Machine Learning for Data Science (CS4786)
Lecture 2

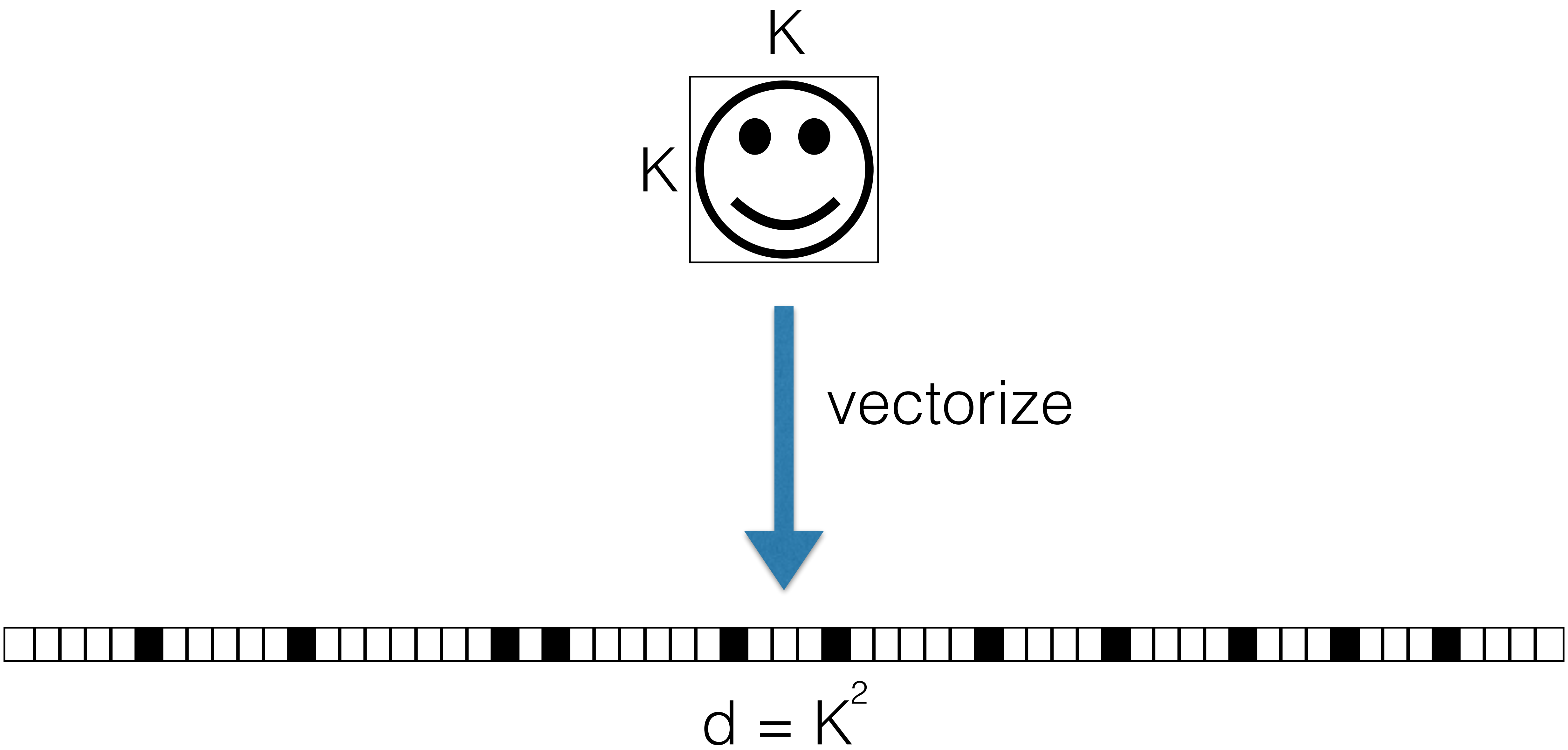
Dimensionality Reduction
&
Principal Component Analysis

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016fa/

Representing Data as Feature Vectors

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

Example: Images



Example: Text (Bag of Words)

Documents:

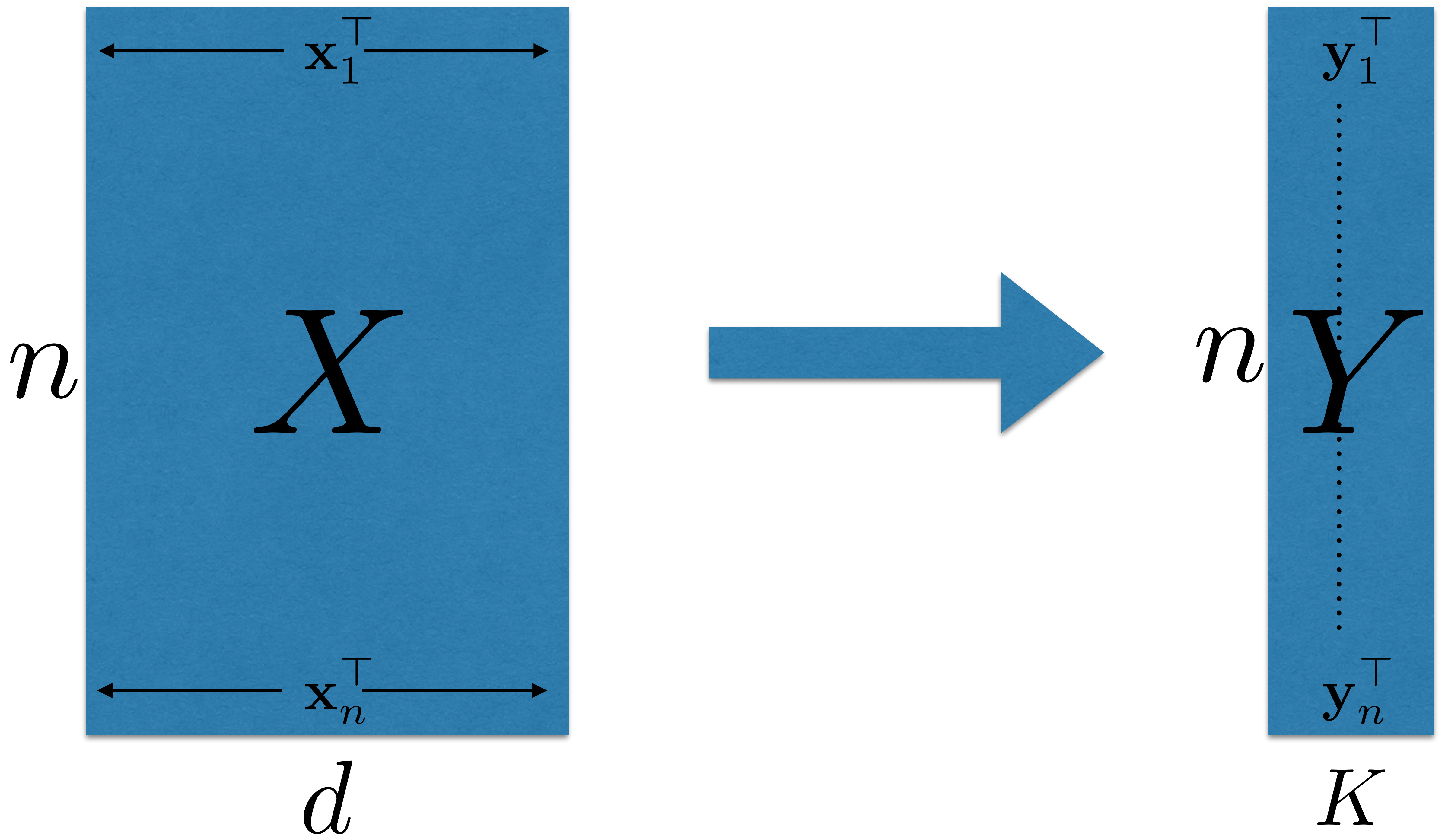
car
engine
hood
tires
truck
trunk

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



Dimensionality Reduction

Given feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, compress the data points into low dimensional representation $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^K$ where $K << d$

Flowers

Iris-Setosa

Iris-versicolor

Iris-virginica

Principal Component Analysis: Demo

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

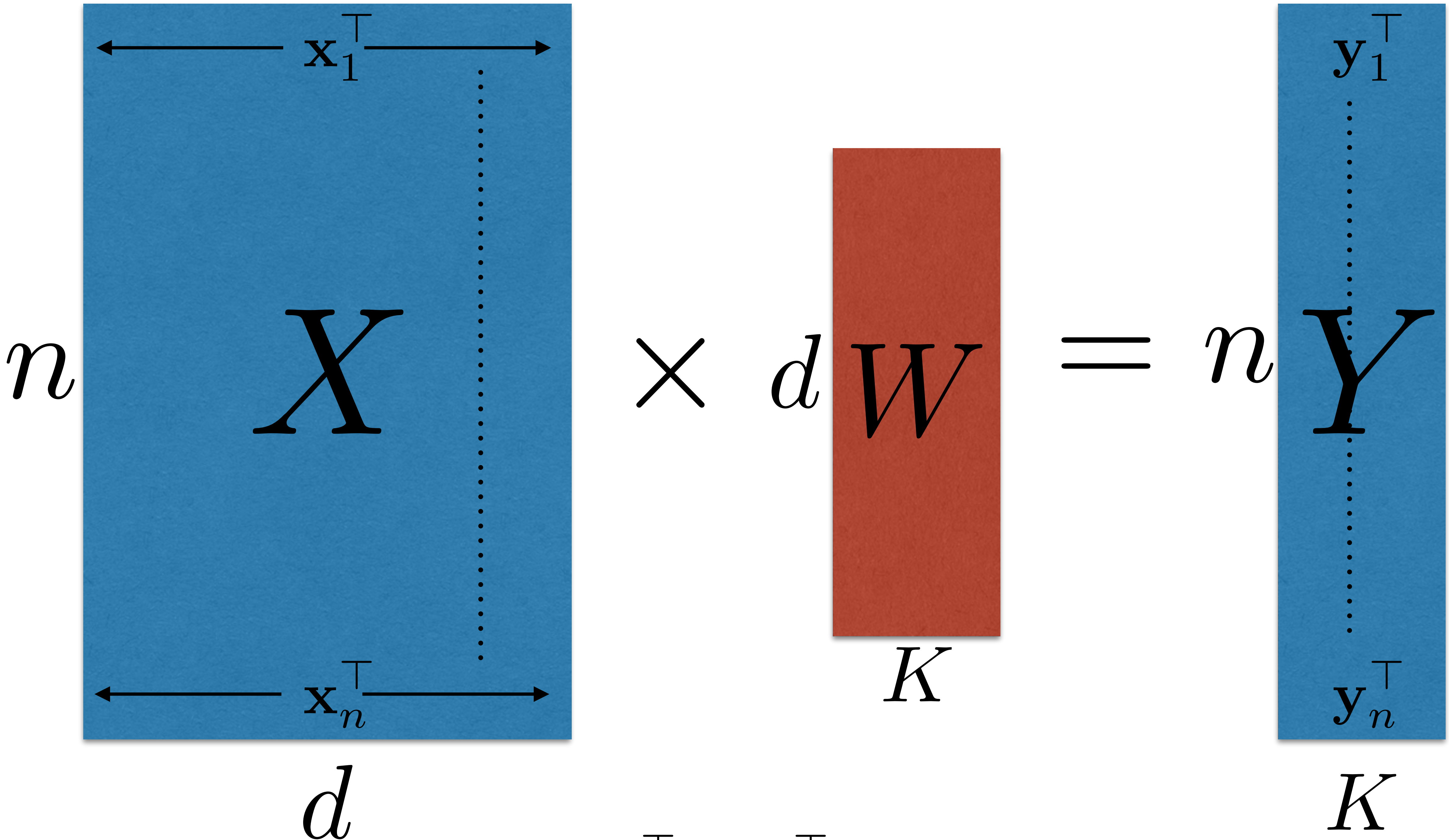
Desired properties:

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

Dim Reduction: Linear Transformation

- Pick a low dimensional subspace
- Project linearly to this subspace
- Subspace retains as much information

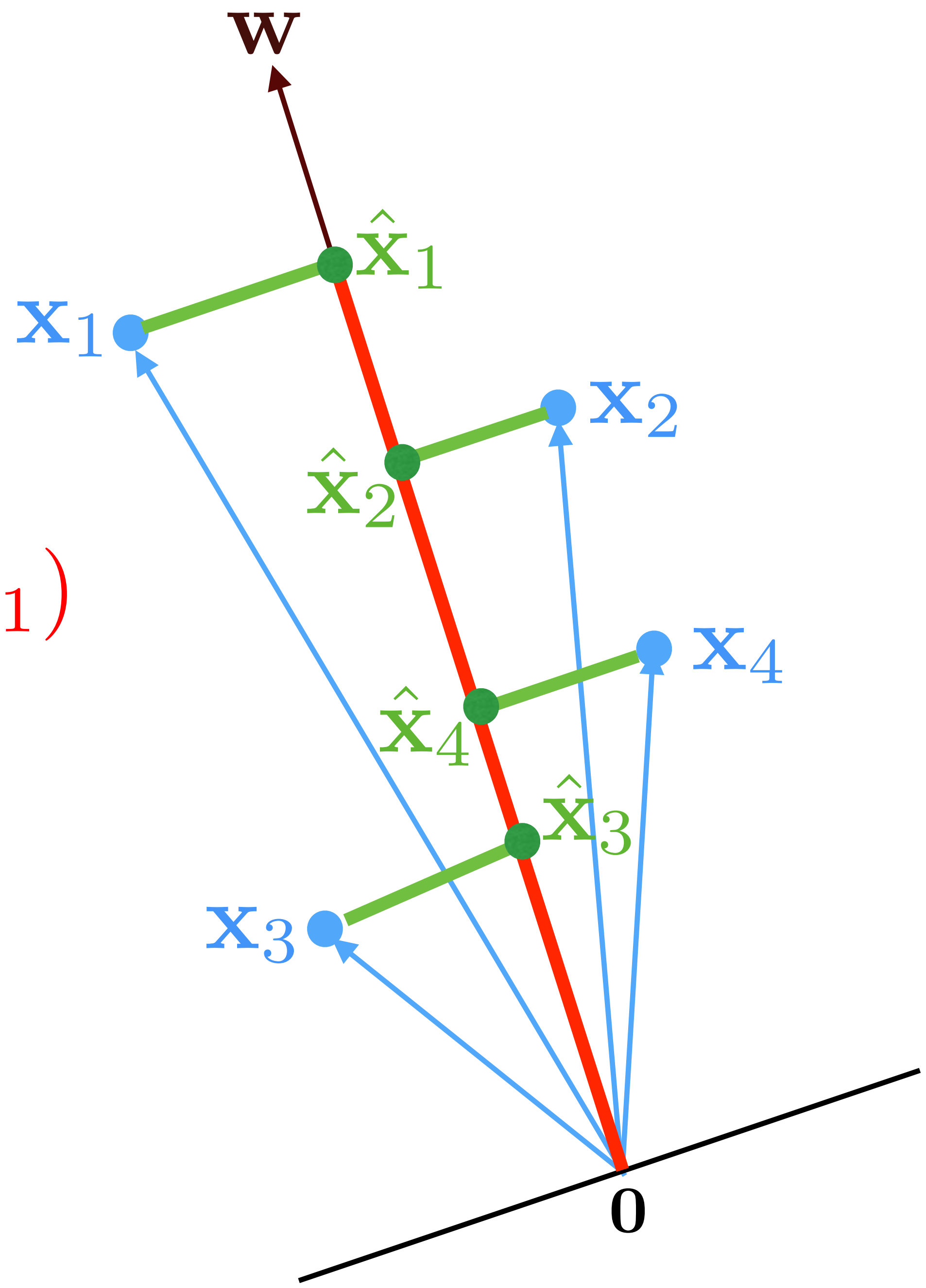
Dim Reduction: Linear Transformation



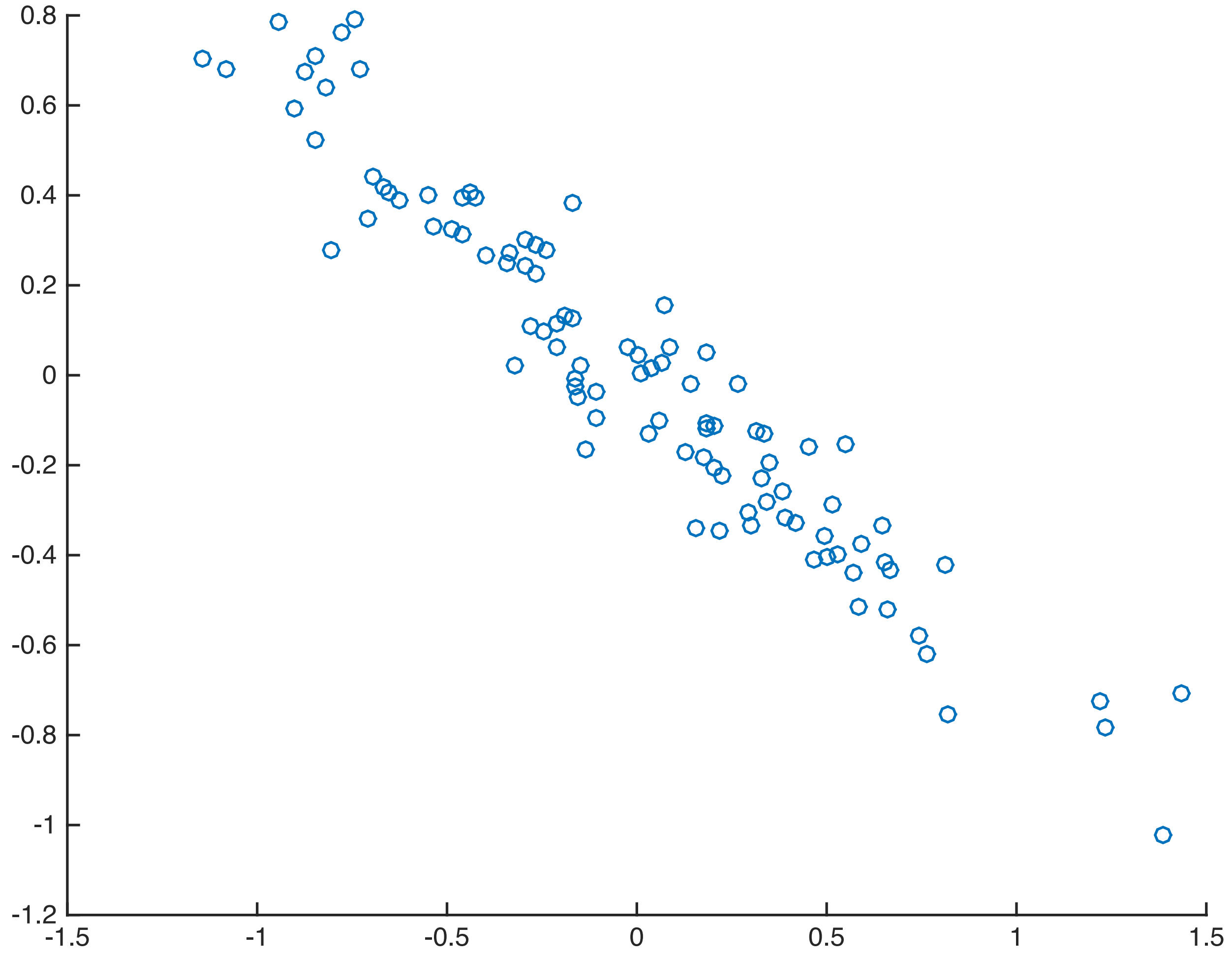
$$\mathbf{y}_i^\top = \mathbf{x}_i^\top W$$

Prelude: reducing to 1 dimension

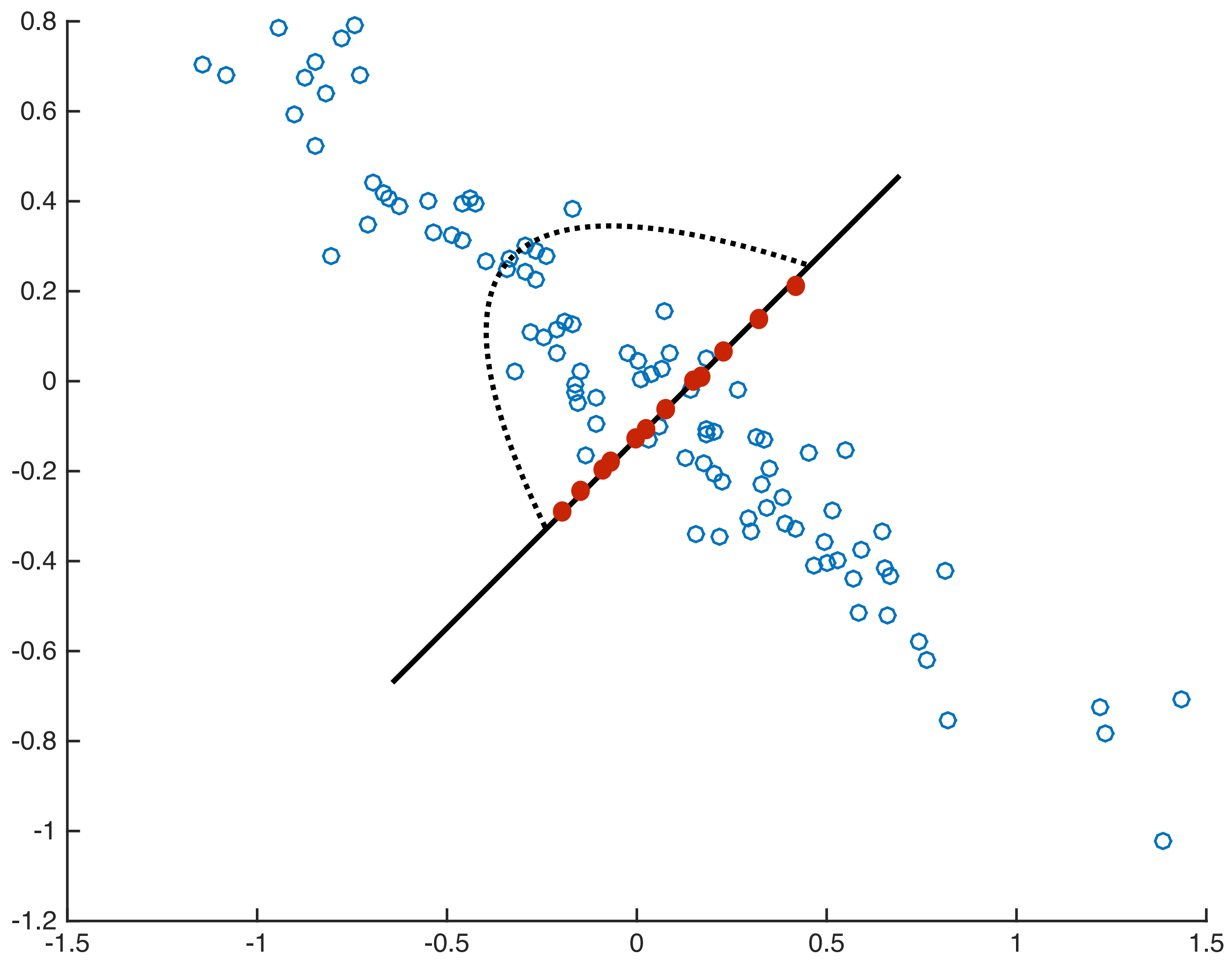
$$\mathbf{y}_1 = \mathbf{w}^\top \mathbf{x}_1 = \|\mathbf{x}_1\| \cos\left(\angle \mathbf{w}\mathbf{x}_1\right)$$



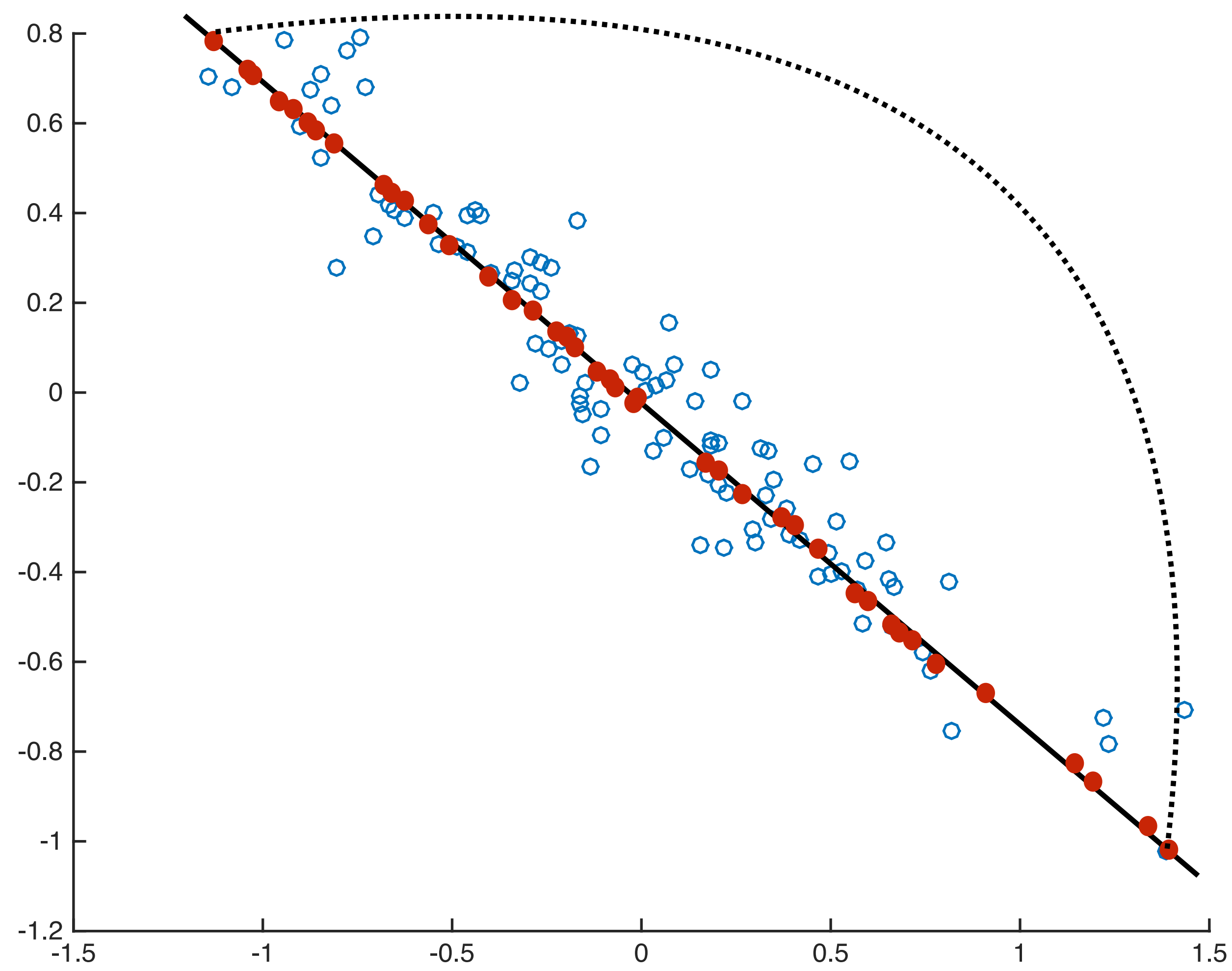
PCA: Variance Maximization



PCA: Variance Maximization



PCA: Variance Maximization



PCA: Variance Maximization

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

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

where Σ is the covariance matrix and $\mu = \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t$

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

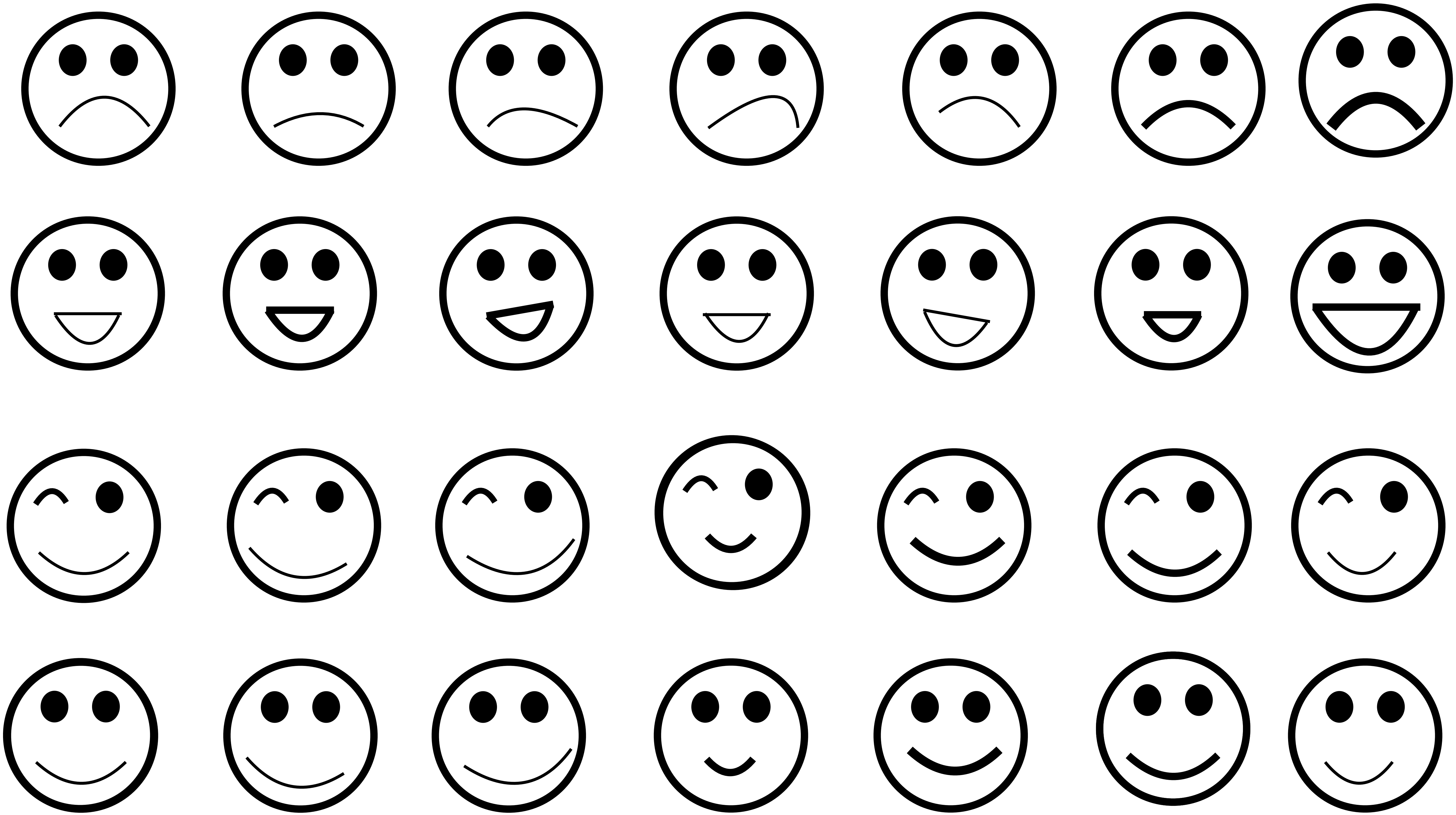
- First principal component:

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

- The solution to the above optimization problem is $\mathbf{w}_1$ is the top Eigen vector of matrix Σ
- Hence in "matlab",

$$S = \mathrm{Cov}(X)$$
$$[W, E] = \mathrm{eigs}(S, 1)$$
$$Y = W * X$$

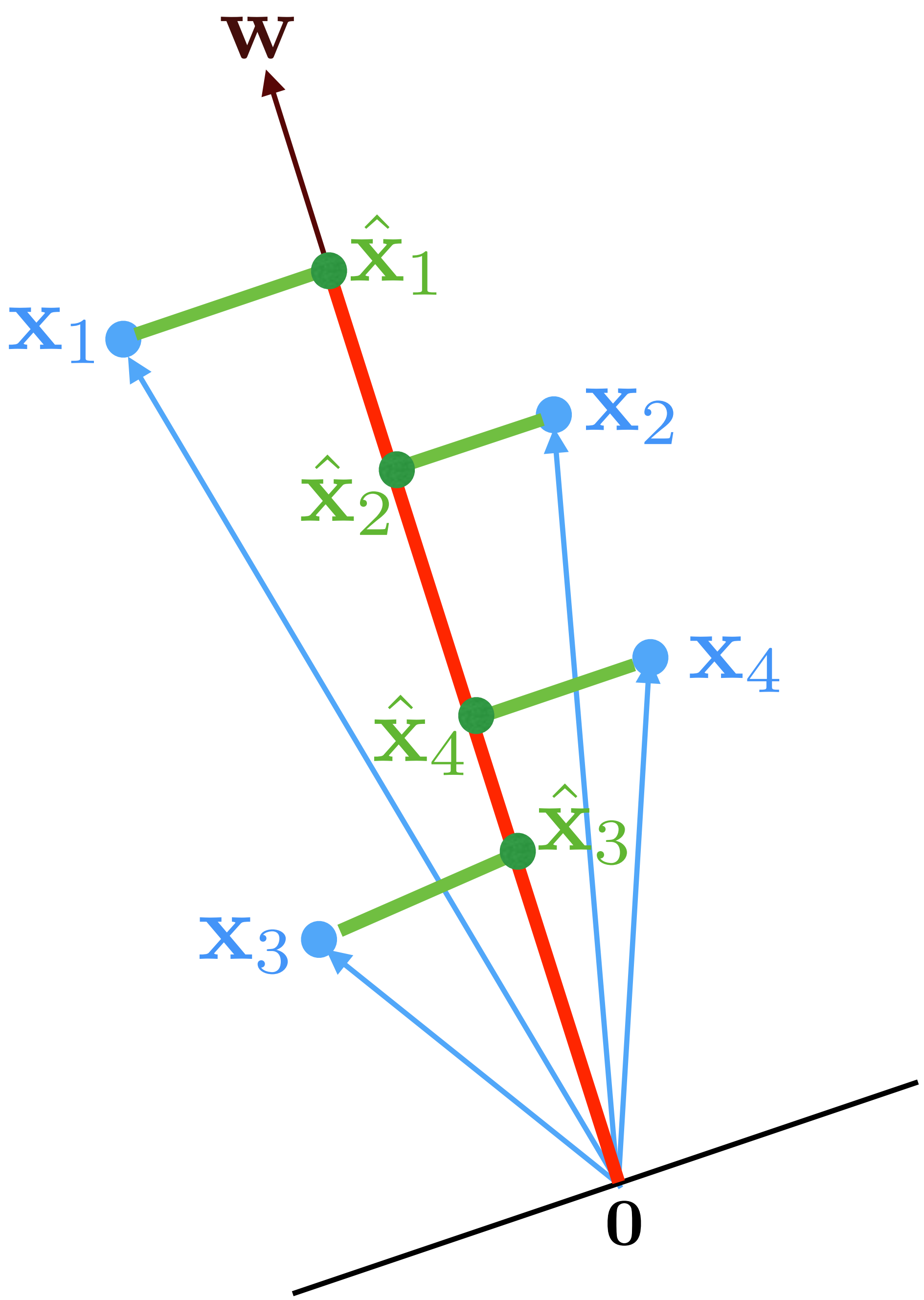
Principal Component Analysis: Demo



Principal Component Analysis ($K > 1$)

What do we do when $K > 1$?

Prelude: reducing to 1 dimension



Orthonormal Projections

- Think of $\mathbf{w}_1, \ldots, \mathbf{w}_K$ as coordinate system for PCA (in a K dimensional subspace)

- $\mathbf{y}$ values provide coefficients in this system

- Without loss of generality, $\mathbf{w}_1, \ldots, \mathbf{w}_K$ can be orthonormal, i.e. $\mathbf{w}_i \perp \mathbf{w}_j$ & $\|\mathbf{w}_i\| = 1$.

- Reconstruction:

$$\hat{\mathbf{x}}_t = \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j$$

- If we take all $\mathbf{w}_1, \ldots, \mathbf{w}_d$, then $\mathbf{x}_t = \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j$. To reduce dimensionality we only consider first K vectors of the basis

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

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

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

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

Principal Component Analysis

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

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

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

Principal Component Analysis: Demo