



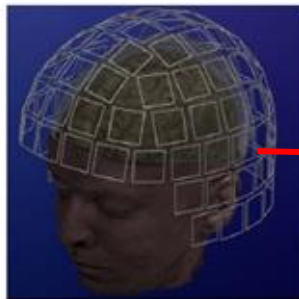
Math Foundation of Data Science/Advanced Math for AI

Principal Component Analysis

Dimensionality Reduction

- Vectors store features. Lots of features!
 - Document classification: thousands of words per doc
 - Netflix surveys: 480189 users x 17770 movies
 - **MEG Brain Imaging**: 120 locations x 500 time points x 20 objects

	movie 1	movie 2	movie 3
Tom	5	?	?
George	?	?	3
Susan	4	3	1
Beth	4	3	?



Dimensionality Reduction

Reduce dimensions

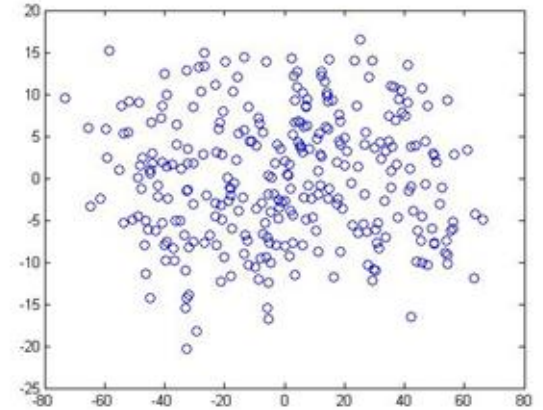
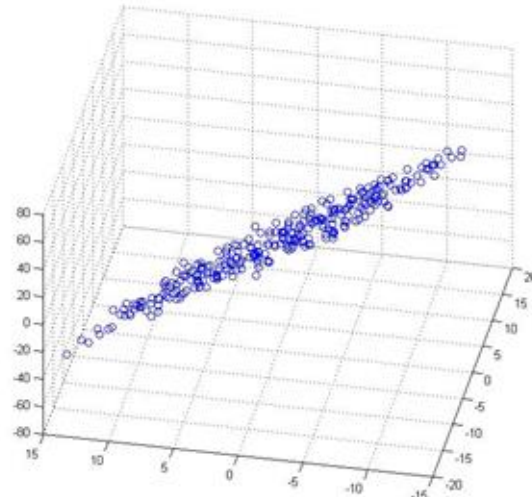
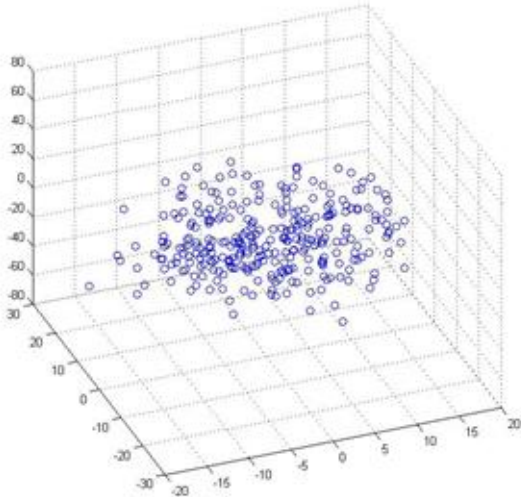
- Why?
 - Lots of features redundant
 - Storage & computation costs
- Goal: take $x \in \mathbb{R}^d \rightarrow x \in \mathbb{R}^r$, for $r \ll d$
 - But, minimize information loss



CreativeBloq

Dimensionality Reduction

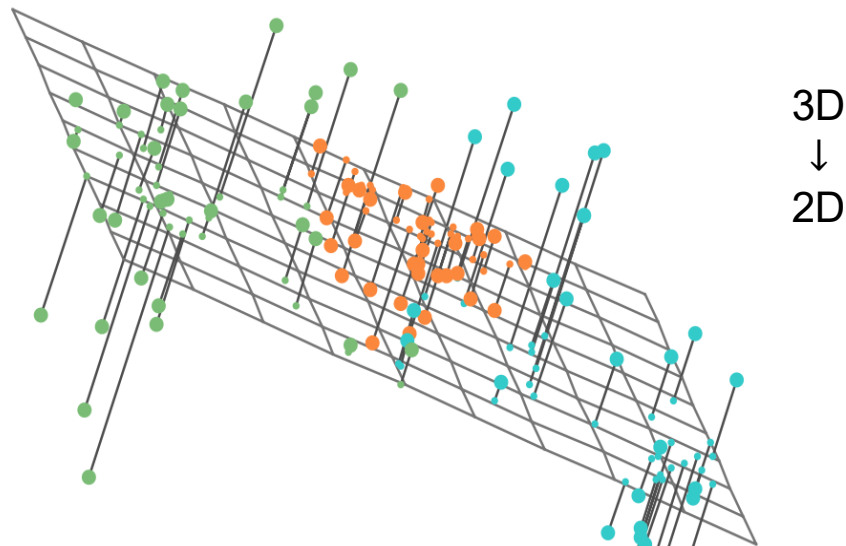
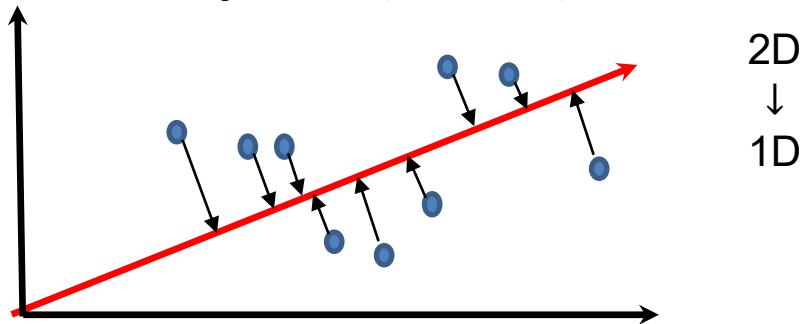
Examples: 3D to 2D



Andrew Ng

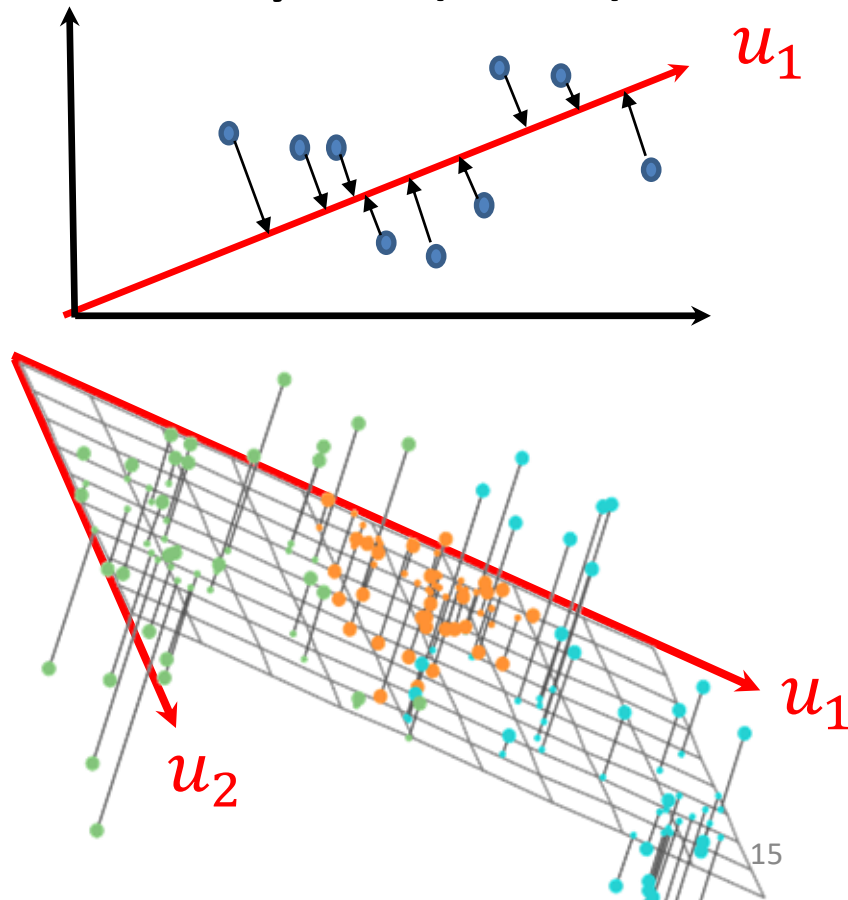
Principal Components Analysis (PCA)

- A type of dimensionality reduction approach
- For when data is **approximately lower dimensional**



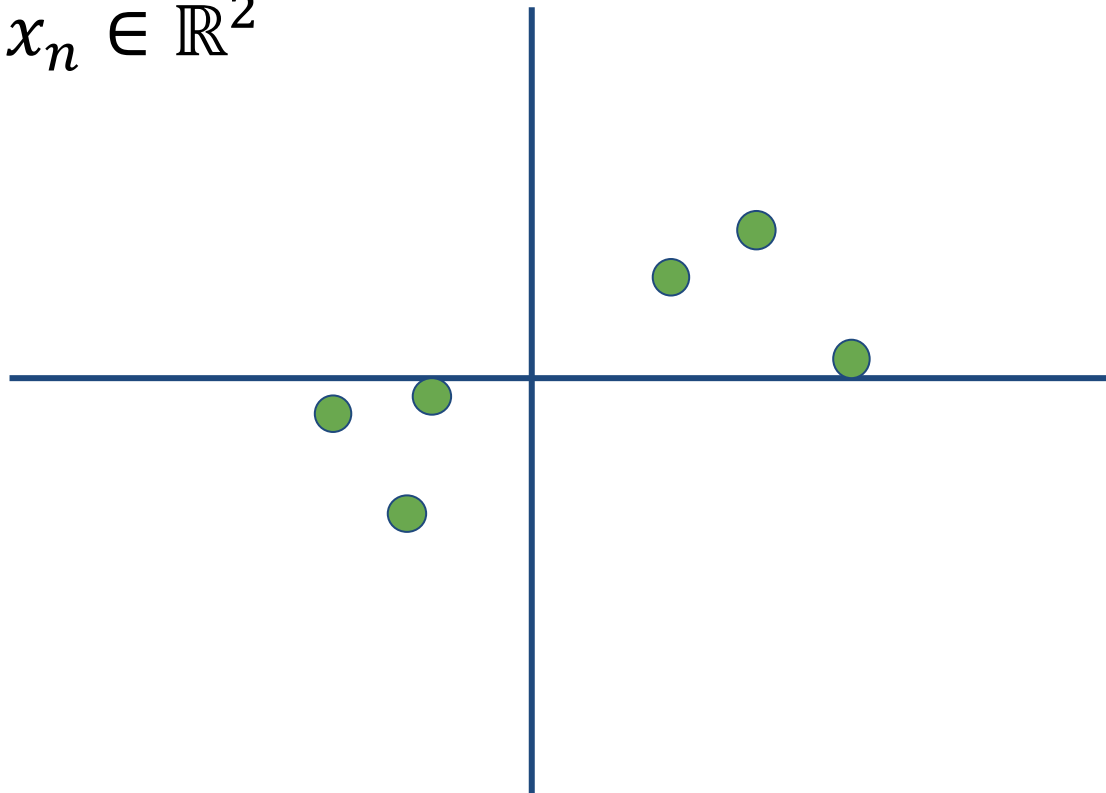
Principal Components Analysis (PCA)

- Find **axes** $u_1, u_2, \dots, u_r \in \mathbb{R}^d$ of a subspace
 - Will project to this subspace
- Want to preserve data
 - minimize projection error
- These vectors are the **principal components**



Projection: An Example

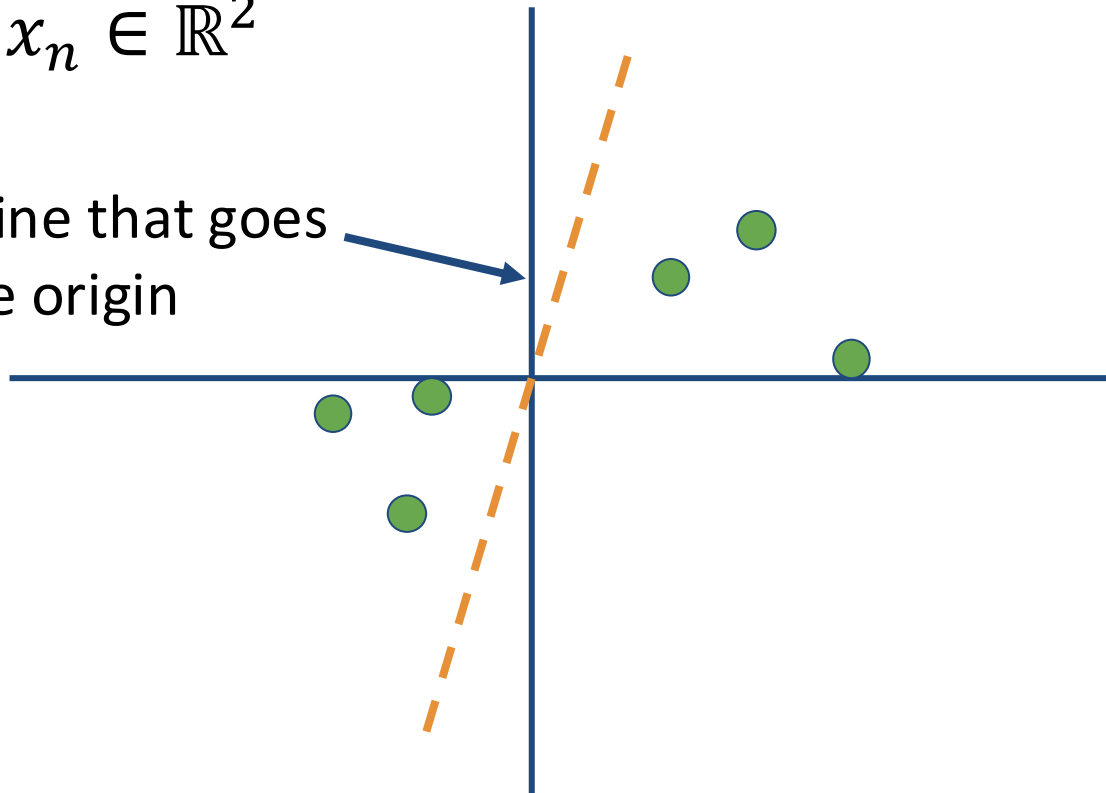
$$x_1, x_2, \dots, x_n \in \mathbb{R}^2$$



Projection: An Example

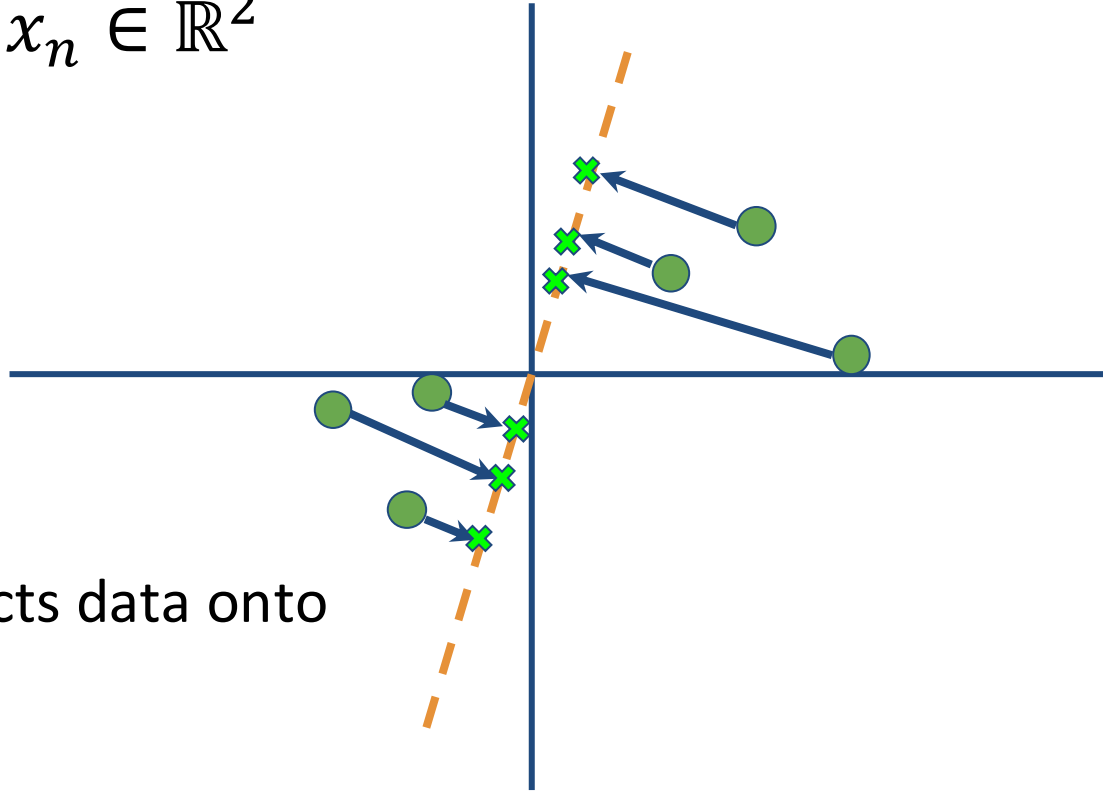
$$x_1, x_2, \dots, x_n \in \mathbb{R}^2$$

A random line that goes
through the origin



Projection: An Example

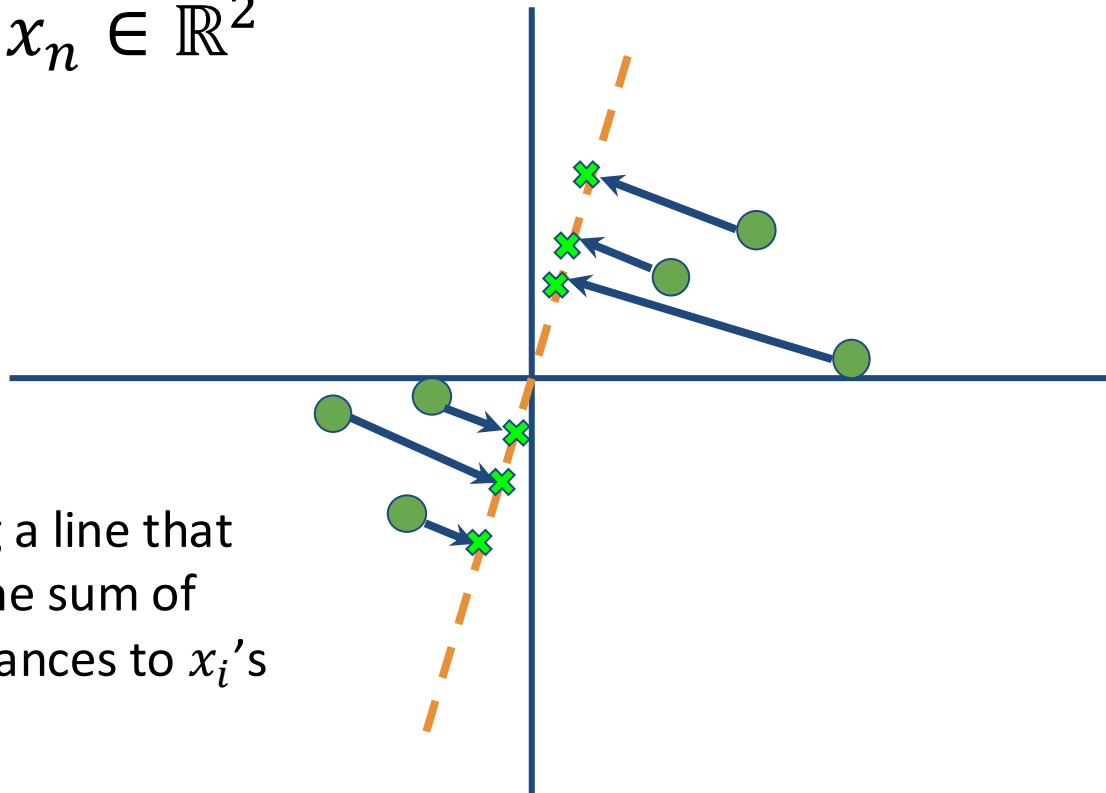
$$x_1, x_2, \dots, x_n \in \mathbb{R}^2$$



PCA projects data onto
this line

Projection: An Example

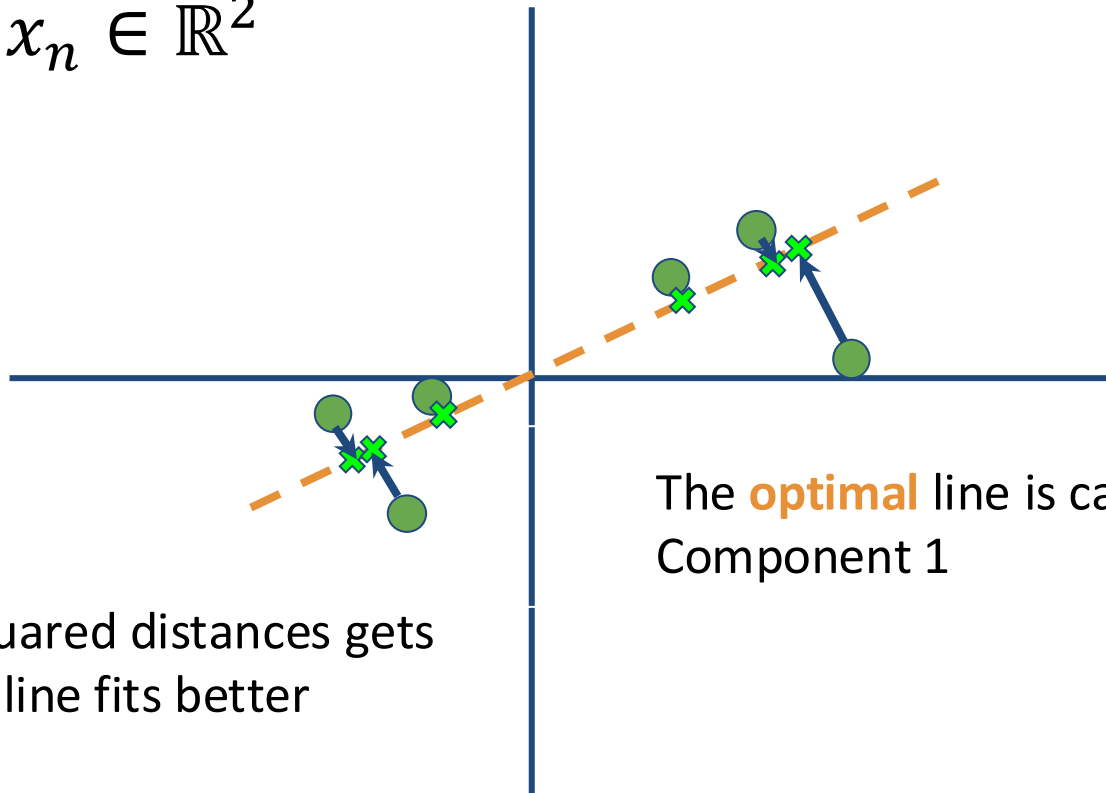
$$x_1, x_2, \dots, x_n \in \mathbb{R}^2$$



Goal: finding a line that
minimizes the sum of
squared distances to x_i 's

Projection: An Example

$$x_1, x_2, \dots, x_n \in \mathbb{R}^2$$

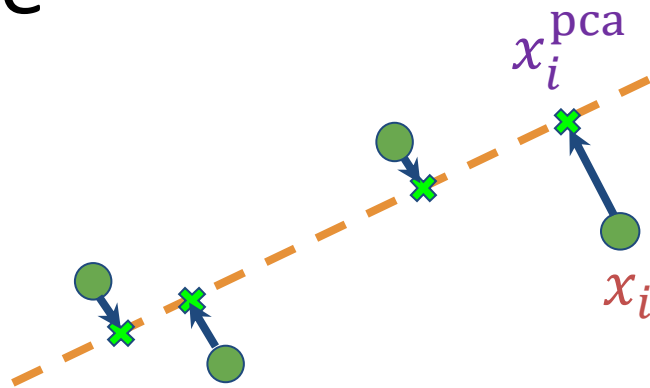


The **optimal** line is called Principal Component 1

The sum of squared distances gets smaller as the line fits better

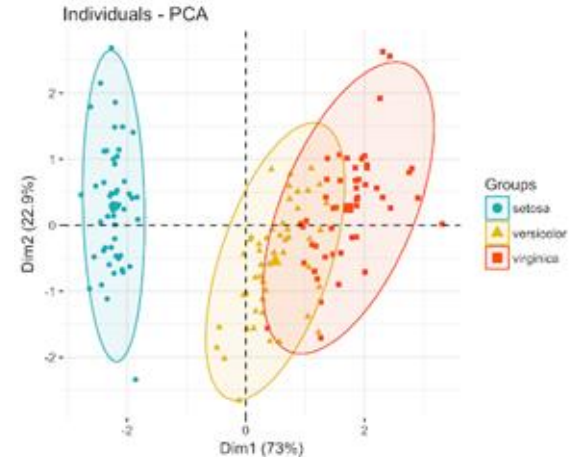
PCA Procedure

- **Inputs:** data $x_1, x_2, \dots, x_n \in \mathbb{R}^d$
 - **Centered data with** $\frac{1}{n} \sum_{i=1}^n x_i = 0$
- **Output:**
principal components $u_1, \dots, u_r \in \mathbb{R}^d$
 - Can show: they are top **eigenvectors** of $S = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top$ (covariance matrix)
 - Each x_i projected to $x_i^{\text{pca}} = \sum_{j=1}^m u_j (u_j^\top x_i)$



Many Variations

- PCA, Kernel PCA, ICA, CCA
 - Extract structure from high dimensional dataset
- Uses:
 - **Visualization**
 - Efficiency
 - Noise removal
 - Downstream machine learning use



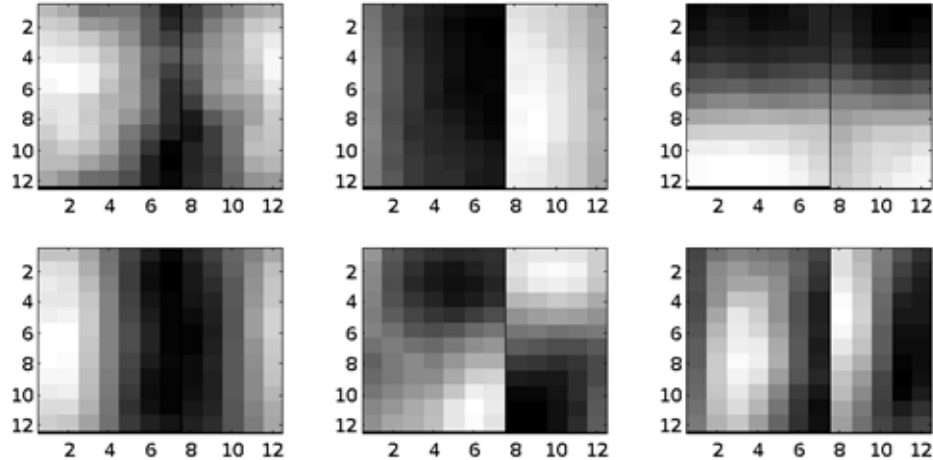
Application: Image Compression

- Start with image; divide into 12x12 patches
 - That is, 144-D vector
 - **Original image:**



Application: Image Compression

- 6 principal components (as an image)



Application: Image Compression

- Project to 6D



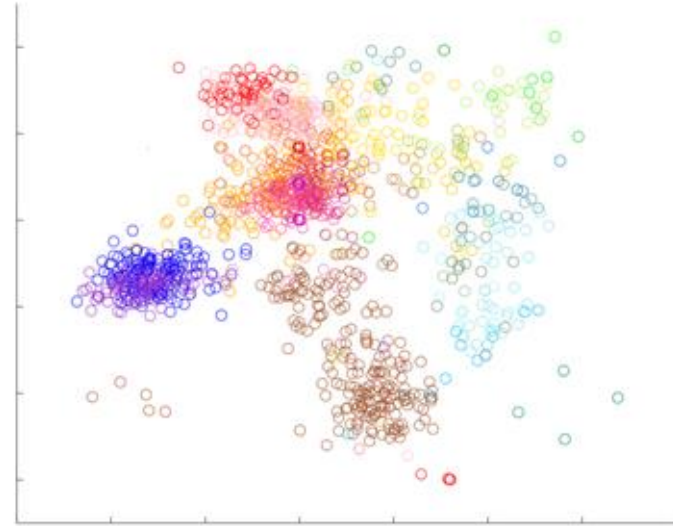
Compressed



Original

Application: Exploratory Data Analysis

- [Novembre et al. '08]: Take top two singular vectors of people x SNP matrix (POPRES)



“Genes Mirror Geography in Europe”