** error terms $\varepsilon_x^2+\varepsilon_y^2$

$$\epsilon_x^2 = ||\mathbf{x} - \alpha\mathbf{z}||^2$$
$$\epsilon_y^2 = ||\mathbf{y} - \beta\mathbf{z}||^2$$

Towards Principal Component Analysis

Find z that best describes x and y





- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$

- Goal: find z that minimizes **both** error terms $\varepsilon_x^2 + \varepsilon_y^2$

$$\epsilon_x^2 = ||\mathbf{x} - \alpha\mathbf{z}||^2$$
$$\epsilon_y^2 = ||\mathbf{y} - \beta\mathbf{z}||^2$$

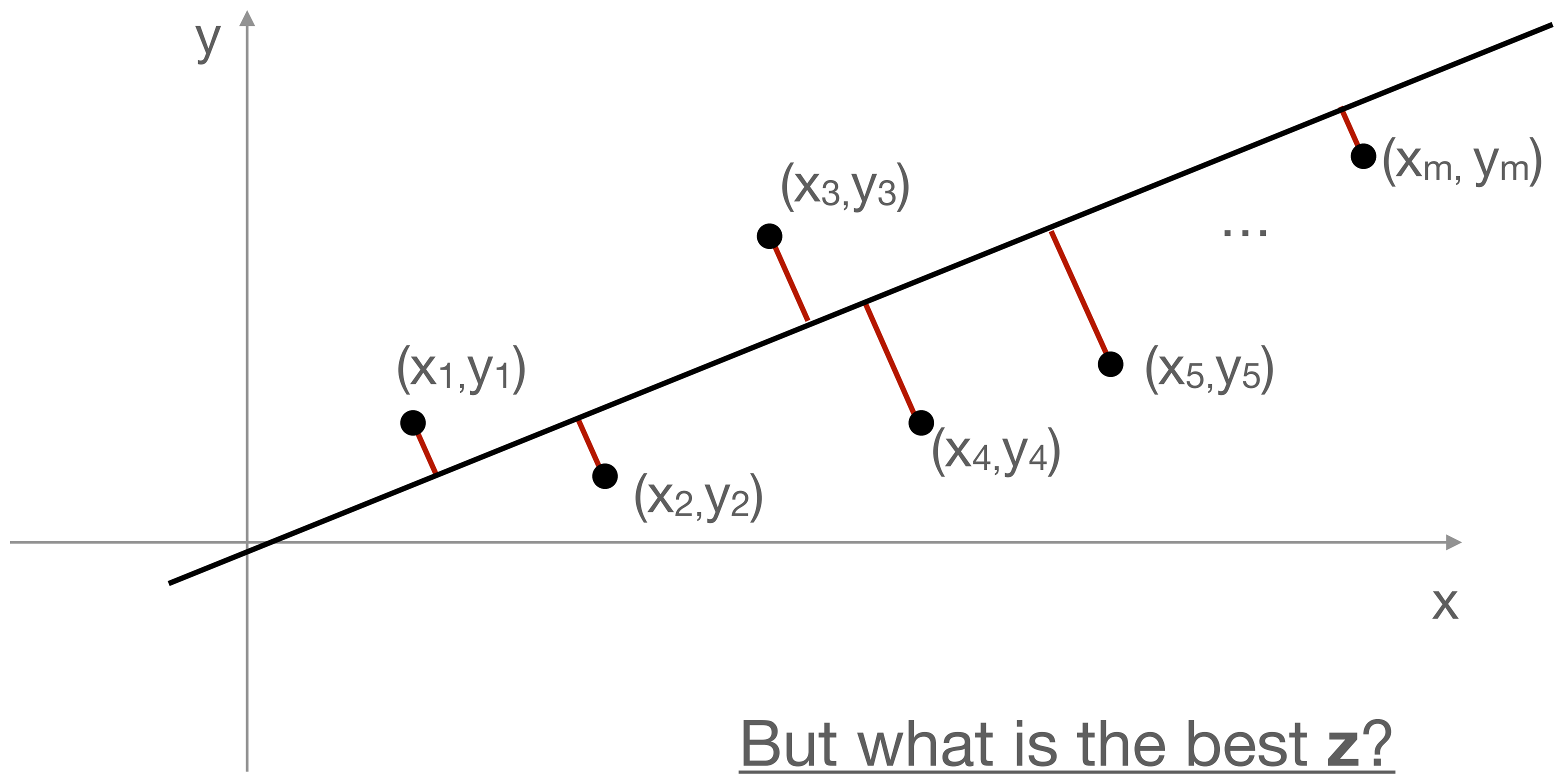
- For a **given z**, we know how to solve for α* and β*

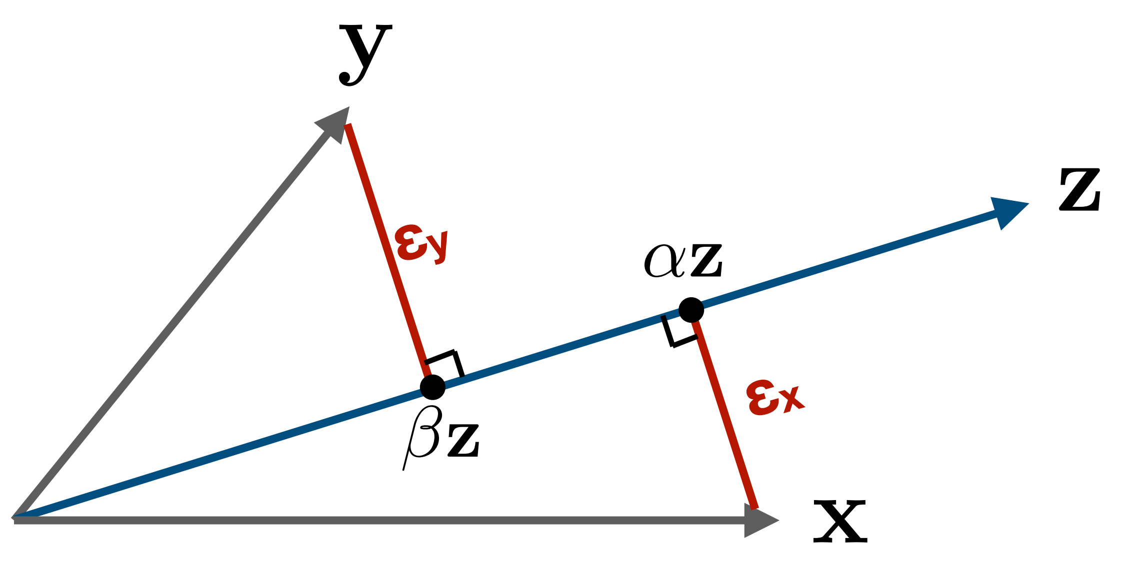
$$\alpha^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{x} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}$$

$$\beta^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{y} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}$$

Towards Principal Component Analysis

Find z that best describes x and y





But what is the best **z**?

$$\min_{\mathbf{z}\in\mathbb{R}^m} \left|\left|\mathbf{x} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2 + \left|\left|\mathbf{y} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2$$

- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$

- Goal: find z that minimizes **both** error terms $\varepsilon_x^2+\varepsilon_y^2$

$$\epsilon_x^2 = ||\mathbf{x} - \alpha\mathbf{z}||^2$$
$$\epsilon_y^2 = ||\mathbf{y} - \beta\mathbf{z}||^2$$

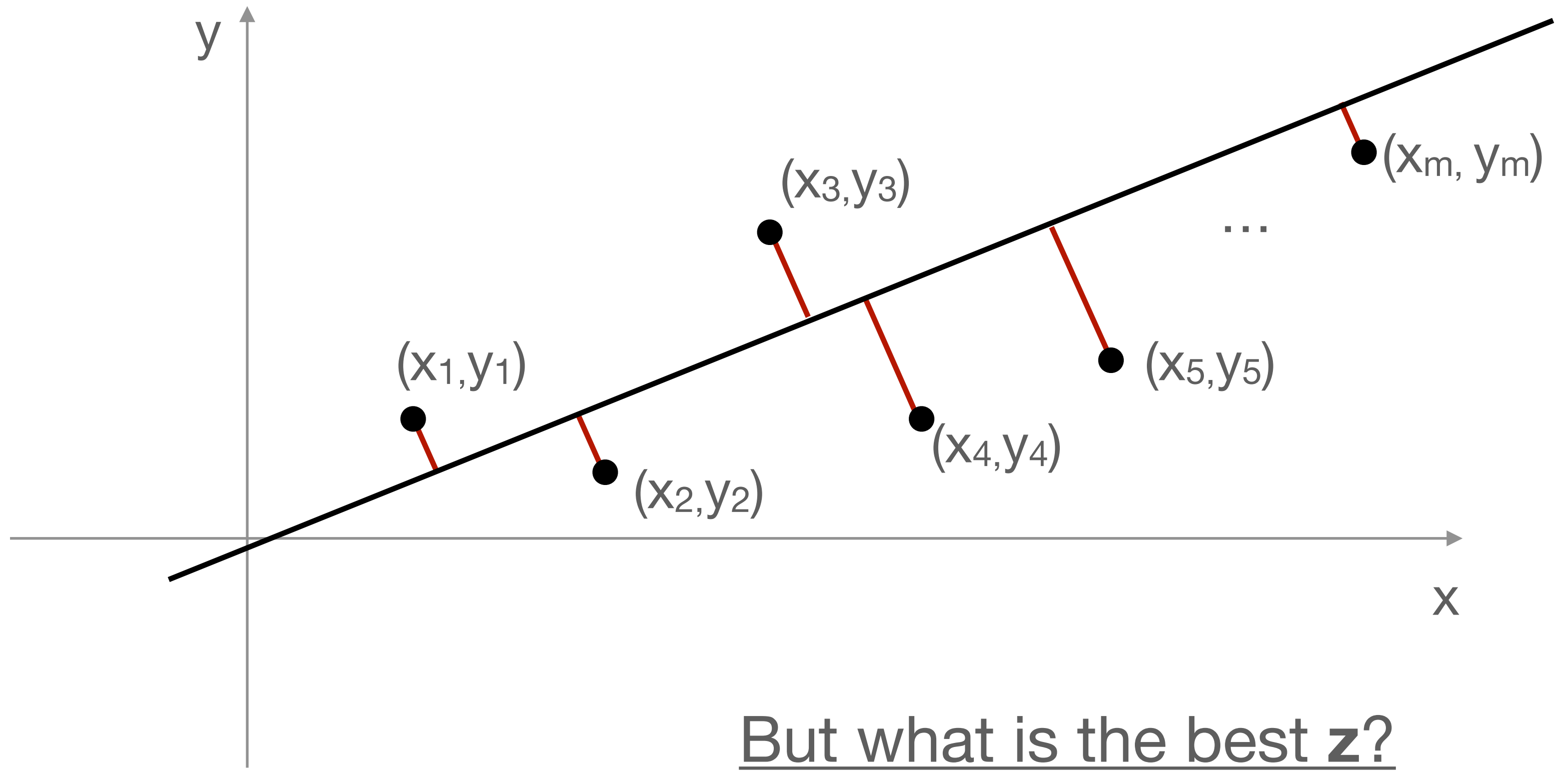
- For a **given z**, we know how to solve for α* and β*

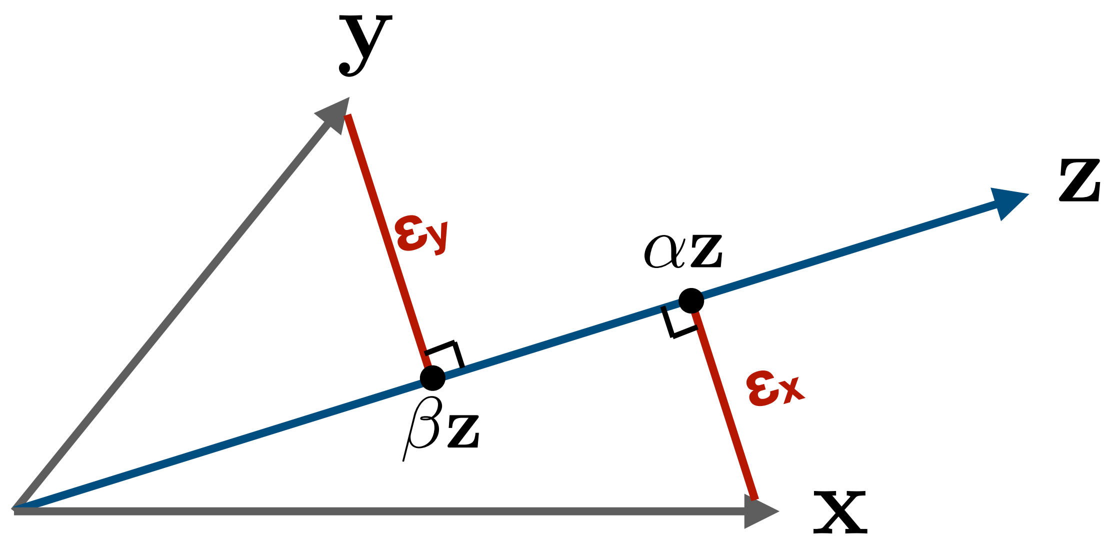
$$\alpha^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{x} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}$$

$$\beta^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{y} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}$$

Towards Principal Component Analysis

Find z that best describes x and y





- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$

- Goal: find z that minimizes **both** error terms $\varepsilon_x^2+\varepsilon_y^2$

$$\epsilon_x^2 = ||\mathbf{x} - \alpha\mathbf{z}||^2$$
$$\epsilon_y^2 = ||\mathbf{y} - \beta\mathbf{z}||^2$$

- For a **given z**, we know how to solve for α* and β*

$$\alpha^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{x} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}$$

$$\beta^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{y} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}$$

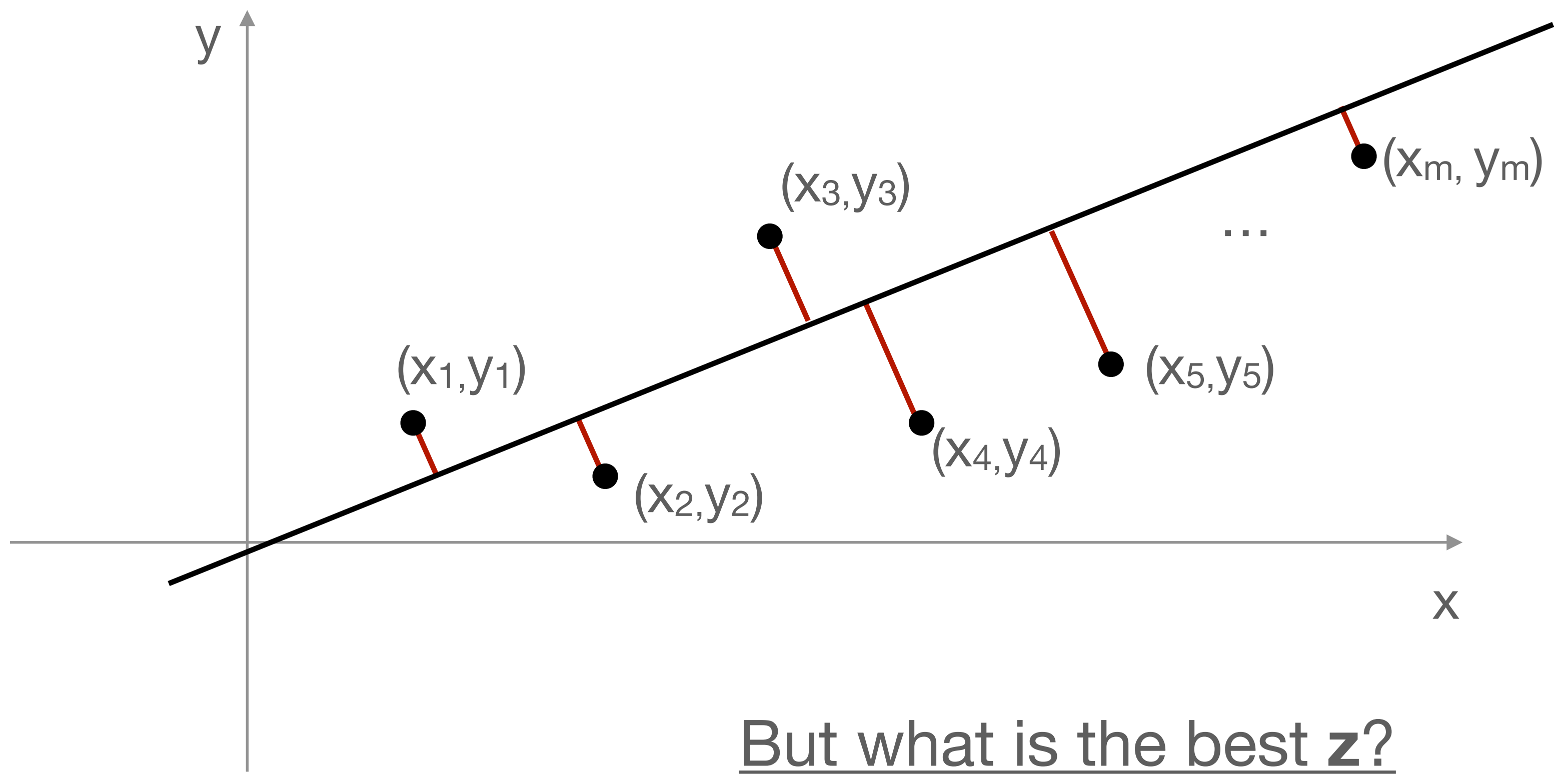
But what is the best **z**?

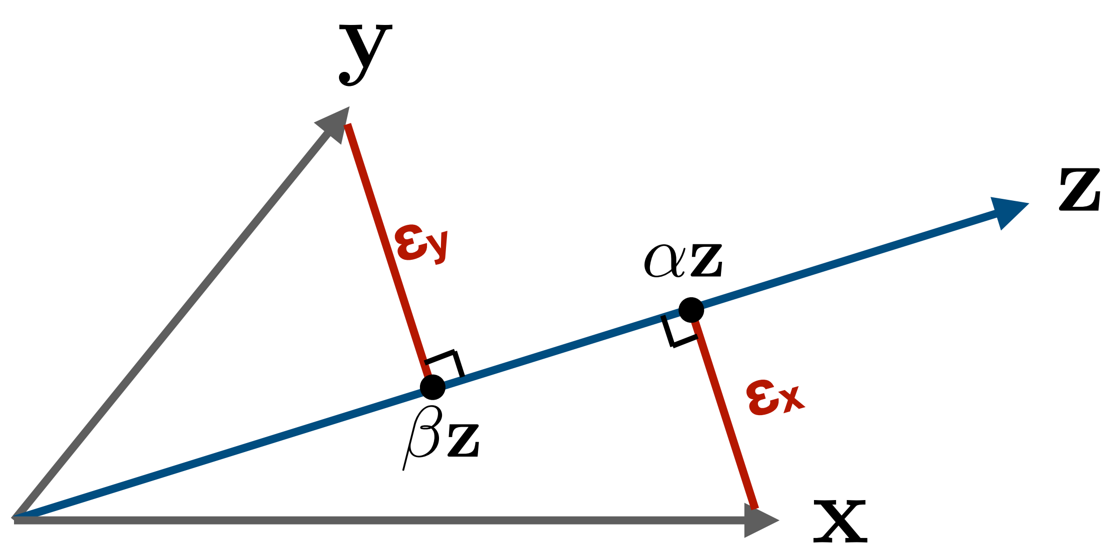
$$\min_{\mathbf{z}\in\mathbb{R}^m} \left|\left|\mathbf{x} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2 + \left|\left|\mathbf{y} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2$$ (the norm of z doesn't affect the error terms)

$$= \min_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{x} - \mathbf{z}^{\mathrm{T}}\mathbf{x}\cdot\mathbf{z}||^2 + ||\mathbf{x} - \mathbf{z}^{\mathrm{T}}\mathbf{y}\cdot\mathbf{z}||^2$$ (we may as well set it to be 1)

Towards Principal Component Analysis

Find z that best describes x and y





- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$

- Goal: find z that minimizes **both** error terms $\varepsilon_x^2+\varepsilon_y^2$

$$\epsilon_x^2 = ||\mathbf{x} - \alpha \mathbf{z}||^2$$
$$\epsilon_y^2 = ||\mathbf{y} - \beta \mathbf{z}||^2$$

- For a **given z**, we know how to solve for α* and β*

$$\alpha^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{x} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}$$

$$\beta^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{y} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}$$

But what is the best **z**?

$$\min_{\mathbf{z}\in\mathbb{R}^m} \left|\left|\mathbf{x} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2 + \left|\left|\mathbf{y} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2$$ (the norm of z doesn't affect the error terms)

$$= \min_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{x} - \mathbf{z}^{\mathrm{T}}\mathbf{x}\cdot\mathbf{z}||^2 + ||\mathbf{x} - \mathbf{z}^{\mathrm{T}}\mathbf{y}\cdot\mathbf{z}||^2$$ (we may as well set it to be 1)

$$= \min_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{x}||^2 - 2\mathbf{z}^{\mathrm{T}}\mathbf{x}\cdot\mathbf{z}^{\mathrm{T}}\mathbf{x} + ||\mathbf{z}^{\mathrm{T}}\mathbf{x}\cdot\mathbf{z}||^2$$
$$+||\mathbf{y}||^2 - 2\mathbf{z}^{\mathrm{T}}\mathbf{y}\cdot\mathbf{z}^{\mathrm{T}}\mathbf{y} + ||\mathbf{z}^{\mathrm{T}}\mathbf{y}\cdot\mathbf{z}||^2$$ (expand the squares)

$$= \min_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{x}||^2 - (\mathbf{z}^{\mathrm{T}}\mathbf{x})^2 + ||\mathbf{y}||^2 - (\mathbf{z}^{\mathrm{T}}\mathbf{y})^2$$ (simplify)

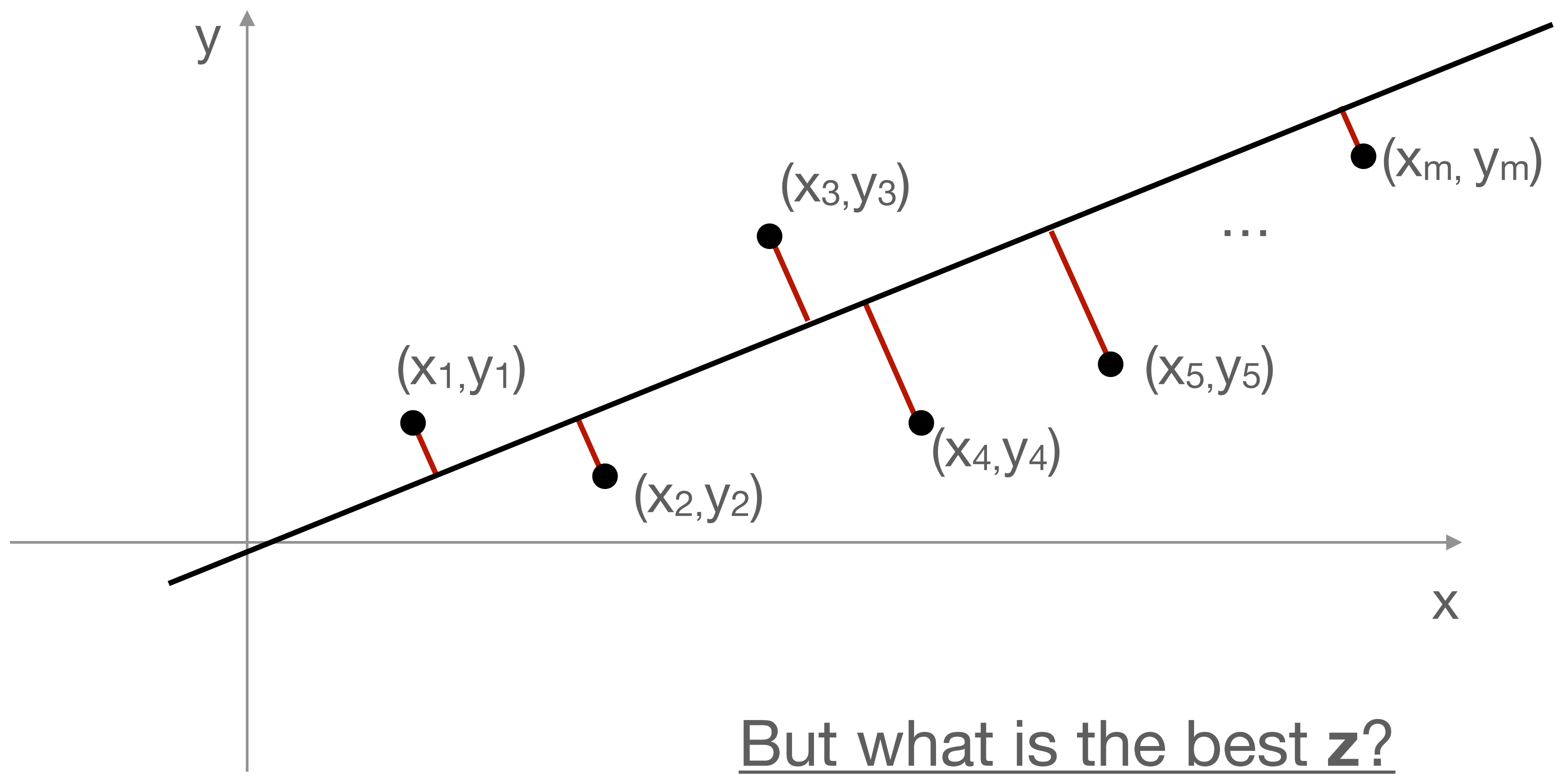
$$= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - \max_{\mathbf{z}:||\mathbf{z}||=1} (\mathbf{z}^{\mathrm{T}}\mathbf{x})^2 + (\mathbf{z}^{\mathrm{T}}\mathbf{y})^2$$ (x and y terms do not depend on z)

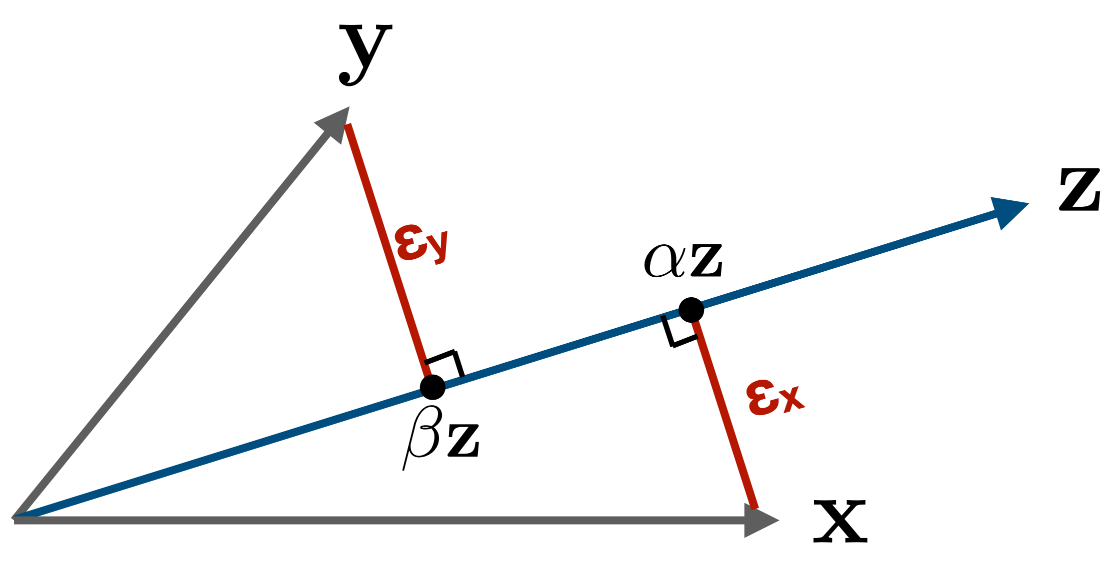
Towards Principal Component Analysis

Find z that best describes x and y

- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$





But what is the best **z**?

- Goal: find z that minimizes **both** error terms $\varepsilon_x^2 + \varepsilon_y^2$

$$\epsilon_x^2 = ||\mathbf{x} - \alpha \mathbf{z}||^2$$
$$\epsilon_y^2 = ||\mathbf{y} - \beta \mathbf{z}||^2$$

- For a **given z**, we know how to solve for α* and β*

$$\alpha^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{x} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}$$

$$\beta^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{y} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}$$

$$\min_{\mathbf{z}\in\mathbb{R}^m} \left|\left|\mathbf{x} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2 + \left|\left|\mathbf{y} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2 \quad \text{(the norm of z doesn't affect the error terms)}$$

$$= \min_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{x} - \mathbf{z}^{\mathrm{T}}\mathbf{x}\cdot\mathbf{z}||^2 + ||\mathbf{x} - \mathbf{z}^{\mathrm{T}}\mathbf{y}\cdot\mathbf{z}||^2 \quad \text{(we may as well set it to be 1)}$$

$$= \min_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{x}||^2 - 2\mathbf{z}^{\mathrm{T}}\mathbf{x}\cdot\mathbf{z}^{\mathrm{T}}\mathbf{x} + ||\mathbf{z}^{\mathrm{T}}\mathbf{x}\cdot\mathbf{z}||^2$$
$$+||\mathbf{y}||^2 - 2\mathbf{z}^{\mathrm{T}}\mathbf{y}\cdot\mathbf{z}^{\mathrm{T}}\mathbf{y} + ||\mathbf{z}^{\mathrm{T}}\mathbf{y}\cdot\mathbf{z}||^2 \quad \text{(expand the squares)}$$

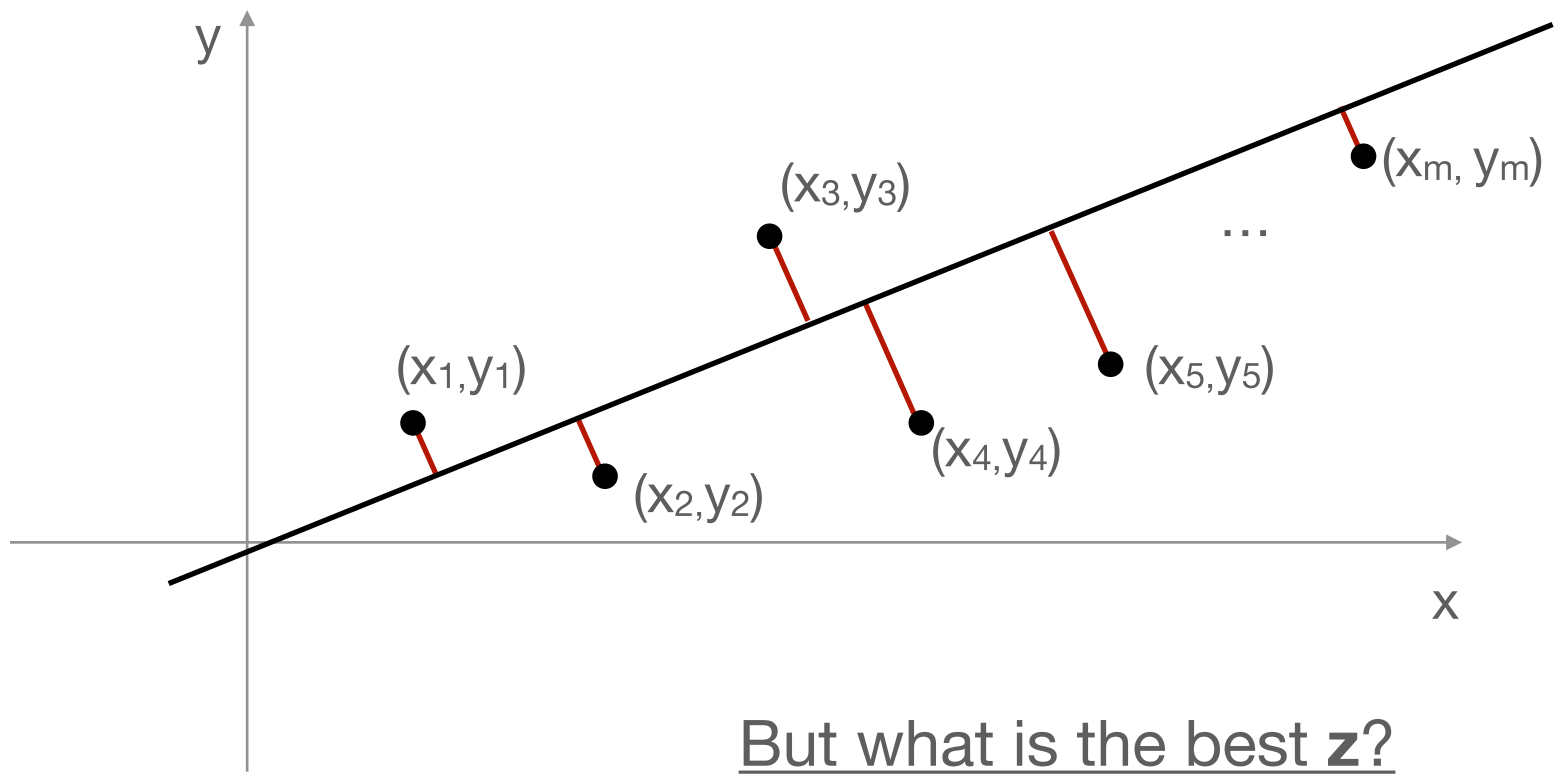
$$= \min_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{x}||^2 - (\mathbf{z}^{\mathrm{T}}\mathbf{x})^2 + ||\mathbf{y}||^2 - (\mathbf{z}^{\mathrm{T}}\mathbf{y})^2 \quad \text{(simplify)}$$

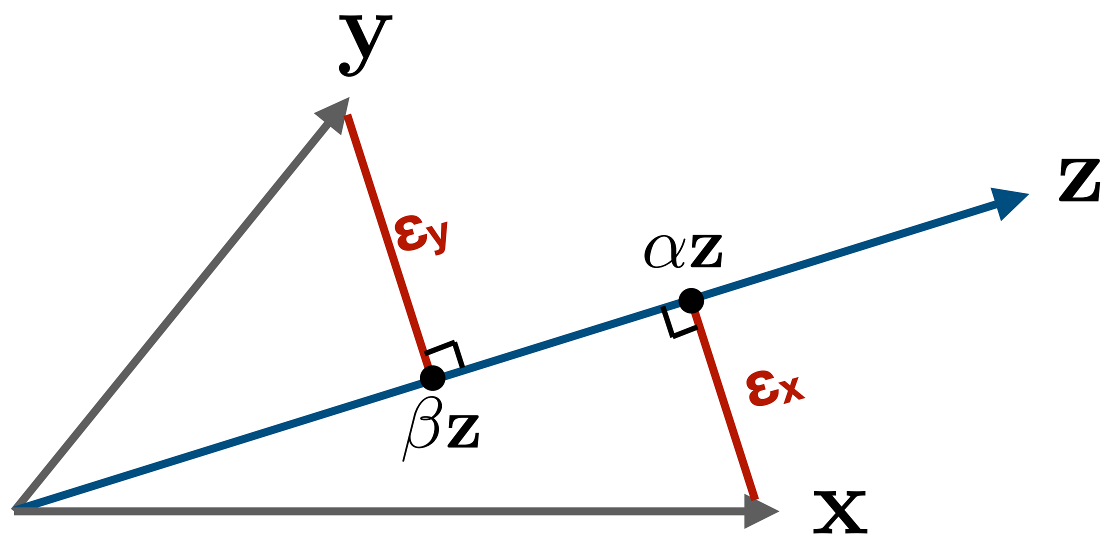
$$= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - \boxed{\max_{\mathbf{z}:||\mathbf{z}||=1} (\mathbf{z}^{\mathrm{T}}\mathbf{x})^2 + (\mathbf{z}^{\mathrm{T}}\mathbf{y})^2} \quad \text{(x and y terms do not depend on z)}$$

- The same z maximizes the magnitude of the projections!

Towards Principal Component Analysis

Find z that best describes x and y





- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$

But what is the best **z**?

- Goal: find z that minimizes **both** error terms $\varepsilon_x^2+\varepsilon_y^2$

$$\epsilon_x^2 = ||\mathbf{x} - \alpha\mathbf{z}||^2$$
$$\epsilon_y^2 = ||\mathbf{y} - \beta\mathbf{z}||^2$$

- For a **given z**, we know how to solve for α* and β*

$$\alpha^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{x} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}$$

$$\beta^* = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{y} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}$$

$$\min_{\mathbf{z}\in\mathbb{R}^m} \left|\left|\mathbf{x} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{x}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2 + \left|\left|\mathbf{y} - \frac{\mathbf{z}^{\mathrm{T}}\mathbf{y}}{||\mathbf{z}||^2}\cdot\mathbf{z}\right|\right|^2$$ (the norm of z doesn't affect the error terms)

$$= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - \boxed{\max_{\mathbf{z}:||\mathbf{z}||=1} (\mathbf{z}^{\mathrm{T}}\mathbf{x})^2 + (\mathbf{z}^{\mathrm{T}}\mathbf{y})^2}$$ (x and y terms do not depend on z)

- The same z maximizes the magnitude of the projections!

$$\max_{\mathbf{z}:||\mathbf{z}||=1} (\mathbf{z}^{\mathrm{T}}\mathbf{x})^2 + (\mathbf{z}^{\mathrm{T}}\mathbf{y})^2$$ (rearrange terms and simplify)

$$= \max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\underbrace{(\mathbf{x}\mathbf{x}^{\mathrm{T}} + \mathbf{y}\mathbf{y}^{\mathrm{T}})}_{\text{mxm symmetric matrix}}\mathbf{z}$$

$$= \max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z}\ ,\ \text{where } \mathbf{M} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{m\times 2}$$ (the DATA matrix!)

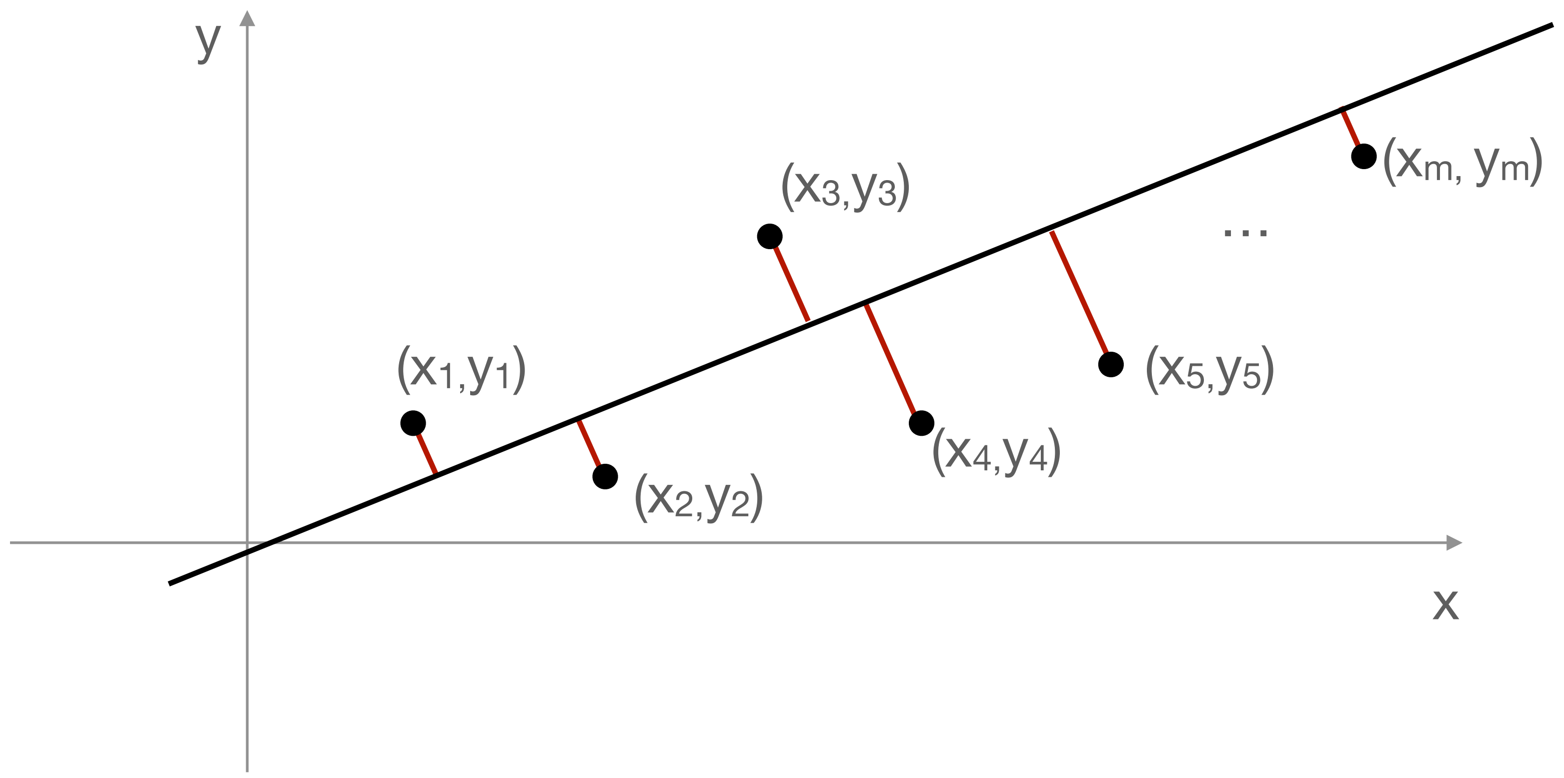
$$= \max_{\mathbf{z}:||\mathbf{z}||=1} ||\mathbf{M}^{\mathrm{T}}\mathbf{z}||^2$$

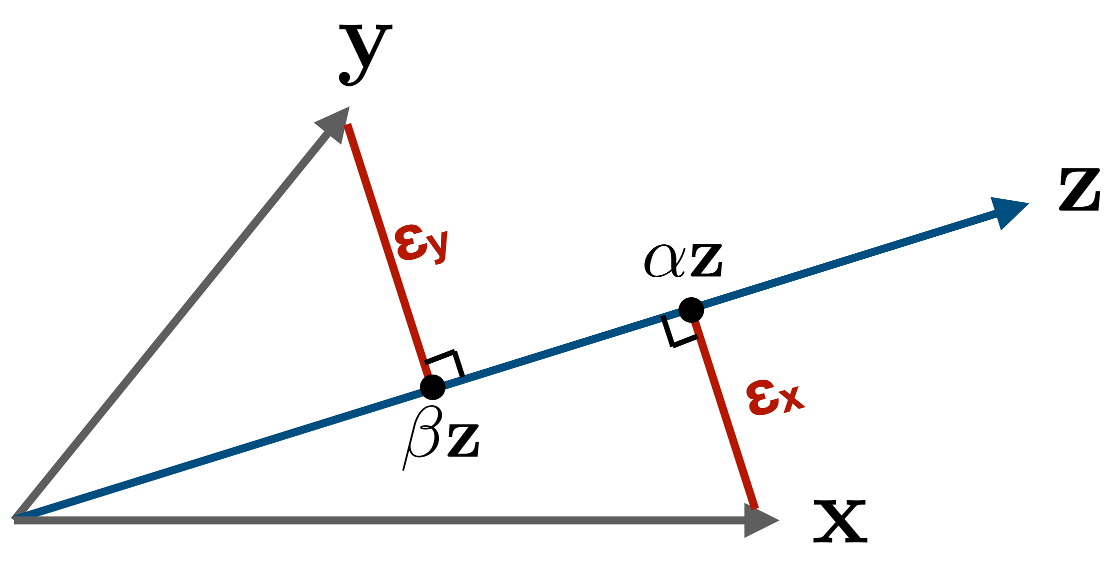
Towards Principal Component Analysis

Find z that best describes x and y

- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \dots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \dots, y_m]^{\mathrm{T}}$$





$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z}\ ,\ \text{where } \mathbf{M} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{m\times 2} \quad \text{(the DATA matrix)}$$

- Maximizing a *quadratic function* subject to constraints is notoriously hard (NP-hard problem)
- In our case above: very **efficient algorithm** exists (elegant solution mathematically):

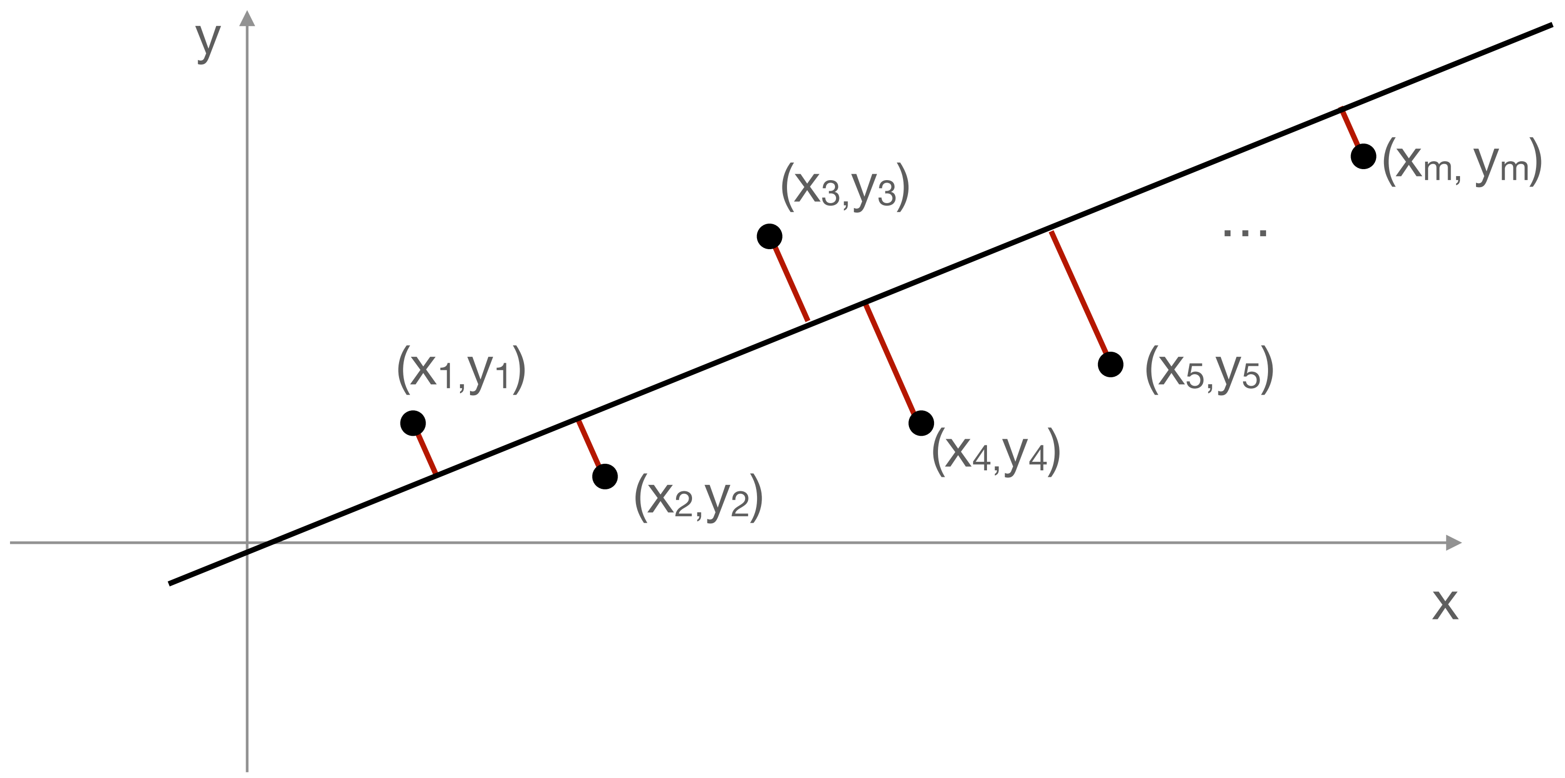
$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z} = \lambda_{max}(\mathbf{M}\mathbf{M}^{\mathrm{T}}),\ \mathbf{z}^* = \text{principal eigenvector of } \mathbf{M}\mathbf{M}^{\mathrm{T}}$$

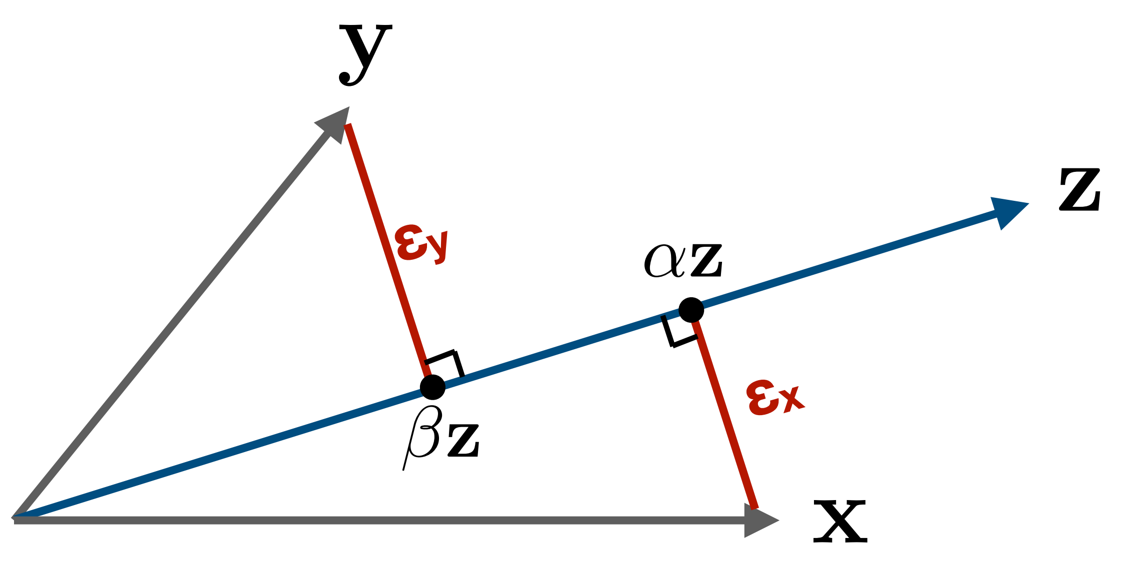
Towards Principal Component Analysis

Find z that best describes x and y

- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$





$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z} \text{ , where } \mathbf{M} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{m\times 2} \quad \text{(the DATA matrix)}$$

- Maximizing a *quadratic function* subject to constraints is notoriously hard (NP-hard problem)
- In our case above: very **efficient algorithm** exists (elegant solution mathematically):

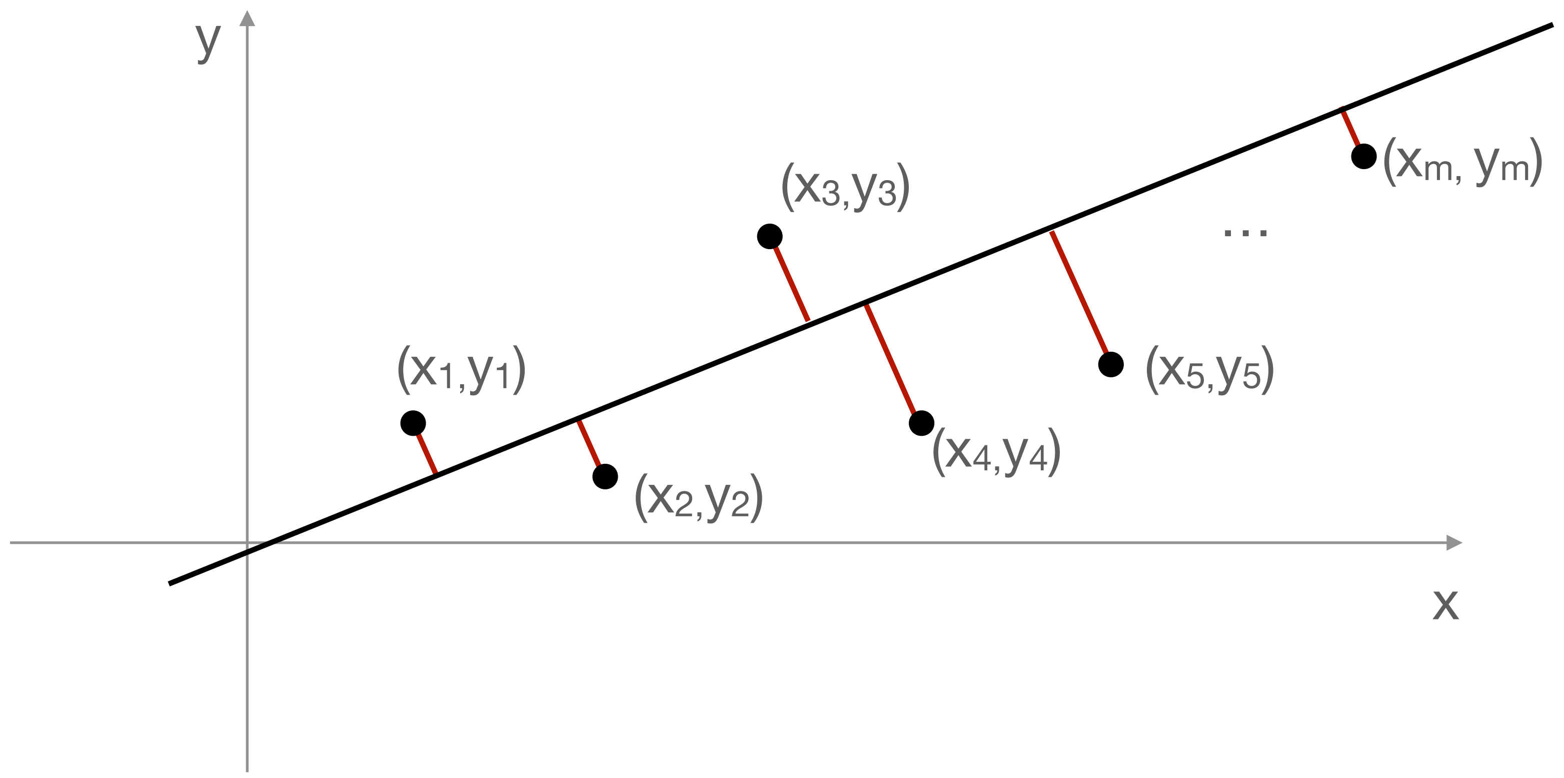
$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z} = \lambda_{max}(\mathbf{M}\mathbf{M}^{\mathrm{T}}), \ \mathbf{z}^* = \text{principal eigenvector of } \mathbf{M}\mathbf{M}^{\mathrm{T}}$$
$$= \max_{\mathbf{w}:||\mathbf{w}||=1} \mathbf{w}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{w}, \ \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$

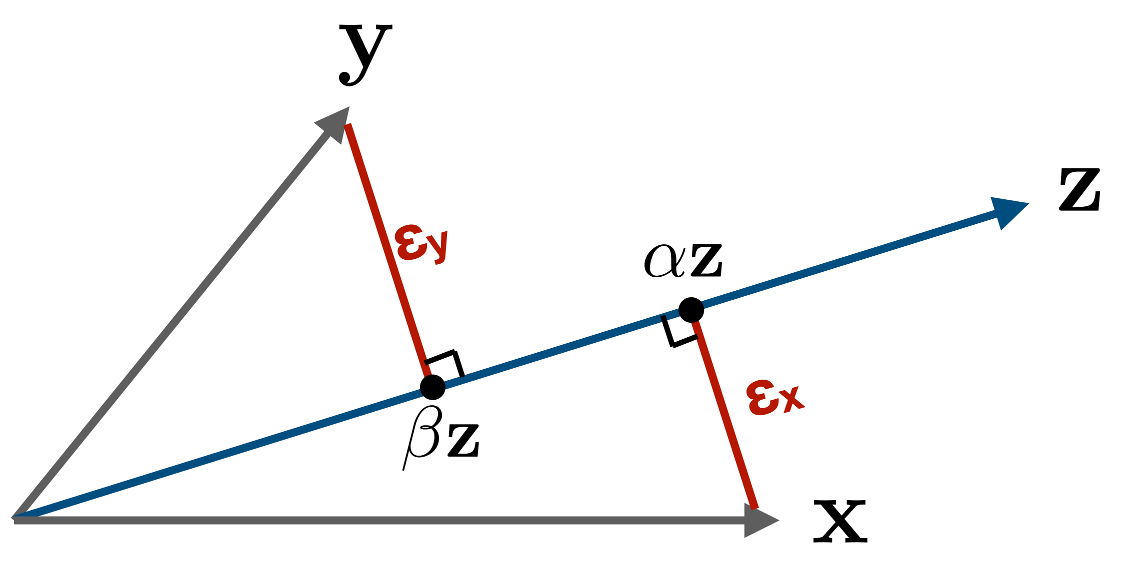
Towards Principal Component Analysis

Find z that best describes x and y

- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$





$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z} \text{ , where } \mathbf{M} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{m\times 2} \quad \text{(the DATA matrix)}$$

- Maximizing a *quadratic function* subject to constraints is notoriously hard (NP-hard problem)
- In our case above: very **efficient algorithm** exists (elegant solution mathematically):

$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z} = \lambda_{max}(\mathbf{M}\mathbf{M}^{\mathrm{T}}),\ \mathbf{z}^* = \text{principal eigenvector of } \mathbf{M}\mathbf{M}^{\mathrm{T}}$$
$$= \max_{\mathbf{w}:||\mathbf{w}||=1} \mathbf{w}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{w},\ \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$

- In our derivation, we have ignored the "intercept". Following exactly the same steps, leads to the same optimization problem but with the **data matrix M** being **zero-centered**: $\mathbf{M} = [\mathbf{x} - \mu_x, \mathbf{y} - \mu_y]$

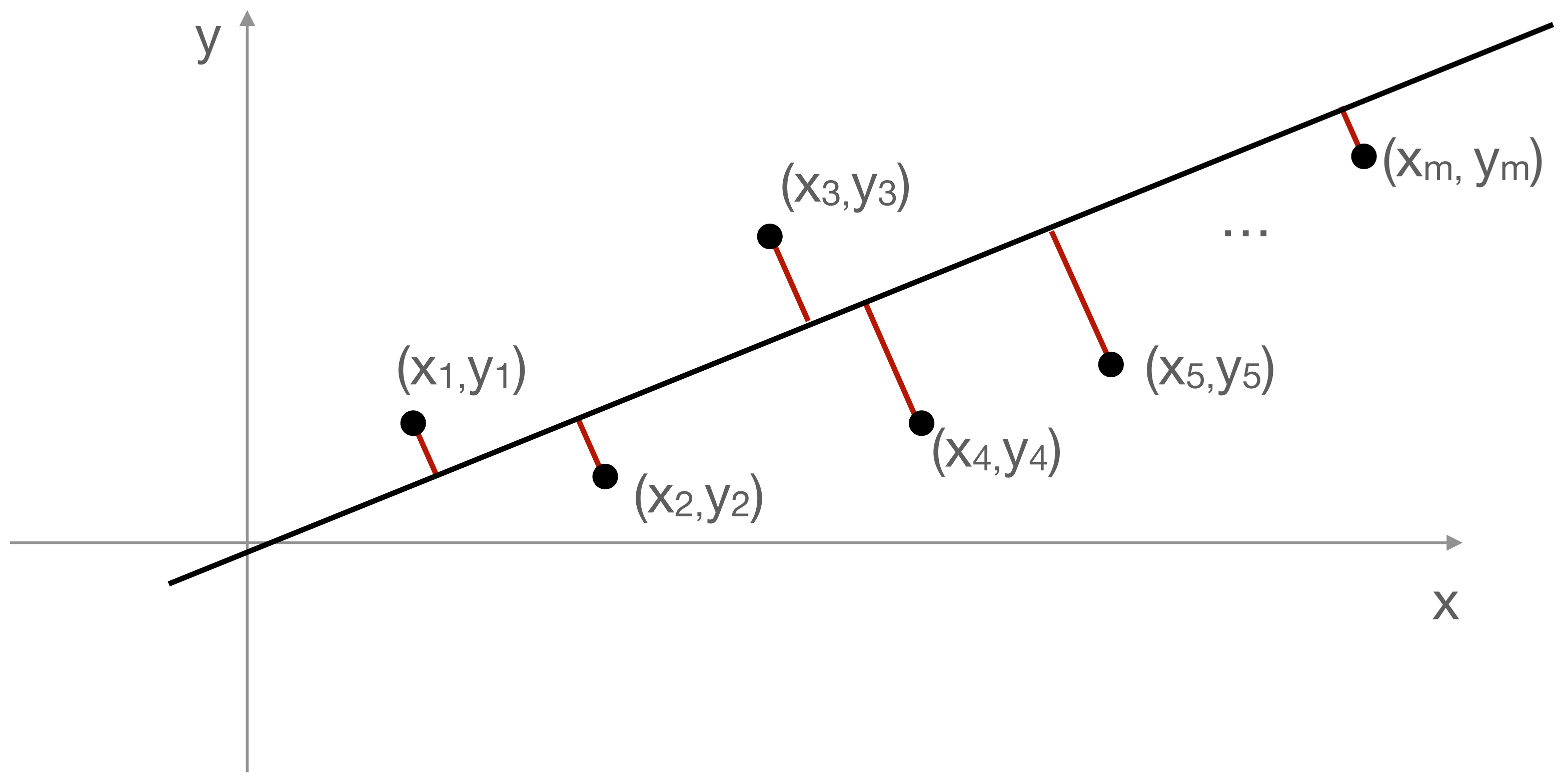
Work this out!

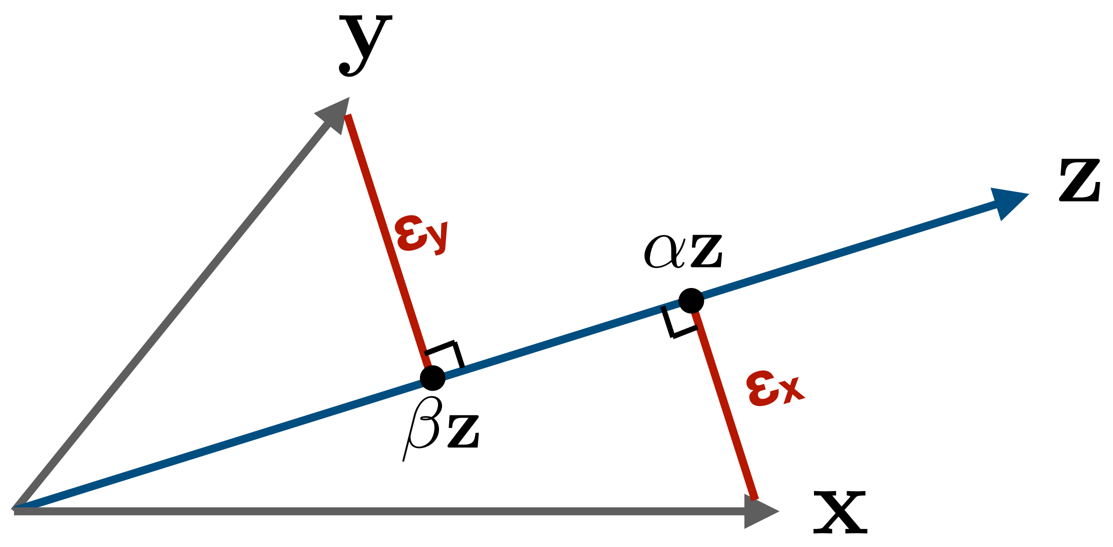
Towards Principal Component Analysis

Find z that best describes x and y

- Points in two dimensions as in previous example

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$$





$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z}\ ,\ \text{where } \mathbf{M} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{m\times 2} \quad \text{(the DATA matrix)}$$

- Maximizing a *quadratic function* subject to constraints is notoriously hard (NP-hard problem)
- In our case above: very **efficient algorithm** exists (elegant solution mathematically):

$$\max_{\mathbf{z}:||\mathbf{z}||=1} \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{z} = \lambda_{max}(\mathbf{M}\mathbf{M}^{\mathrm{T}}),\ \mathbf{z}^* = \text{principal eigenvector of } \mathbf{M}\mathbf{M}^{\mathrm{T}}$$
$$= \max_{\mathbf{w}:||\mathbf{w}||=1} \mathbf{w}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{w},\ \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$

- In our derivation, we have ignored the "intercept". Following exactly the same steps, leads to the same optimization problem but with the **data matrix M** being **zero-centered**: $\mathbf{M} = [\mathbf{x} - \mu_x, \mathbf{y} - \mu_y]$

- $\frac{1}{m-1}\mathbf{M}^{\mathrm{T}}\mathbf{M}$ is the empirical **covariance matrix**!

$\mathbf{w}^*$ is the first principal component

$\mathbf{M}\mathbf{w}^* = \mathbf{z}^*\sqrt{\lambda_{max}}$ gives the coordinates of the projected points

$\frac{\lambda_{max}}{m-1}$ is the "explained variance"

Principal Component Analysis — General Case

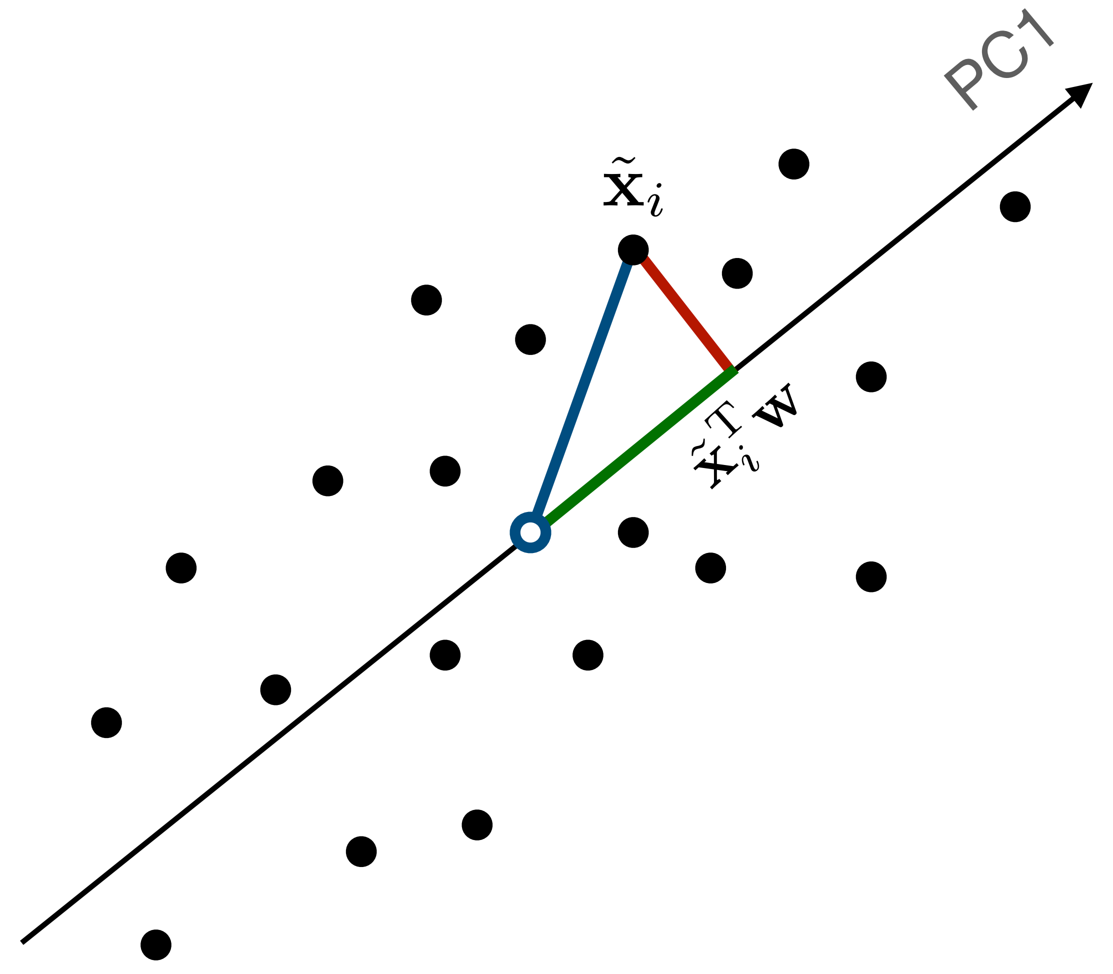
- A collection of **n features** across **m samples**; datapoints: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_m \in \mathbb{R}^n$
  - zero-centered **data matrix**: $\mathbf{M} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_m]^{\mathrm{T}} \in \mathbb{R}^{m \times n}$, where $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{m}\sum_k \mathbf{x}_k$

(PC1) $$\max_{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}|| = 1} ||\mathbf{M}\mathbf{w}||^2, \;\; \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$

Principal Component Analysis — General Case

- A collection of **n features** across **m samples**; datapoints: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_m \in \mathbb{R}^n$
  - zero-centered **data matrix**: $\mathbf{M} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_m]^{\mathrm{T}} \in \mathbb{R}^{m \times n}$, where $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{m}\sum_k \mathbf{x}_k$

(PC1) $$\max_{\mathbf{w}\in\mathbb{R}^n : ||\mathbf{w}||=1} ||\mathbf{M}\mathbf{w}||^2, \;\; \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$



PCA minimizes the "error" from points to projection

PCA finds the "direction" of **maximum variance**

- **Projection** along **w*** reduces the dimensions from n to 1:

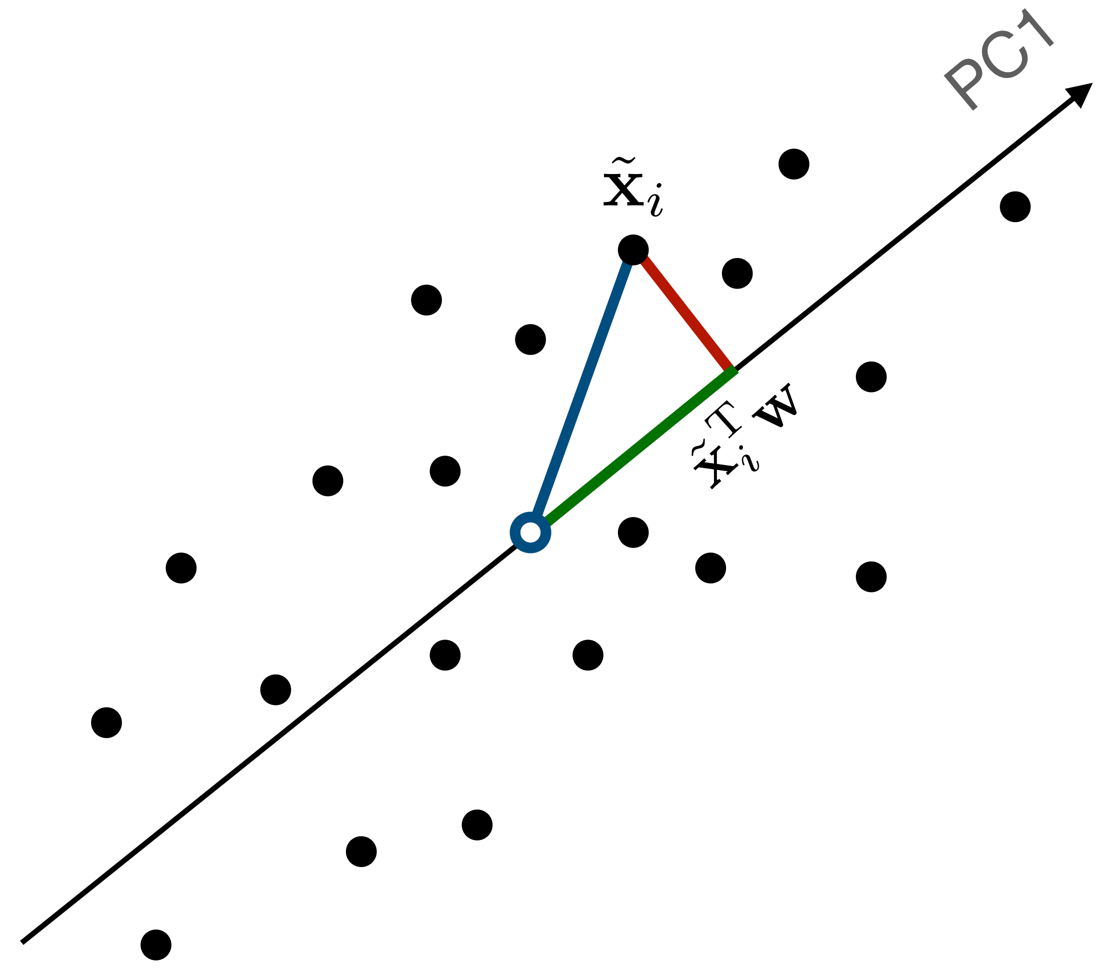
— projected datapoints: $\tilde{\mathbf{x}}_1^{\mathrm{T}}\mathbf{w}^*, \tilde{\mathbf{x}}_2^{\mathrm{T}}\mathbf{w}^*, \ldots, \tilde{\mathbf{x}}_m^{\mathrm{T}}\mathbf{w}^* \in \mathbb{R}$

— variance: $\frac{1}{m-1}\sum_i (\tilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{w}^*)^2 = \frac{||\mathbf{M}\mathbf{w}^*||^2}{m-1} = \frac{\lambda_{\max}}{m-1}$

Principal Component Analysis — General Case

- A collection of **n features** across **m samples**; datapoints: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_m \in \mathbb{R}^n$
  - zero-centered **data matrix**: $\mathbf{M} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_m]^{\mathrm{T}} \in \mathbb{R}^{m \times n}$, where $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{m}\sum_k \mathbf{x}_k$

(PC1) $$\max_{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}|| = 1} ||\mathbf{M}\mathbf{w}||^2, \ \ \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$



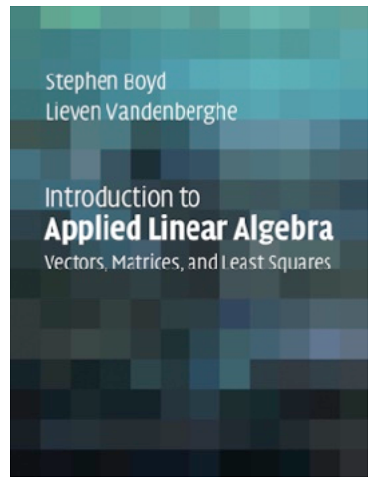
PCA minimizes the "error" from points to projection

PCA finds the "direction" of **maximum variance**

- **Projection** along **w*** reduces the dimensions from n to 1:
  - — projected datapoints: $\tilde{\mathbf{x}}_1^{\mathrm{T}}\mathbf{w}^*, \tilde{\mathbf{x}}_2^{\mathrm{T}}\mathbf{w}^*, \ldots, \tilde{\mathbf{x}}_m^{\mathrm{T}}\mathbf{w}^* \in \mathbb{R}$
  - — variance: $\frac{1}{m-1}\sum_i (\tilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{w}^*)^2 = \frac{||\mathbf{M}\mathbf{w}^*||^2}{m-1} = \frac{\lambda_{\max}}{m-1}$

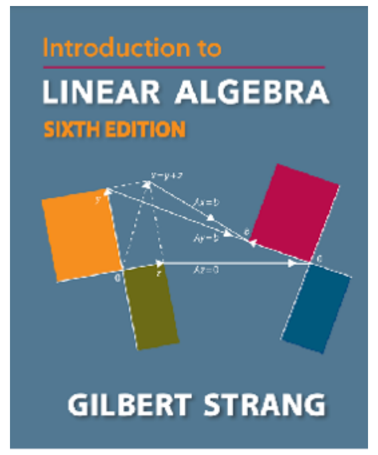
  - (next lecture)
    - — Core Algorithm: Singular Value Decomposition (SVD)
    - — Python Examples / Code (MDS, pPCA, EM)

Resources



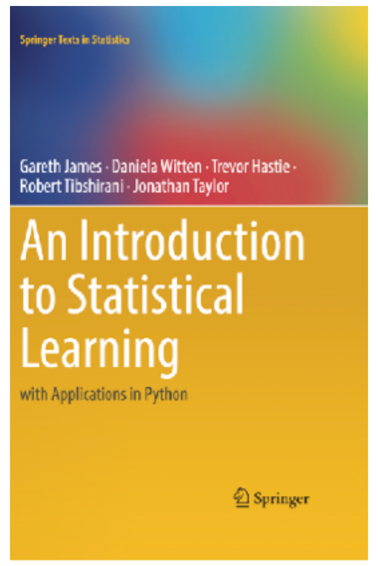
Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares
Stephen Boyd and Lieven Vandenberghe

https://web.stanford.edu/~boyd/vmls/vmls.pdf



Introduction to Linear Algebra
Gilbert Strang

https://math.mit.edu/~gs/linearalgebra/ila6/indexila6.html

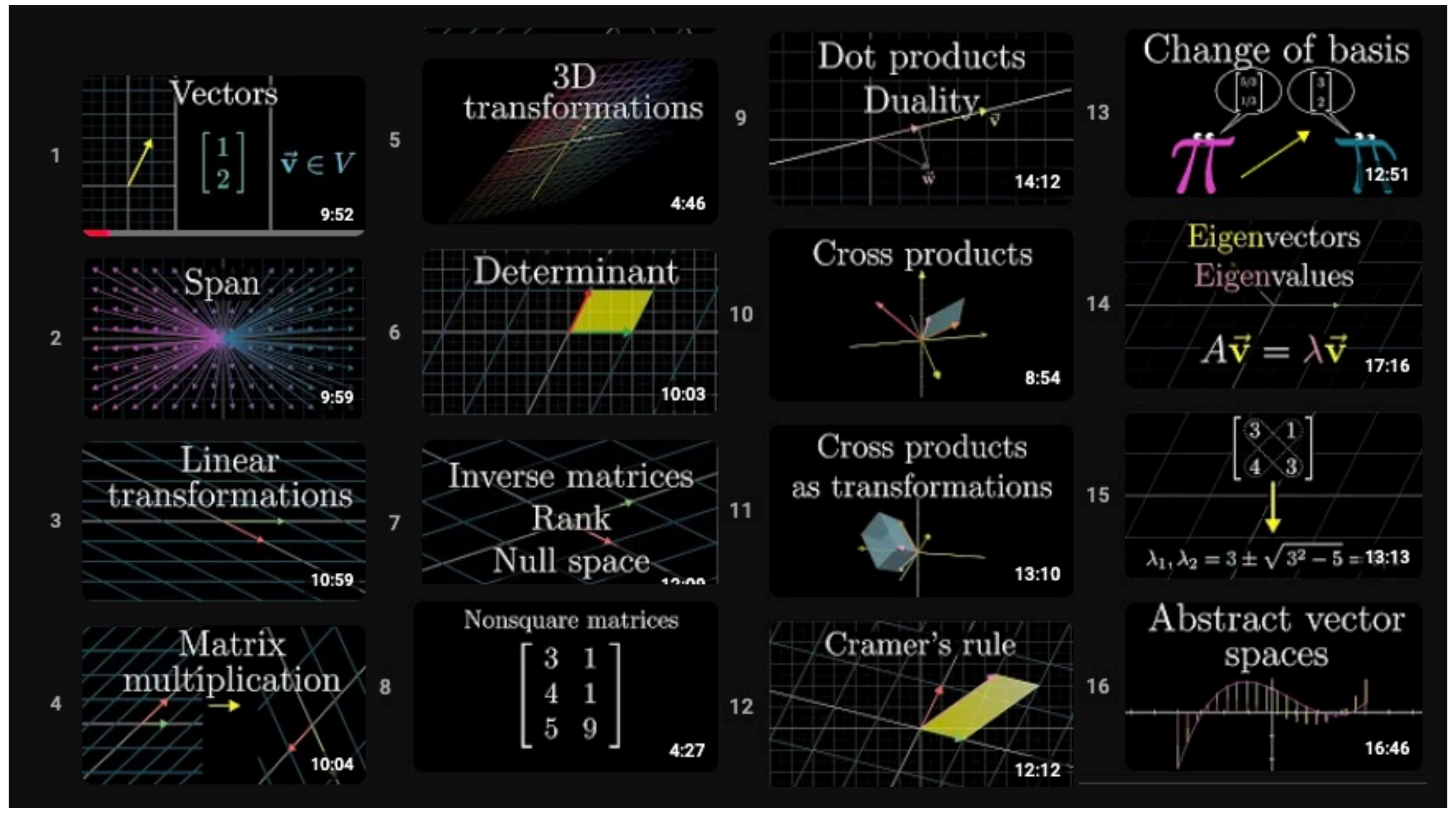


Introduction to Statistical Learning
James, Witten, Hastie, Tibshirani, Taylor

https://www.statlearning.com/

3Blue1Brown

Essence of Linear Algebra



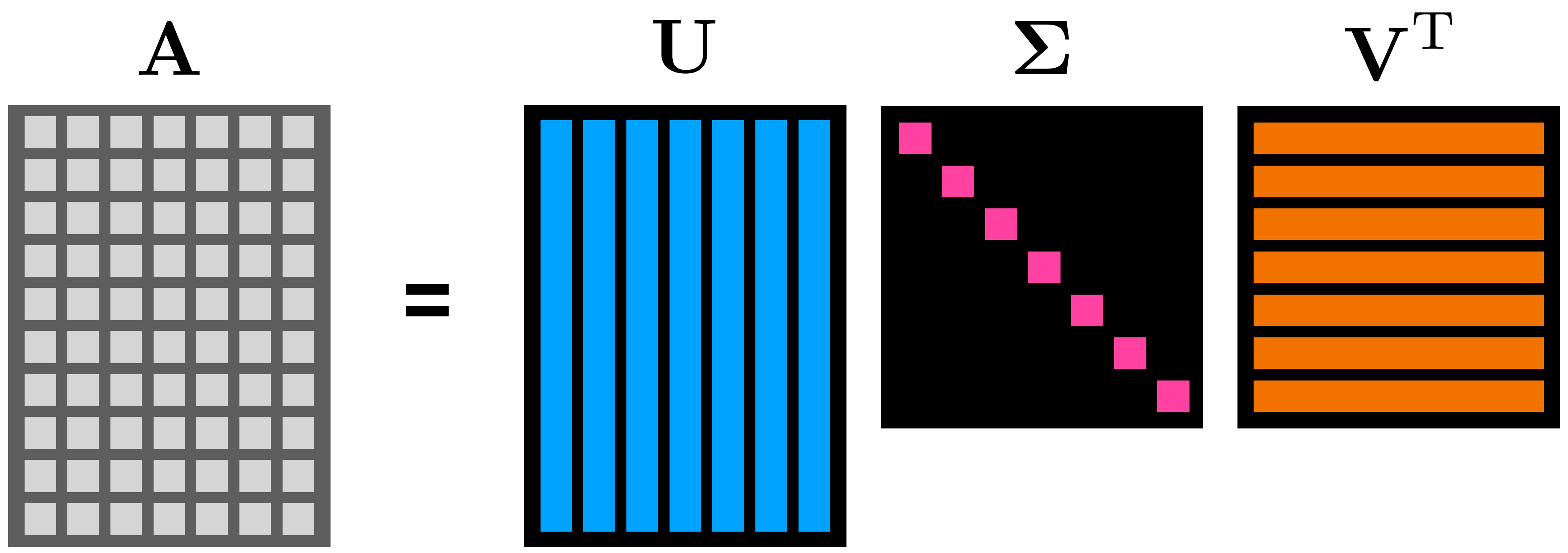
BMI 206

Singular Value Decomposition

Vasilis Ntranos 2024

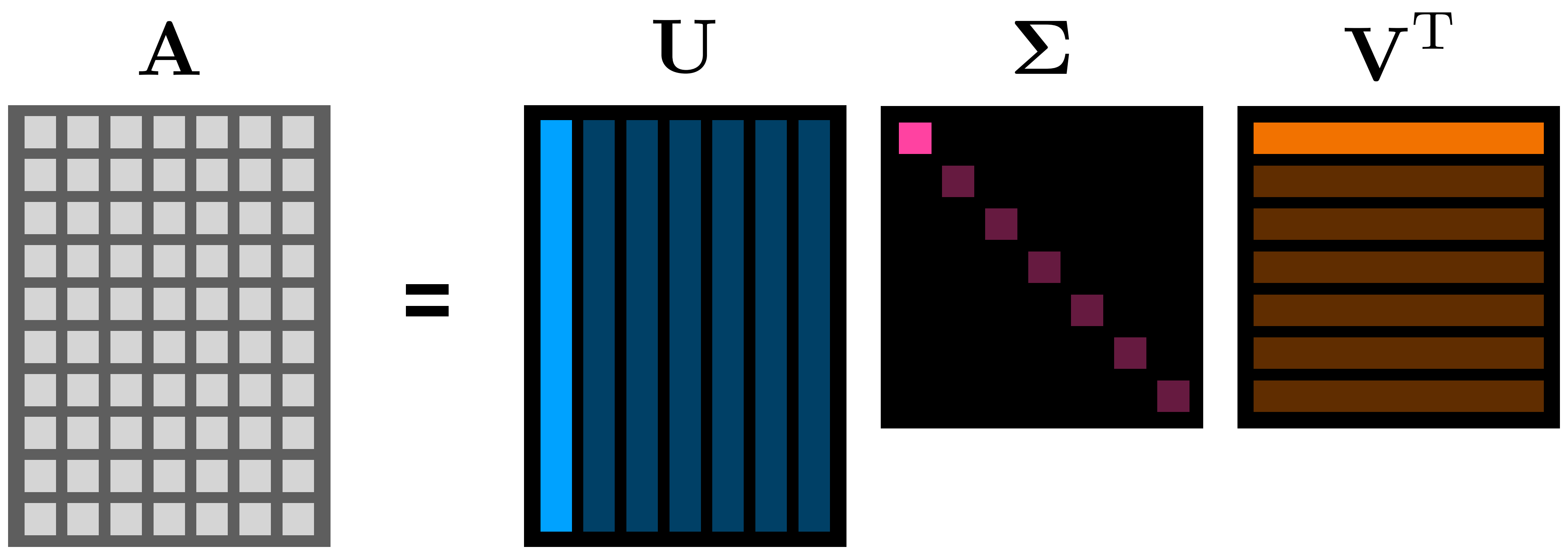
Singular Value Decomposition

- Any m×n matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V**$^{\mathbf{T}}$ with the following properties:
    - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
    - $\mathbf{\Sigma}$ is diagonal.



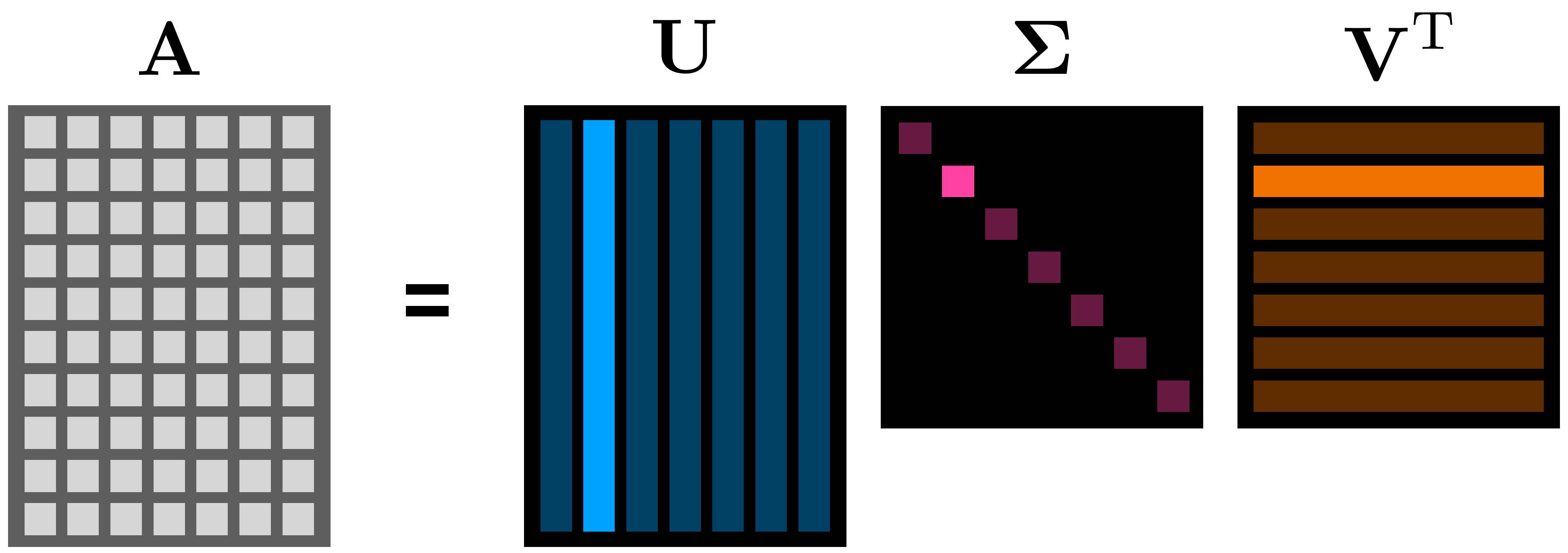
Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.



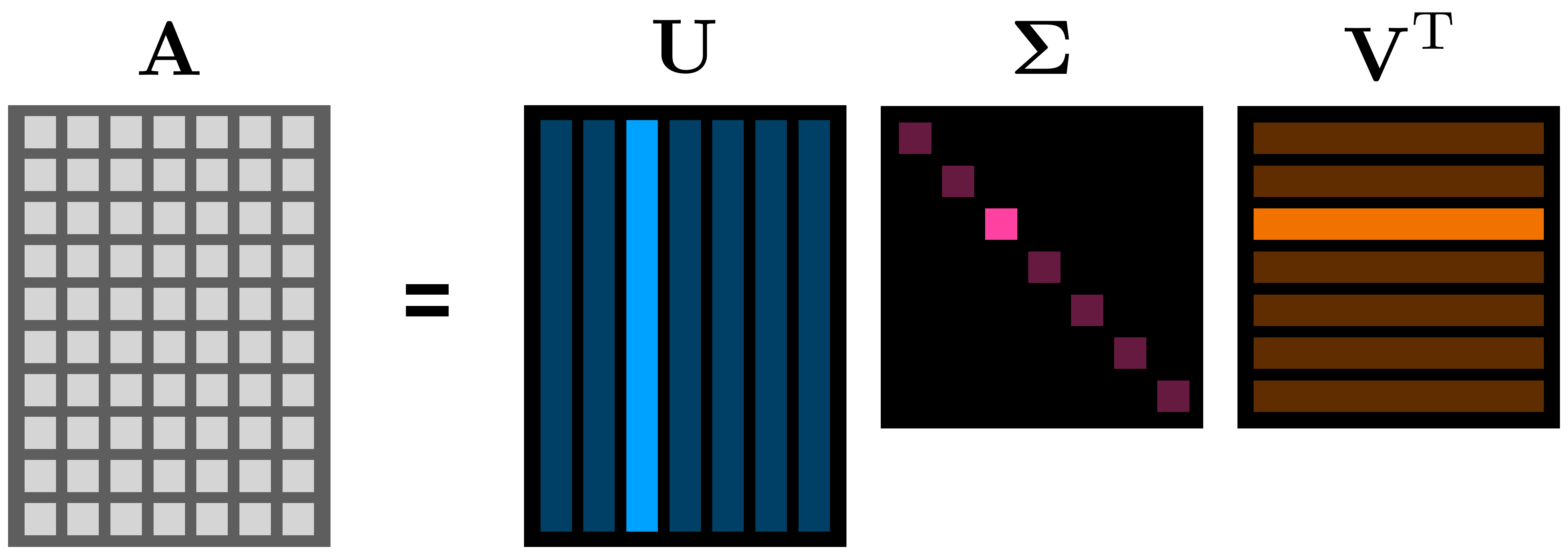
Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and $\mathbf{V}^\mathrm{T}$ with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^\mathrm{T}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\mathrm{T}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.



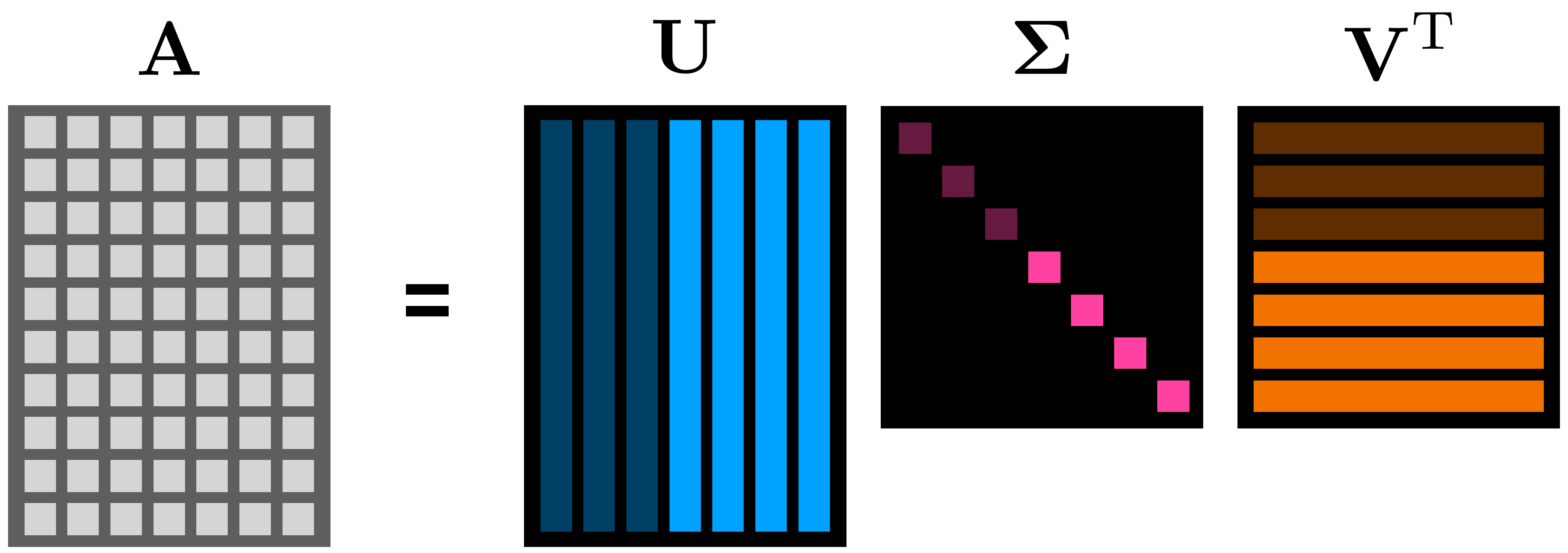
Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and $\mathbf{V}^\mathrm{T}$ with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$ are orthonormal, (i.e., $\mathbf{U}^\mathrm{T}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\mathrm{T}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.



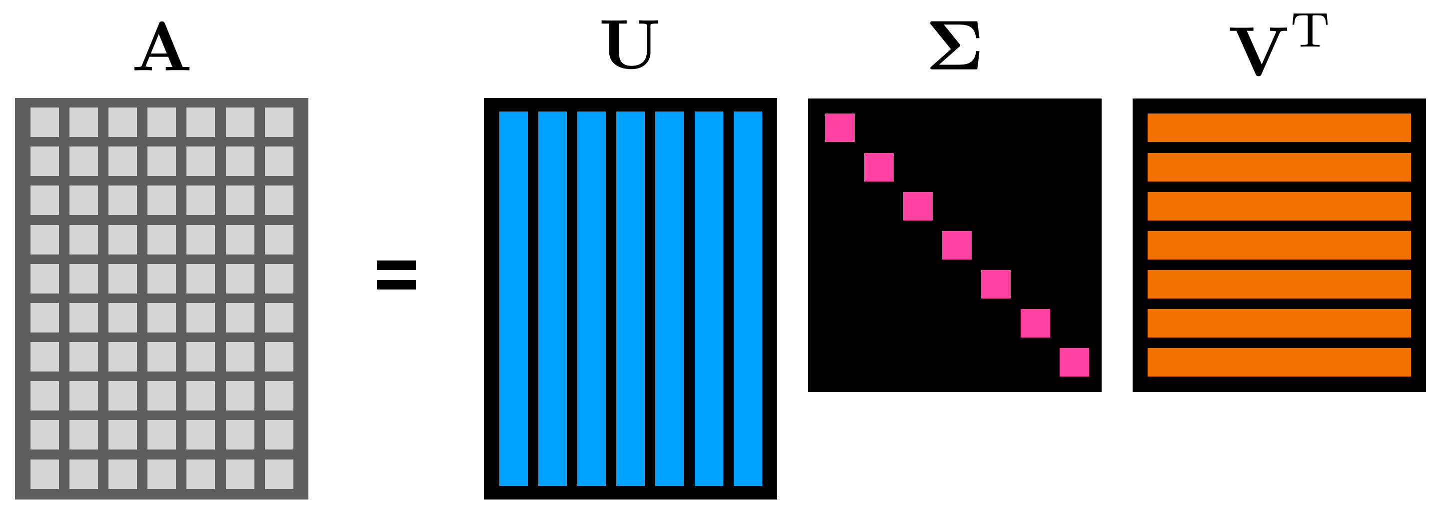
Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and $\mathbf{V}^T$ with the following properties:
    - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
    - $\mathbf{\Sigma}$ is diagonal.



Singular Value Decomposition

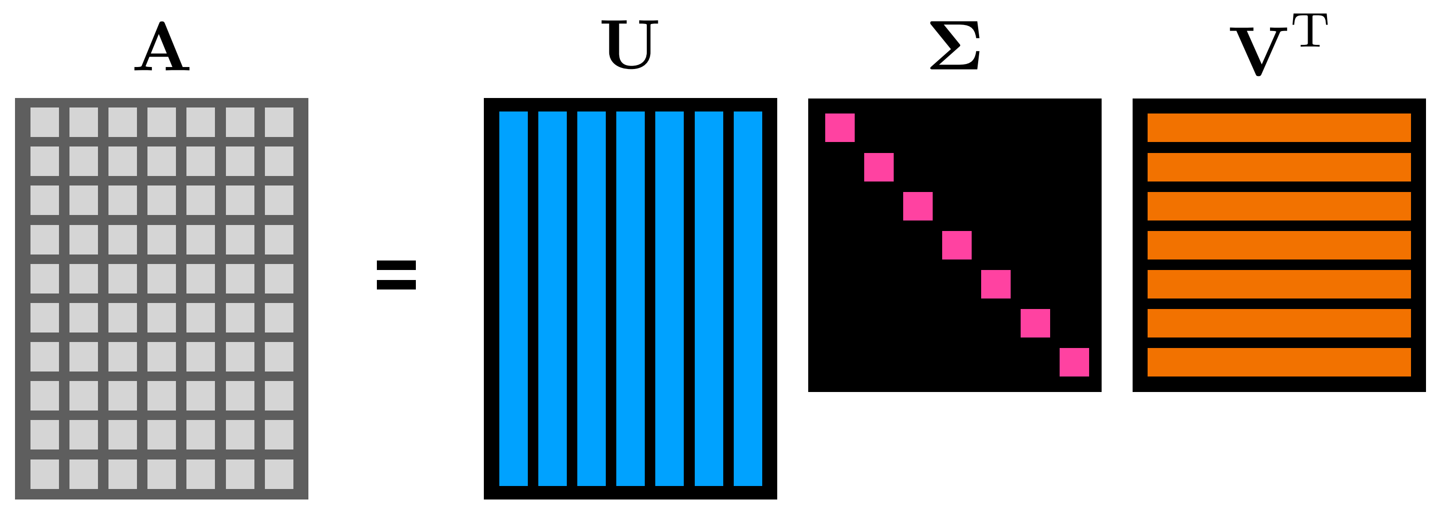
- Any m×n matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V**$^{\mathbf{T}}$ with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.



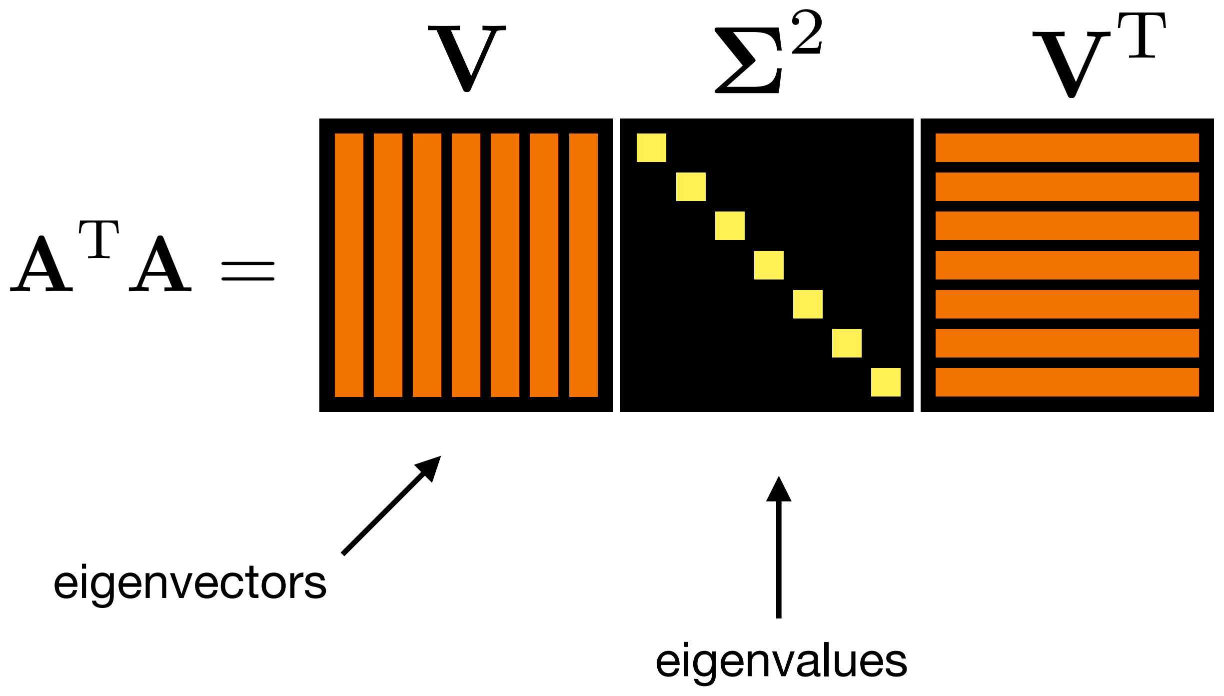
$$
\begin{aligned}
\mathbf{A}^{\mathrm{T}}\mathbf{A} &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}})^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \\
&= (\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}})\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \\
&= \mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\underbrace{\mathbf{U}^{\mathrm{T}}\mathbf{U}}_{=\mathbf{I}}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \\
&= \mathbf{V}\underbrace{\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}}_{=\mathbf{\Sigma}^2}\mathbf{V}^{\mathrm{T}} \\
&= \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^{\mathrm{T}}
\end{aligned}
$$

Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.

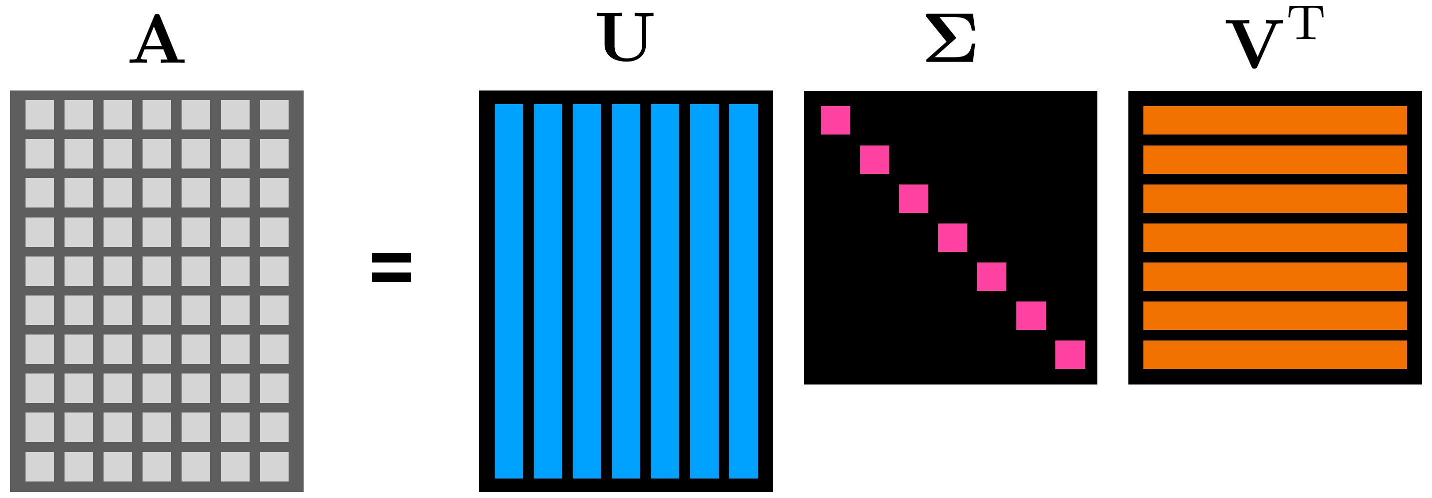


$$\begin{aligned}
\mathbf{A}^{\mathrm{T}}\mathbf{A} &= (\mathbf{U\Sigma V}^{\mathrm{T}})^{\mathrm{T}}\mathbf{U\Sigma V}^{\mathrm{T}} \\
&= (\mathbf{V\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}})\mathbf{U\Sigma V}^{\mathrm{T}} \\
&= \mathbf{V\Sigma}^{\mathrm{T}}\underbrace{\mathbf{U}^{\mathrm{T}}\mathbf{U}}_{=\mathbf{I}}\mathbf{\Sigma V}^{\mathrm{T}} \\
&= \mathbf{V}\underbrace{\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}}_{=\mathbf{\Sigma}^2}\mathbf{V}^{\mathrm{T}} \\
&= \mathbf{V\Sigma}^2\mathbf{V}^{\mathrm{T}}
\end{aligned}$$

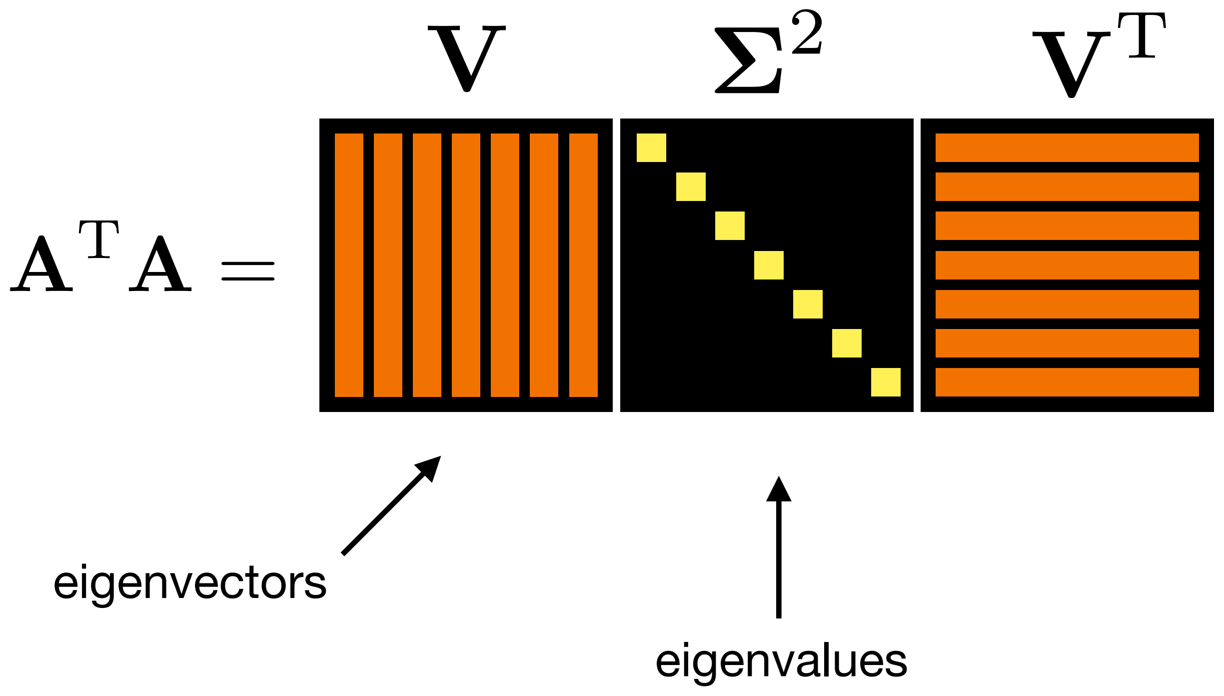


Singular Value Decomposition

- Any m×n matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.

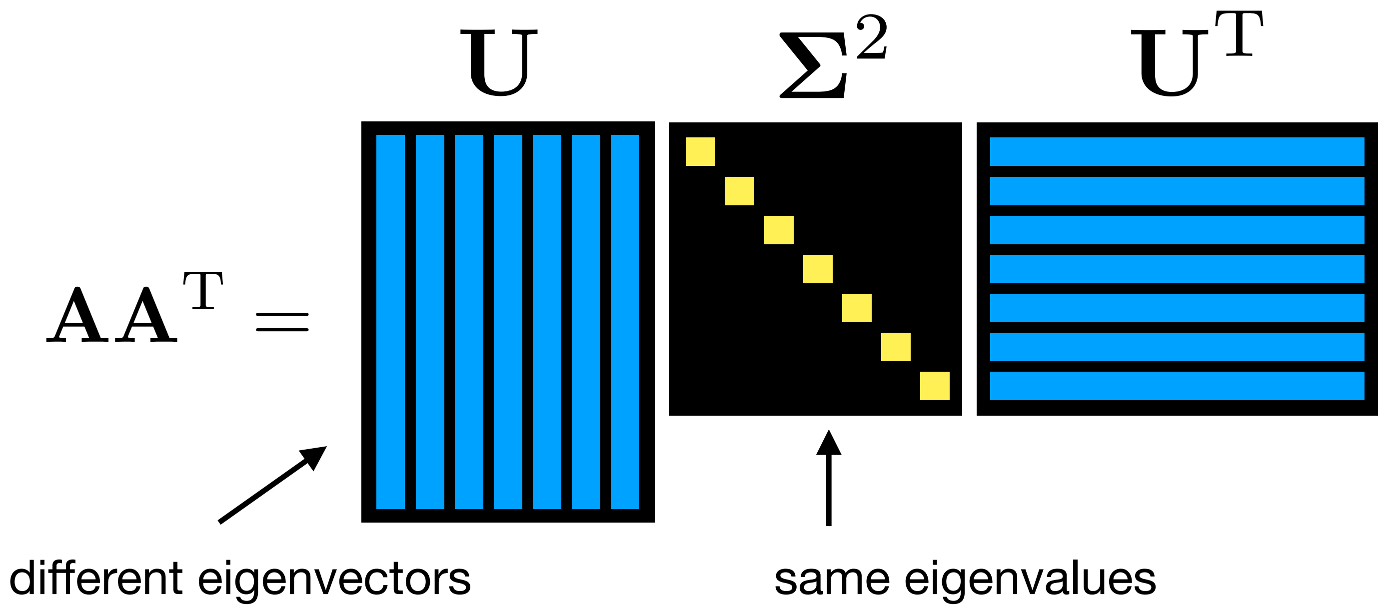


$$
\begin{aligned}
\mathbf{A}^{\mathrm{T}}\mathbf{A} &= (\mathbf{U\Sigma V}^{\mathrm{T}})^{\mathrm{T}}\mathbf{U\Sigma V}^{\mathrm{T}} \\
&= (\mathbf{V\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}})\mathbf{U\Sigma V}^{\mathrm{T}} \\
&= \mathbf{V\Sigma}^{\mathrm{T}}\underbrace{\mathbf{U}^{\mathrm{T}}\mathbf{U}}_{=\mathbf{I}}\mathbf{\Sigma V}^{\mathrm{T}} \\
&= \mathbf{V}\underbrace{\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}}_{=\mathbf{\Sigma}^2}\mathbf{V}^{\mathrm{T}} \\
&= \mathbf{V\Sigma}^2\mathbf{V}^{\mathrm{T}}
\end{aligned}
$$



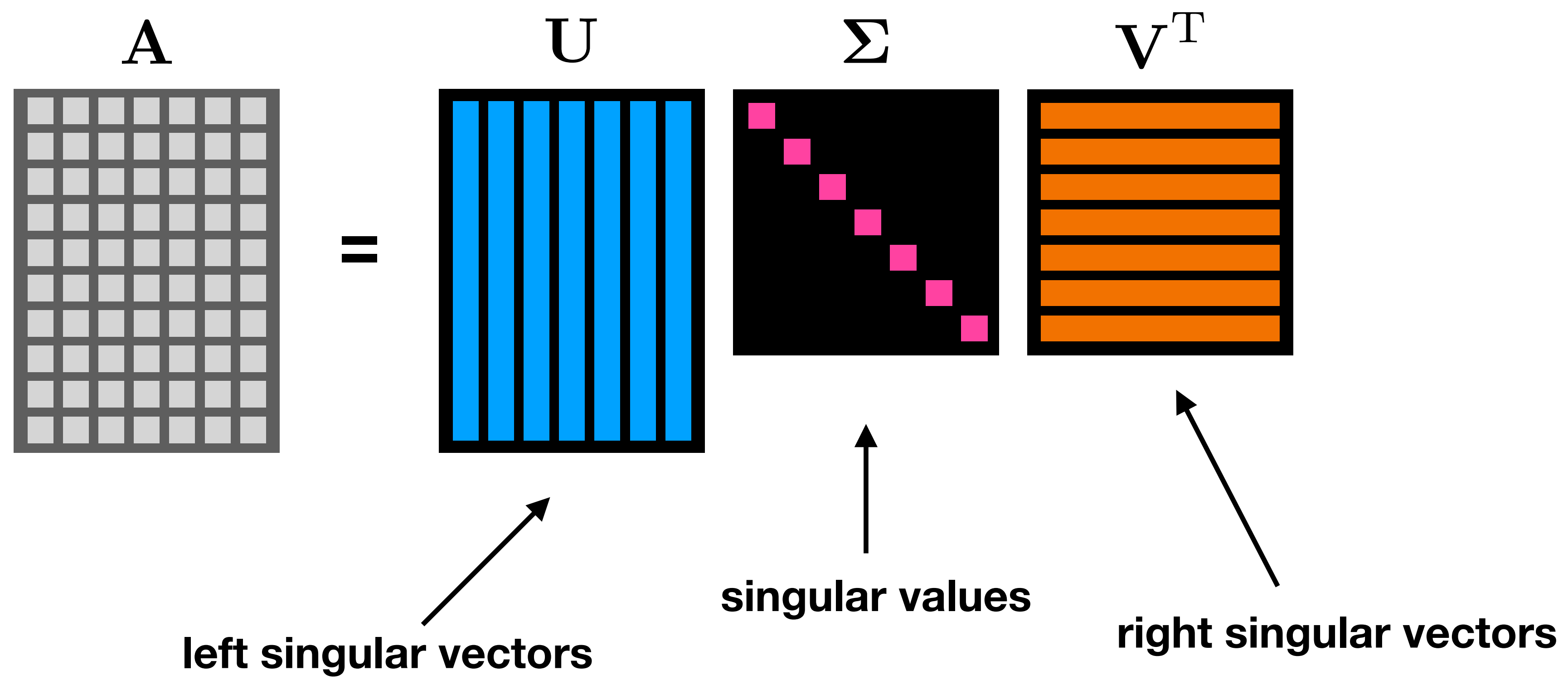
- Similarly,

$$\mathbf{AA}^{\mathrm{T}} = \mathbf{U\Sigma}^2\mathbf{U}^{\mathrm{T}}$$

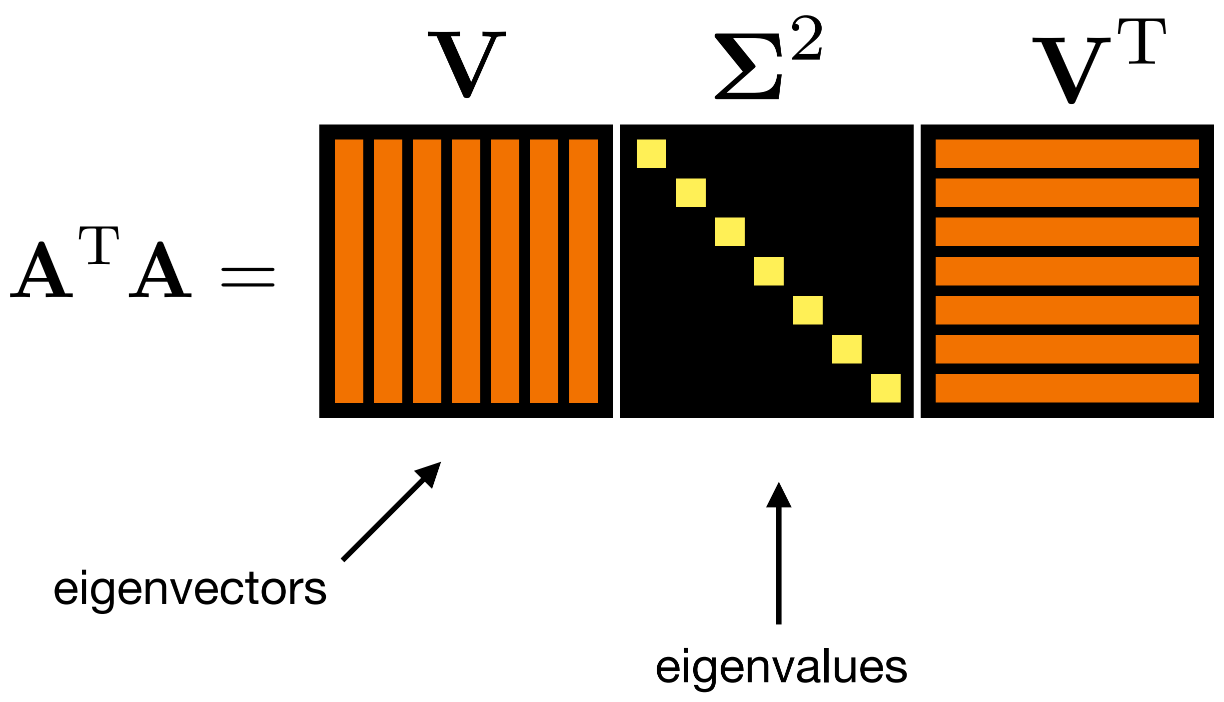


Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^\mathrm{T}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\mathrm{T}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.

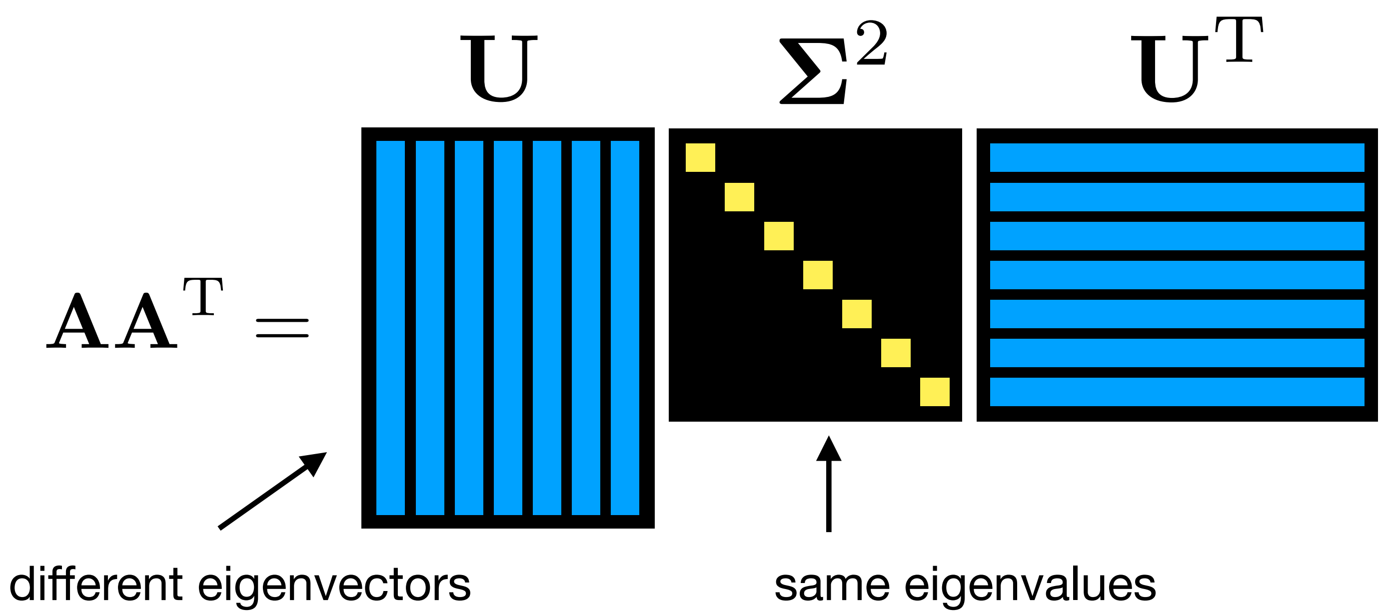


$$\begin{aligned}
\mathbf{A}^\mathrm{T}\mathbf{A} &= (\mathbf{U\Sigma V}^\mathrm{T})^\mathrm{T}\mathbf{U\Sigma V}^\mathrm{T} \\
&= (\mathbf{V\Sigma}^\mathrm{T}\mathbf{U}^\mathrm{T})\mathbf{U\Sigma V}^\mathrm{T} \\
&= \mathbf{V\Sigma}^\mathrm{T}\underbrace{\mathbf{U}^\mathrm{T}\mathbf{U}}_{=\mathbf{I}}\mathbf{\Sigma V}^\mathrm{T} \\
&= \mathbf{V}\underbrace{\mathbf{\Sigma}^\mathrm{T}\mathbf{\Sigma}}_{=\mathbf{\Sigma}^2}\mathbf{V}^\mathrm{T} \\
&= \mathbf{V\Sigma}^2\mathbf{V}^\mathrm{T}
\end{aligned}$$



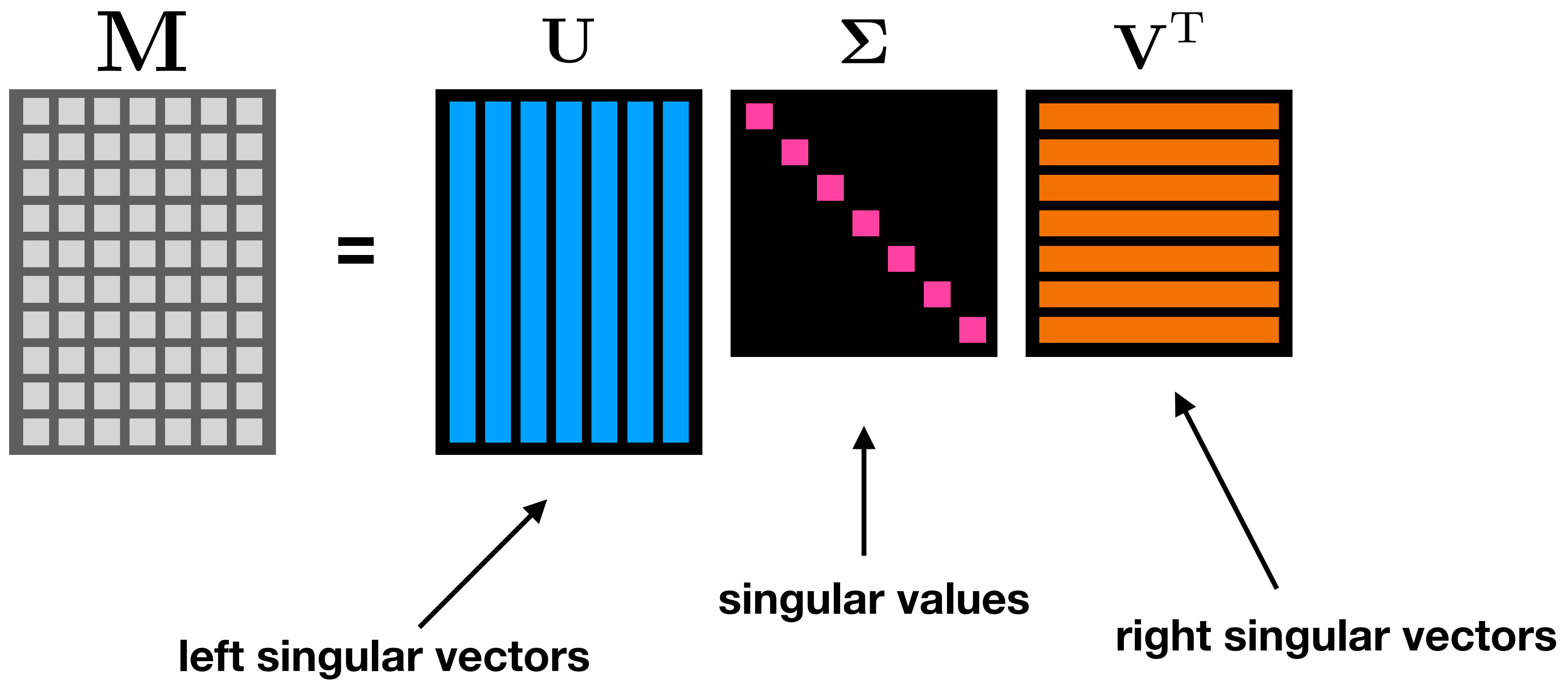
- Similarly,

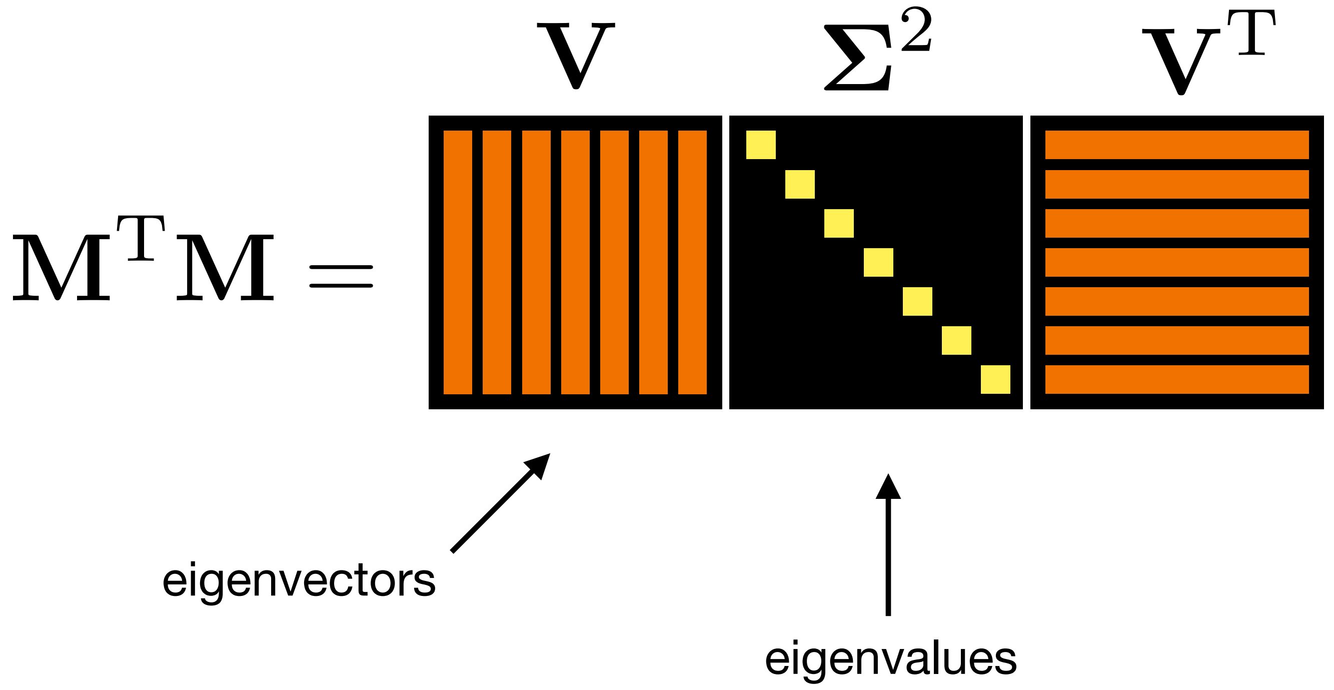
$$\mathbf{AA}^\mathrm{T} = \mathbf{U\Sigma}^2\mathbf{U}^\mathrm{T}$$



Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.



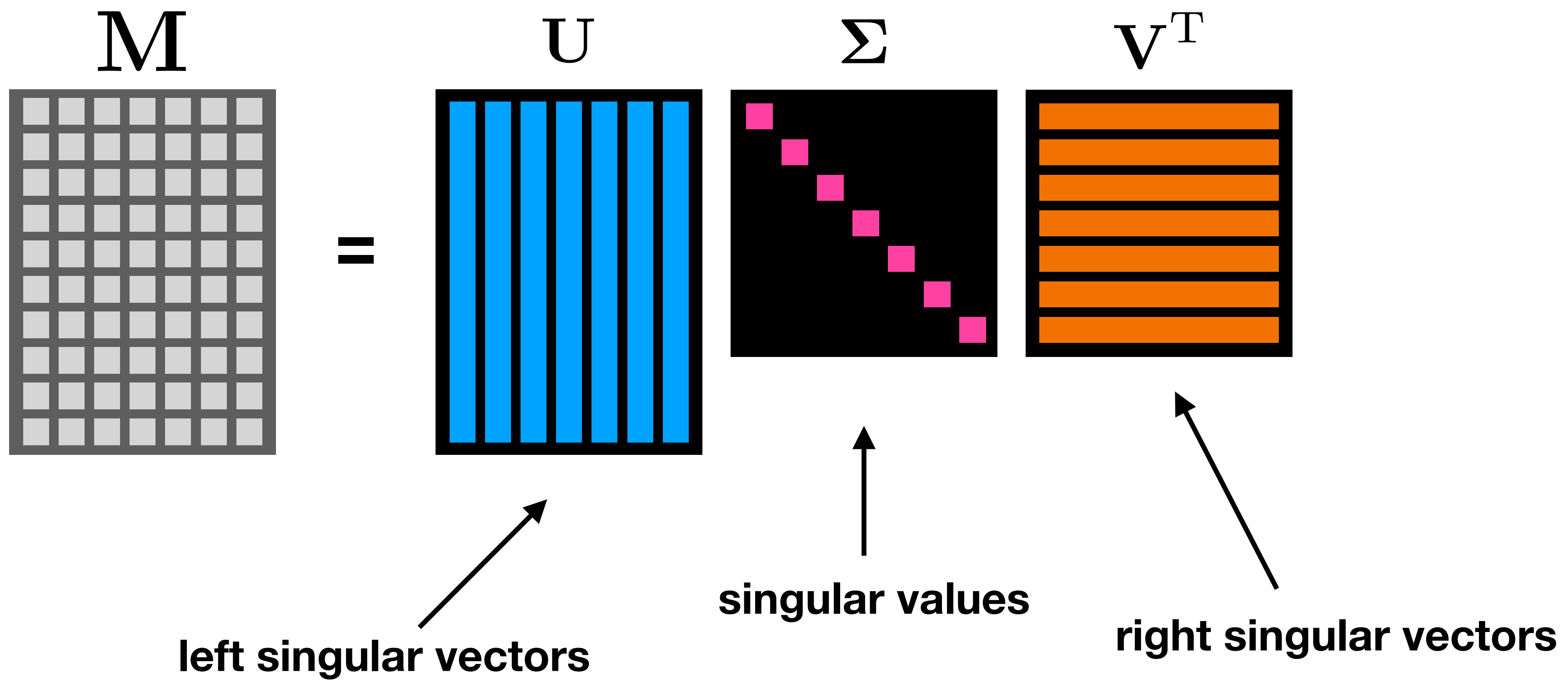


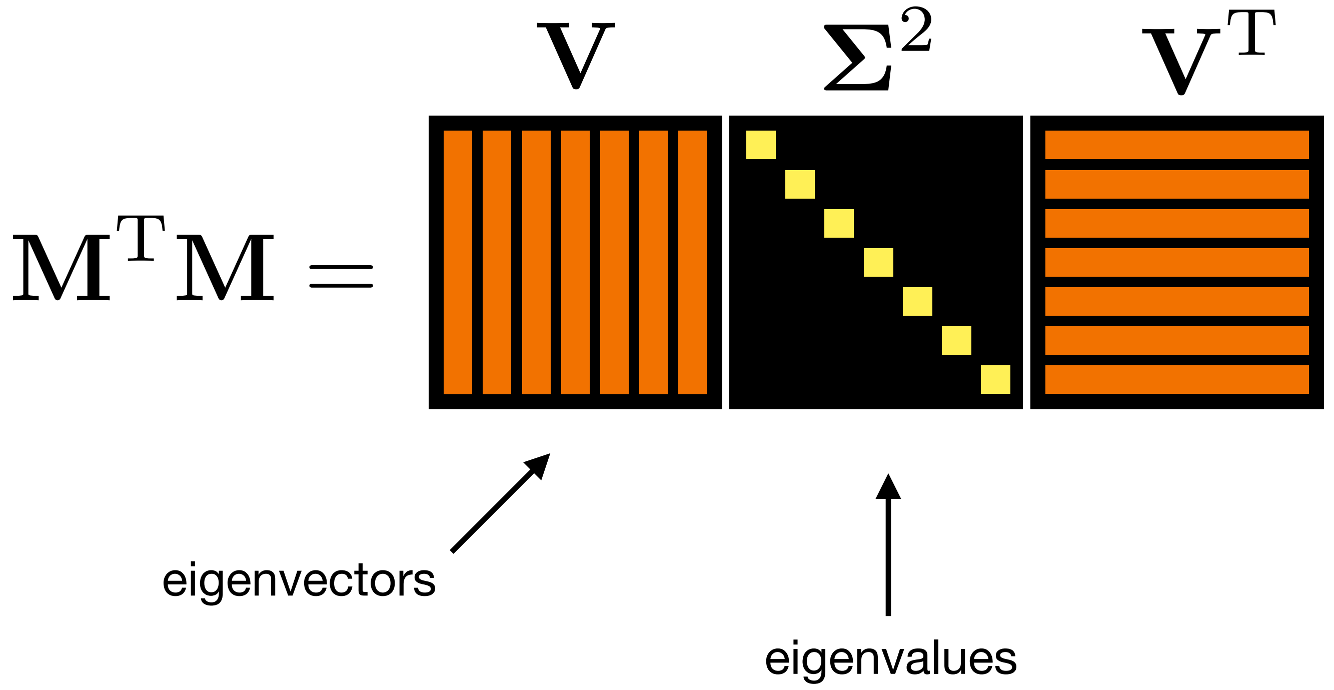
PCA. Recall that:

(PC1) $\max_{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||=1} ||\mathbf{M}\mathbf{w}||^2, \;\; \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$

Singular Value Decomposition

- Any m×n matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.



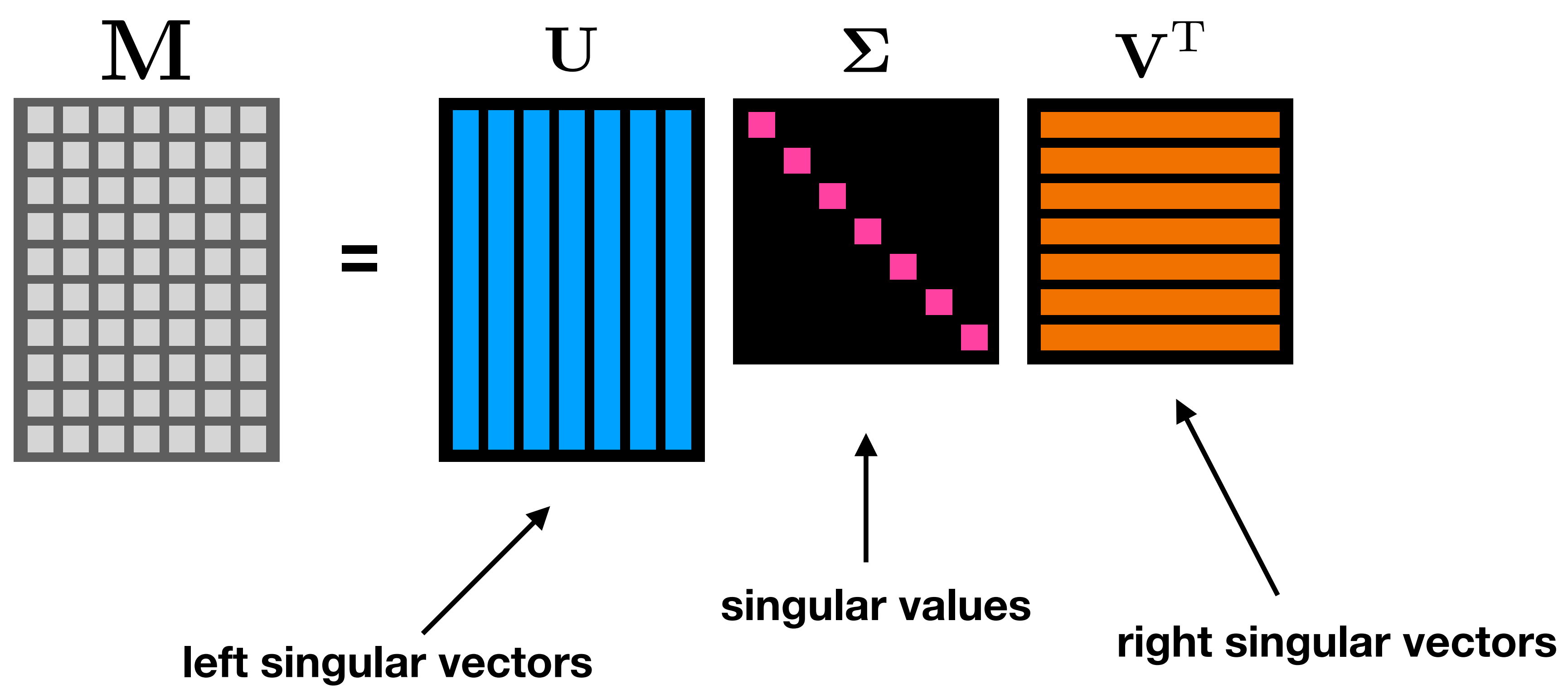


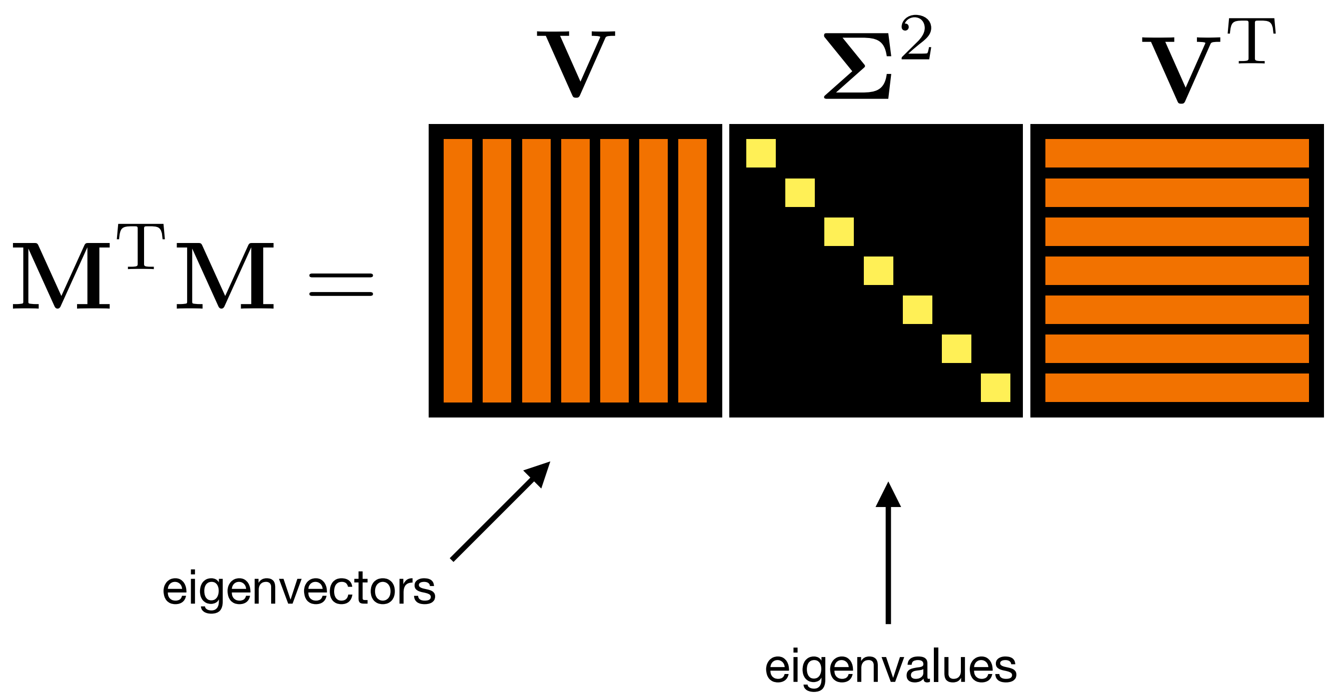
PCA. Recall that:

(PC1) $$\max_{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||=1} ||\mathbf{M}\mathbf{w}||^2, \quad \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$
$$= \text{first (right) singular vector of } \mathbf{M}$$

Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.





PCA. Recall that:

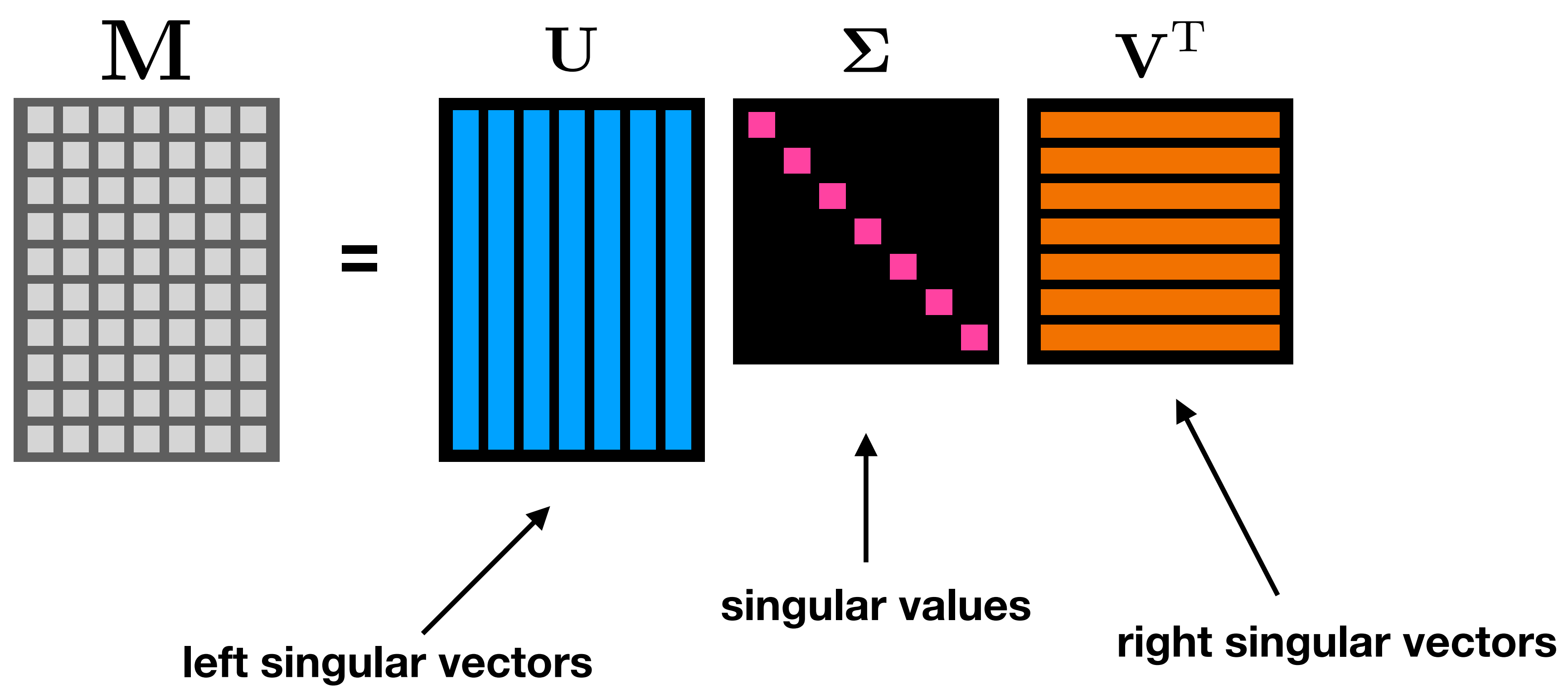
(PC1) $\max_{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||=1} ||\mathbf{M}\mathbf{w}||^2, \quad \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$

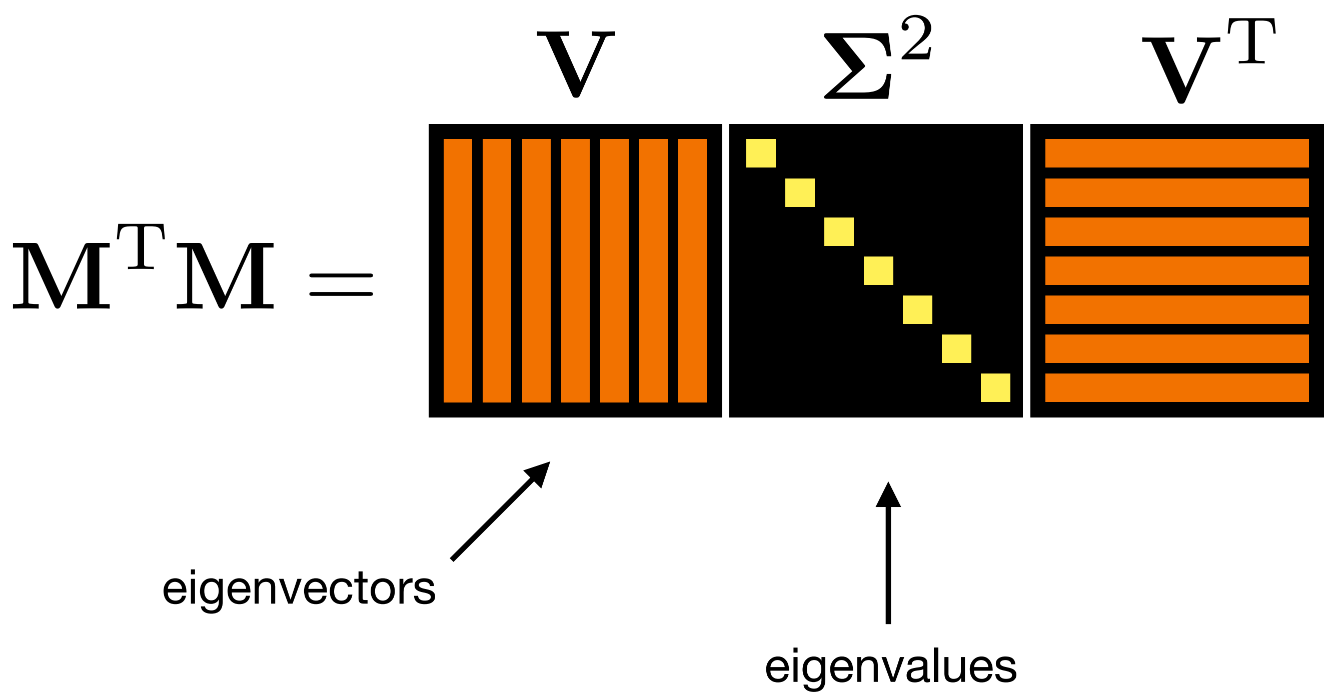
$= \text{first (right) singular vector of } \mathbf{M}$

$\mathbf{M}\mathbf{w}^* = \mathbf{z}^* \sqrt{\lambda_{max}}$ gives the coordinates of the projected points

Singular Value Decomposition

- Any mxn matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.





PCA. Recall that:

(PC1) $$\max_{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||=1} ||\mathbf{M}\mathbf{w}||^2, \quad \mathbf{w}^* = \text{principal eigenvector of } \mathbf{M}^{\mathrm{T}}\mathbf{M}$$
$$= \text{first (right) singular vector of } \mathbf{M}$$

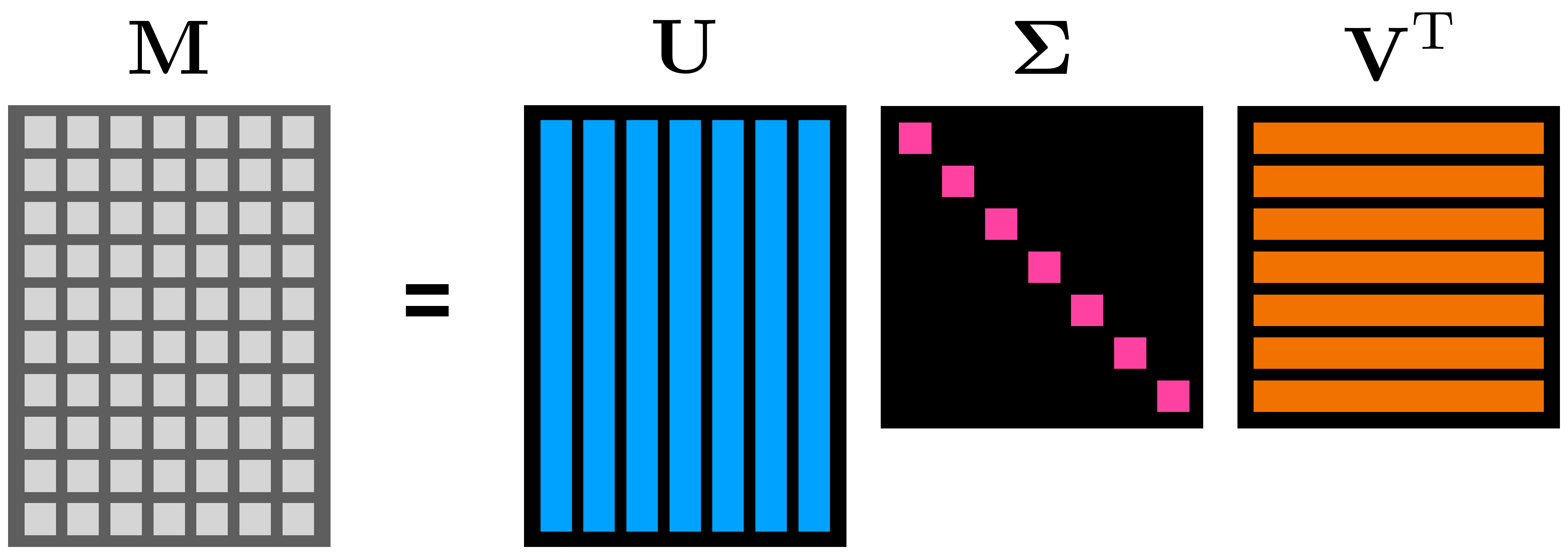
$\mathbf{M}\mathbf{w}^* = \mathbf{z}^* \sqrt{\lambda_{max}}$ gives the coordinates of the projected points

$\mathbf{M}\mathbf{v}_1 = \mathbf{u_1} \cdot \sigma_1$ gives the coordinates of the projected points

Singular Value Decomposition

- Any m×n matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.

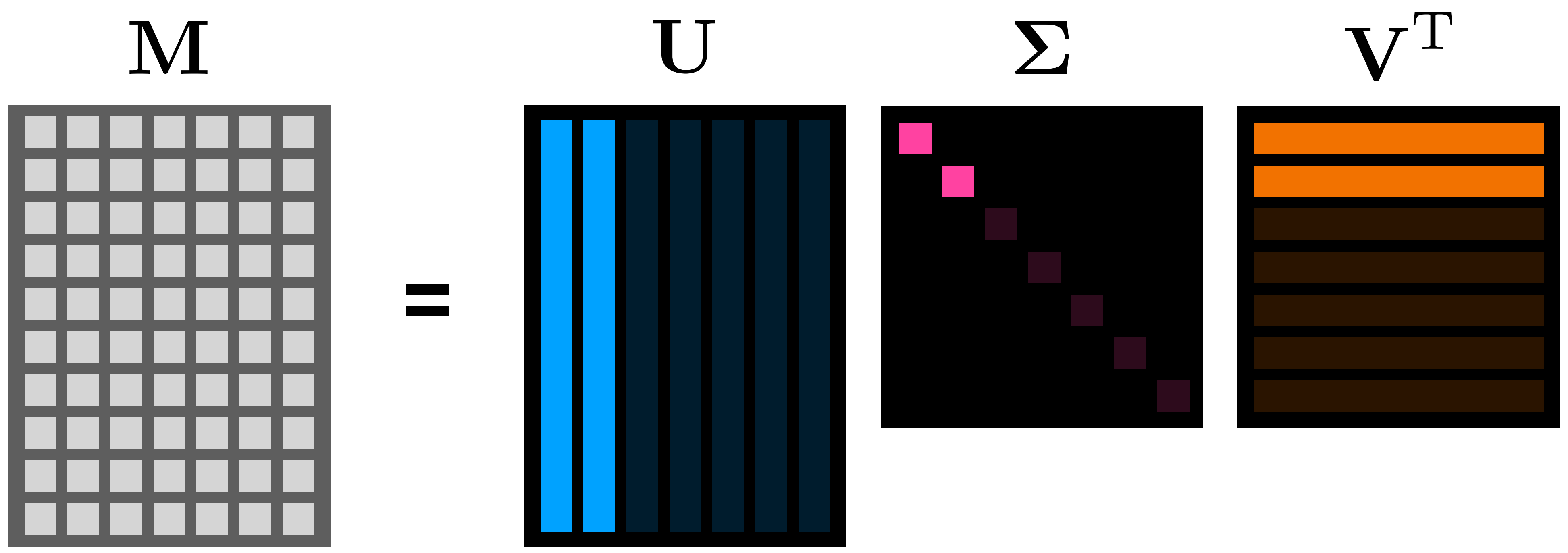
2-dimensional PCA



Singular Value Decomposition

- Any m×n matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V^T** with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.

2-dimensional PCA



Singular Value Decomposition

- Any m×n matrix **A** can be written as a product of three matrices (decomposition) **U**, **Σ**, and **V**$^{\mathbf{T}}$ with the following properties:
  - $\mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal, (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$)
  - $\mathbf{\Sigma}$ is diagonal.

2-dimensional PCA

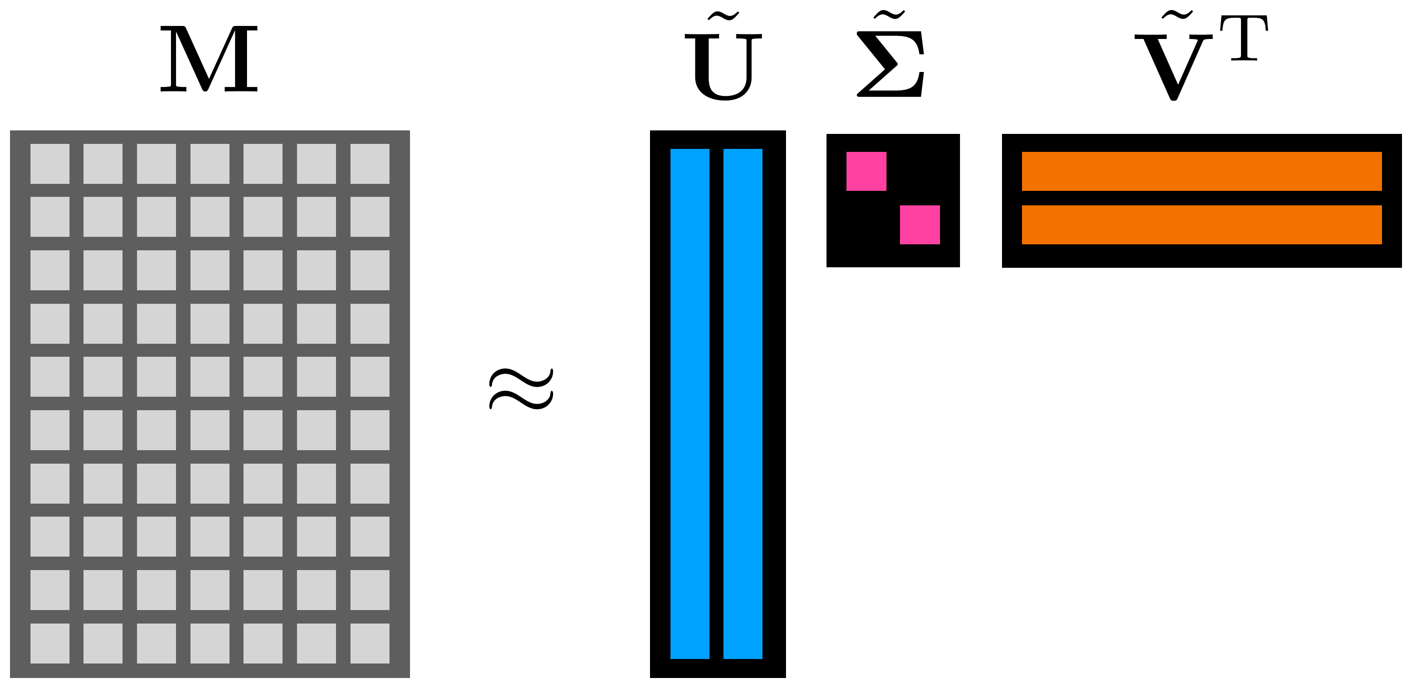
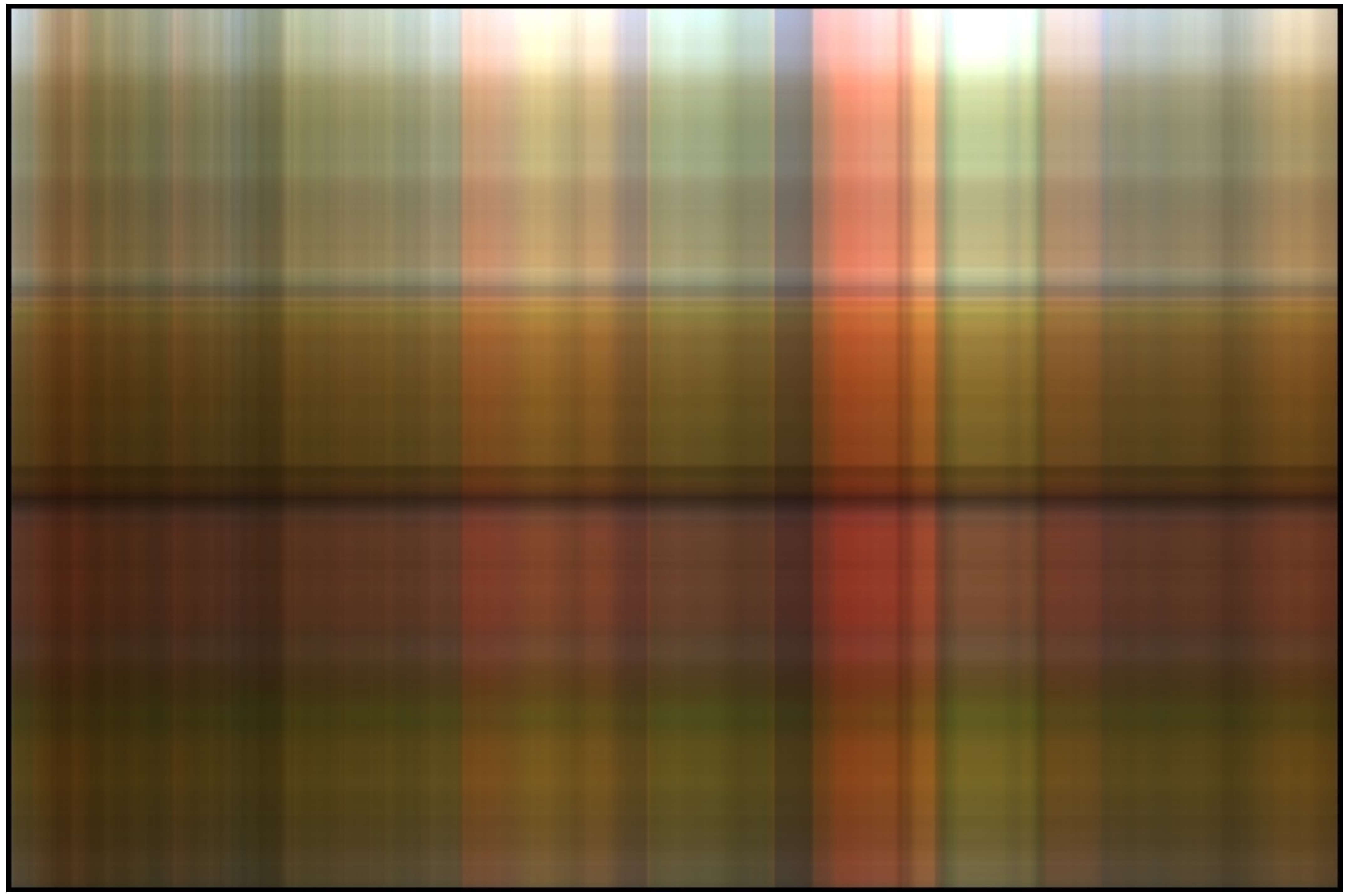


Image reconstruction with d-dimensional SVD

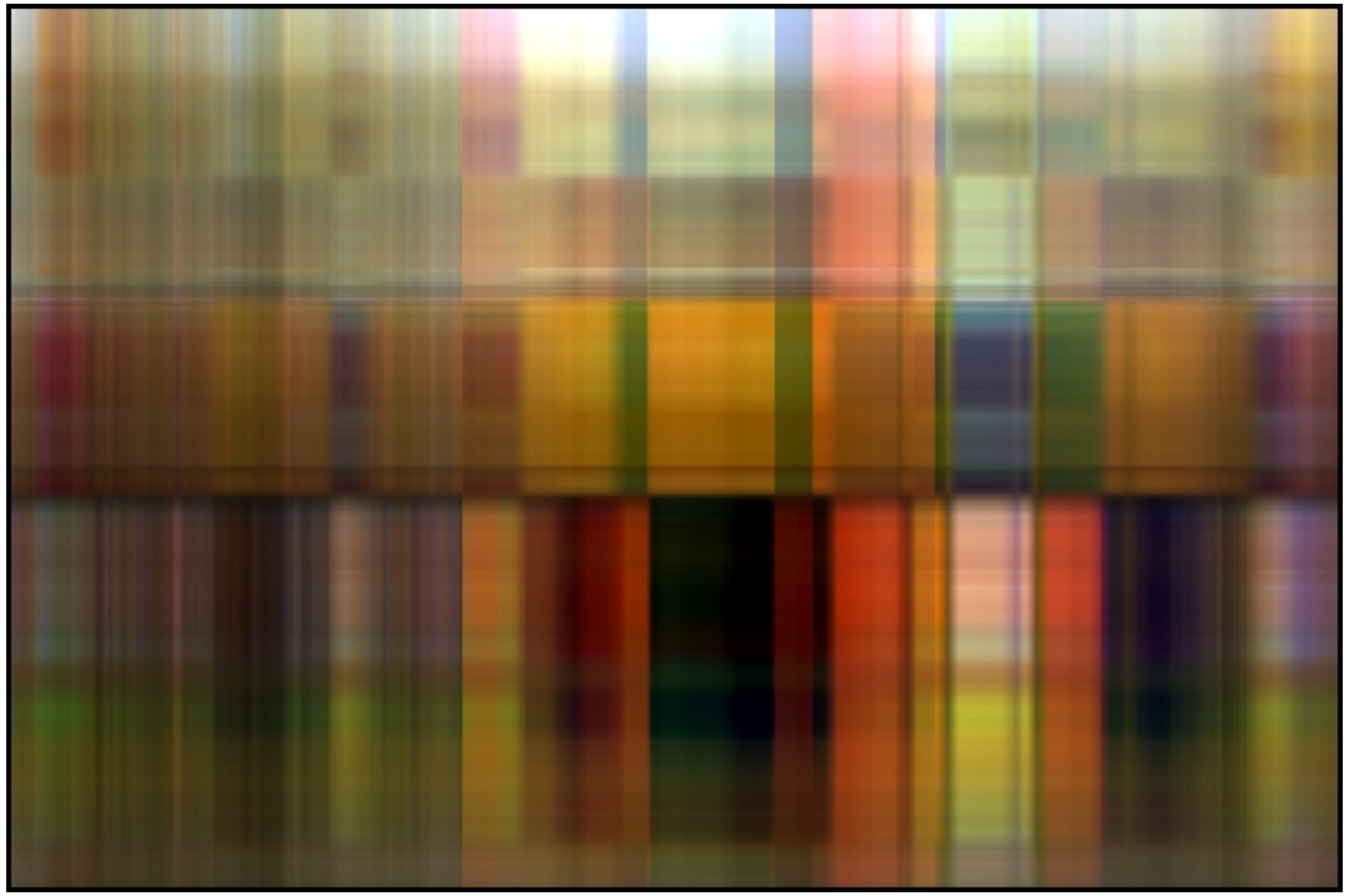
d = 1



https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

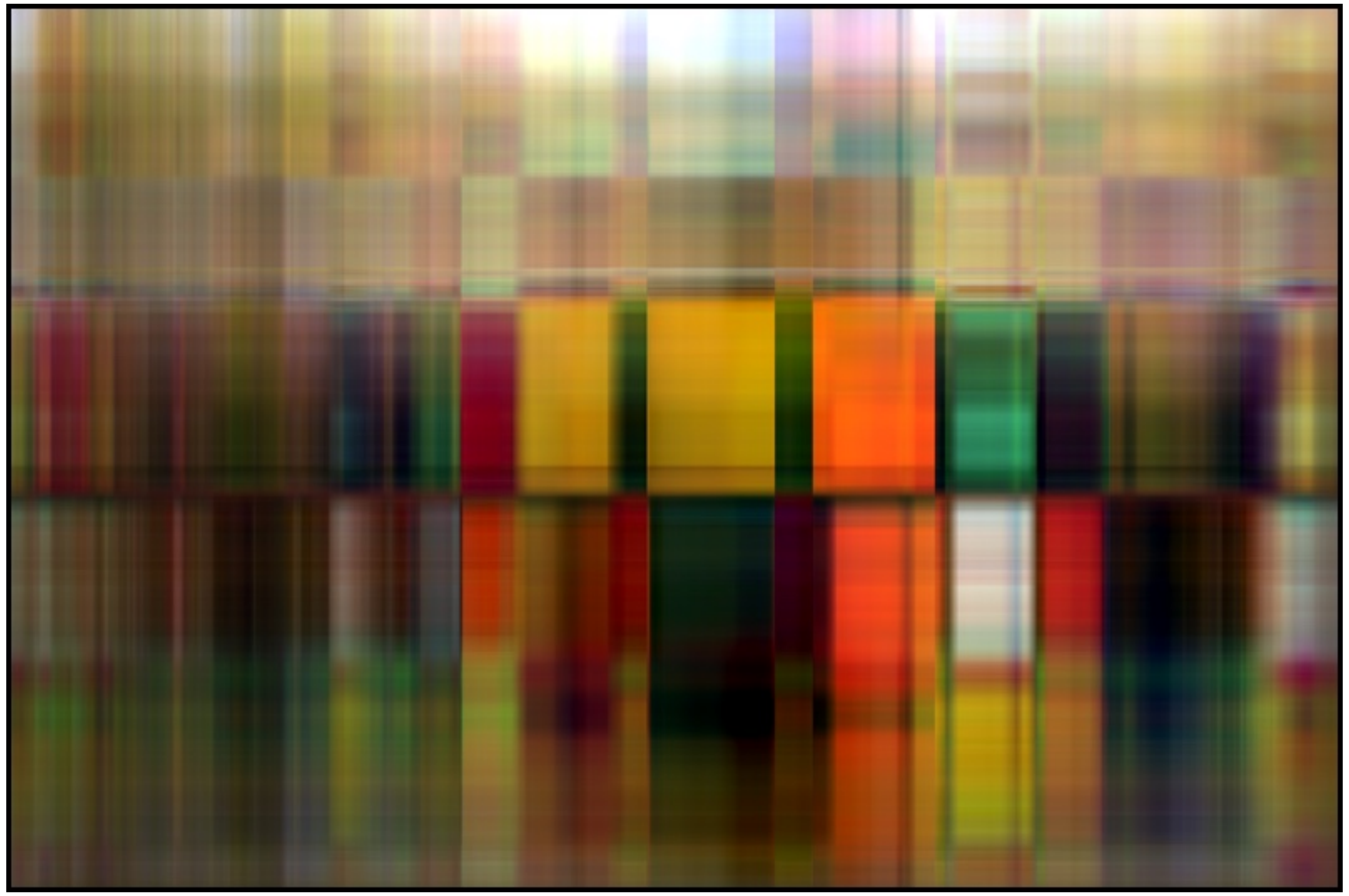
d = 2



https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

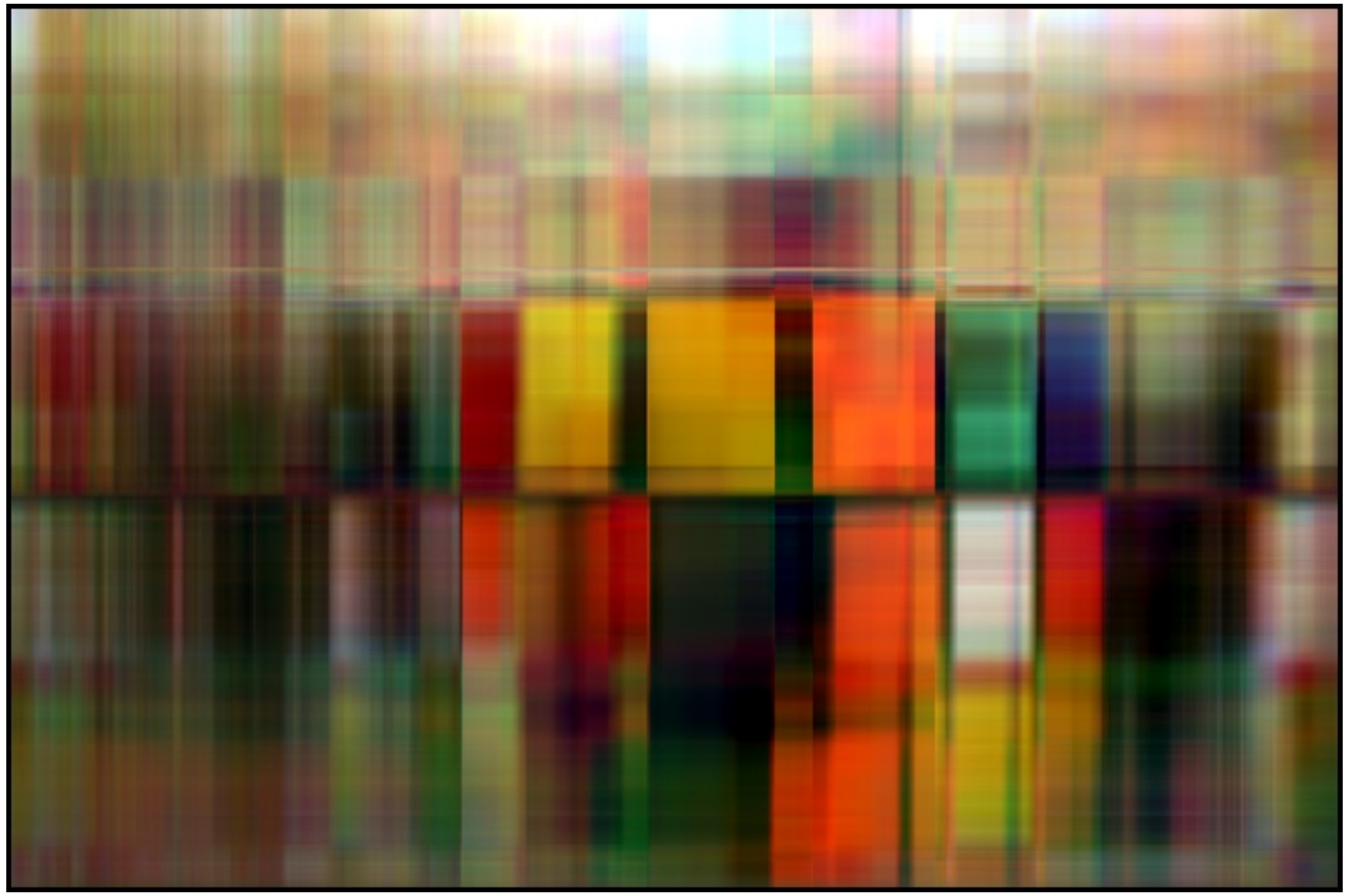
d = 3



https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

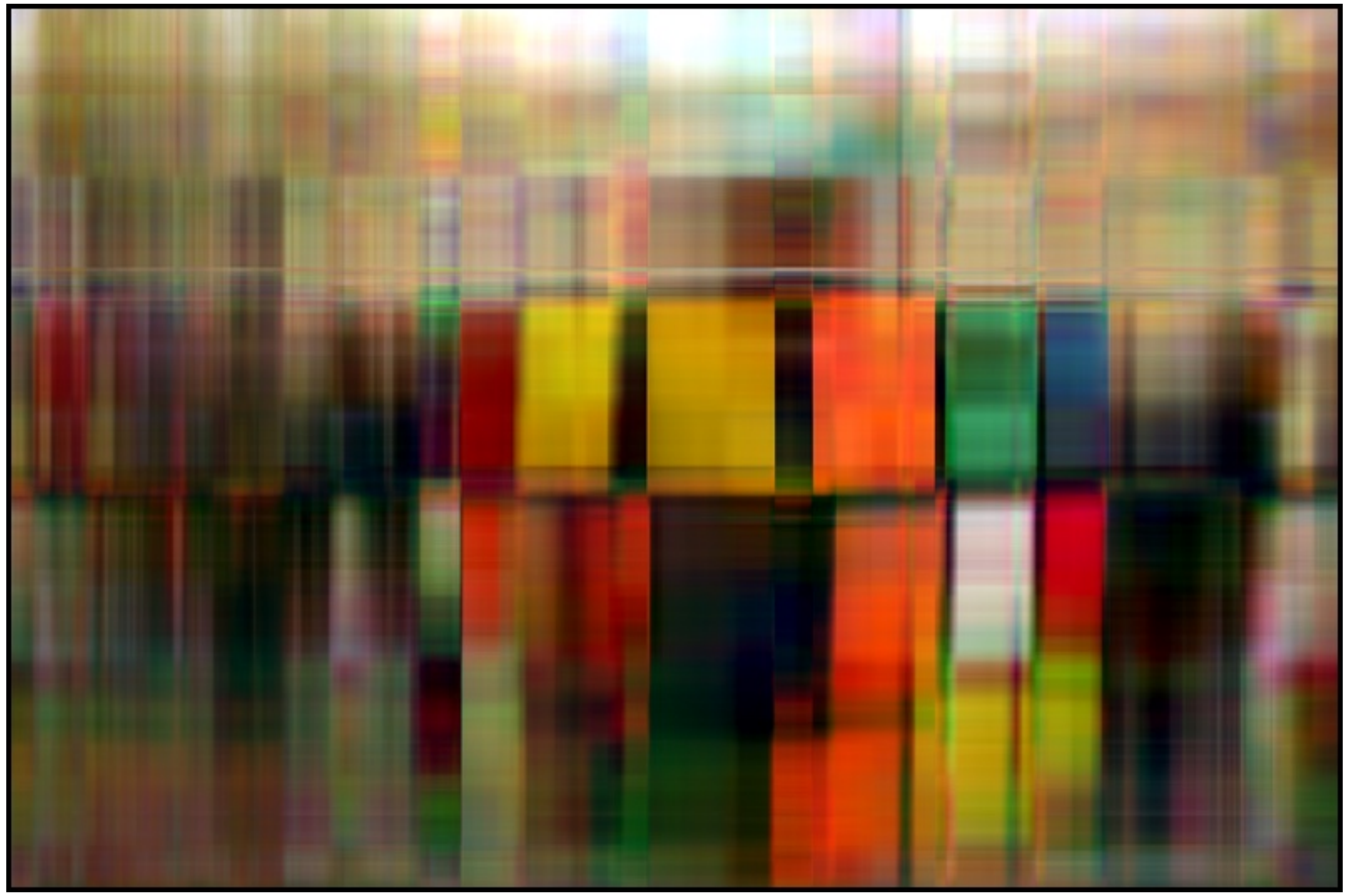
d = 4



https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

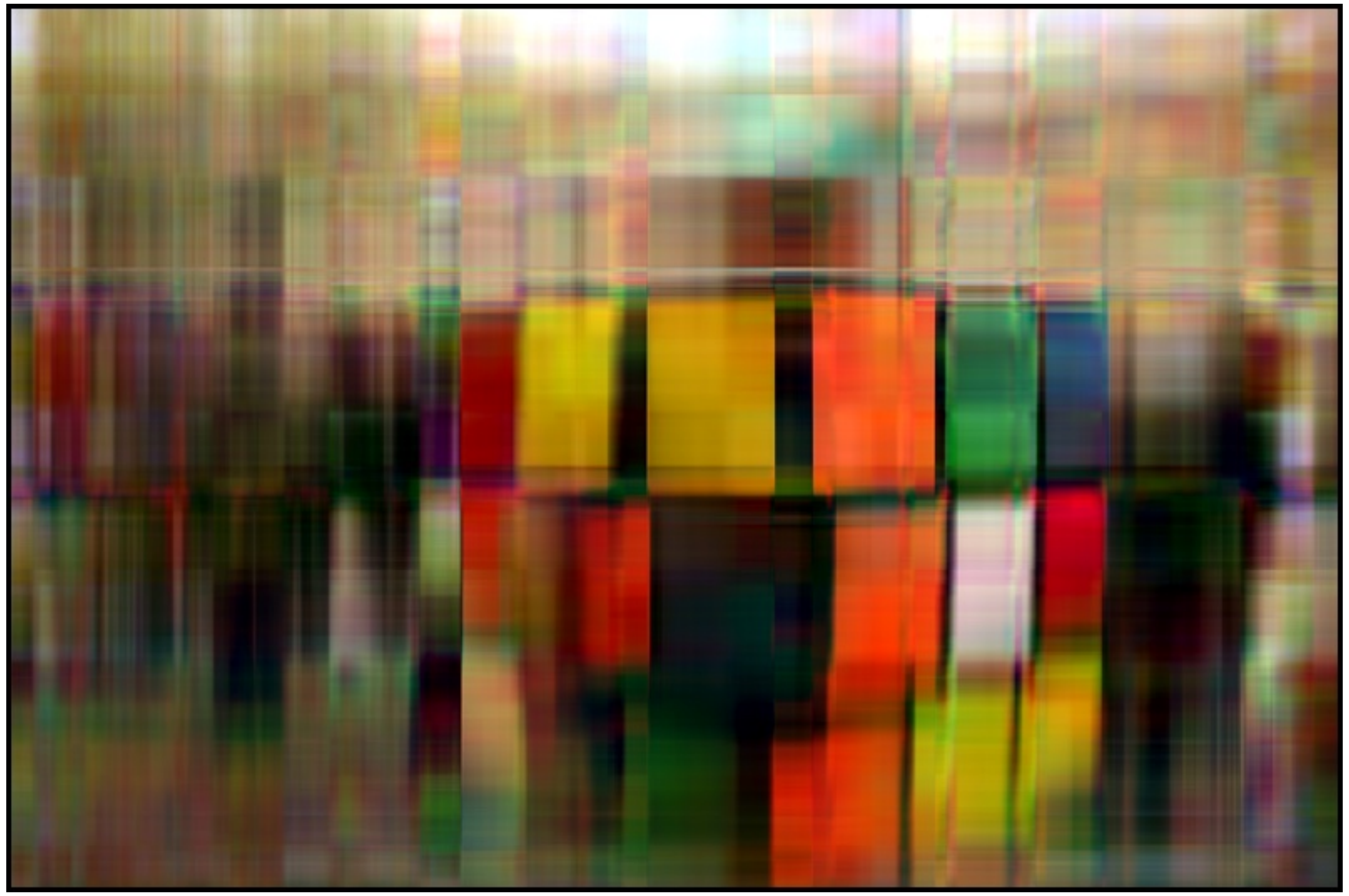
d = 5



https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

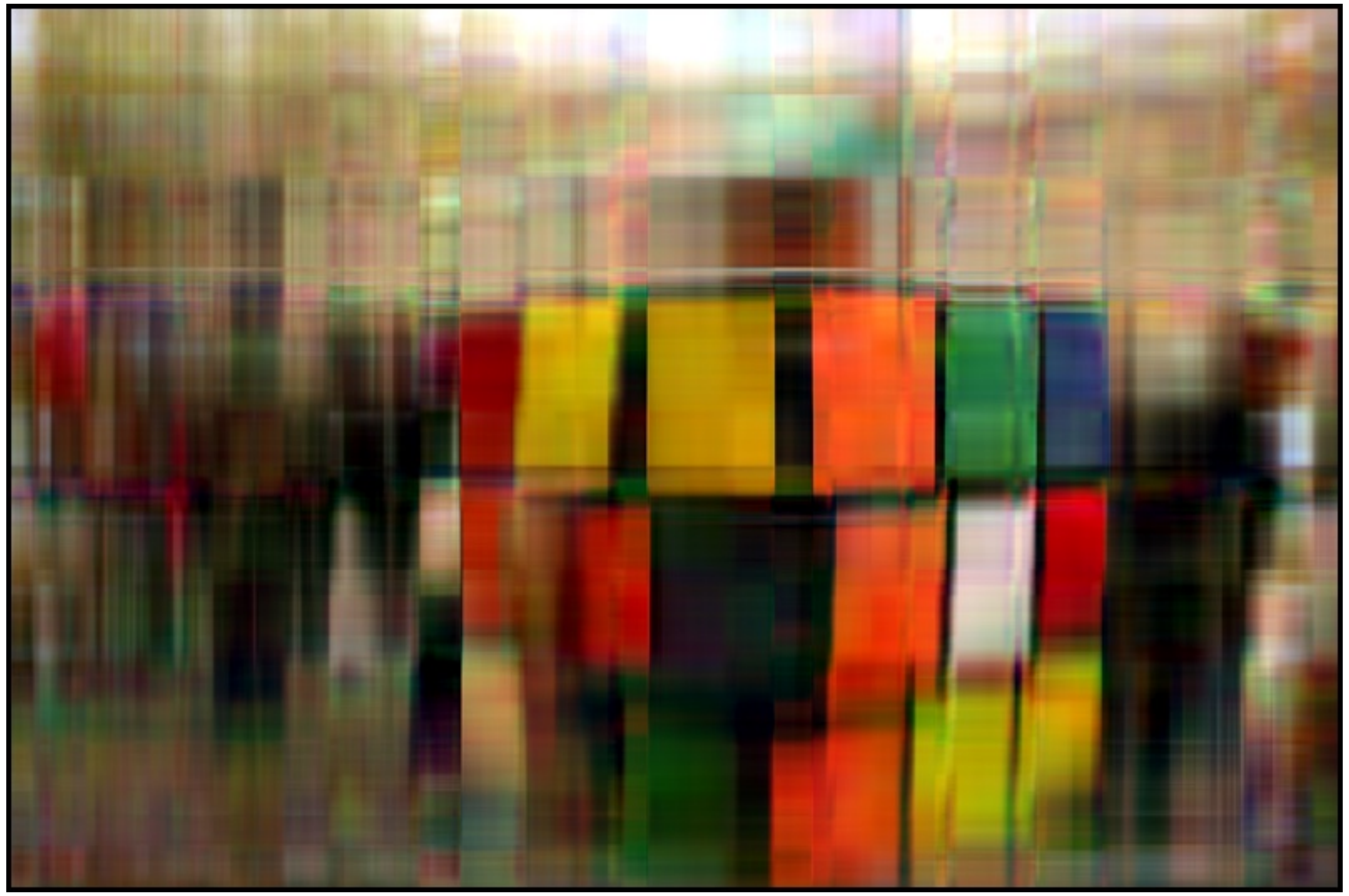
d = 6



https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

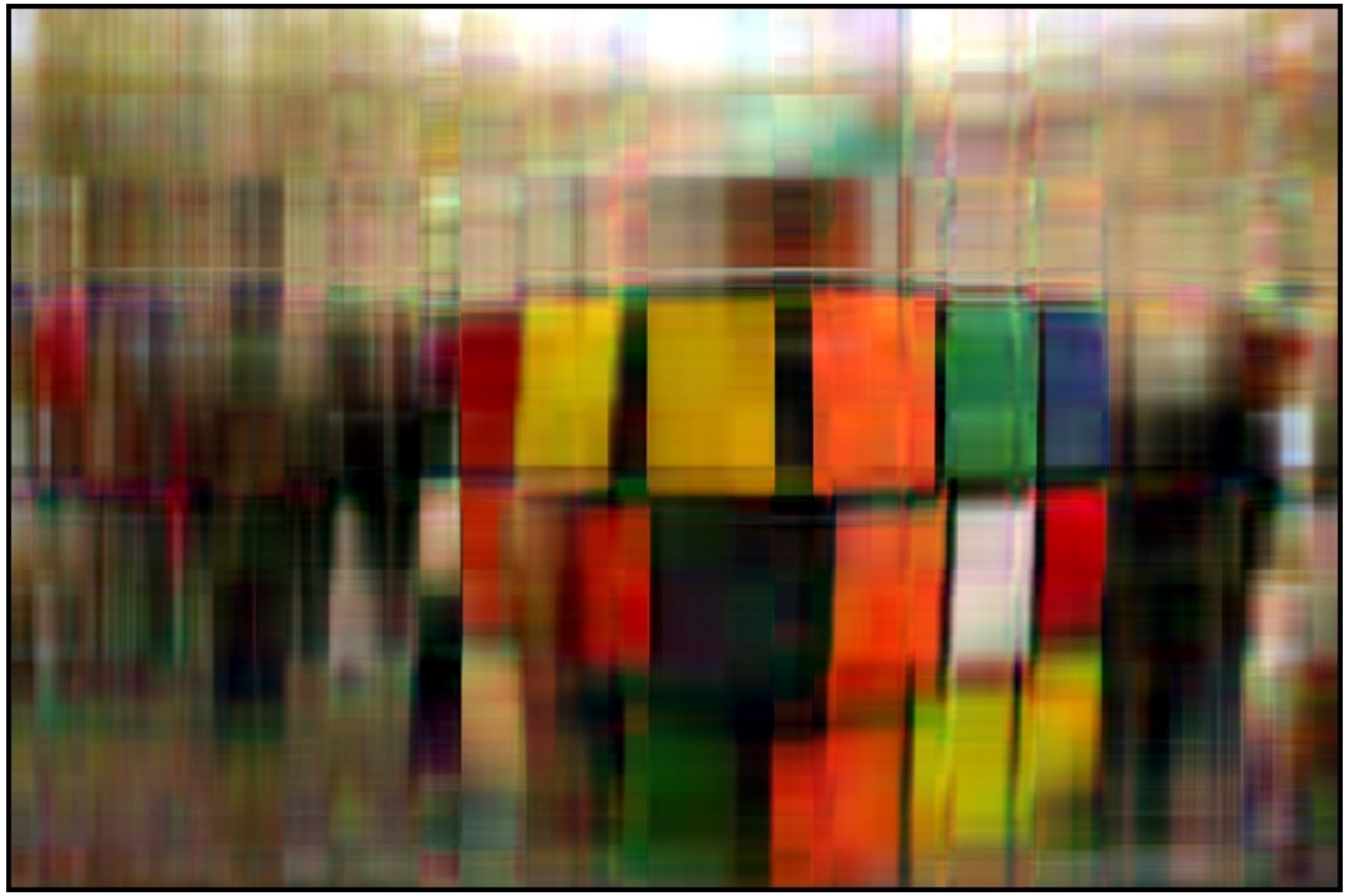
d = 7



https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 7



GPT-4o: describe what you see in this image

The image appears to show a large Rubik's Cube-like structure in an outdoor setting, surrounded by blurred figures of people, likely indicating a public or event space. The cube has vividly colored squares on each side, similar to a traditional Rubik's Cube, with colors like orange, yellow, red, green, blue, and white visible. The image is heavily distorted with a pixelated or glitch-like effect, making the details somewhat unclear. The background seems to depict an outdoor environment, possibly with trees or structures in the distance.

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 8

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 9

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 10

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 12

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 14

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 14

GPT-4o: describe what you see in this image

The image appears to show a person dressed as a giant Rubik's Cube, with colorful squares covering the cube-shaped costume. The background seems to be outdoors, with other people and possibly a fence or wall visible in a slightly blurred and distorted manner. The overall quality of the image includes digital noise or artifacts, which may distort the fine details.

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 16

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 18

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 20

https://timbaumann.info/svd-image-compression-demo/

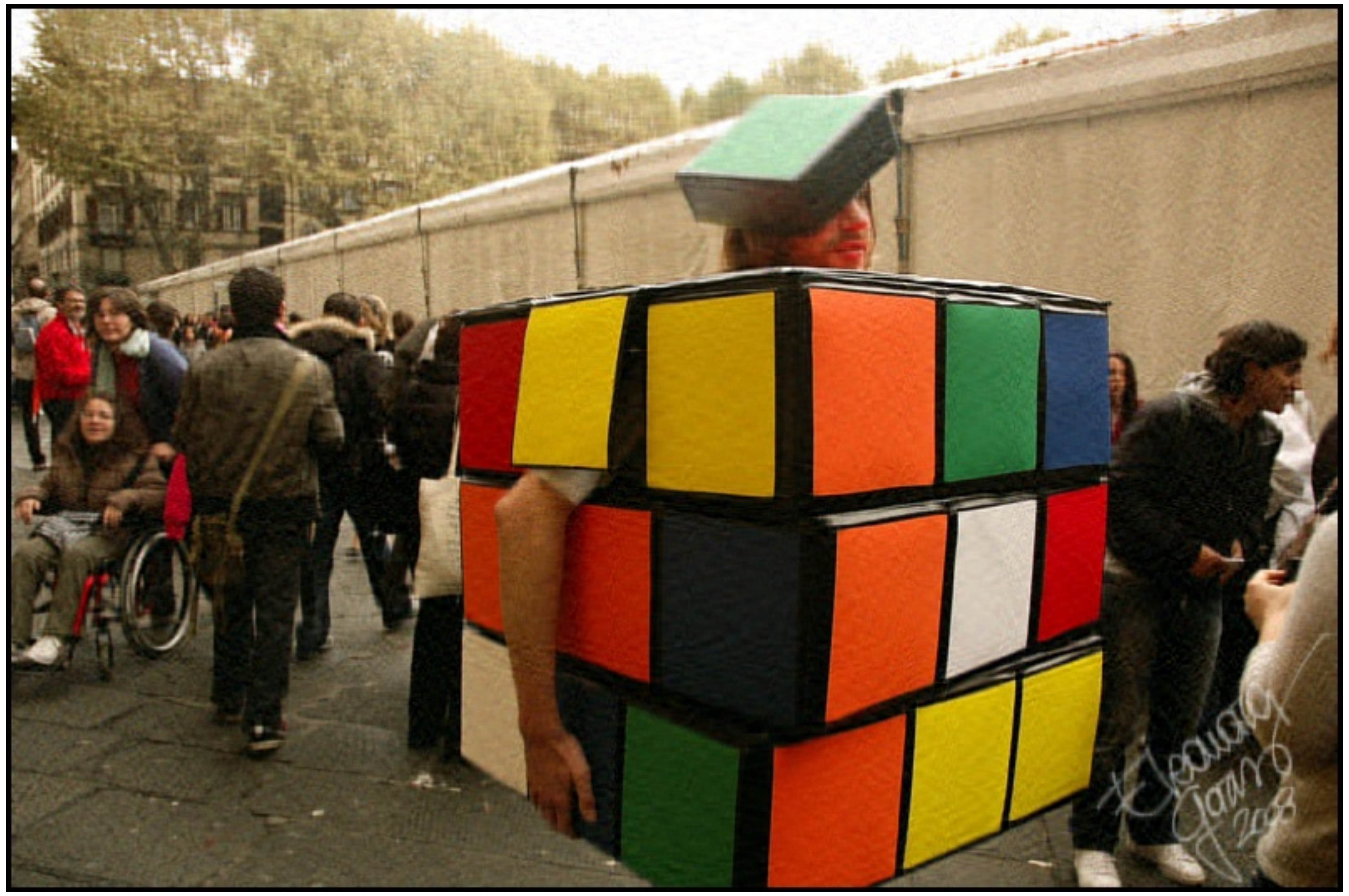
Image reconstruction with d-dimensional SVD

d = 50

https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 100

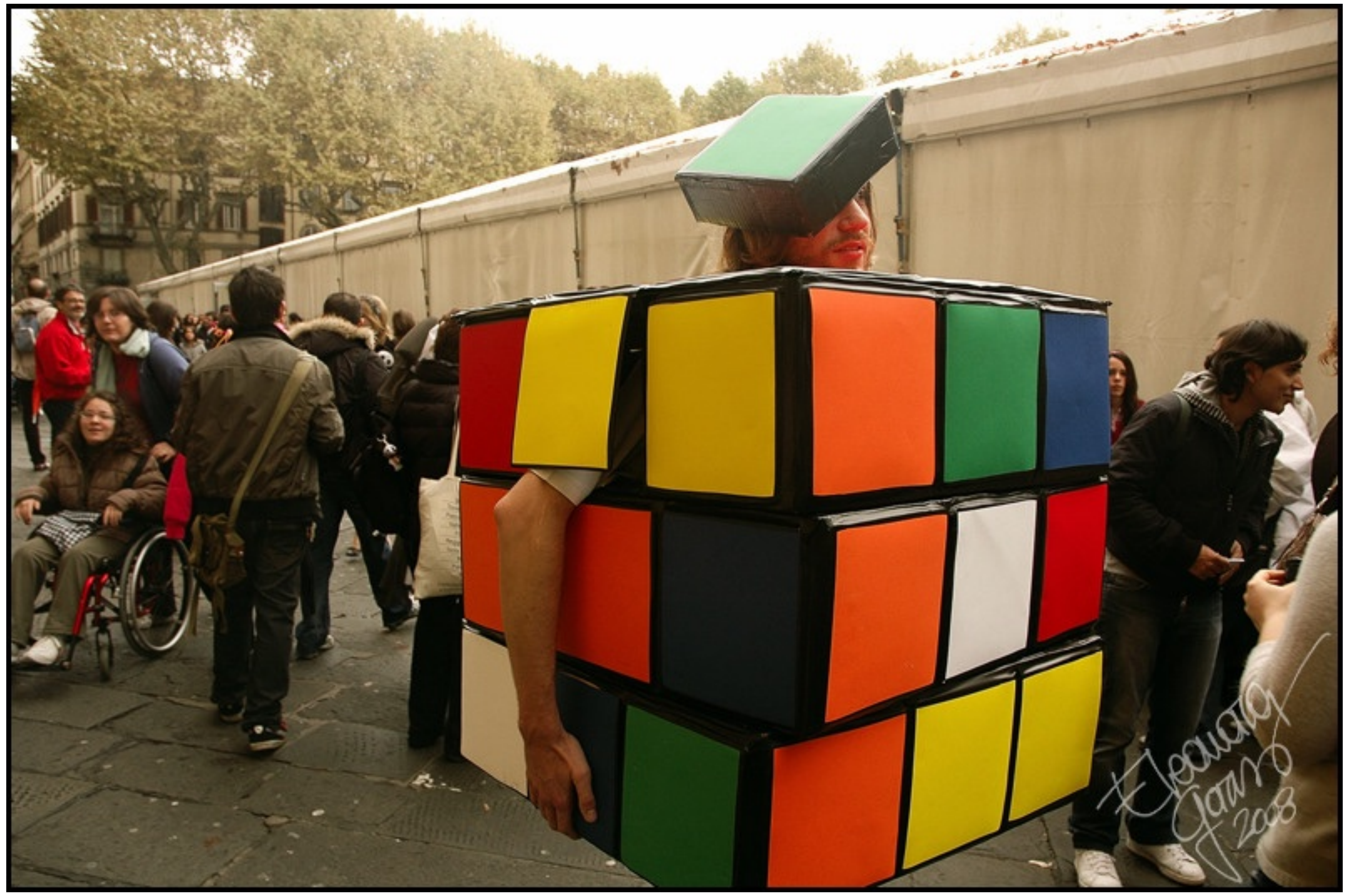


https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 532

(original size)

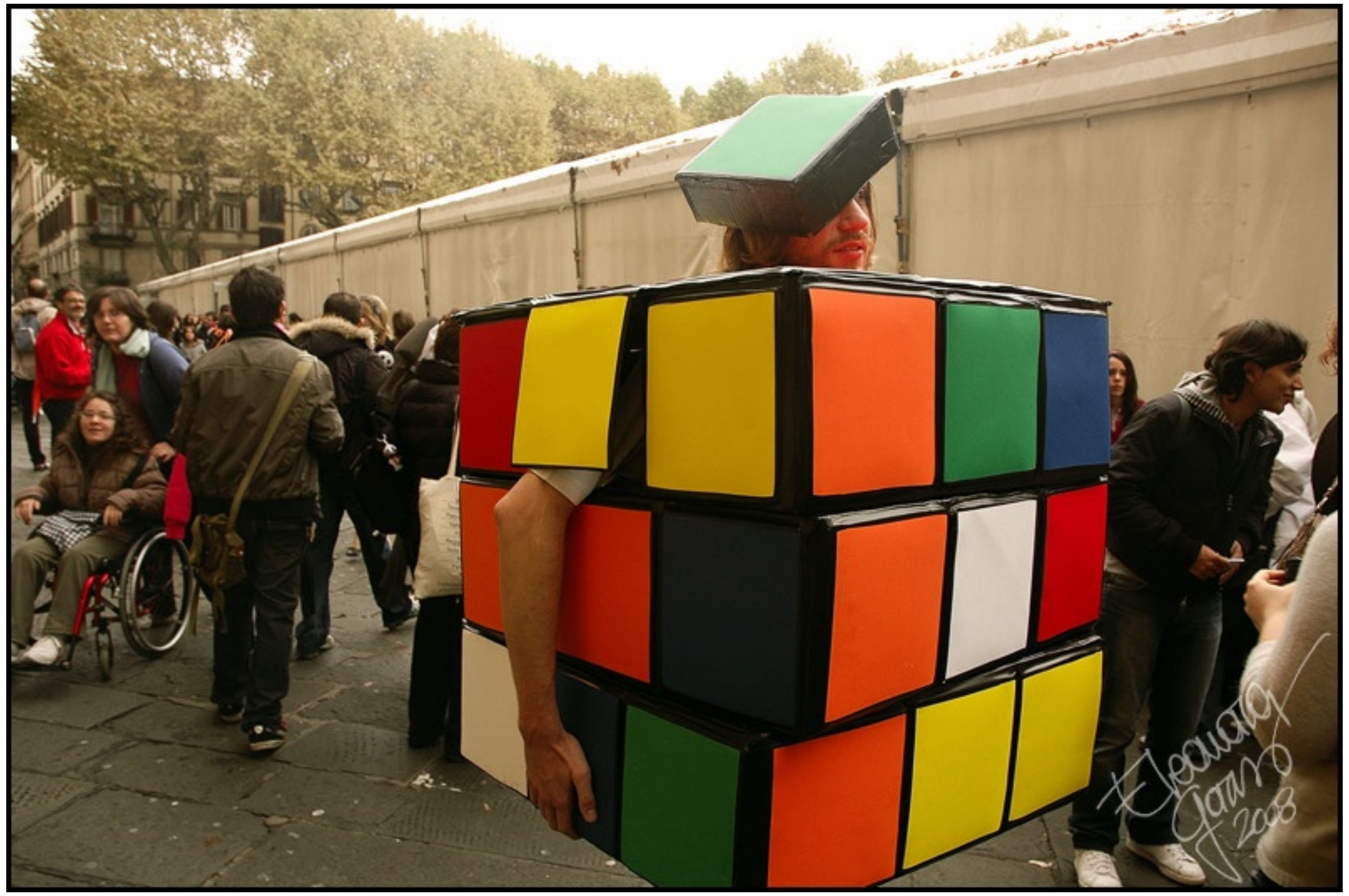


https://timbaumann.info/svd-image-compression-demo/

Image reconstruction with d-dimensional SVD

d = 532

(original size)

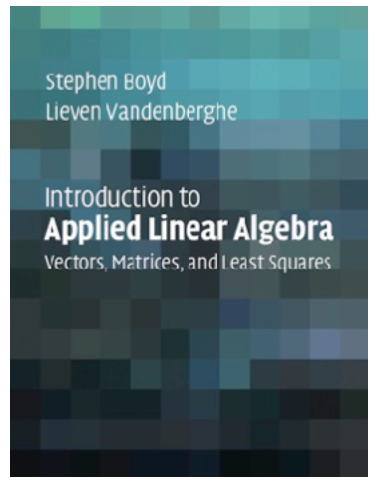


GPT-4o: describe what you see in this image

The image shows a person dressed in a creative and colorful costume resembling a Rubik's Cube. The costume is made to look like a large, three-dimensional Rubik's Cube, with panels in various colors such as red, yellow, green, orange, blue, and white. The person's arms and face are visible, and the top of the cube has a section open, with part of the costume extending upwards, possibly representing a corner of the cube. The individual appears to be in a public space, surrounded by other people, some of whom are standing, and one person in a wheelchair. The background features trees and a long white tent-like structure, suggesting this might be an outdoor event or festival. The image is signed with "Eleonora Jans 2008" in the bottom right corner.

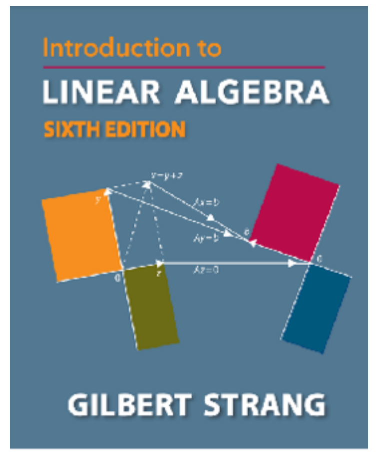
https://timbaumann.info/svd-image-compression-demo/

Resources



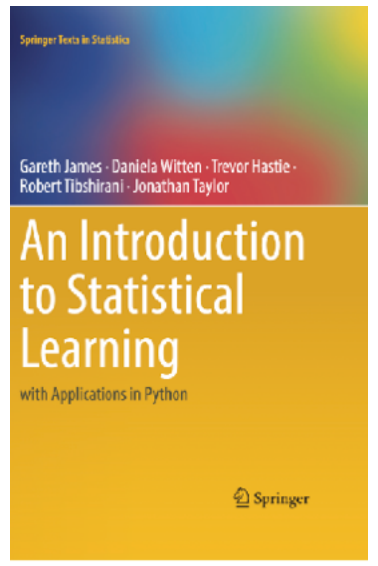
Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares
Stephen Boyd and Lieven Vandenberghe

https://web.stanford.edu/~boyd/vmls/vmls.pdf



Introduction to Linear Algebra
Gilbert Strang

https://math.mit.edu/~gs/linearalgebra/ila6/indexila6.html



Introduction to Statistical Learning
James, Witten, Hastie, Tibshirani, Taylor

https://www.statlearning.com/

Essence of Linear Algebra