Clustering

Virginia de Sa
desa at cogsci

Visualizing High-Dimensional Data

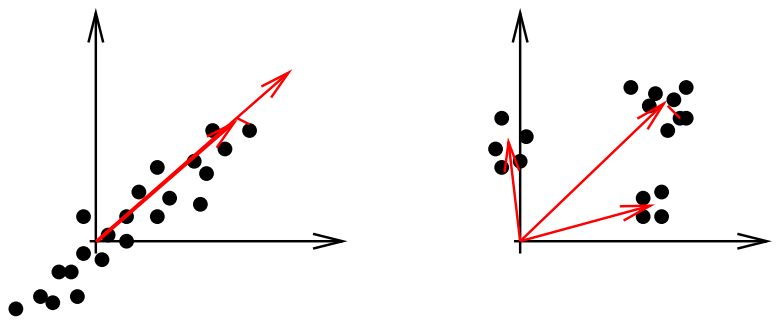
In general we might want to reduce the dimensionality of our data for better visualization. (Note this is very useful for machine learning algorithms too and also for efficient storage and transmission of information (gzip, jpg).)

We want to do this while preserving as much as the *useful* information as possible. (Of course how *useful* is determined is critical).

Clustering and **PCA** are different methods of dimensionality reduction that can aid visualization.

Here is a nice example to show you the benefit to clustering data. clustering code and output

PCA and Clustering



PCA represents a point using a fewer number of dimensions. The directions are the directions of greatest variance in the data

Clustering represents a point using prototype points.

K-means

a simple but effective clustering algorithm

partitions the data in to K disjoint sets (clusters)

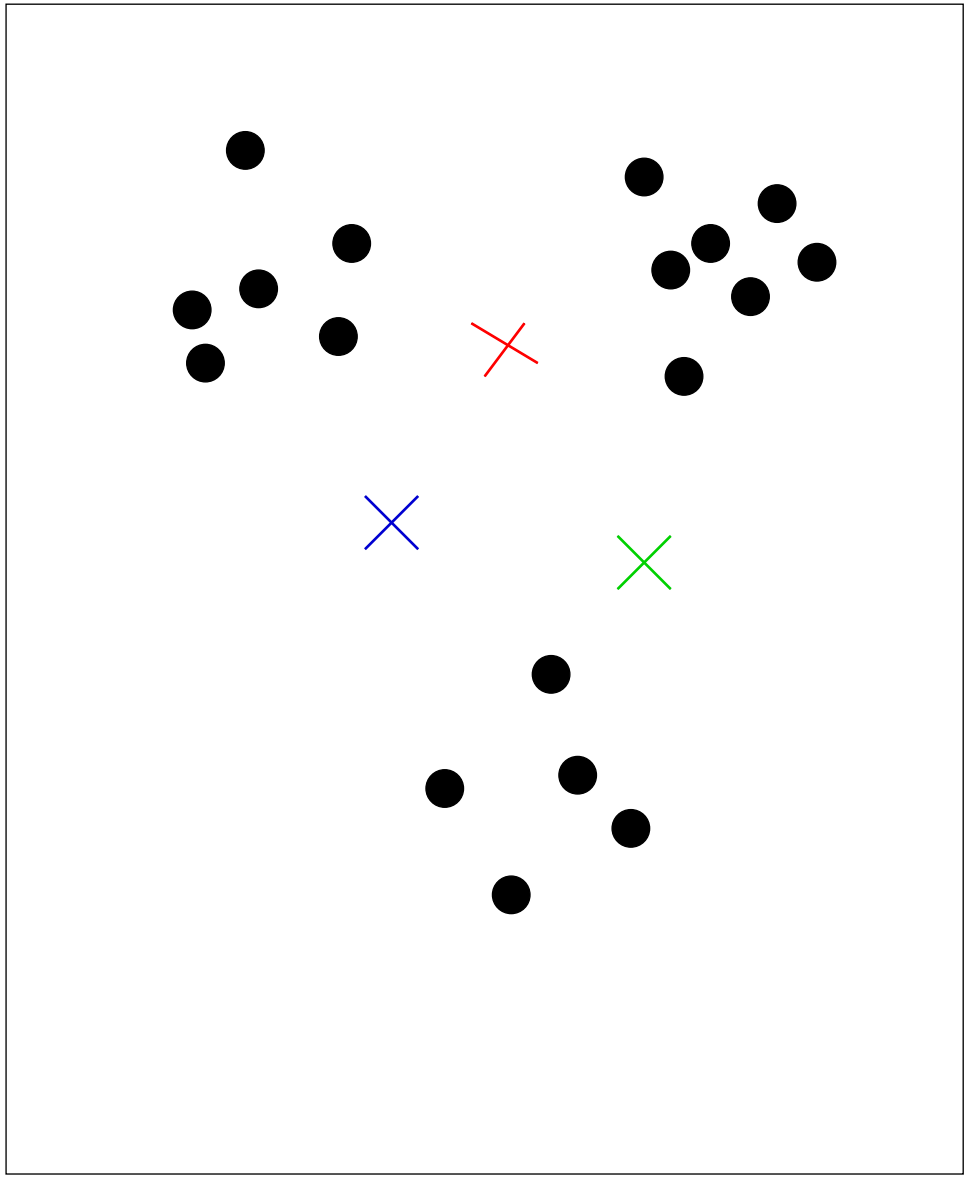
iterative batch algorithm

- Start with initial guess of k centers
- $S^{(j)}$ is all points closest to $\mu^{(j)}$
- Update

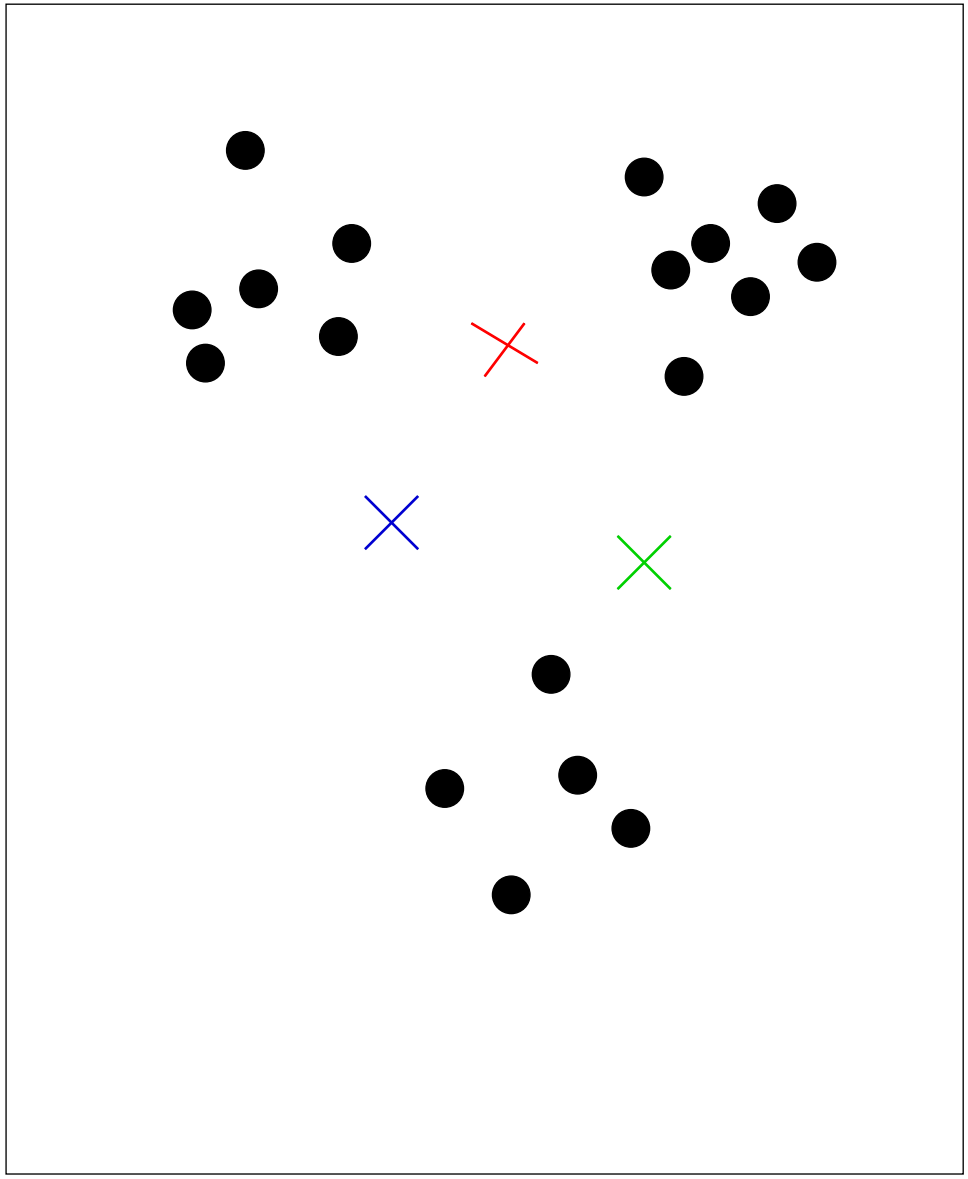
$$\mu^{(j)} = 1/N_j \sum_{n \in S^{(j)}} \mathbf{x}^{(n)}$$

- until no change in the means

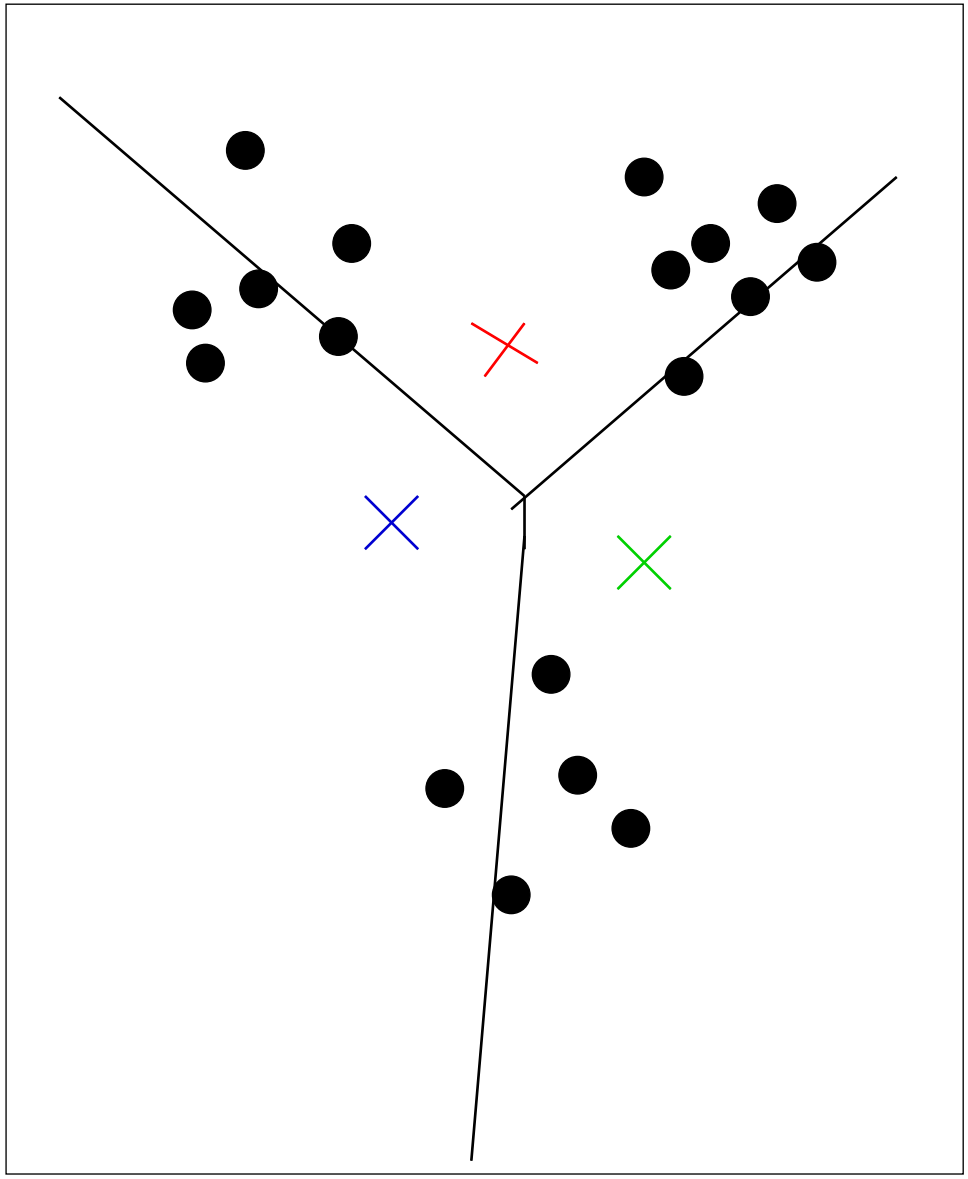
K-means



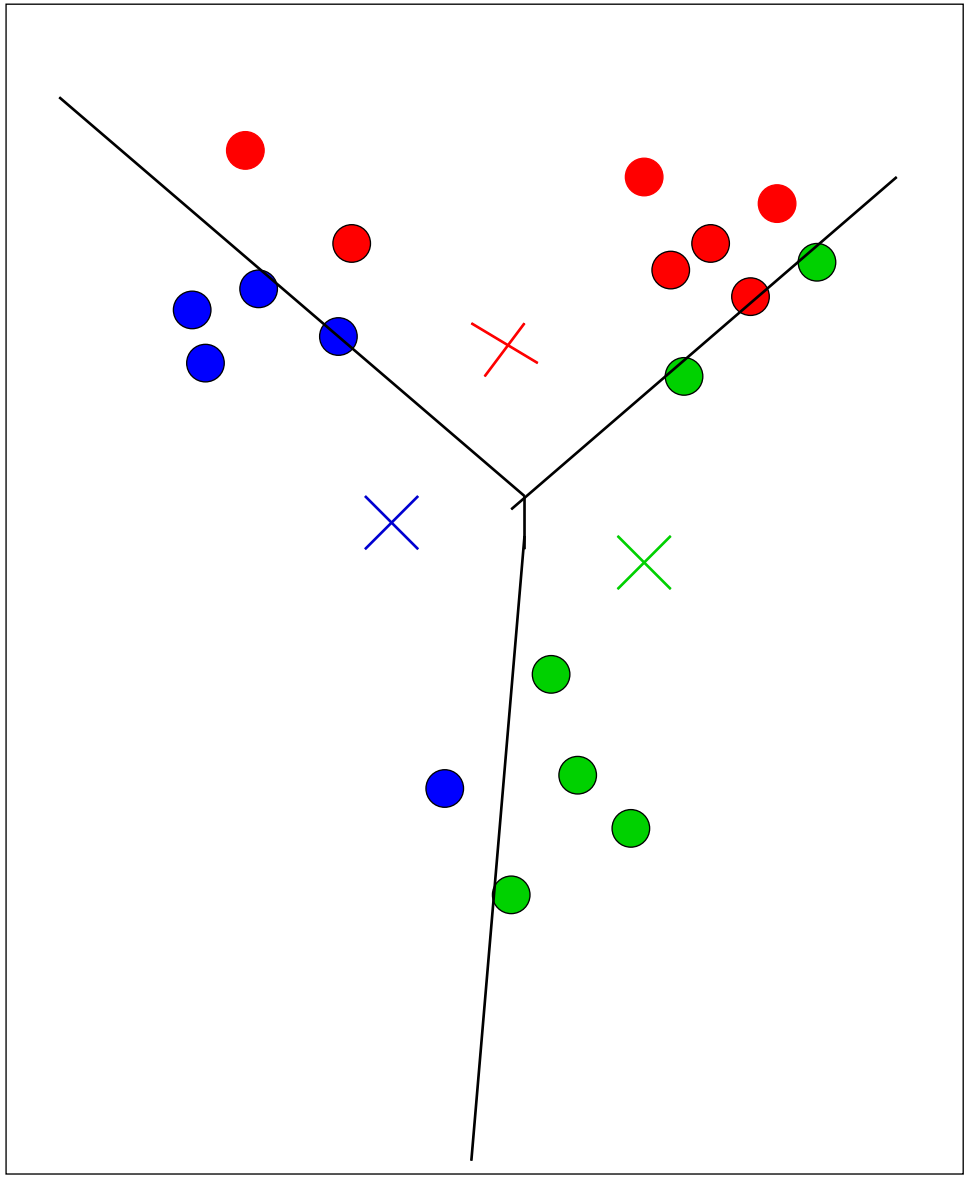
K-means



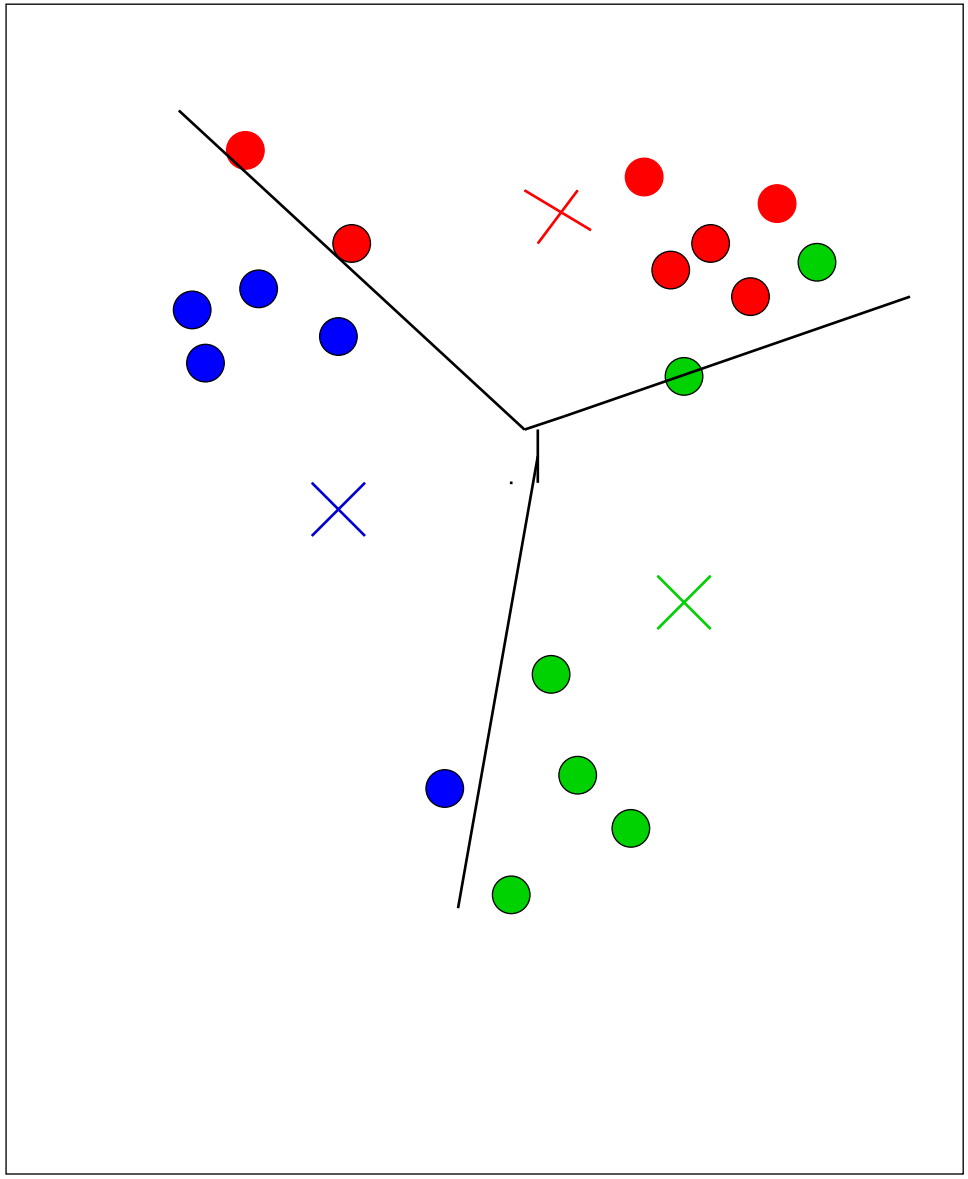
K-means



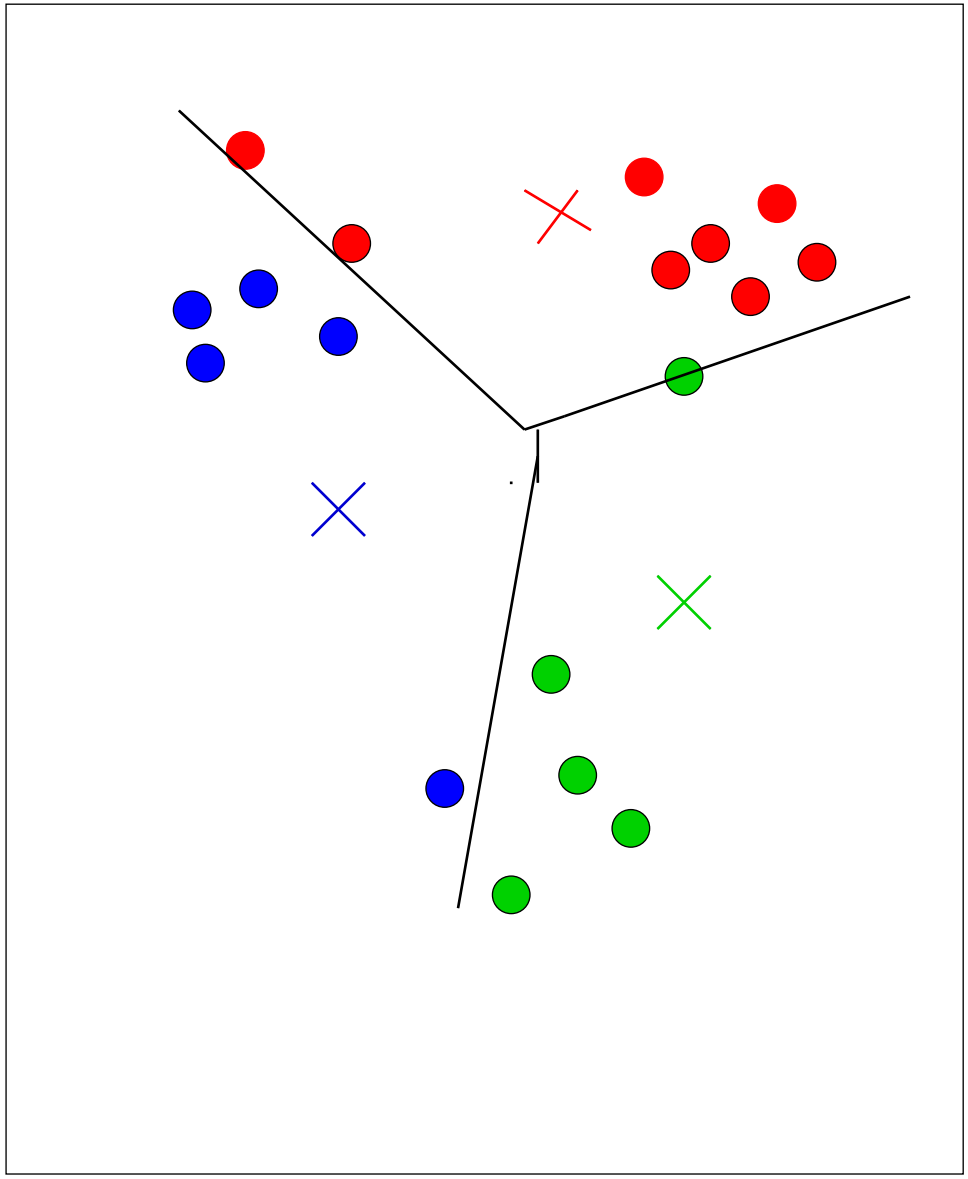
K-means



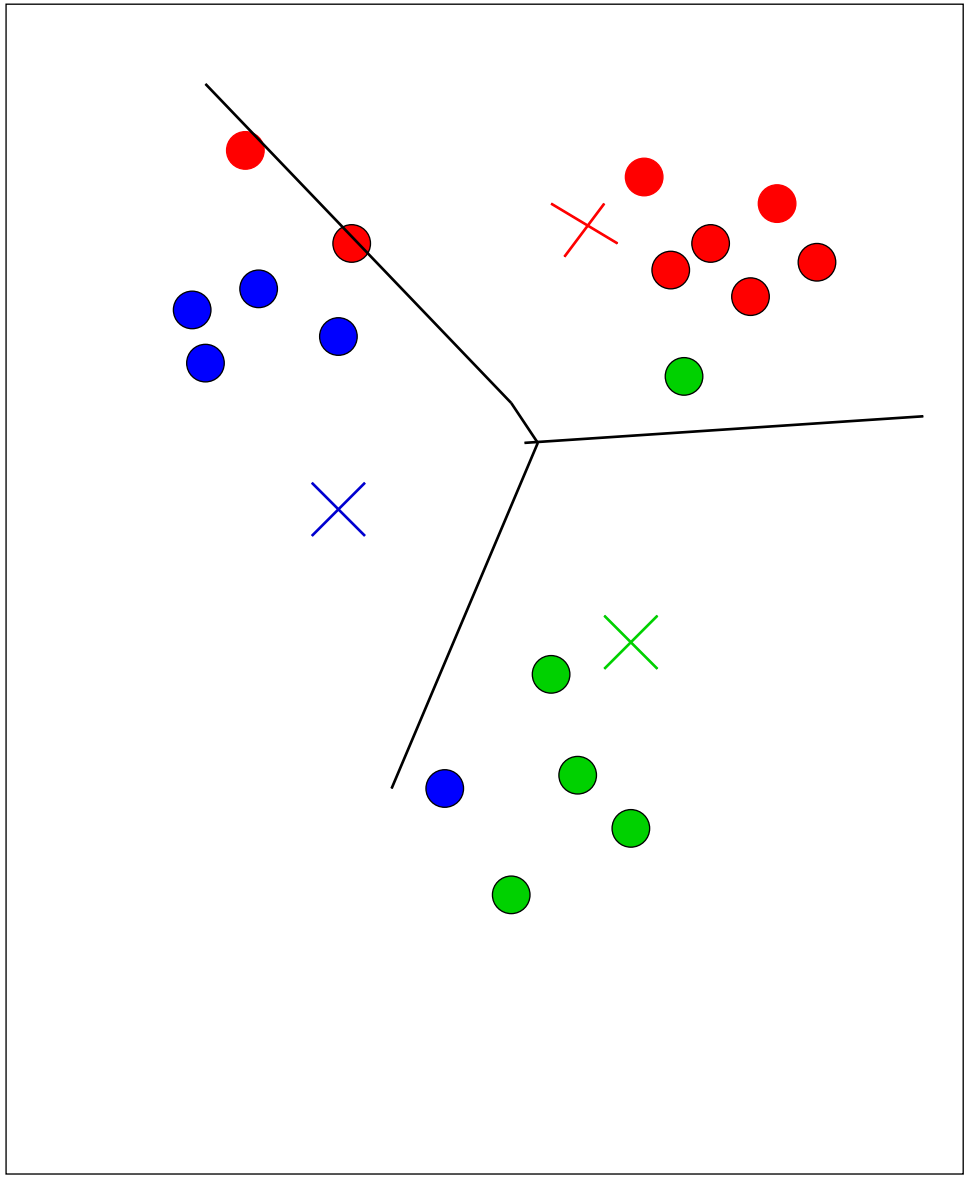
K-means



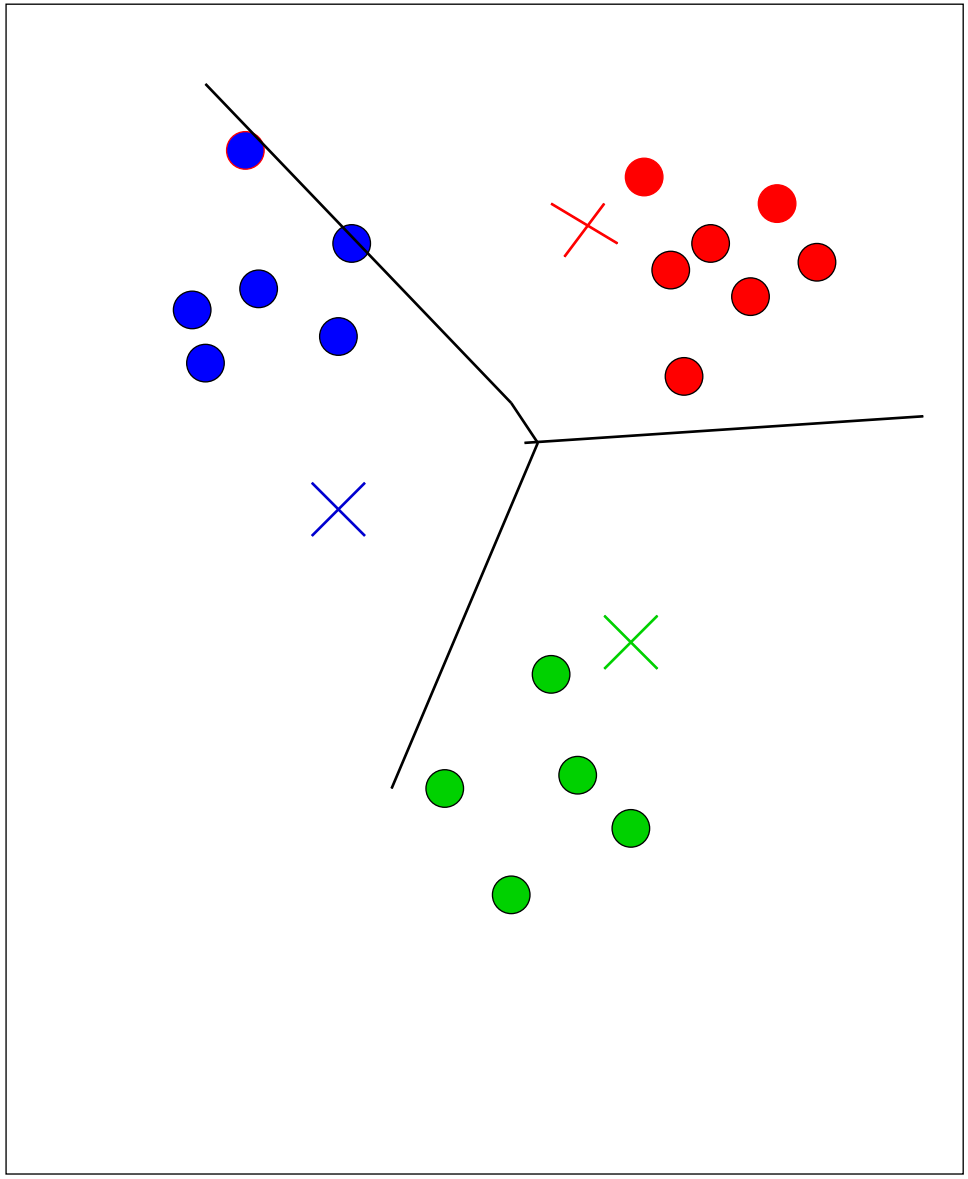
K-means



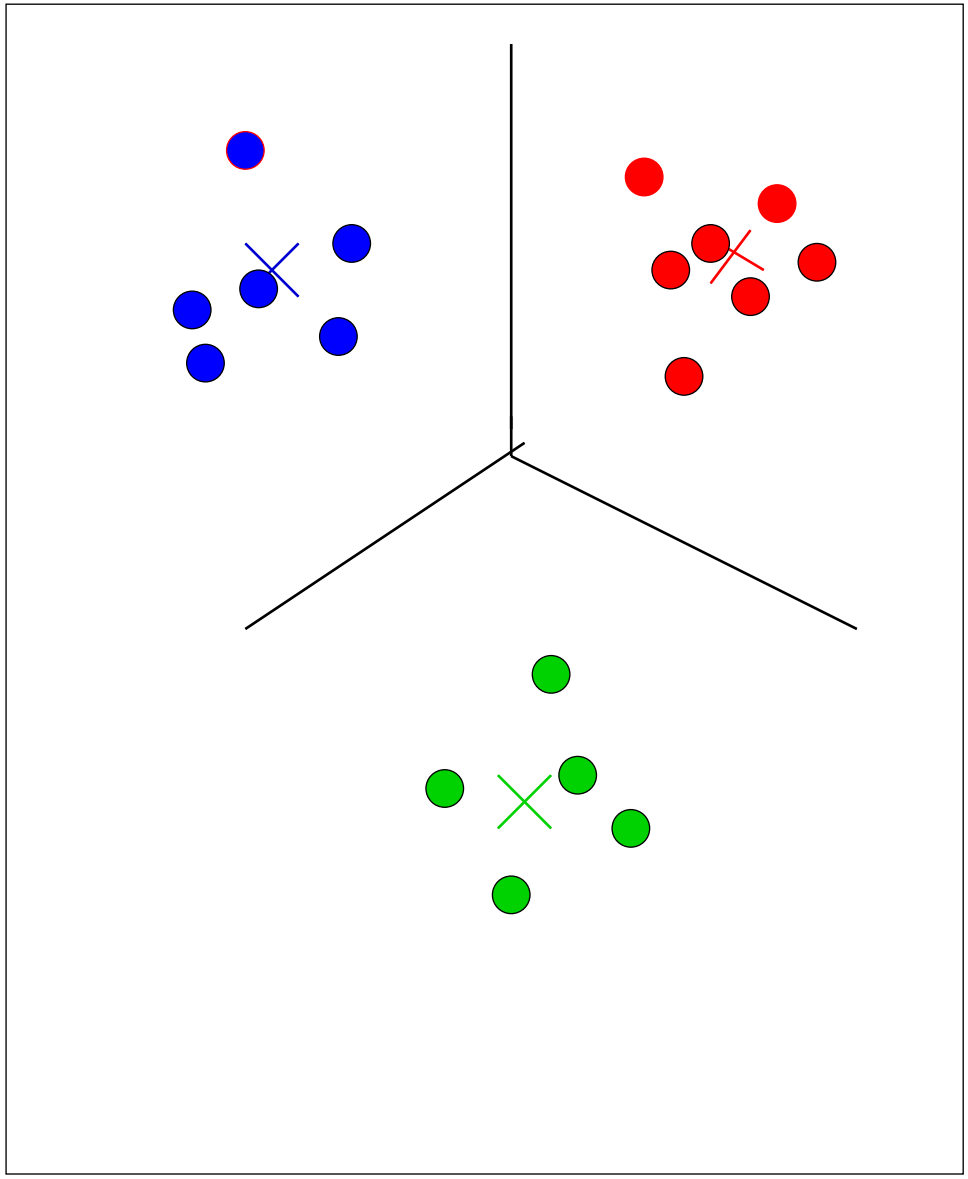
K-means



K-means



K-means



K-means

a simple but effective clustering algorithm

partitions the data in to K disjoint sets (clusters)

iterative batch algorithm

- Start with initial guess of k centers
- $S^{(j)}$ is all points closest to $\mu^{(j)}$
- Update

$$\mu^{(j)} = 1/N_j \sum_{n \in S^{(j)}} \mathbf{x}^{(n)}$$

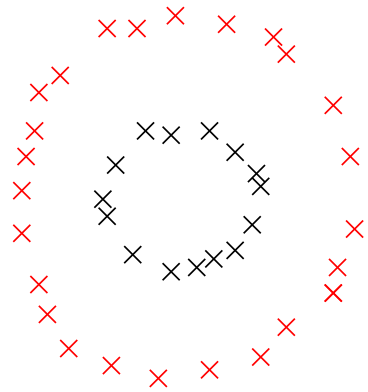
- until no change in the means

There are many more complex clustering schemes

- hierarchical clustering allows hierarchy of clusters
- mixture of Gaussians allows elongated clusters
- spectral clustering allows non-compact clusters

Spectral clustering

- Can cluster data of this form



- is very efficient to program and run in Matlab

Image segmentation results (Fowlkes,Martin,Malik) using spectral clustering

http://www.cs.berkeley.edu/ fowlkes/BSE/cvpr-segs/

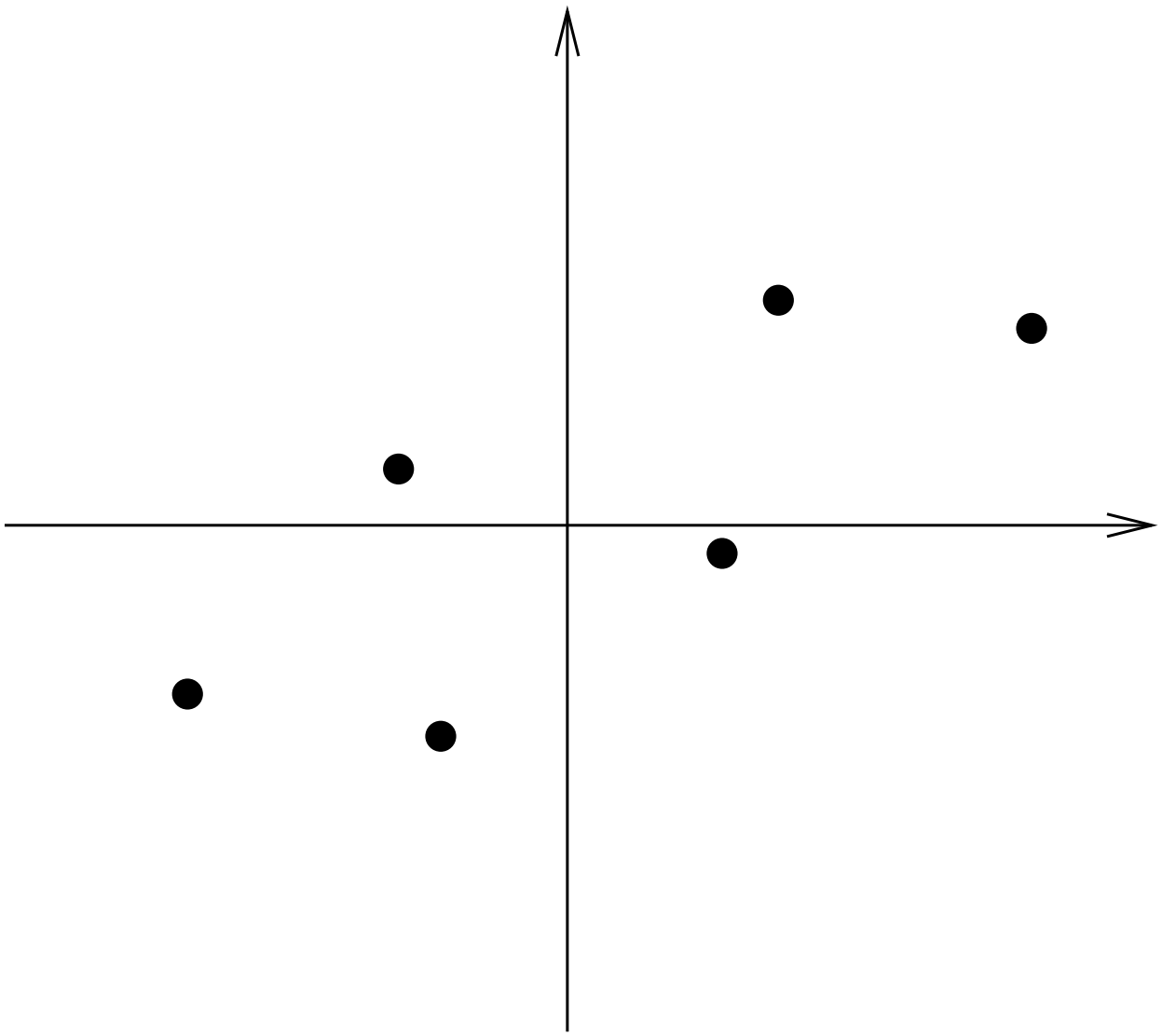
(work by Fowlkes, Martin, Malik (UC Berkeley) – uses sophisticated learning of pairwise similarities)

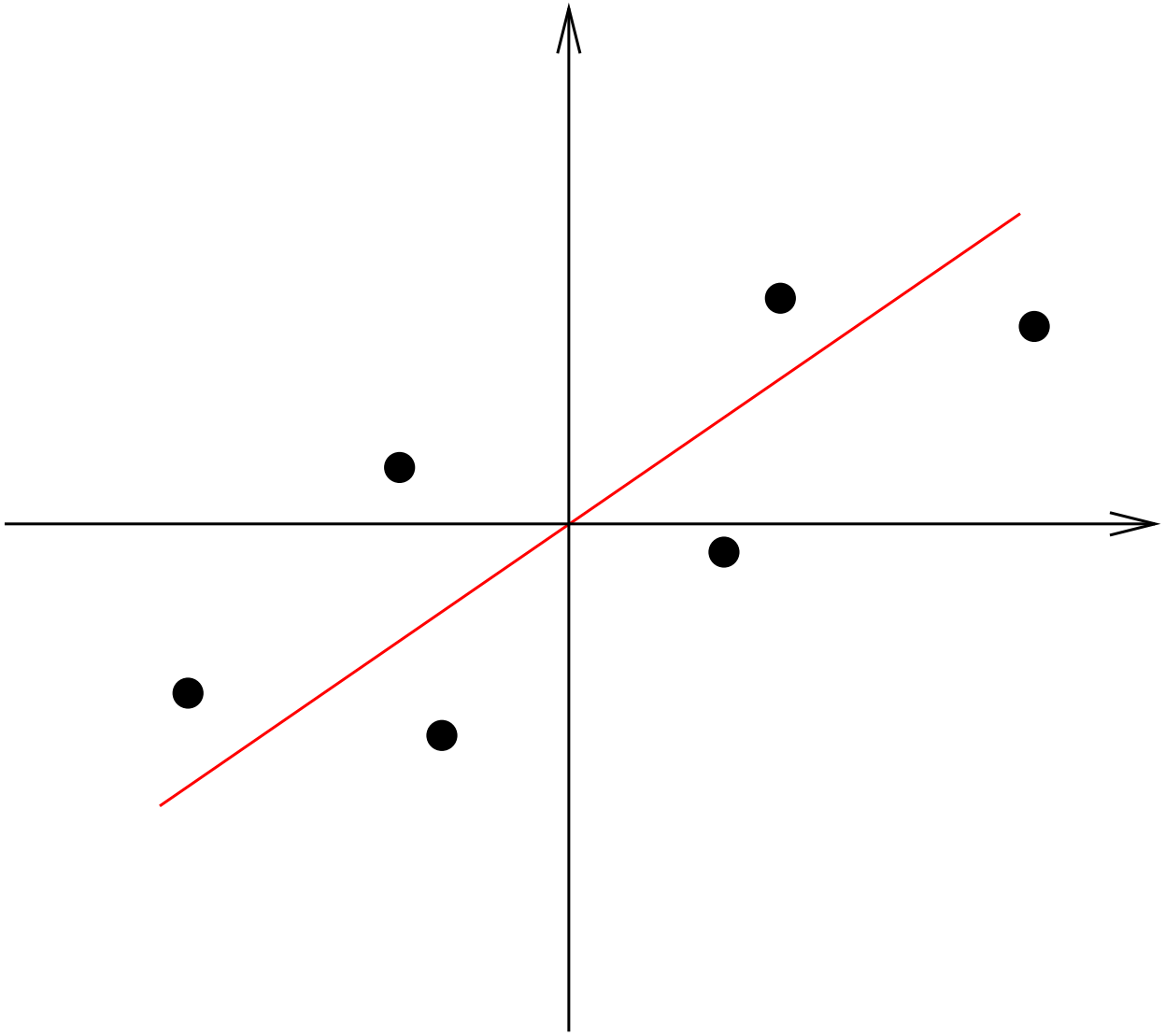
PCA overview

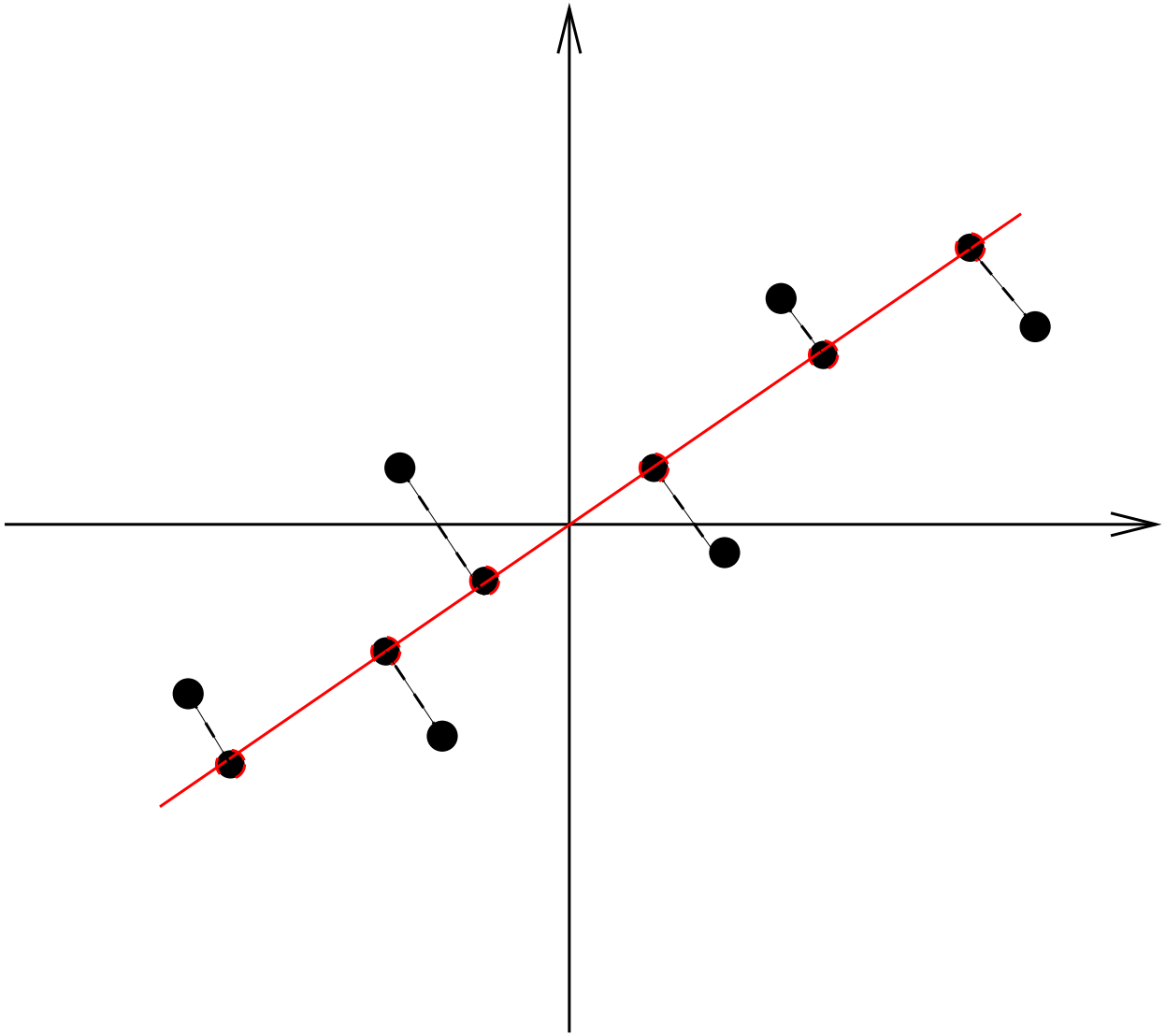
PCA finds the linear subspace with the greatest projected variance.

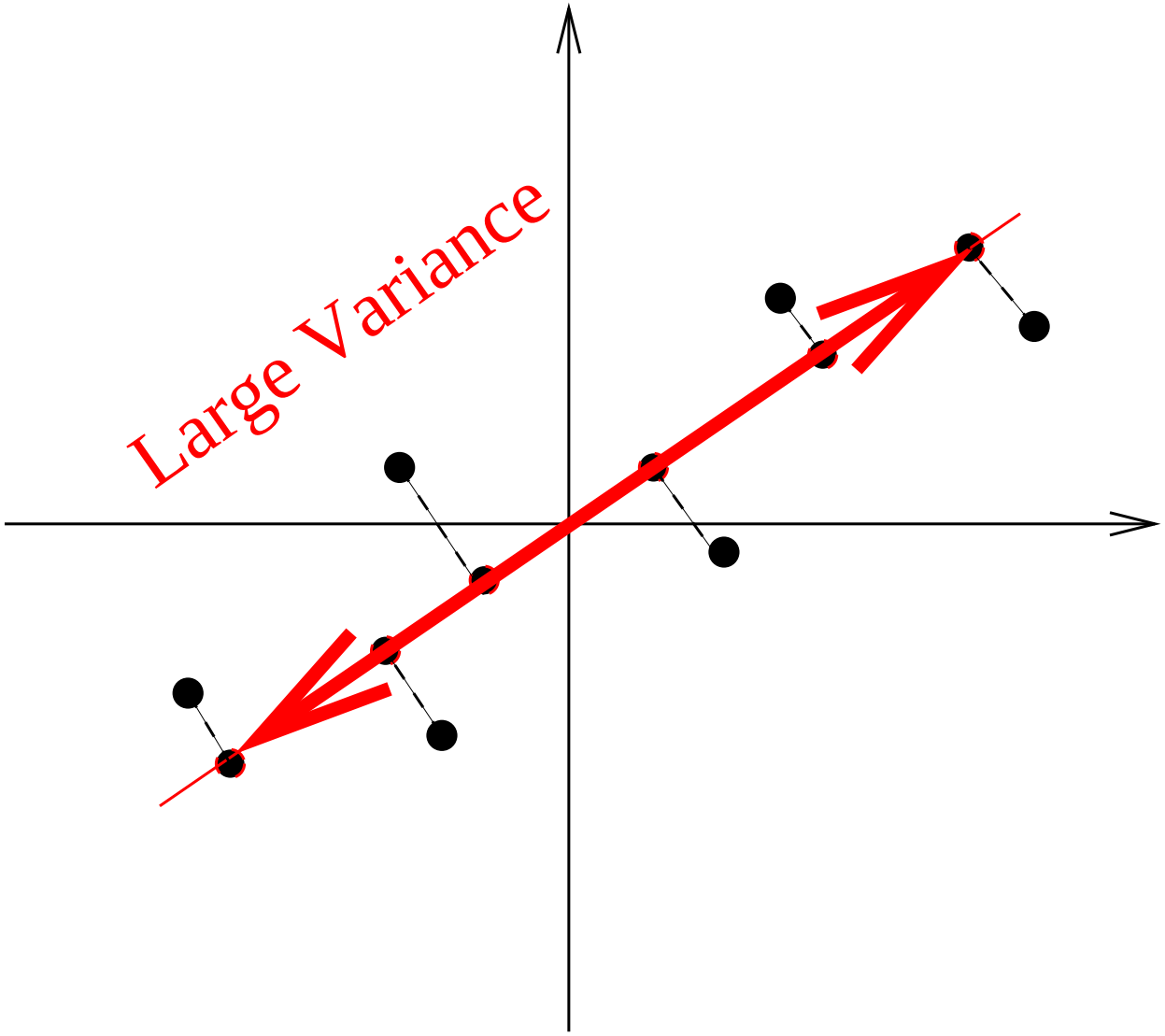
It is commonly used for dimensionality reduction – project to a reduced subspace that contains most of the variance.

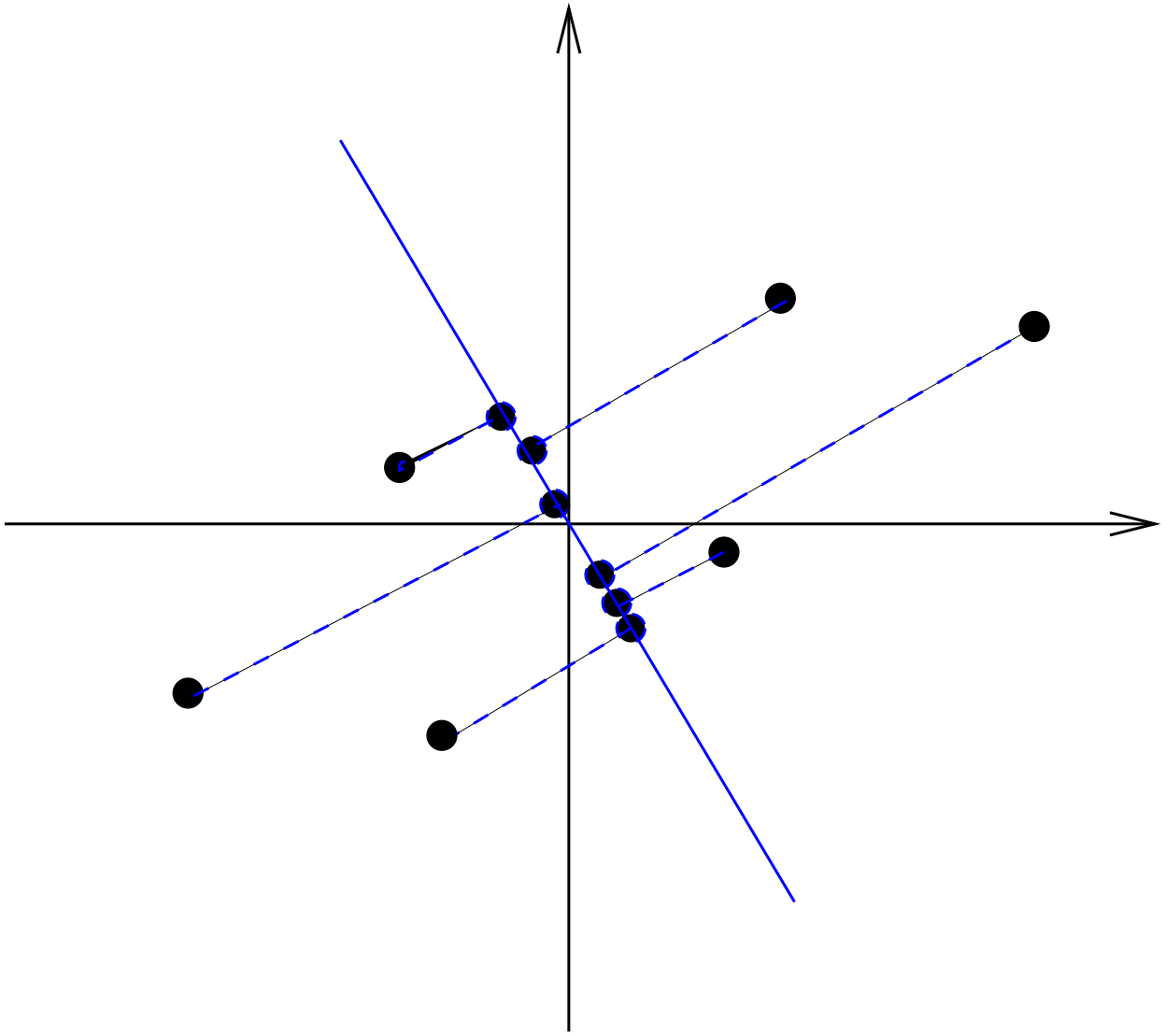
Once the desired dimensionality subspace is found, the data are represented as the projected point in that subspace.

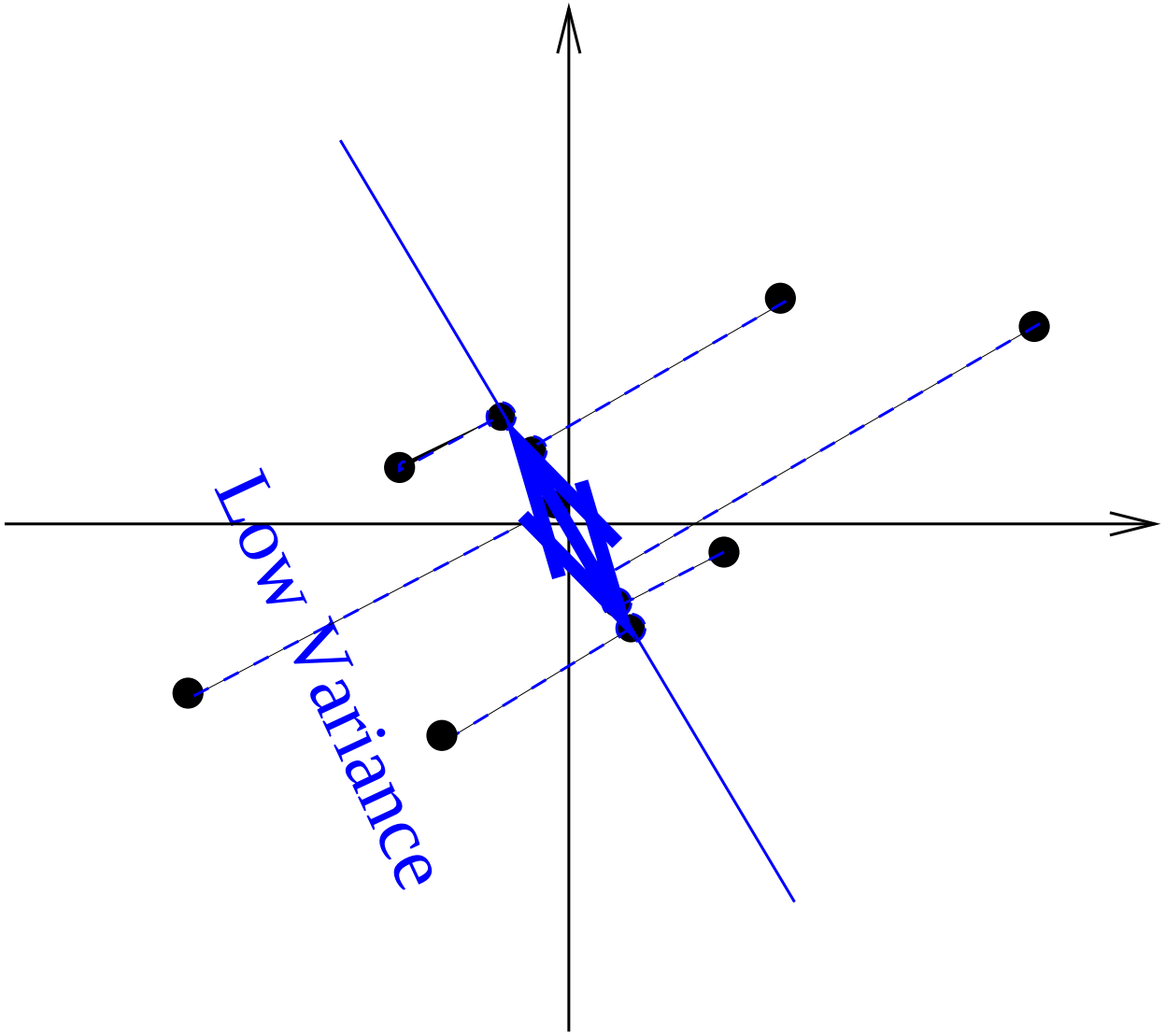







Large Variance




Low Variance