CropS_545 - Statistical Genomics

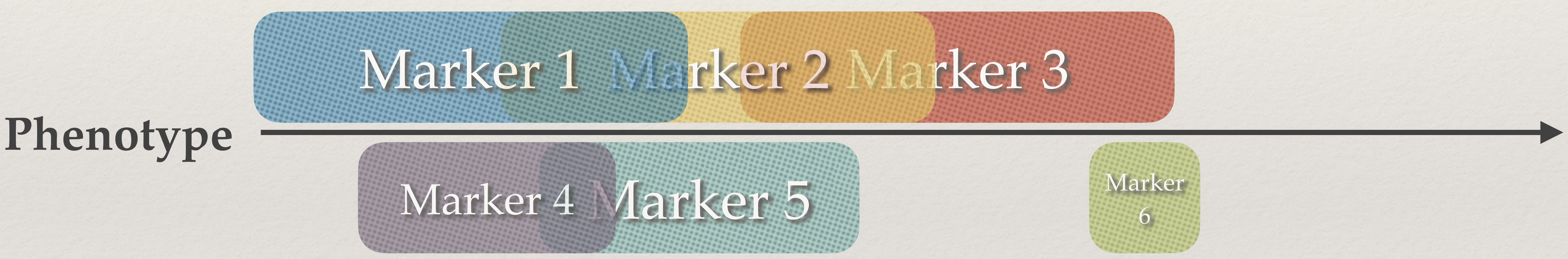
Principle Component Analysis (PCA)

James Chen

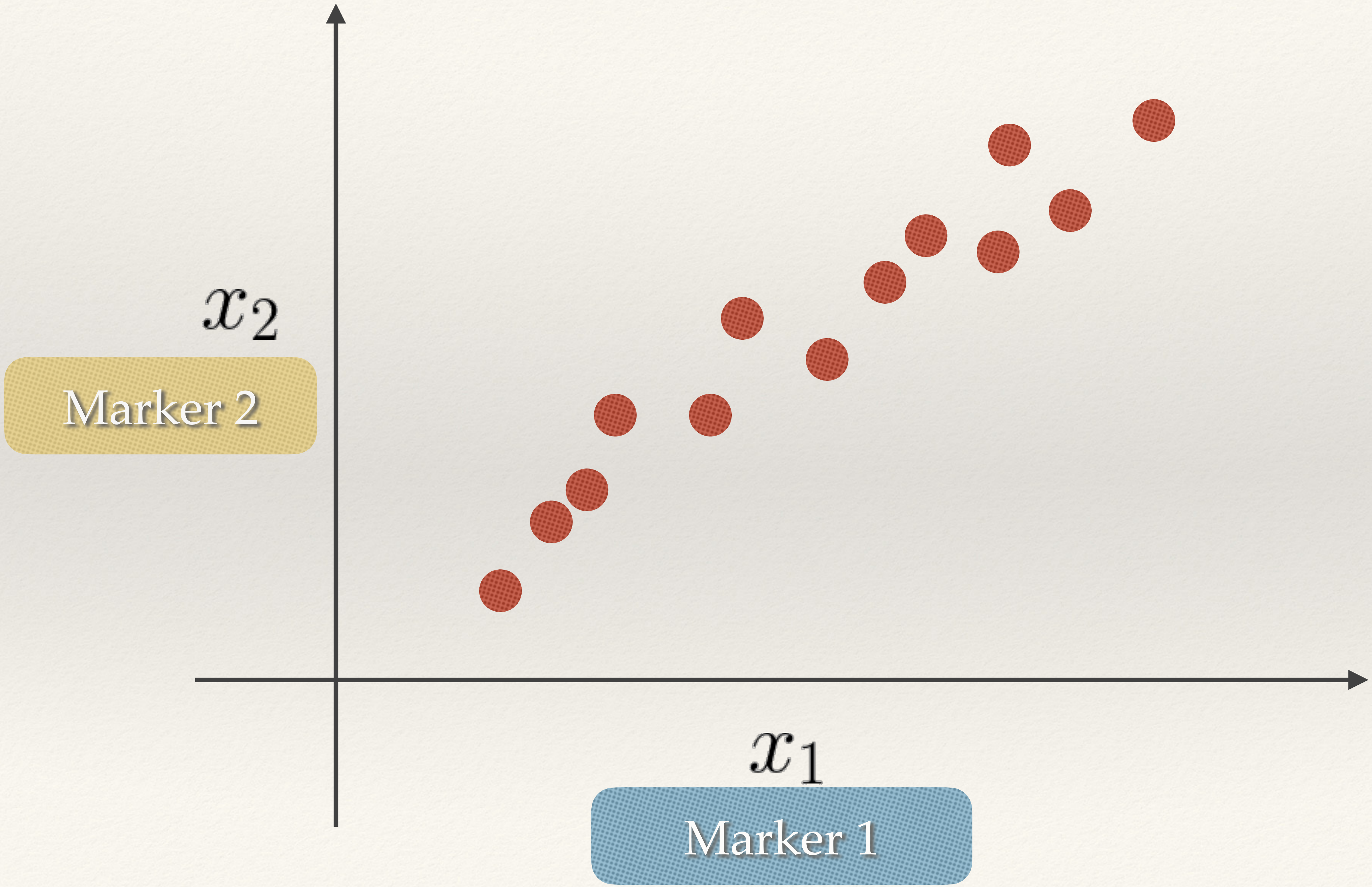
Why PCA?

- Imaging you have a dataset with 5k markers and 200 individuals, how can we model it?
  - Overfitting
    - Too many estimated parameters
    - Low degree of freedom
    - Some effect of markers confound each other
  - Too many dimensions
    - Unable to visualize it

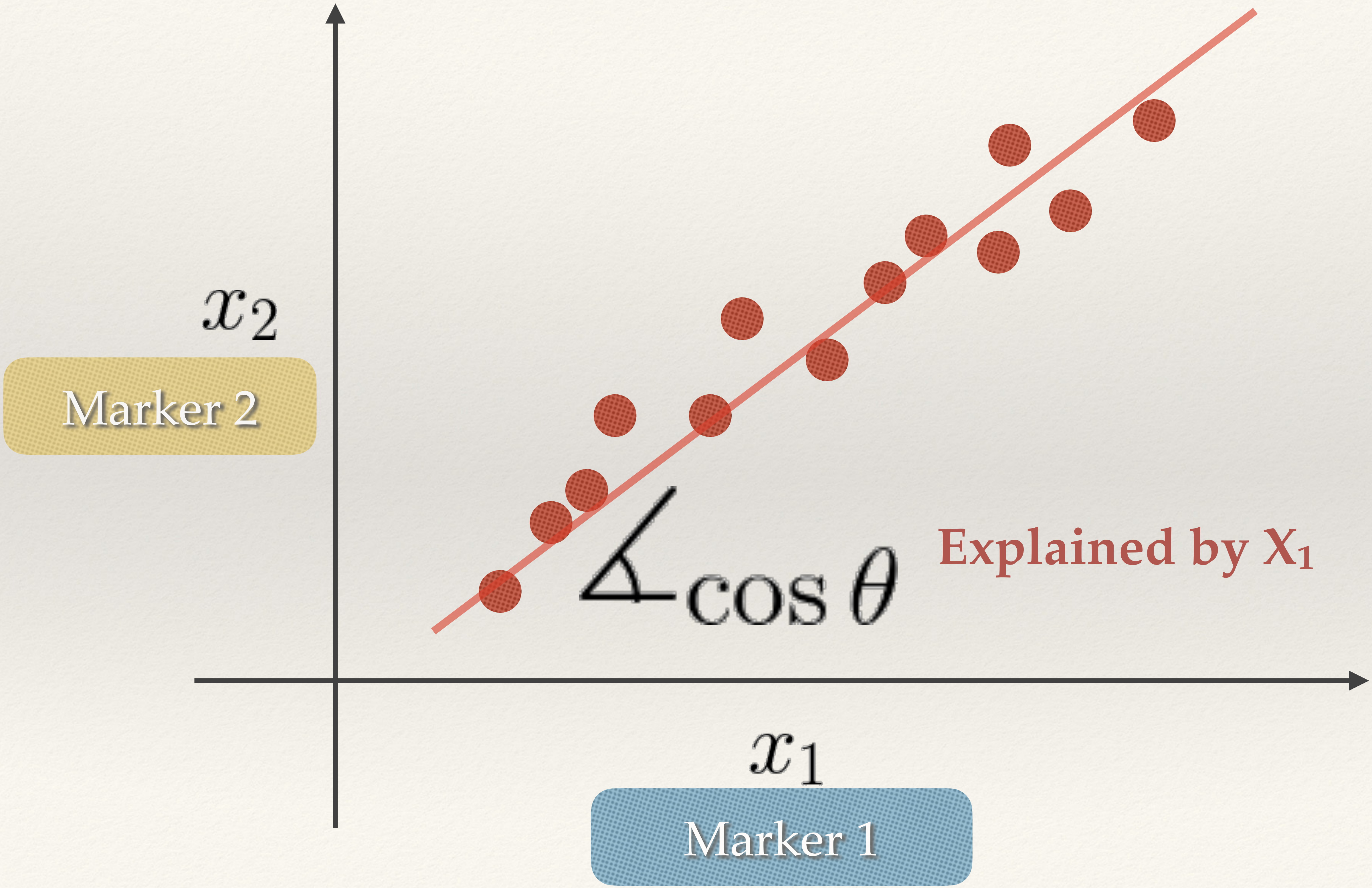
Variation of Phenotype



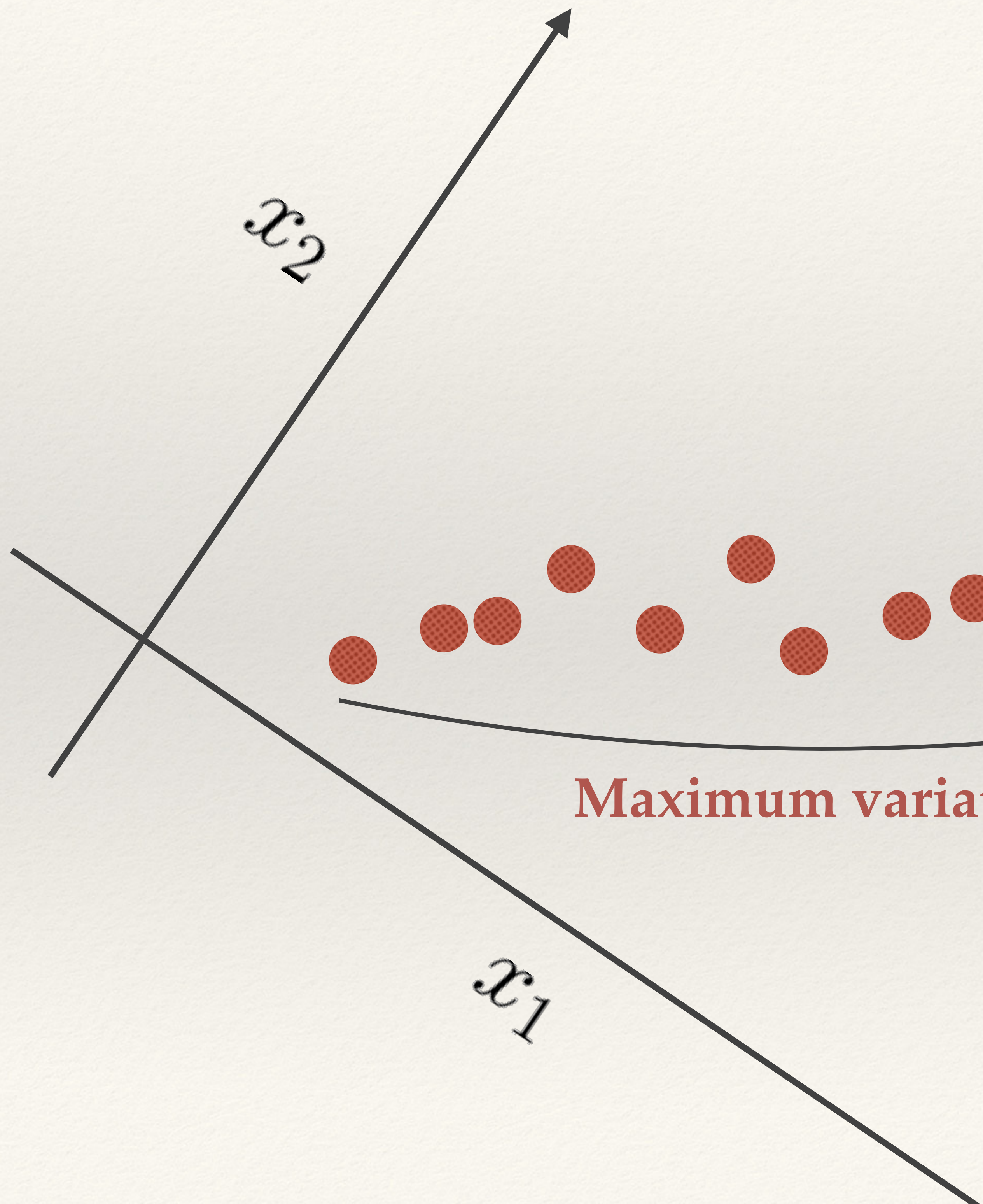
Transformation



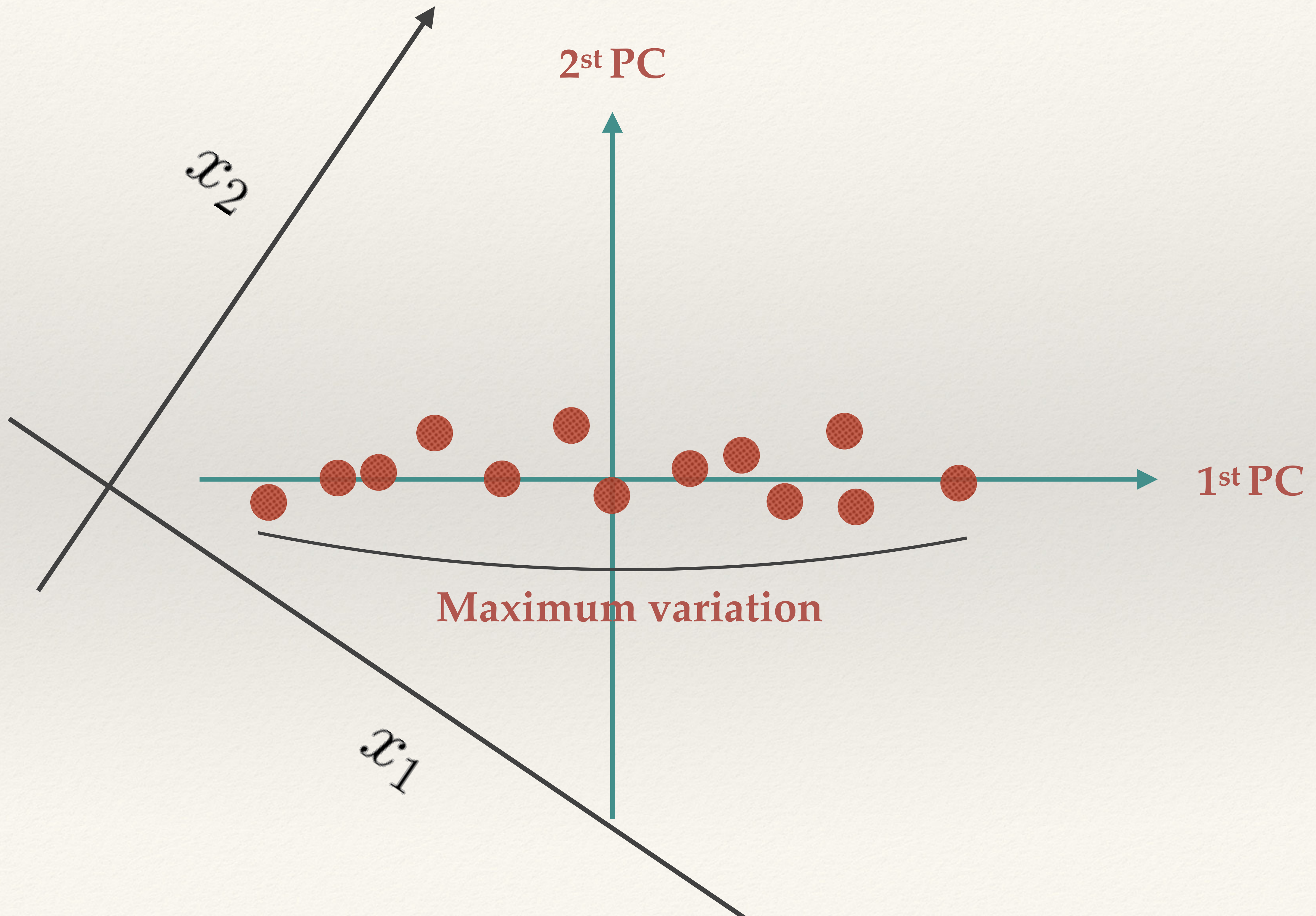
Transformation



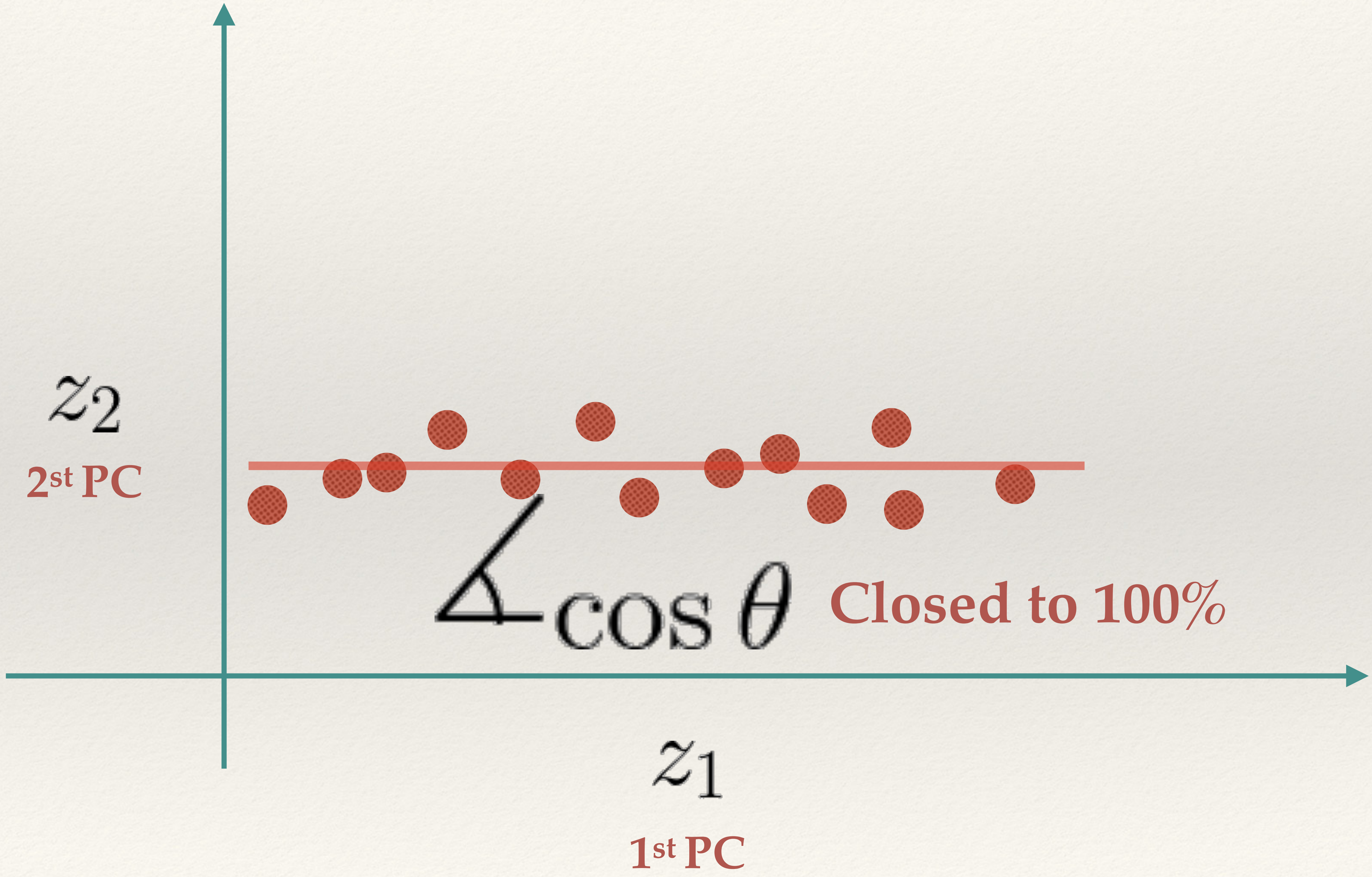
Transformation



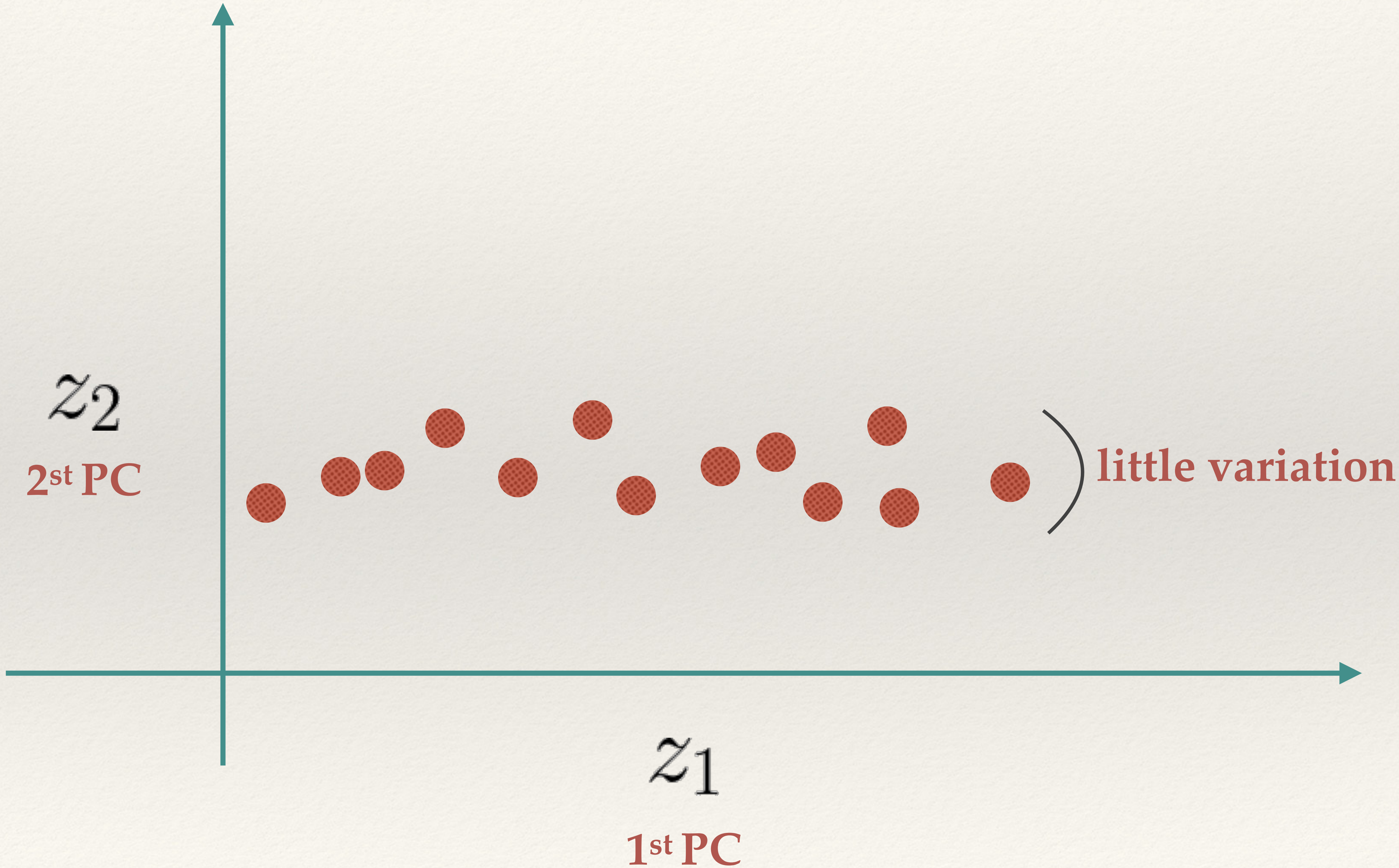
Transformation



Transformation



Transformation



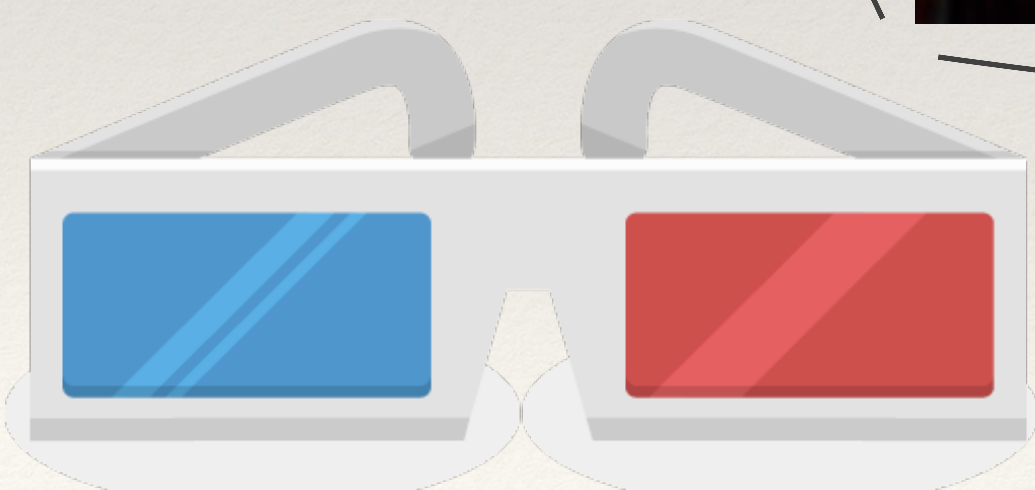
Transformation

Dimension: 2 -> 1

Does the depth matter?

x_2

x_1



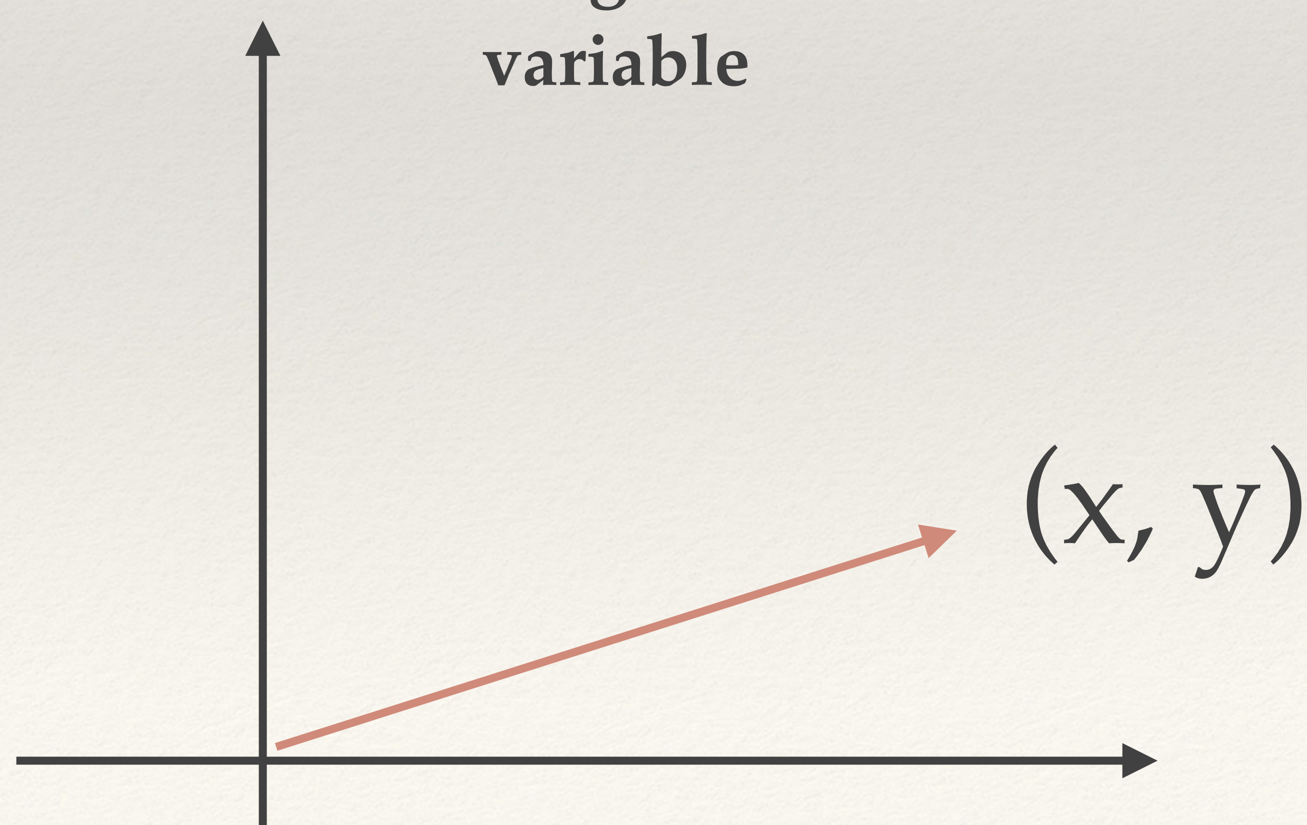
$x_3 : Depth$

Pulp fiction (1994)

Transformation

$$\begin{bmatrix} x \\ y \end{bmatrix}$$

Original variable



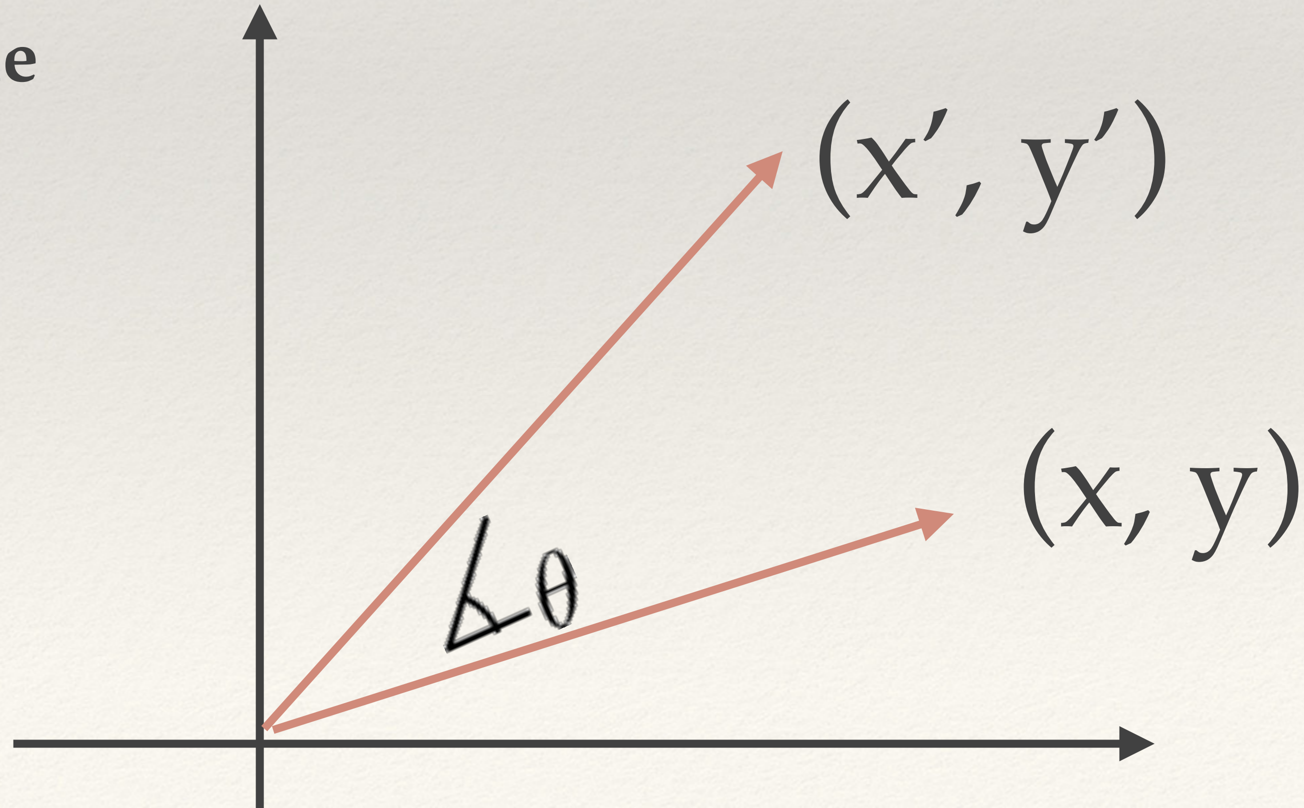
Transformation

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Transformed variable

Project matrix

Original variable



Transformation

$$z = u^T x \quad \forall \;\; E(x) = 0; E(z) = 0$$

Transformed variable (Principle Component)

Projection Matrix

Original variable

Maximum variance

$$z = u^T x \quad \forall \; E(x) = 0; \boxed{E(z) = 0}$$

$$\max_u \Sigma_z = \max_u E(zz^T)$$

Maximum variance

$$\boxed{z = u^T x} \quad \forall \;\; E(x) = 0; E(z) = 0$$

$$\max_u \Sigma_z = \max_u E(zz^T)$$

$$= \max_u E(u^T x x^T u)$$

Maximum variance

$$z = u^T x \quad \forall \;\; E(x) = 0; E(z) = 0$$

$$\max_u \Sigma_z = \max_u E(zz^T)$$
$$= \max_u E(u^T x x^T u)$$
$$= \max_u u^T \Sigma_x u$$

Maximum variance

$$\max_u \Sigma_z = \max_u E(zz^T)$$
$$= \max_u E(u^T x x^T u)$$
$$= \max_u u^T \Sigma_x u$$

Lagrange Multiplier

$$f(x) \; \forall g(x)$$
$$= f(x) + \lambda \, g(x)$$

Where g(x) is the **constrain** for X and λ is a **constant**

Maximum variance

$$\max_u \Sigma_z = \max_u E(zz^T)$$
$$= \max_u E(u^T x x^T u)$$
$$= \max_u u^T \Sigma_x u$$

Constrain

$$u^T u = 1$$
$$g(u) = u^T u - 1$$

Lagrange Multiplier

$$f(x) \, \forall g(x)$$
$$= f(x) + \lambda \, g(x)$$

Where g(x) is the **constrain** for X and λ is a **constant**

Maximum variance

$$\max_u \Sigma_z = \max_u E(zz^T)$$

$$= \max_u E(u^T x x^T u)$$

$$= \max_u u^T \Sigma_x u$$

$$= \max_u \underbrace{u^T \Sigma_x u}_{\text{f(u)}} - \lambda \underbrace{(u^T u - 1)}_{\text{g(u)}}$$

Lagrange Multiplier

$$f(x) \, \forall g(x)$$

$$= f(x) + \lambda \, g(x)$$

Maximum variance

$$\max_u \Sigma_z = \max_u E(zz^T)$$

$$= \max_u E(u^T x x^T u)$$

$$= \max_u u^T \Sigma_x u$$

$$= \max_u \underbrace{u^T \Sigma_x u - \lambda(u^T u - 1)}_{\text{Z(u)}}$$

$$\frac{\partial Z(u)}{\partial u} = 0$$

Maximum variance

$$= \max_u \underbrace{u^T \Sigma_x u - \lambda(u^T u - 1)}_{\text{Z(u)}}$$

$$\frac{\partial Z(u)}{\partial u} = 0$$

$$= \Sigma_x u - \lambda u \longrightarrow \Sigma_x u = \lambda u$$

Eigenvector

Eigenvector

Eigenvalue

Eigen structure

$$\Sigma_x u = \lambda u$$

$$\Sigma_x u - \lambda u$$
$$= (\Sigma_x - \lambda I)u = 0$$

$$det(\Sigma_x - \lambda I) = 0$$ **Find** λ

Eigen structure

$$\Sigma_x u = \lambda u$$

Eigenvector

Covariance of
the original variables

Eigenvalue

Eigen structure

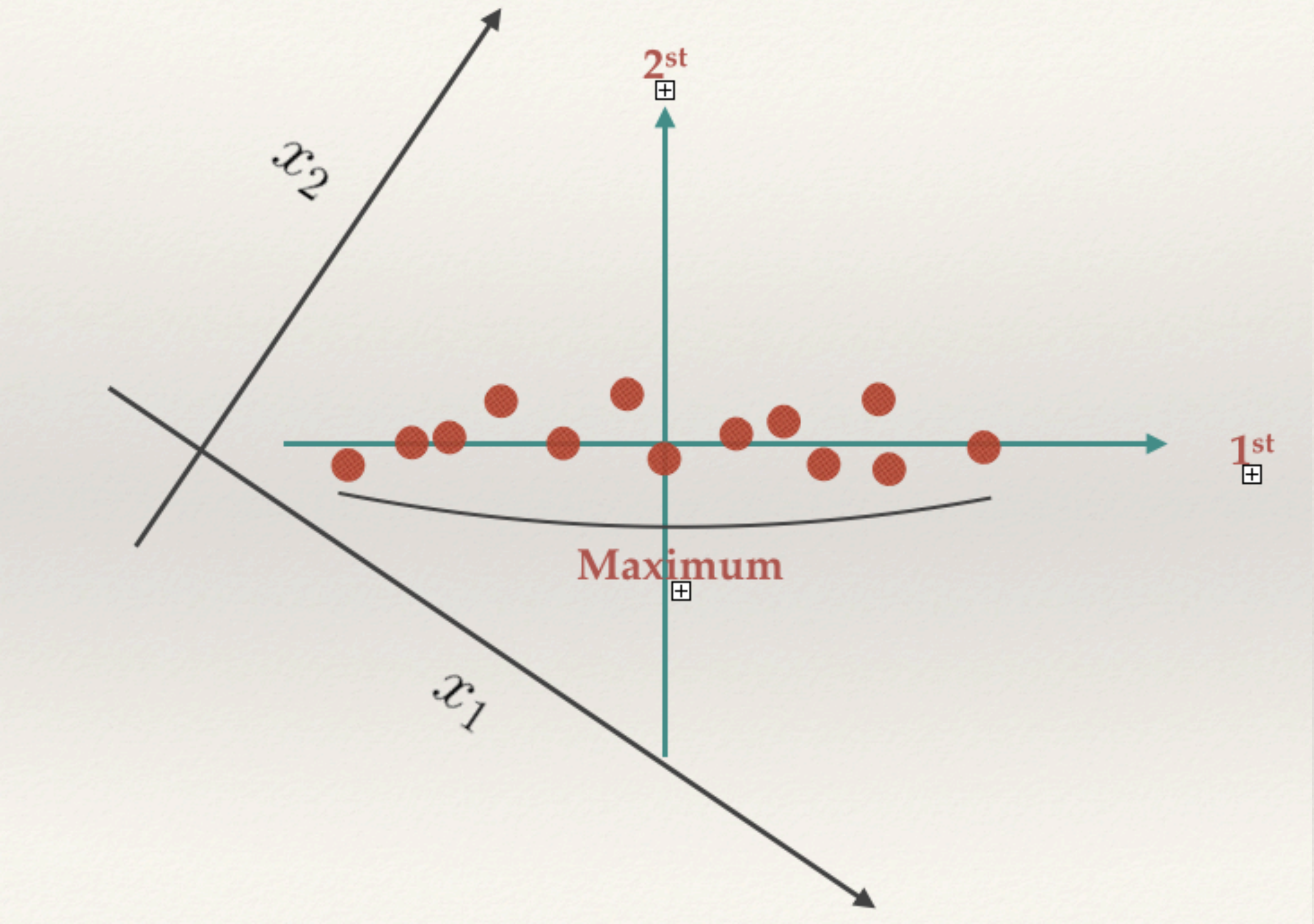
$$\Sigma_x u = \lambda u$$

Covariance of
the original variables

Eigenvector

$\because \Sigma_x$ is a symmetric matrix,
its eigenvectors would be orthogonal between each other

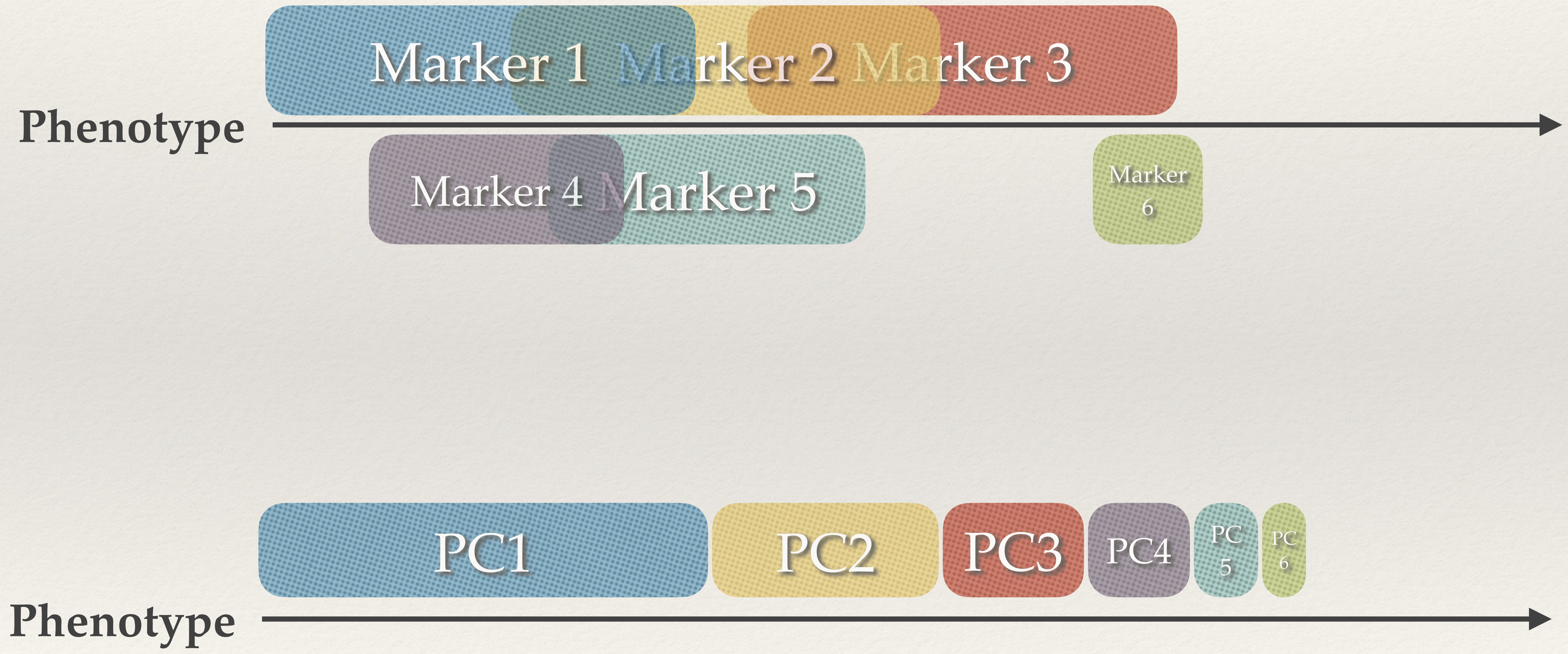
Eigen structure



$\because \Sigma_x$ **is a symmetric matrix,**

its eigenvectors would be orthogonal between each other

Variation of Phenotype



Eigen structure

$$\Sigma_x u = \lambda u$$

$$\Sigma_z = u^T \Sigma_x u$$

Eigen structure

$$\boxed{\Sigma_x u = \lambda u}$$

$$\Sigma_z = u^T \Sigma_x u$$

$$= u^T \lambda u = I\lambda$$

Eigen structure

$$\Sigma_x u = \lambda u$$

$$\Sigma_z = u^T \Sigma_x u$$

$$= u^T \lambda u = \underbrace{I}_{u^T u = 1} \lambda$$

Eigen structure

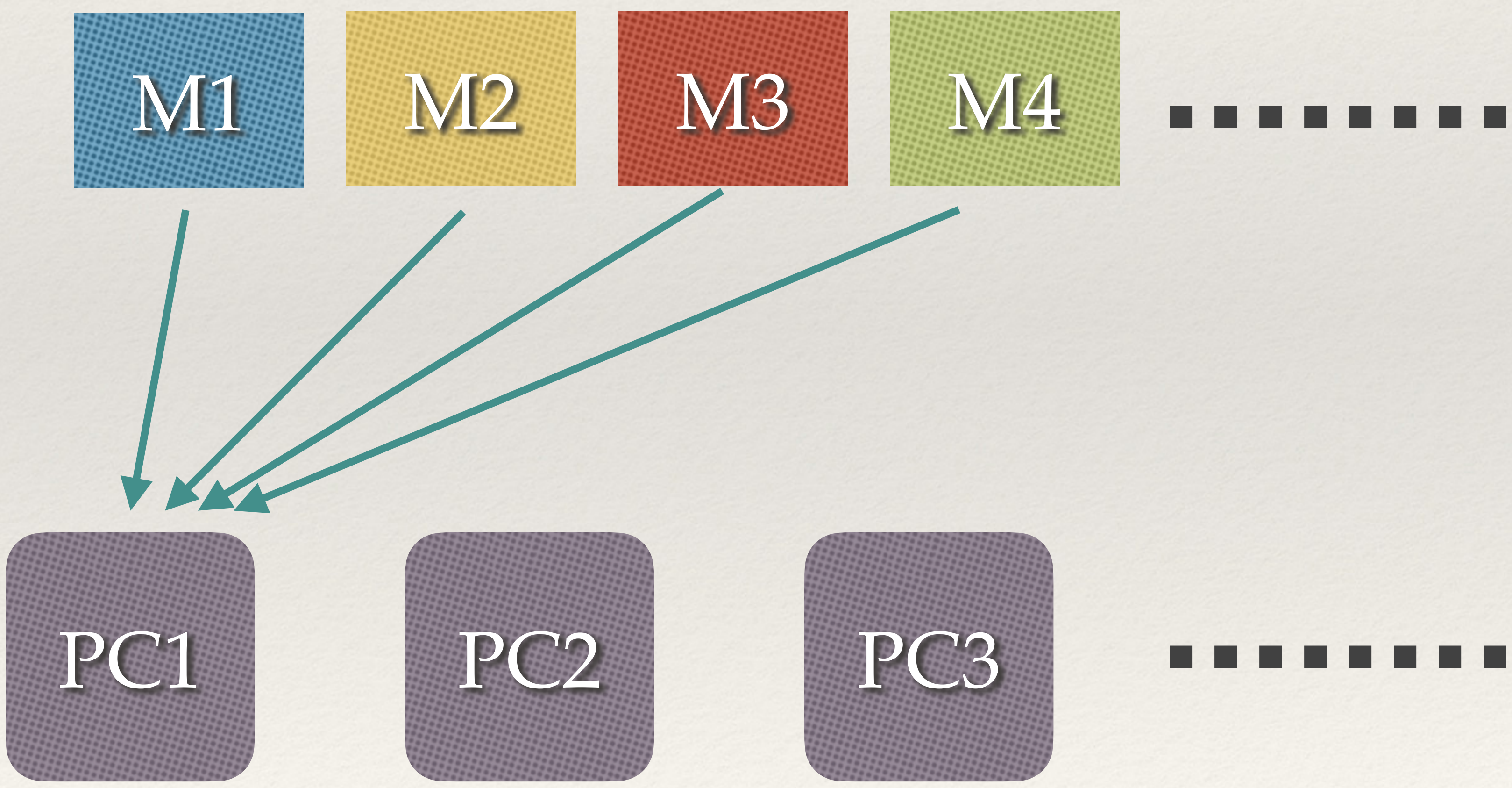
$$\Sigma_x u = \lambda u$$

$$\Sigma_z = \lambda$$

λ **(Eigenvalue) = Variance of the principle component (PC)**

Summary

- We transform variables to aggregate variance into principle components



Summary

- We transform variables to aggregate variance into principle components
- 1st PC would always has the largest variance, and each PC is independent

Summary

- We transform variables to aggregate variance into principle components
- 1st PC would always has the largest variance, and each PC is independent
- Use the covariance matrix of original data to compute eigenvectors

$$\frac{\partial Z(u)}{\partial u} = 0$$

$$= \Sigma_x u - \lambda u \longrightarrow \Sigma_x u = \lambda u$$

Summary

- We transform variables to aggregate variance into principle components
- 1st PC would always has the largest variance, and each PC is independent
- Use the covariance matrix of original data to compute eigenvectors
- Eigenvalue = Variance of the PC

$$\Sigma_z = u^T \Sigma_x u$$
$$= u^T \lambda u = I\lambda$$