Machine Learning for Data Science (CS4786)
Lecture 2

Dimensionality Reduction
&
Principal Component Analysis

Quiz

- Let Σ be the empirical covariance matrix of n points in d dimensions

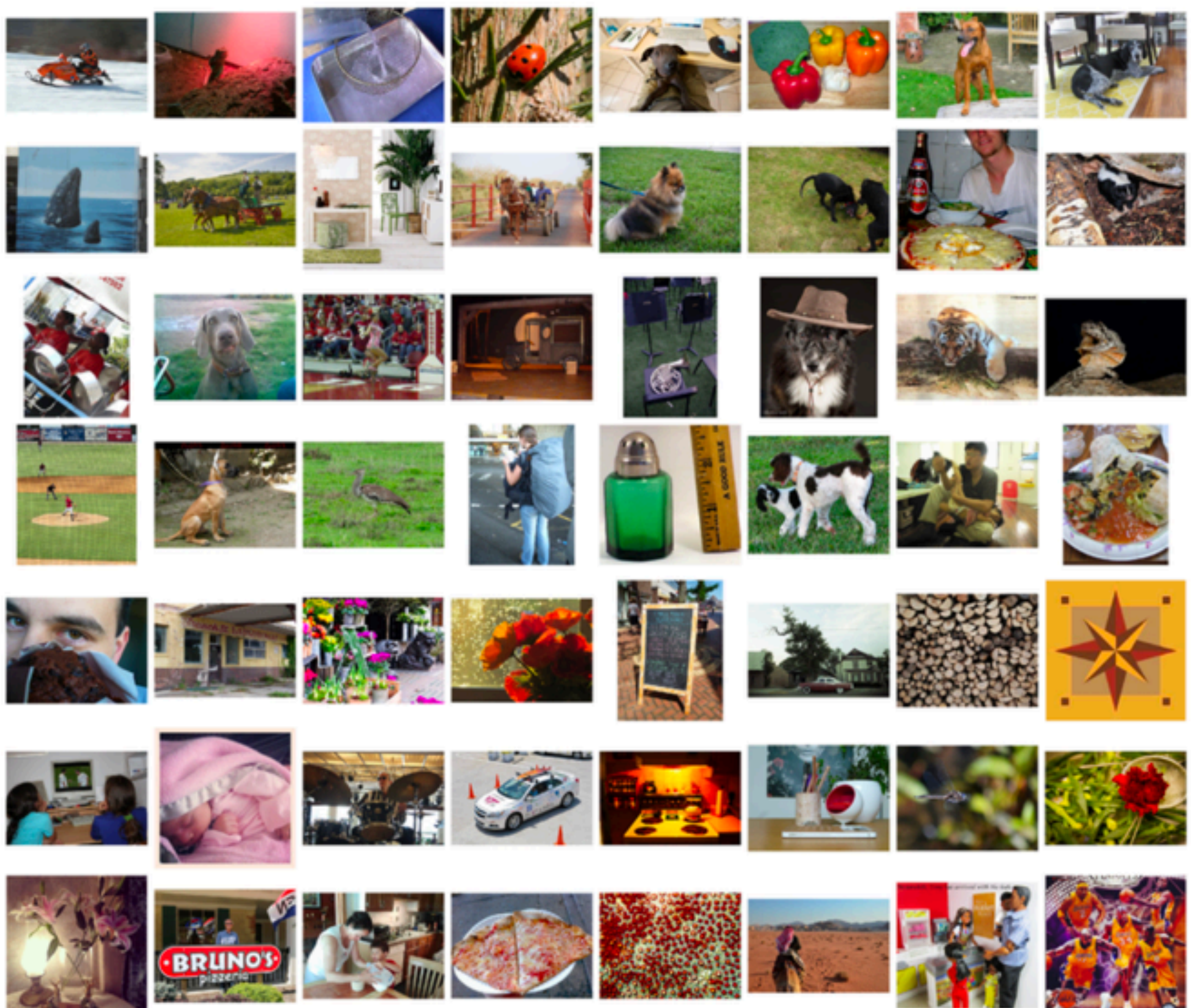
A. Σ is an n x n matrix

B. Σ is a d x d matrix

C. Σ is a m x m matrix where m is the underlying dimensionality of the n points (which can be at most d)

D. rank(Σ) is m where m is the underlying dimensionality of the n points

We can compress the following images using JPEG?

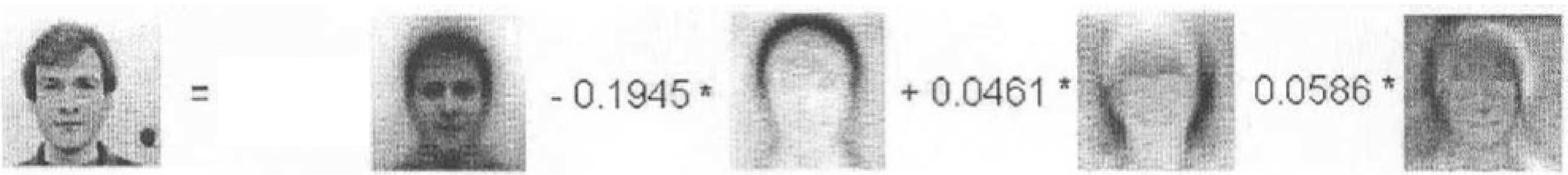


What if our dataset looked like this?

Principal Component Analysis (PCA)

Turk & Pentland'91

Eigen Face:

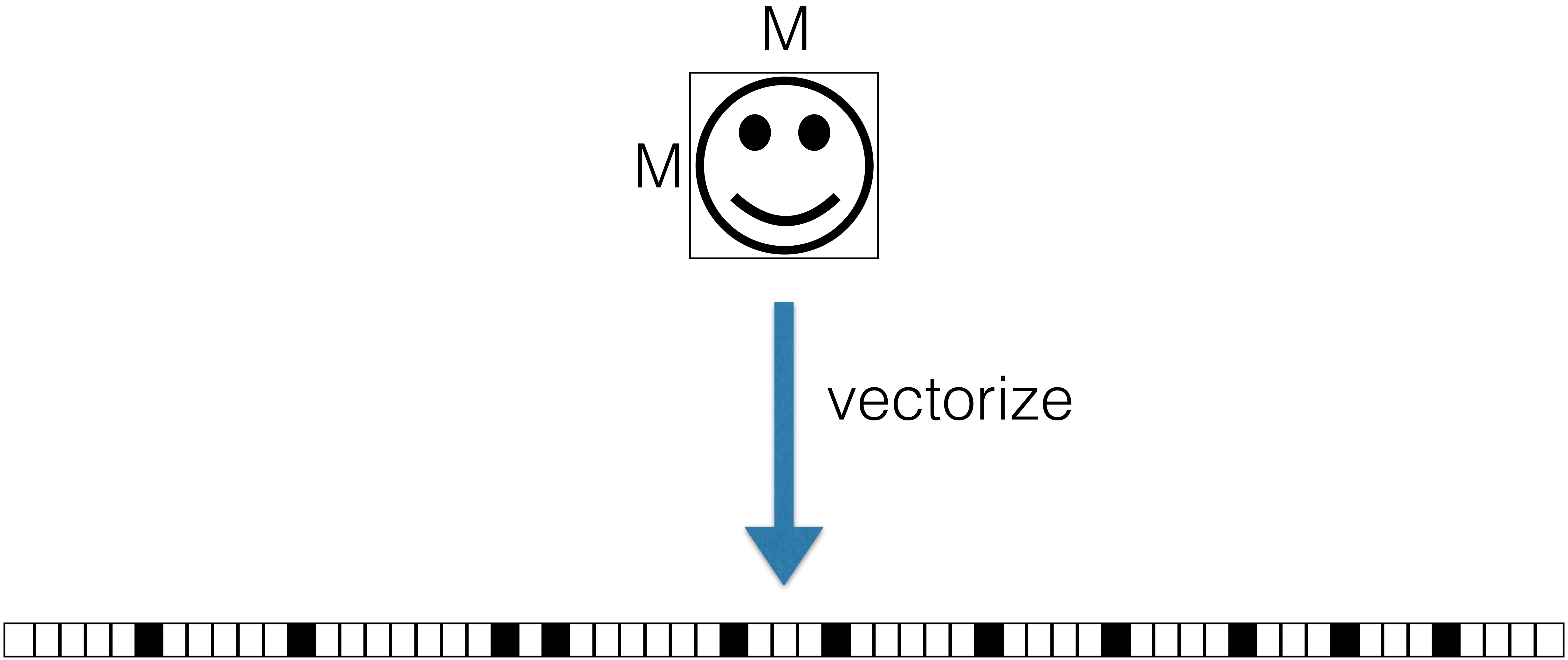


- Write down each data point as a linear combination of small number of basis vectors
- Data specific compression scheme
- One of the early successes: in face recognition: classification based on nearest neighbor in the reduced dimension space

Representing Data as Feature Vectors

- How do we represent data?
- Each data-point often represented as vector referred to as feature vector

Example: Images



$d = M^2$

Example: Text (Bag of Words)

Documents:

car
engine
hood
tires
truck
trunk

car
emissions
hood
make
model
trunk

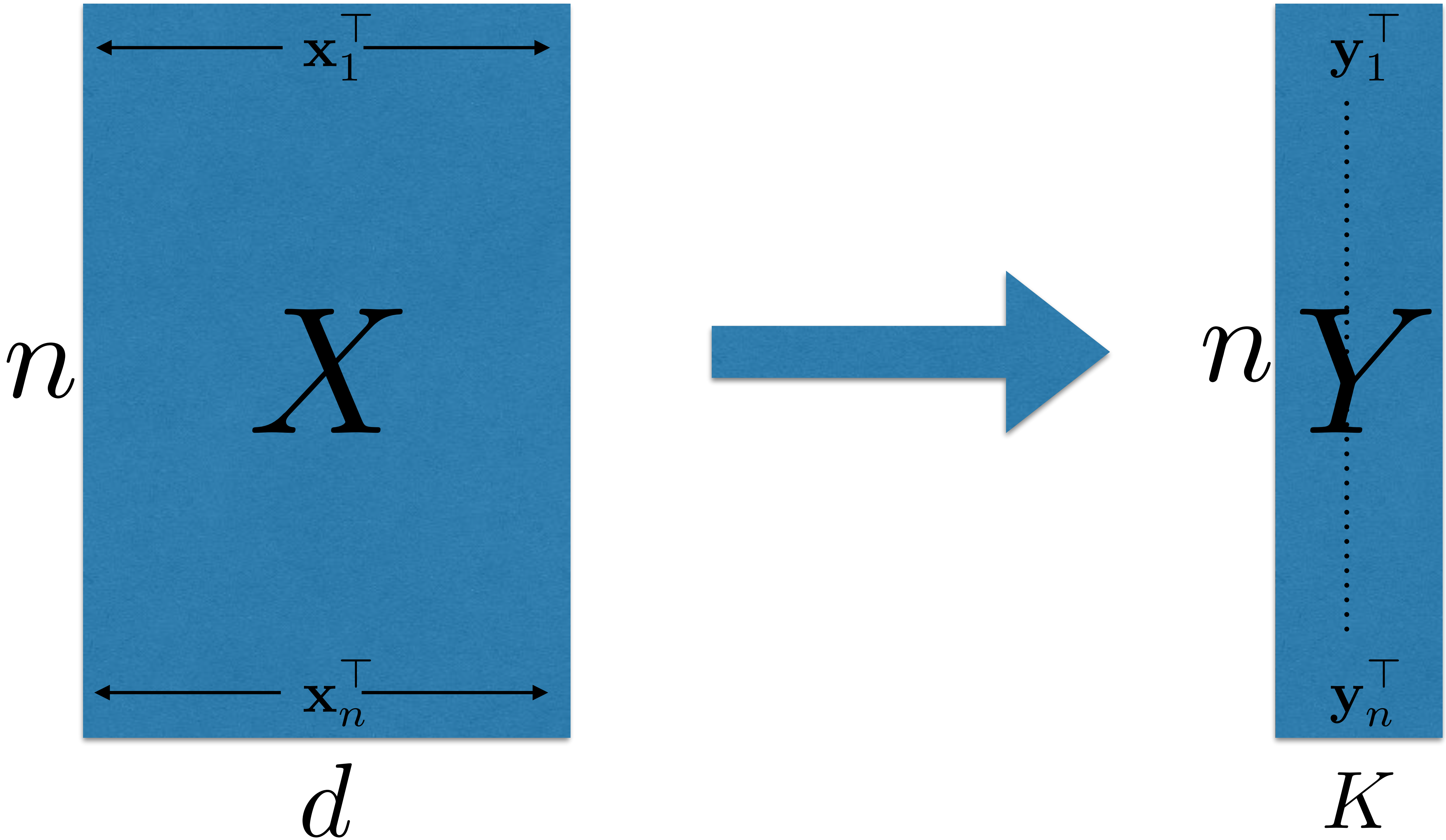
Chomsky
corpus
noun
parsing
tagging
wonderful

car	Chomsky	corpus	emissions	engine	hood	make	model	noun	parsing	tagging	tires	truck	trunk	wonderful
1	0	0	0	1	1	0	0	0	0	0	1	1	1	0
1	0	0	1	0	1	1	1	0	0	0	0	0	1	0
0	1	1	0	0	0	0	0	1	1	1	0	0	0	1

Dimensionality Reduction

Given n data points in high-dimensional space, compress them into corresponding n points in lower dimensional space.

Dimensionality Reduction



Why dimensionality reduction?

- For computational ease
  - As input to supervised learning algorithm
  - Before clustering to remove redundant information and noise
- Data compression & Noise reduction
- Data visualization

Dimensionality Reduction

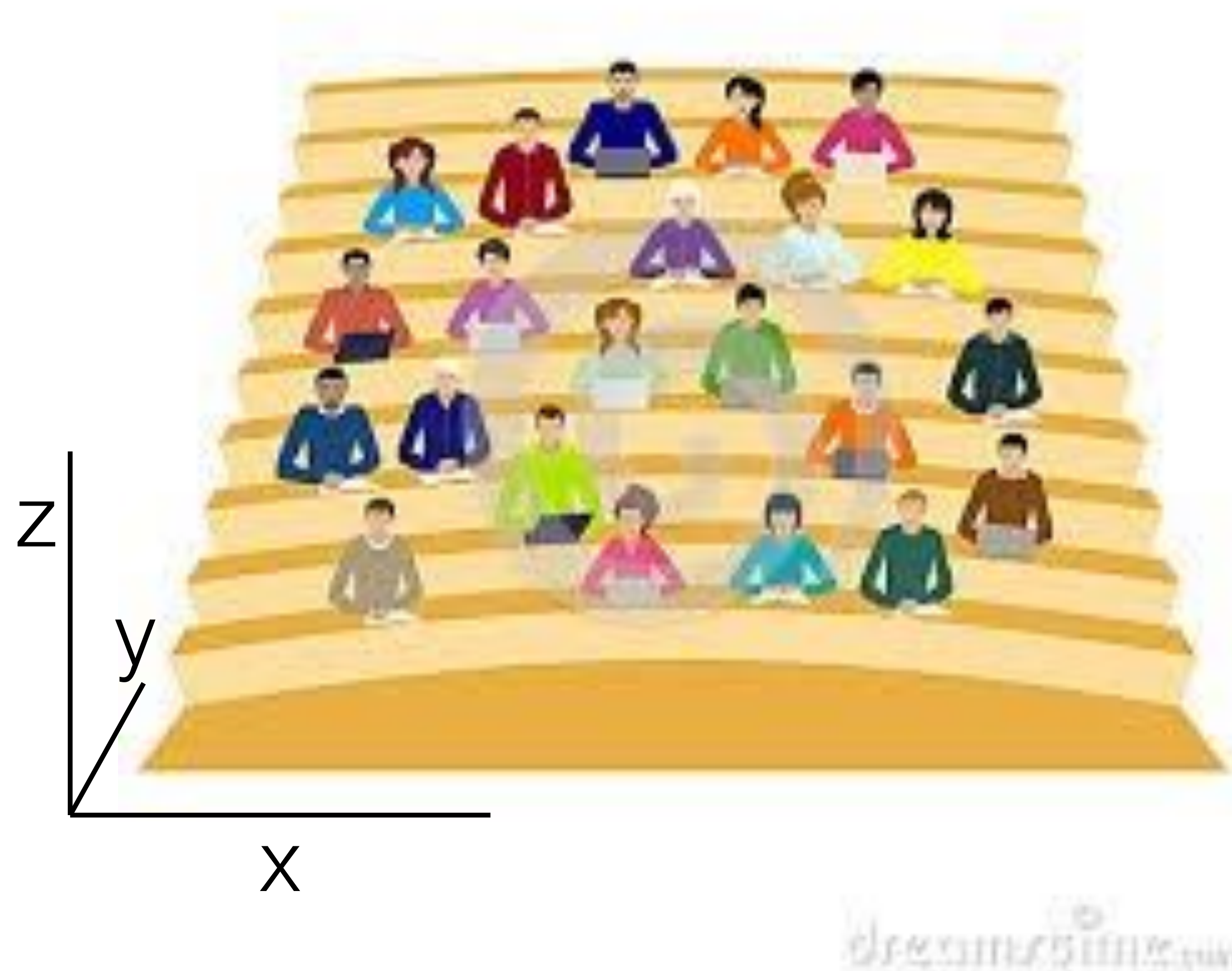
Desired properties:

1. Original data can be (approximately) reconstructed
2. Preserve distances between data points
3. “Relevant” information is preserved
4. Noise is reduced

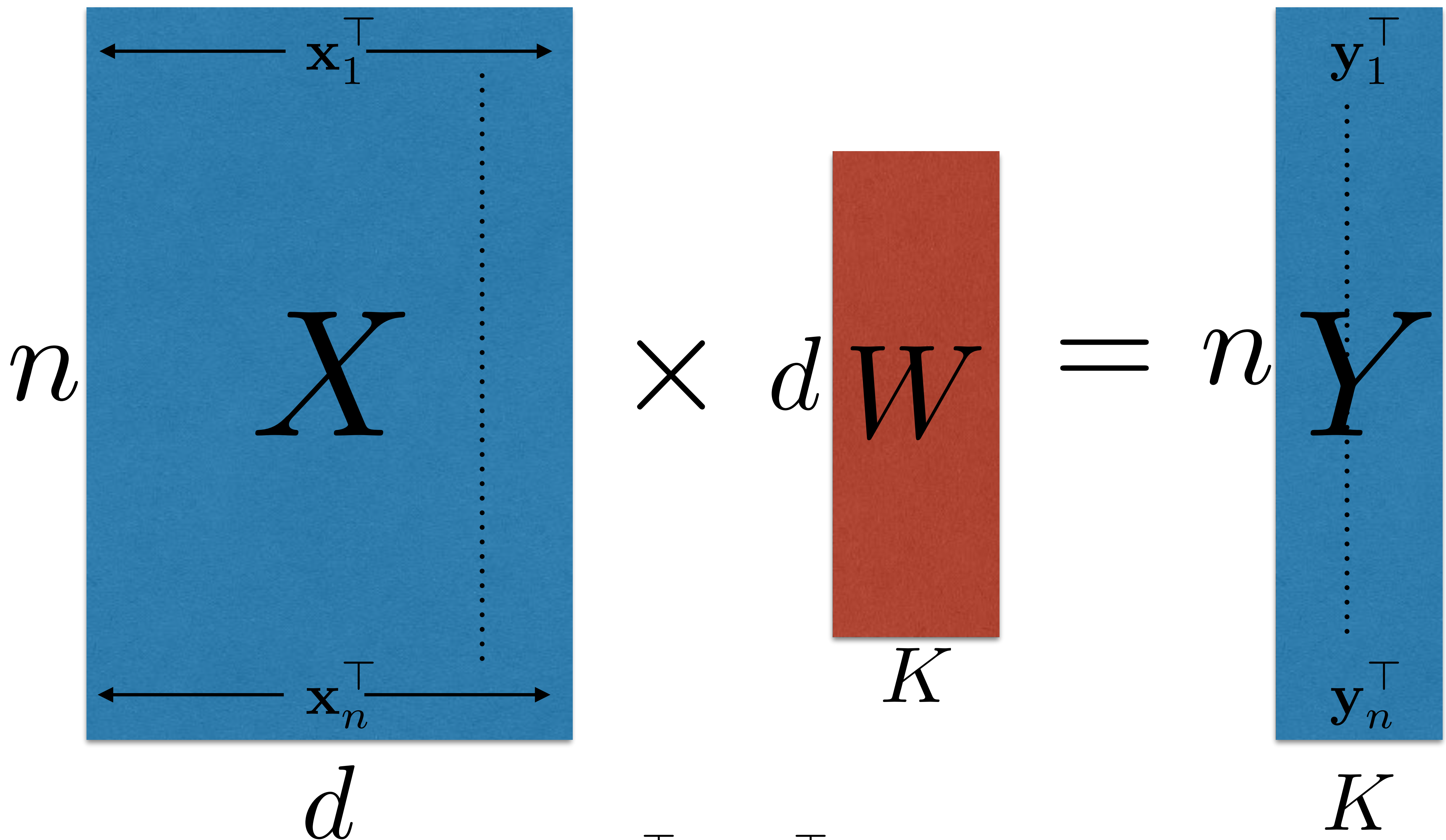
Can we reduce to 1 dim?

0.95225911	-1.90451821	2.85677732
0.60681578	-1.21363156	1.82044733
0.76419773	-1.52839546	2.29259318
0.44430217	-0.88860435	1.33290652
0.98425485	-1.9685097	2.95276456
0.04590113	-0.09180227	0.1377034
0.52408131	-1.04816263	1.57224394
0.2887897	-0.5775794	0.8663691
0.4289135	-0.857827	1.2867405
0.23877452	-0.47754905	0.71632357
0.50031855	-1.00063711	1.50095566
0.7155322	-1.43106441	2.14659661
0.19638816	-0.39277632	0.58916448
0.06743744	-0.13487488	0.20231232
0.18019499	-0.36038997	0.54058496
0.68941225	-1.37882451	2.06823676
0.51882043	-1.03764087	1.5564613
0.71398952	-1.42797904	2.14196857

Example:
Students in classroom



DIM REDUCTION: LINEAR TRANSFORMATION

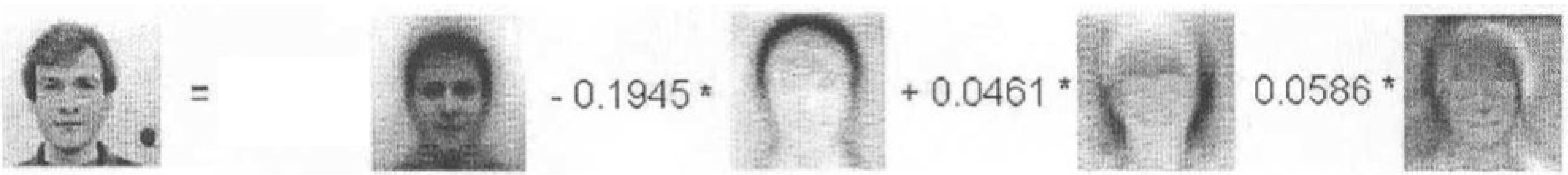


$$\mathbf{y}_i^\top = \mathbf{x}_i^\top W$$

PRINCIPAL COMPONENT ANALYSIS (PCA)

Turk & Pentland'91

Eigen Face:



- Each x_t (each row of X) is a face image (vectorized version)
- Each y_t is the set of coefficients we multiply to the eigen face
- Each column of W is an Eigenface

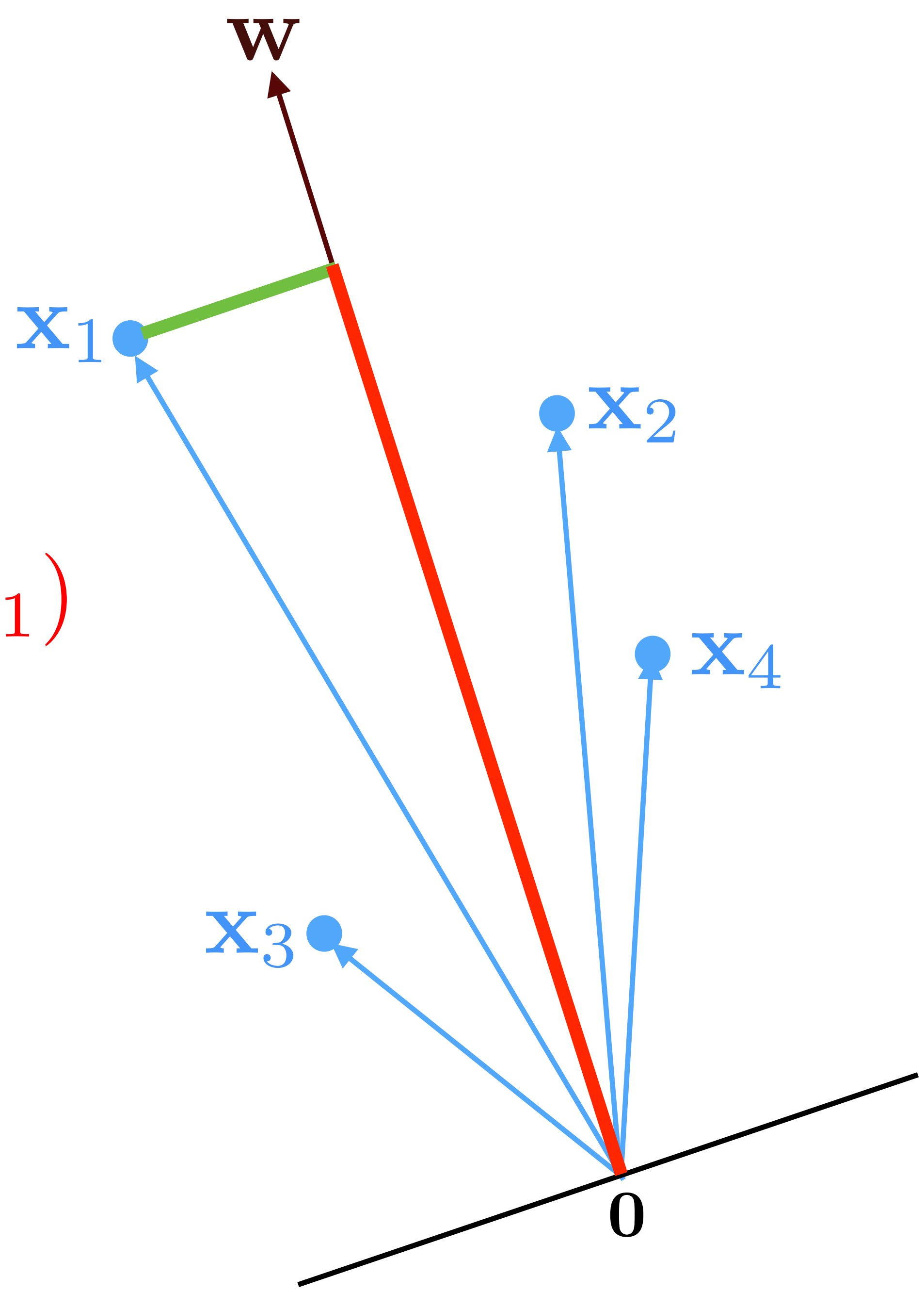
Prelude: Reducing to 1 Dim

- W is a d x 1 matrix (d dimensional vector)
- Each data point is compressed to a single number
- How do we pick this W?

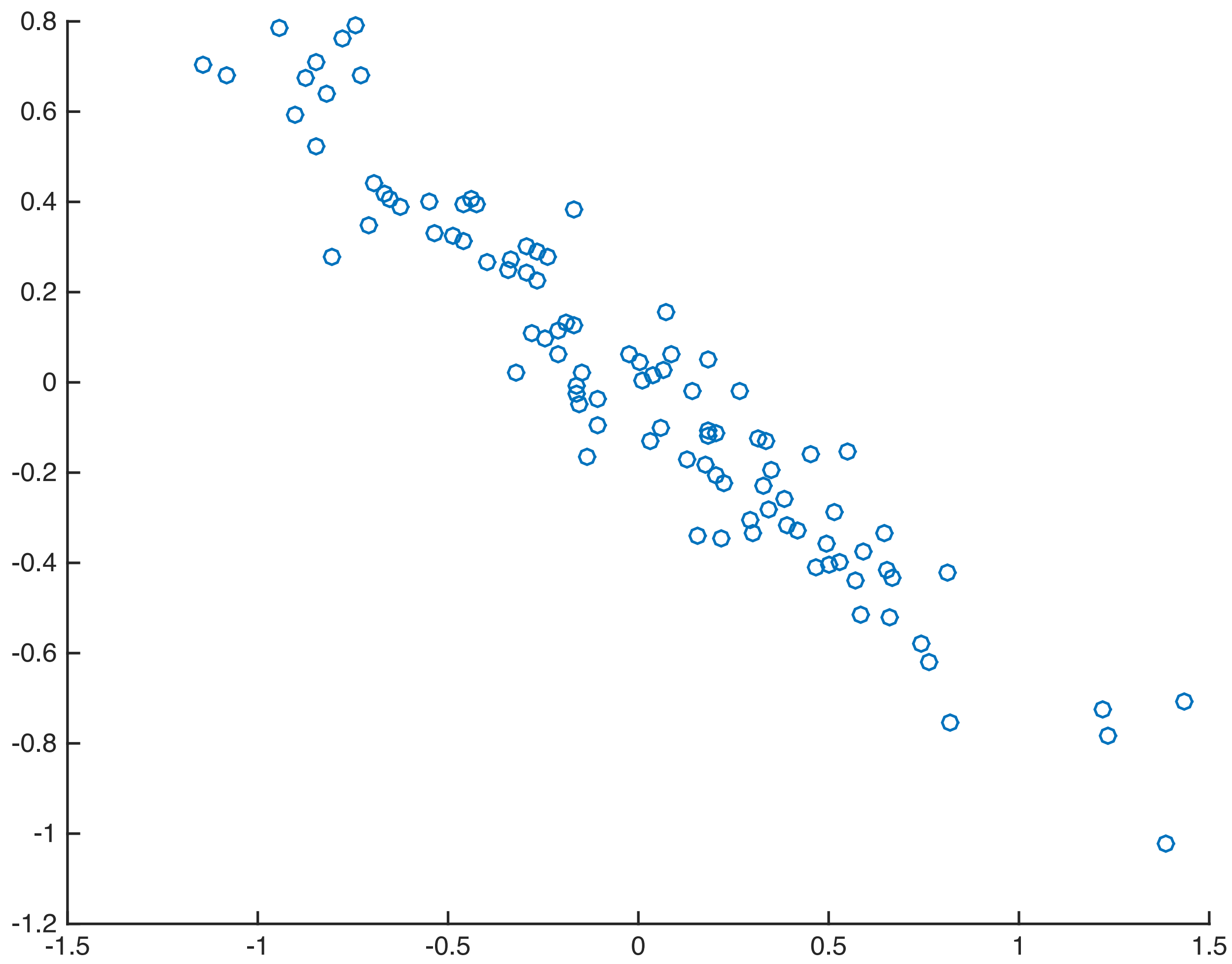
Prelude: reducing to 1 dimension

$$\mathbf{y}_1 = \mathbf{w}^\top \mathbf{x}_1 = \|\mathbf{x}_1\| \cos\left(\angle \mathbf{w}\mathbf{x}_1\right)$$

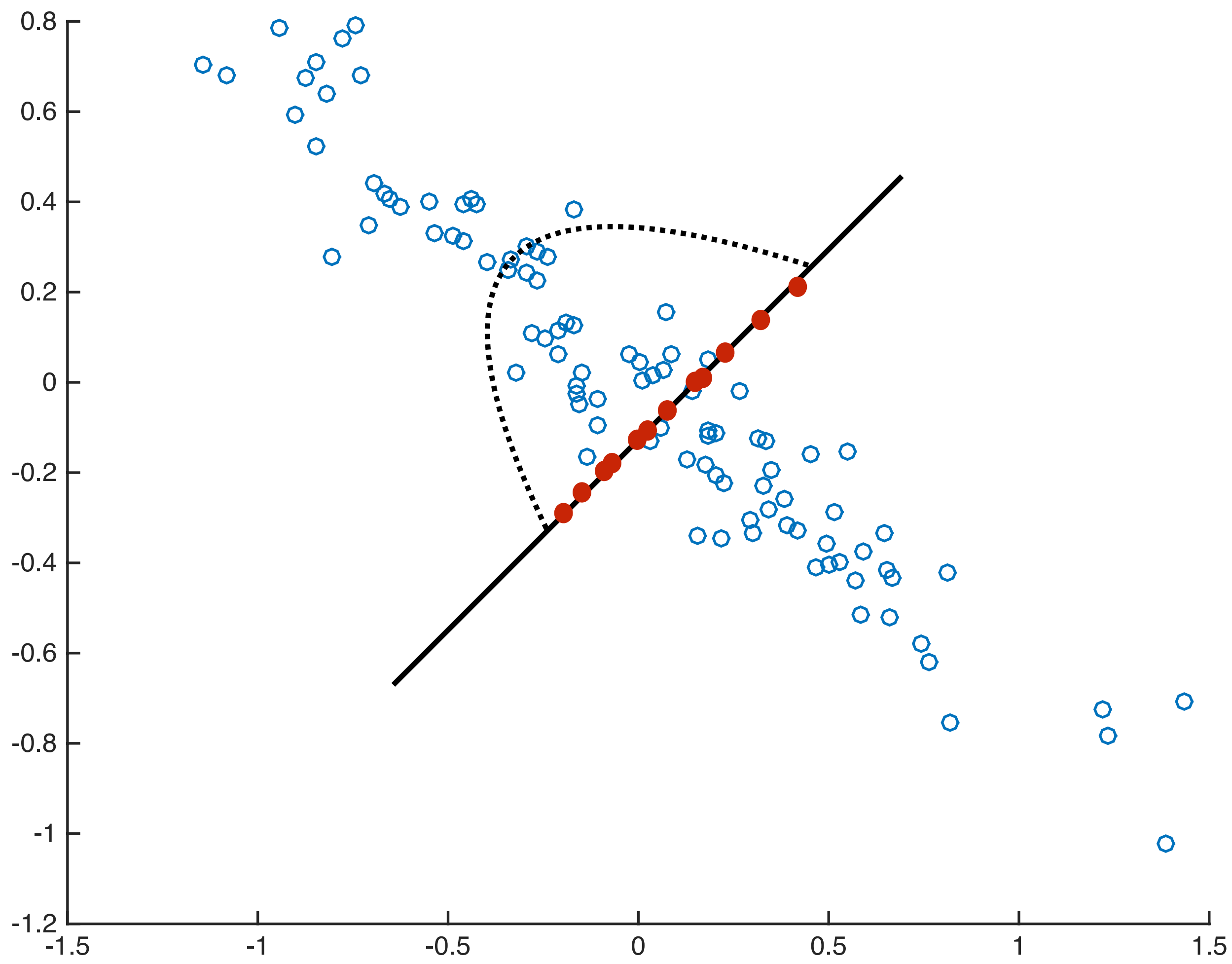
Only direction matters, assume without loss of generality that ||w|| = 1



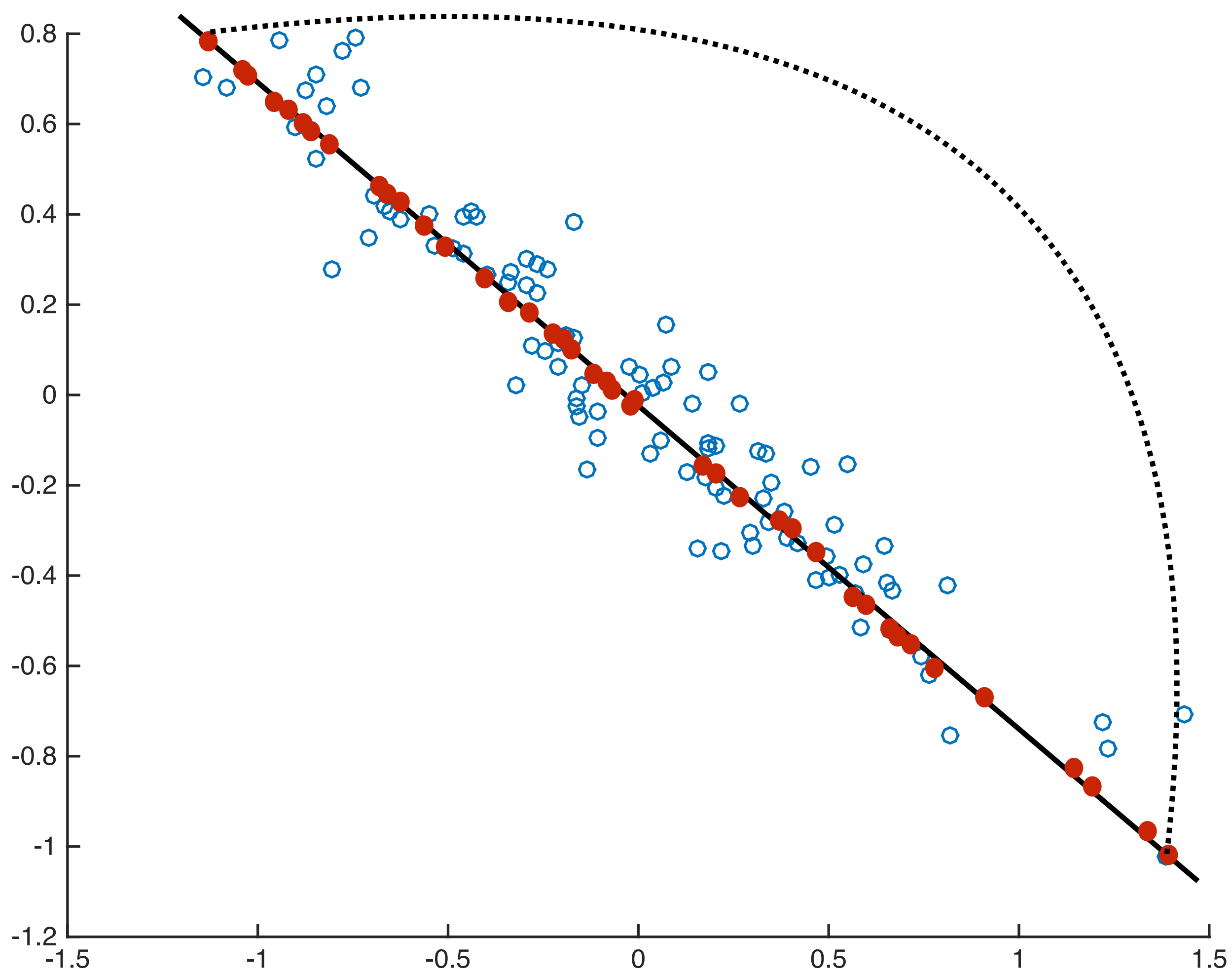
PCA: Variance Maximization



PCA: Variance Maximization



PCA: Variance Maximization

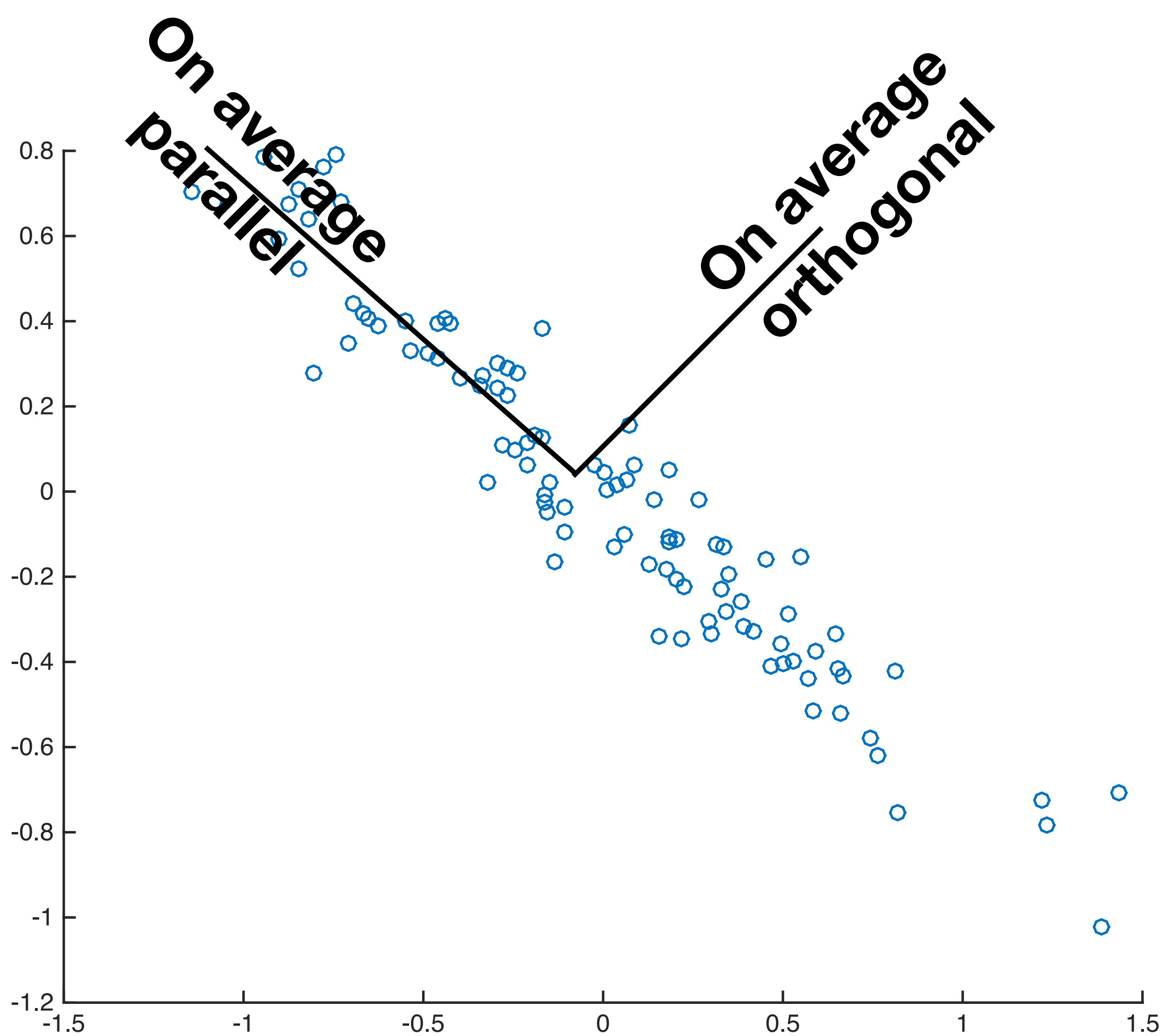


PCA: Variance Maximization

- Pick directions along which data varies the most

$$
\begin{aligned}
\text{Variance} &= \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{1}{n}\sum_{s=1}^{n} y_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{s=1}^{n} \mathbf{w}^\top \mathbf{x}_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \mathbf{w}^\top\left(\frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s\right)\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top(\mathbf{x}_t - \mu)\right)^2 \\
&= \ \text{average squared inner product}
\end{aligned}
$$

Which Direction?



$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^{\top}(\mathbf{x}_t-\mu)\right)^2 = \frac{1}{n}\sum_{t=1}^{n}\|\mathbf{x}_t-\mu\|^2\text{cosine}(w, x_t-\mu)$$

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$
\begin{aligned}
\mathbf{w}_1 &= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top\mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top\mathbf{x}_t\right)^2 \\
&= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top(\mathbf{x}_t - \mu)\right)^2 \\
&= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top\mathbf{w} \\
&= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top\Sigma\mathbf{w}
\end{aligned}
$$

Σ is the covariance matrix

Covariance Matrix

- Its a $d \times d$ matrix, $\Sigma[i,j]$ measures "covariance" of features i and j

$$\Sigma[i,j] = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t[i] - \mu[i])(\mathbf{x}_t[j] - \mu[j])$$

PCA: Variance Maximization

Covariance matrix:

$$\Sigma = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top$$

- Its a $d \times d$ matrix, $\Sigma[i, j]$ measures "covariance" of features i and j

$$\Sigma[i, j] = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t[i] - \mu[i])(\mathbf{x}_t[j] - \mu[j])$$

PCA: Variance Maximization

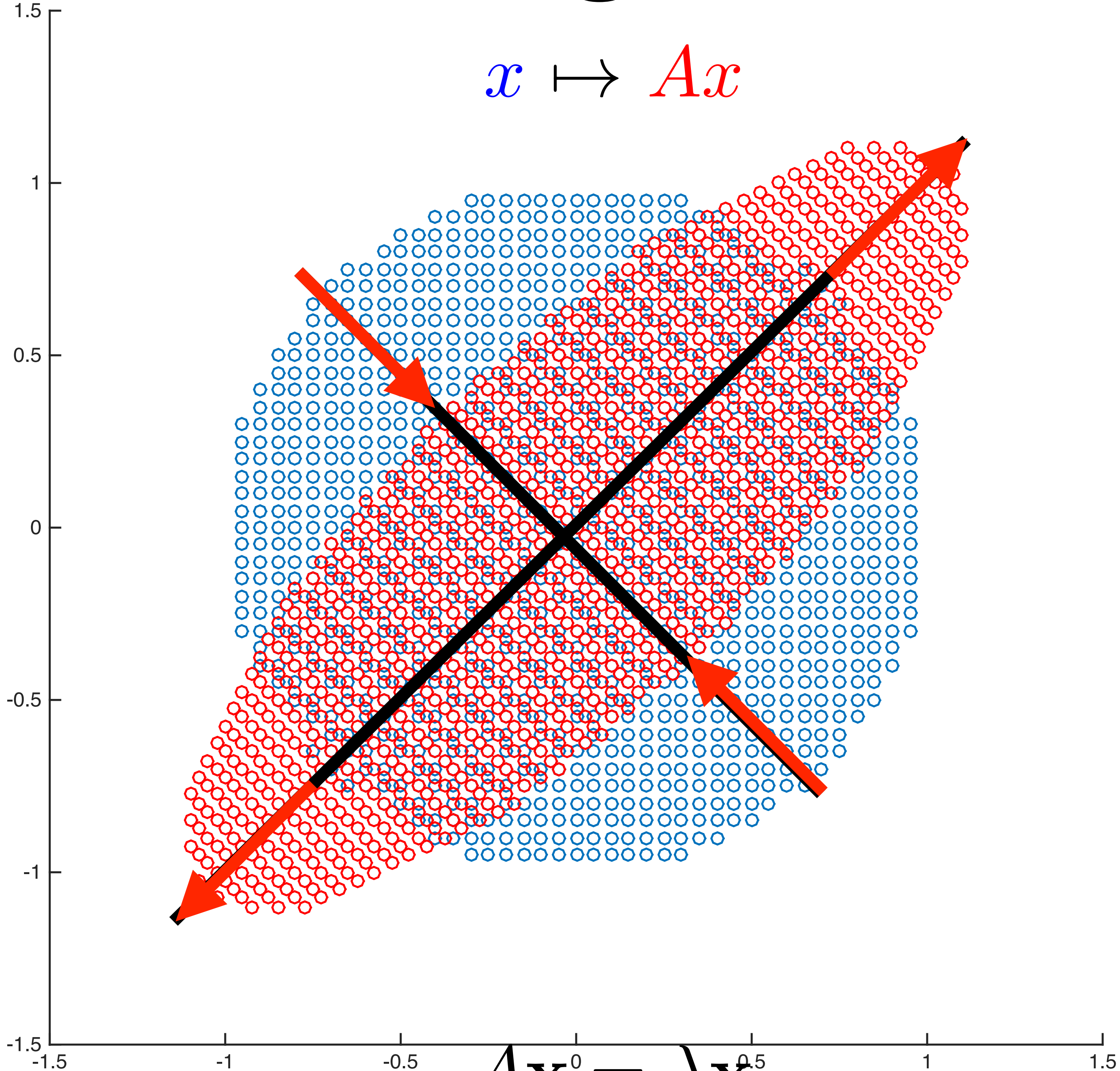
- Pick directions along which data varies the most
- First principal component:

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Sigma \mathbf{w}$$

Σ is the covariance matrix

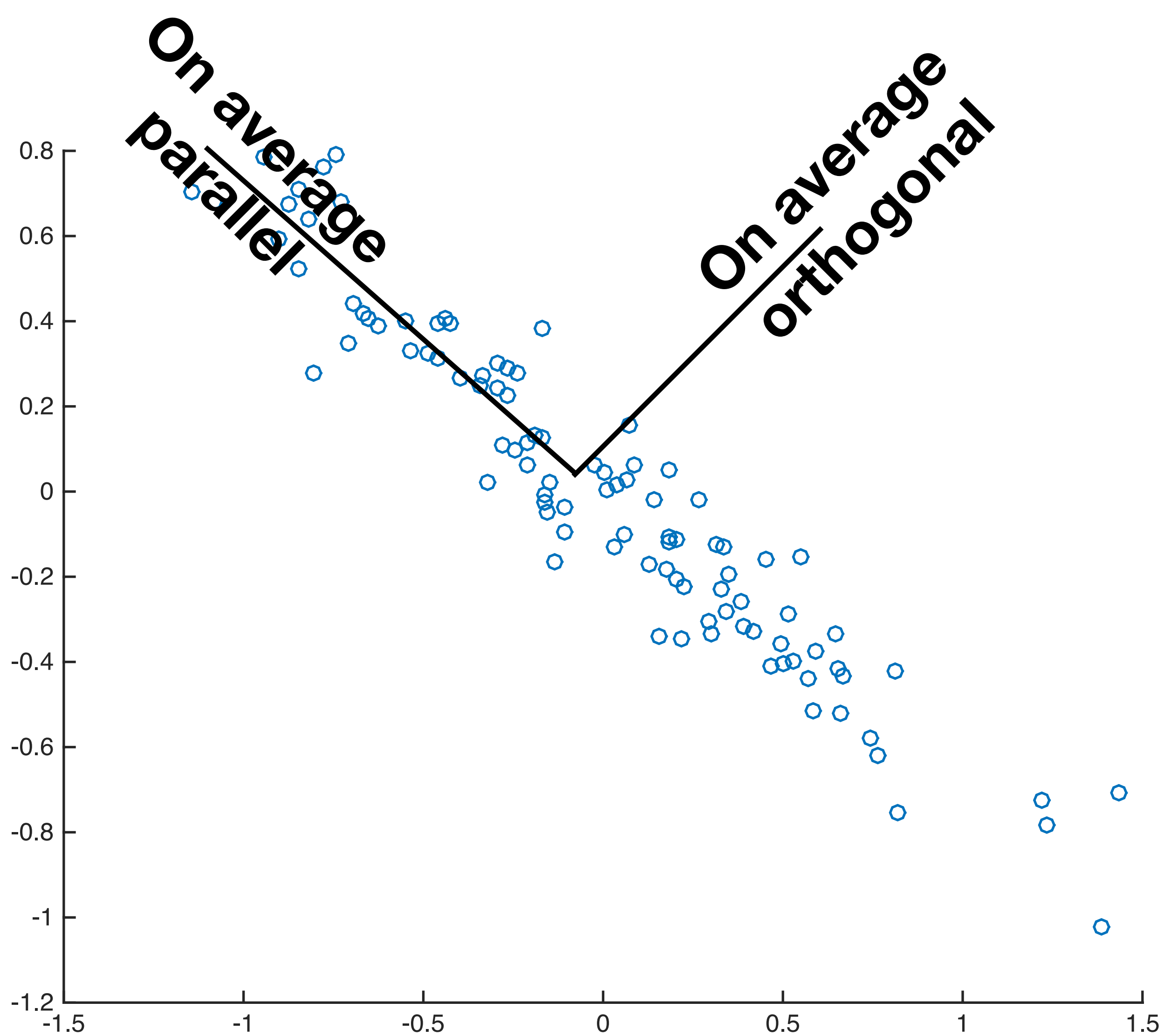
Solution: $\mathbf{w}_1 =$ Largest Eigenvector of Σ

What are Eigen Vectors?



$$A\mathbf{x} = \lambda\mathbf{x}$$

Which Direction?



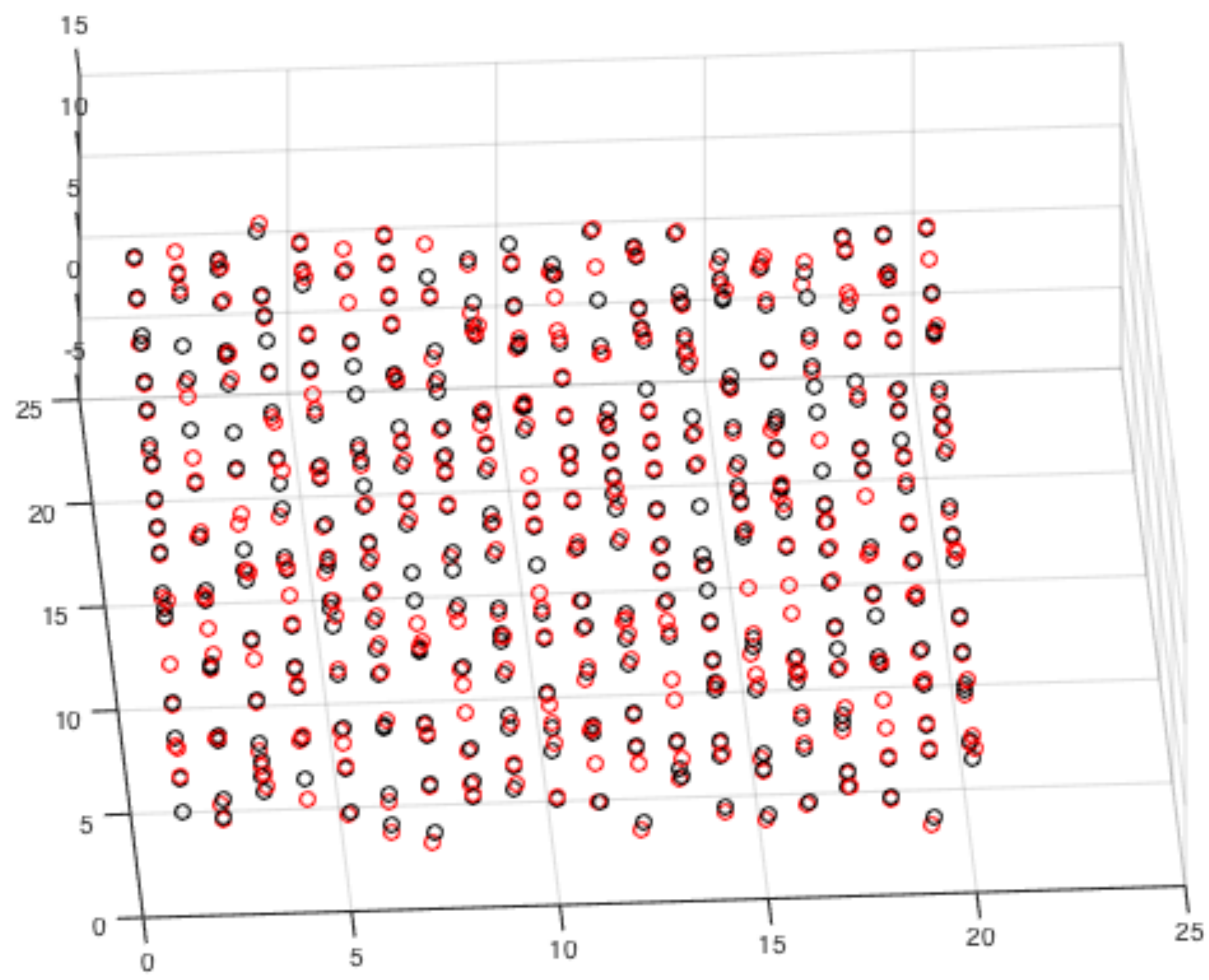
Top Eigenvector of covariance matrix

- What if we want more than one number for each data point?
- That is we want to reduce to K > 1 dimensions?

PCA: Variance Maximization

- How do we find the K components?

Ans: Maximize sum of spread in the K directions



PCA: Variance Maximization

- How do we find the K components?

- We are looking for orthogonal directions that maximize total spread in each direction

- Find orthonormal W that maximizes $\sum_{k=1}^{d} \mathbf{w}_i[k]\mathbf{w}_j[k] = 0 \;\; \& \;\; \sum_{k=1}^{d} \mathbf{w}_i[k] = 1$

$$\sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{y}_t[j] - \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t[j] \right)^2 = \sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_j^\top \left(\mathbf{x}_t - \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \right) \right)^2$$

$$= \sum_{j=1}^{K} \mathbf{w}_j^\top \Sigma \mathbf{w}_j$$

Intuition: Remove top direction, now reduce dimension for remaining d-1 dimensions

- This solutions is given by W = Top K eigenvectors of Σ

Principal Component Analysis

1. $$\Sigma = \mathrm{cov}\left(X \right)$$

2. $$W = \mathrm{eigs}(\Sigma, K)$$

3. $$Y = X \times W$$

Can we reconstruct the original data points?