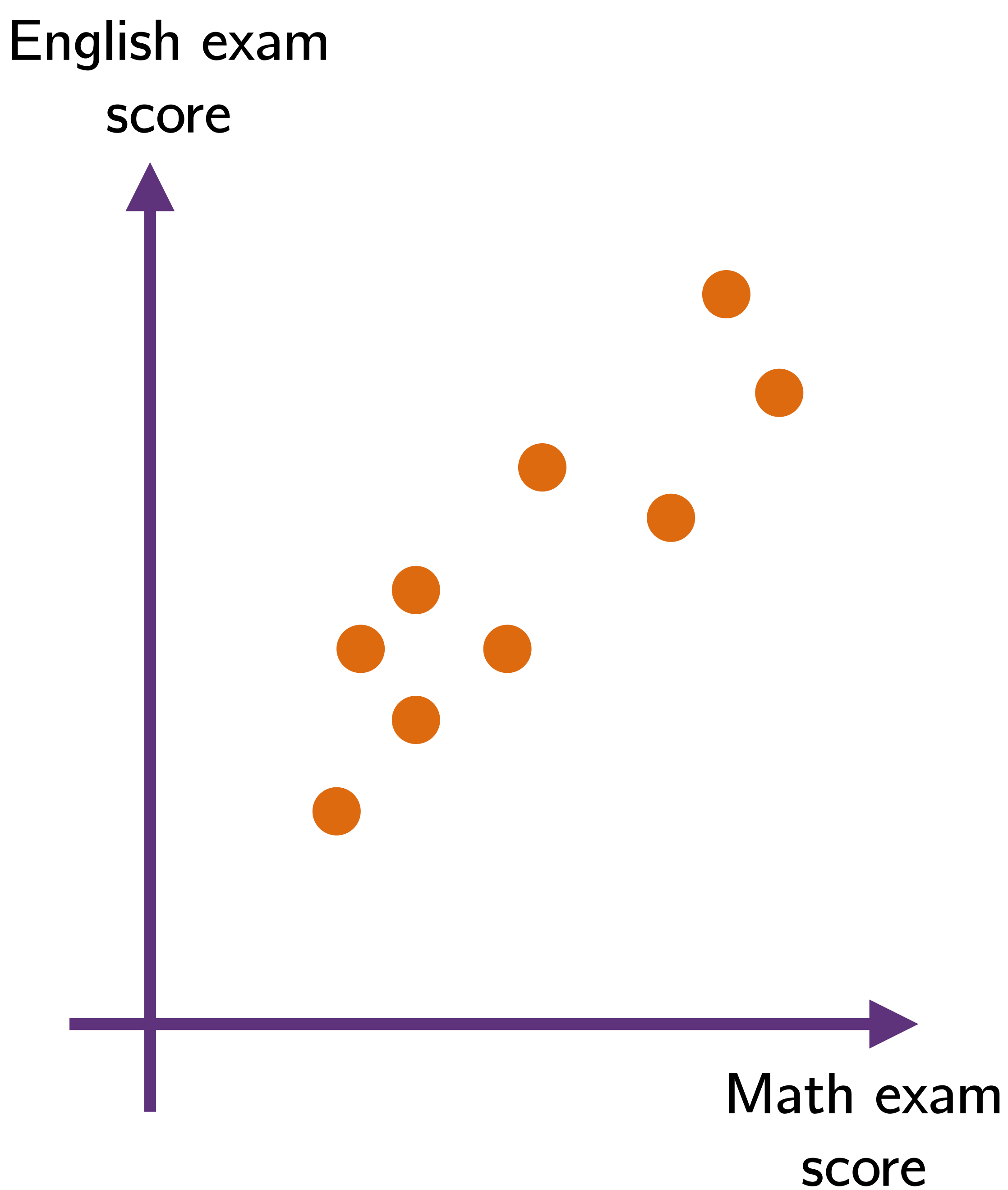
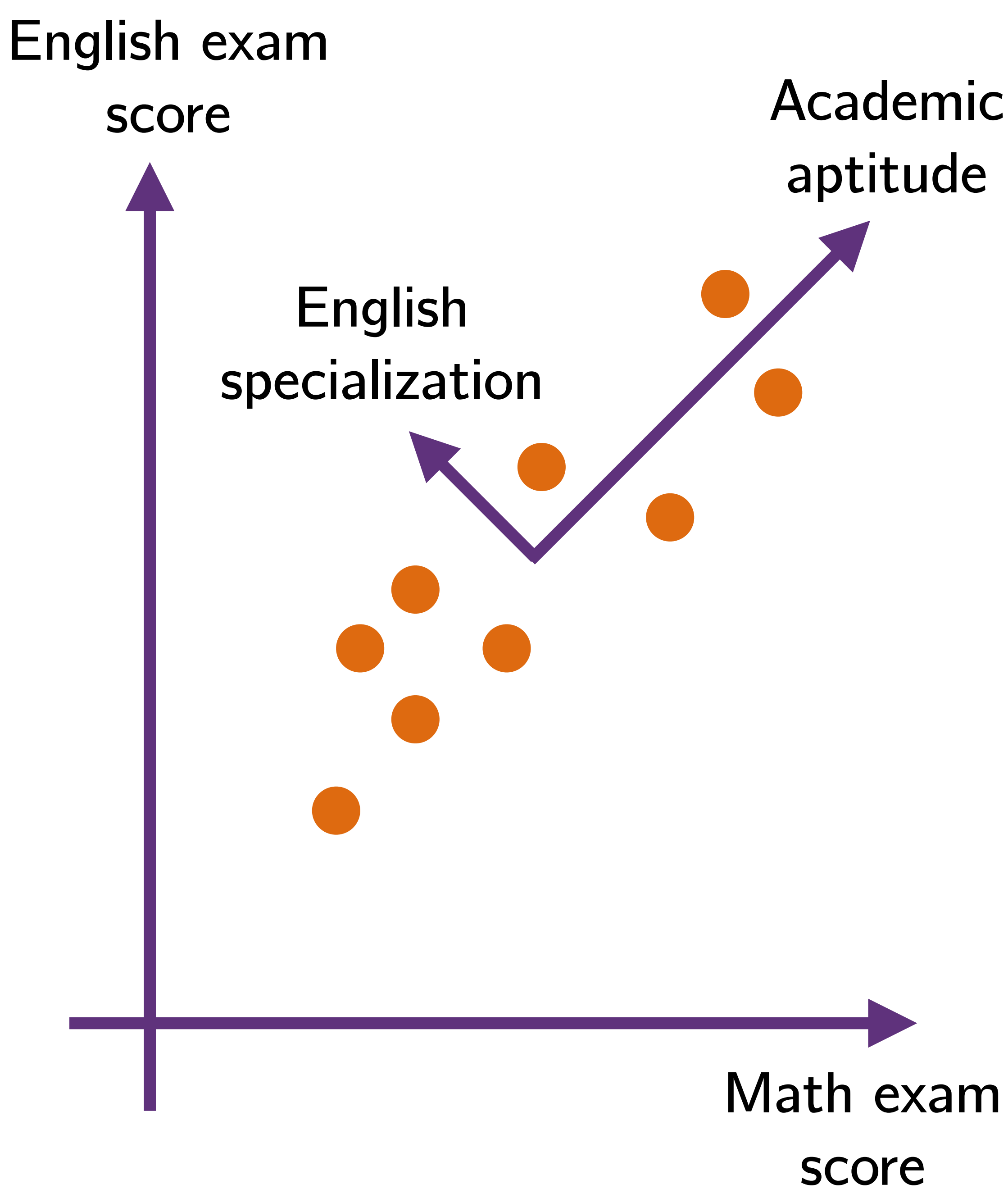
Principal component analysis



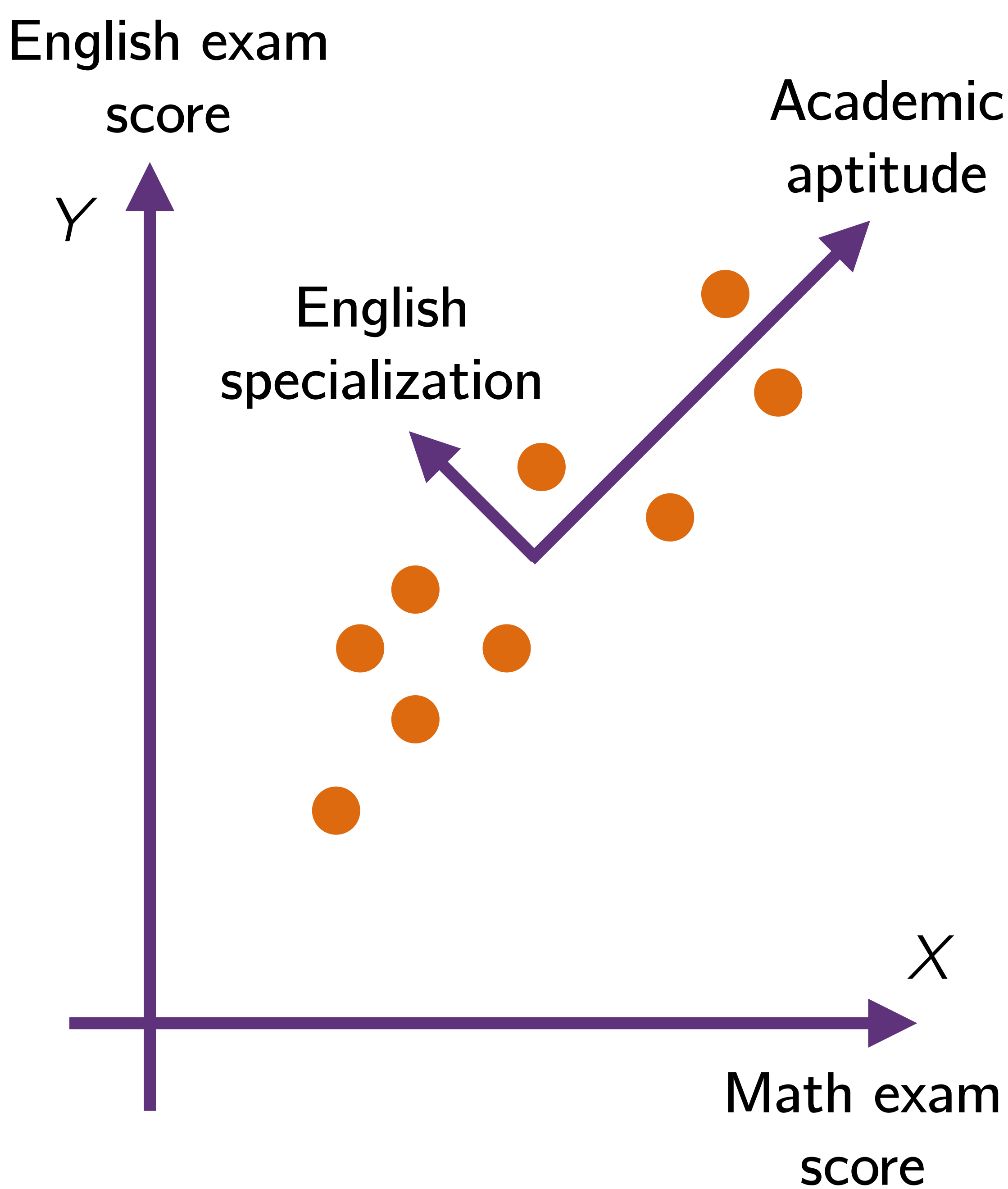
- Suppose a number of students take math and english tests
- Typically, we'd expect that the scores for each individual are highly correlated, and roughly lie along a diagonal line
- Given this correlation, is there a better way to view this data?

Principal component analysis



- Introduce new rotated coordinate system
- Academic aptitude is most important coordinate: it explains the majority of the variance in the students
- If we just stored academic aptitude for each student, that would capture most of what the data says

Principal component analysis



- To automatically find the coordinates, first construct the covariance matrix

$$M = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix}$$

- Symmetric positive-definite matrix, so eigenvalues are positive

- Order eigenvalues

$$\lambda_1 \leq \lambda_2 \leq \dots$$

- Take components as corresponding eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, ...

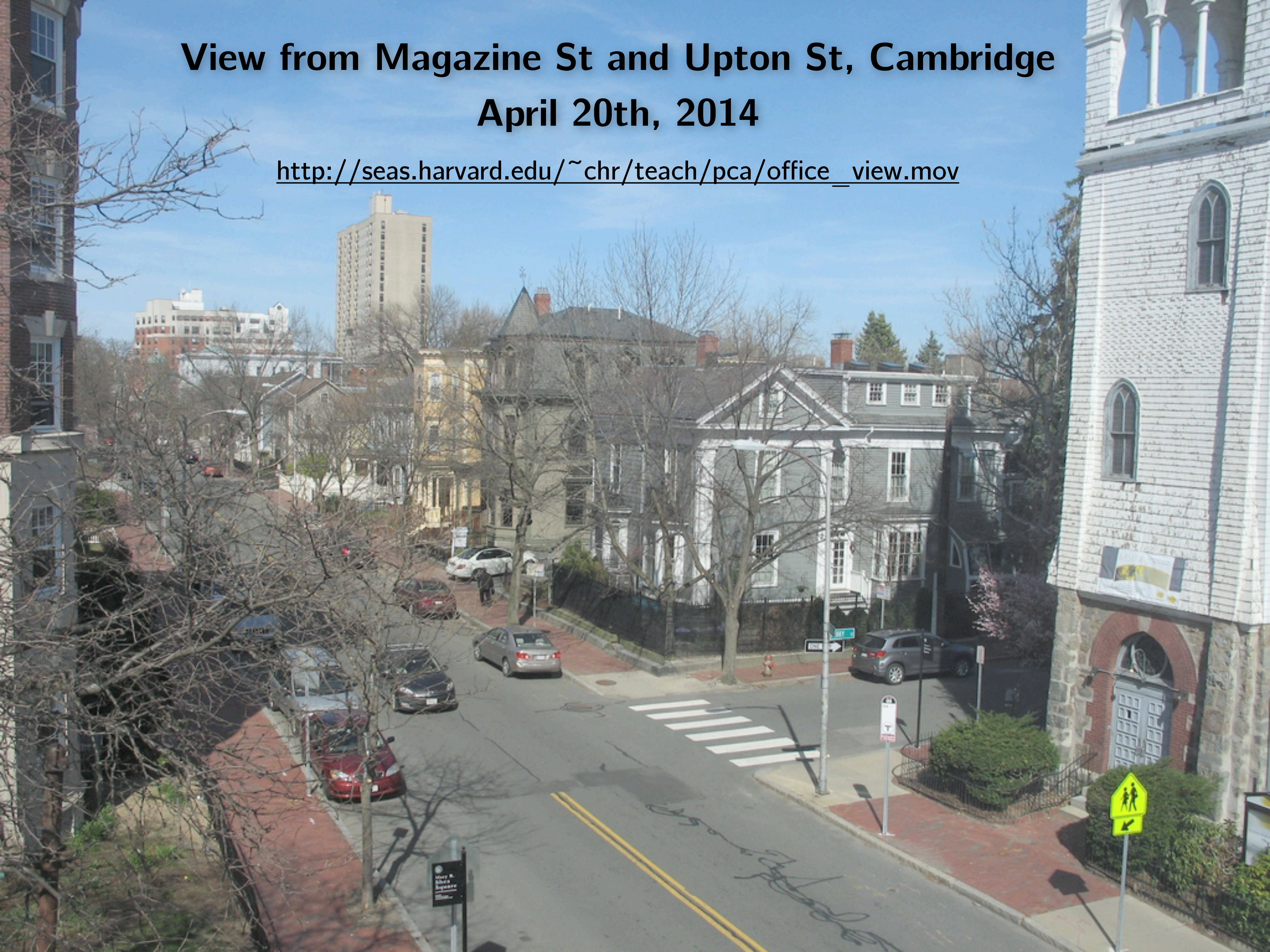
PCA for image analysis

- An $M \times N$ image can be thought of as an MN-dimensional vector $(v_1, v_2, \ldots, v_{MN})$
- Suppose we have many images of similar things, then the points are likely to be on a much lower-dimensional surface within the MN-dimensional space
- Principal components capture the majority of information from the images

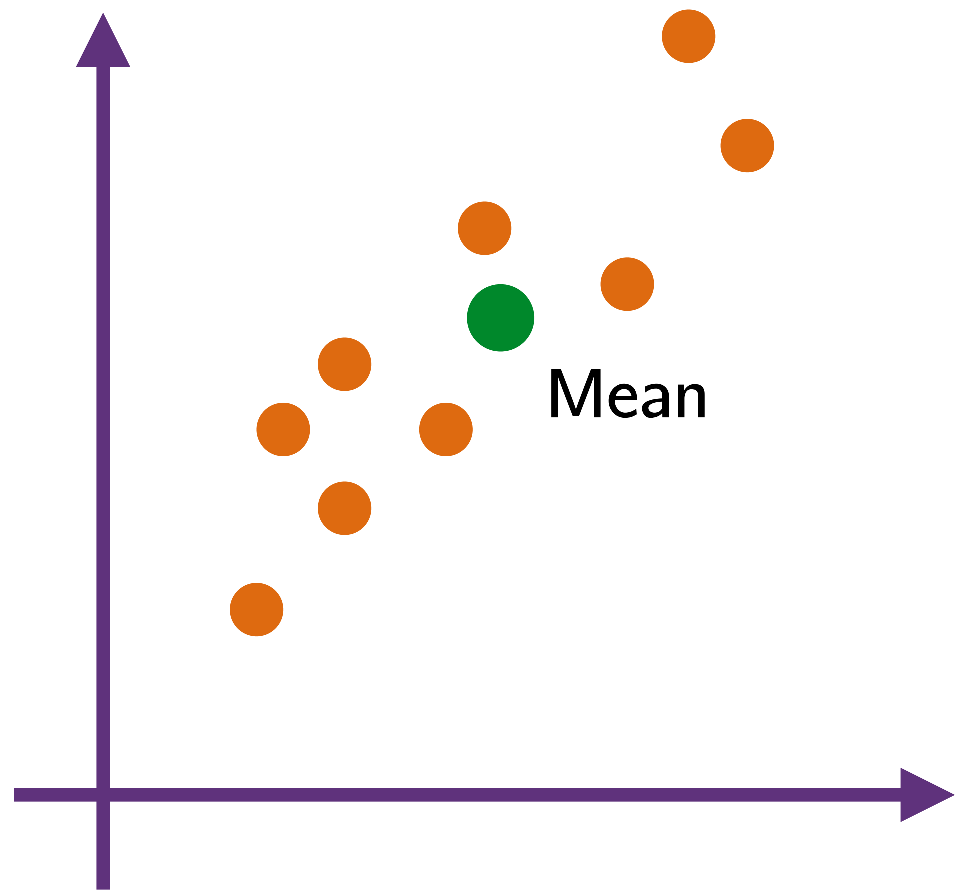
View from Magazine St and Upton St, Cambridge

April 20th, 2014

http://seas.harvard.edu/~chr/teach/pca/office_view.mov



Mean image



Each pixel is the mean color over the 72 frames in the movie

First component

Positive contribution

Negative contribution

The component has a negative part and a positive part, so it's hard to visualize. Hence split the component into two images showing the positive and negative parts separately.

Variance in this component: 85.99% of total

Second component

Positive contribution

Negative contribution

Variance in this component: 7.80% of total

Third component

Positive contribution

Negative contribution

Variance in this component: 2.34% of total

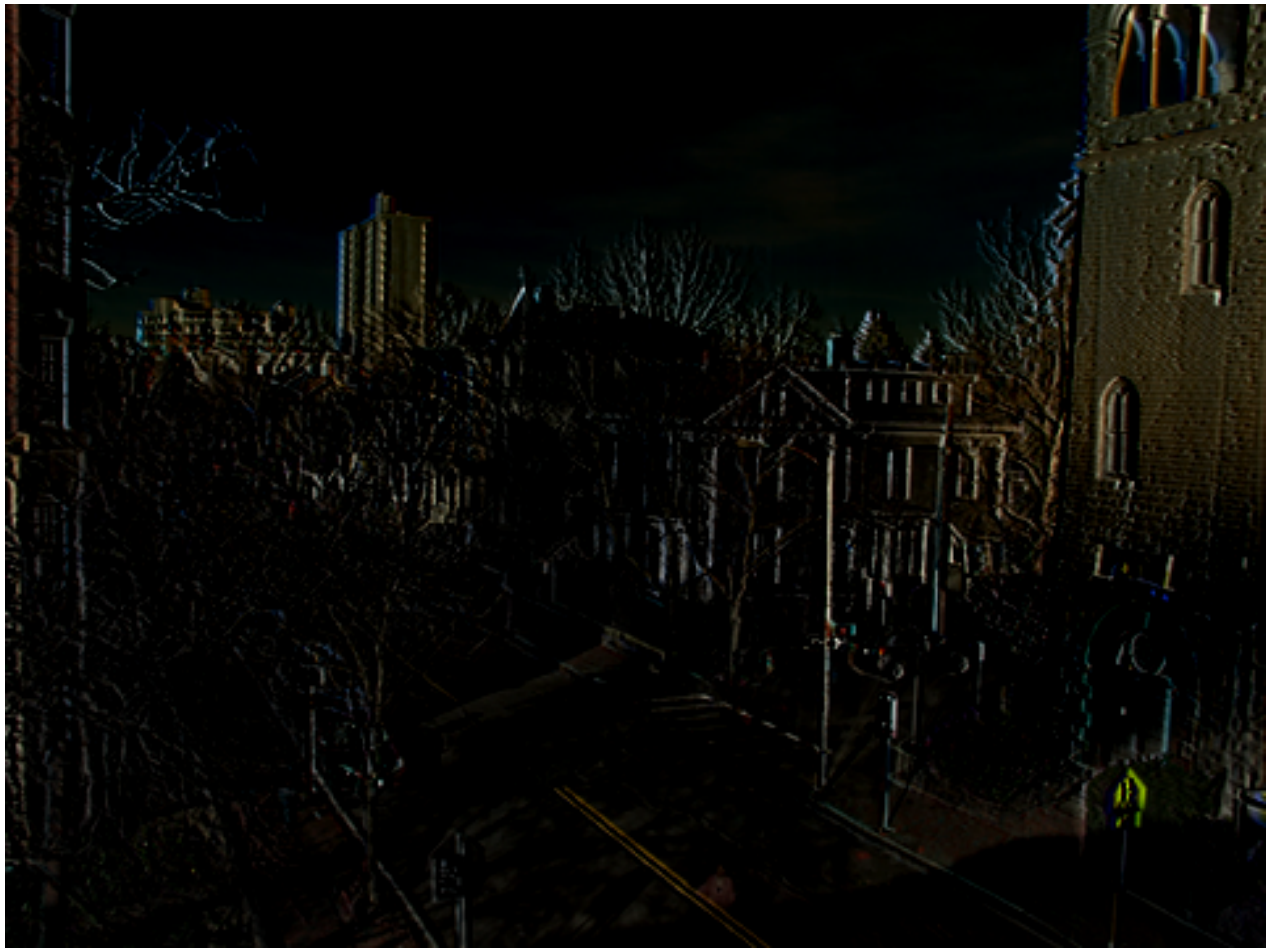
Fourth component

Positive contribution

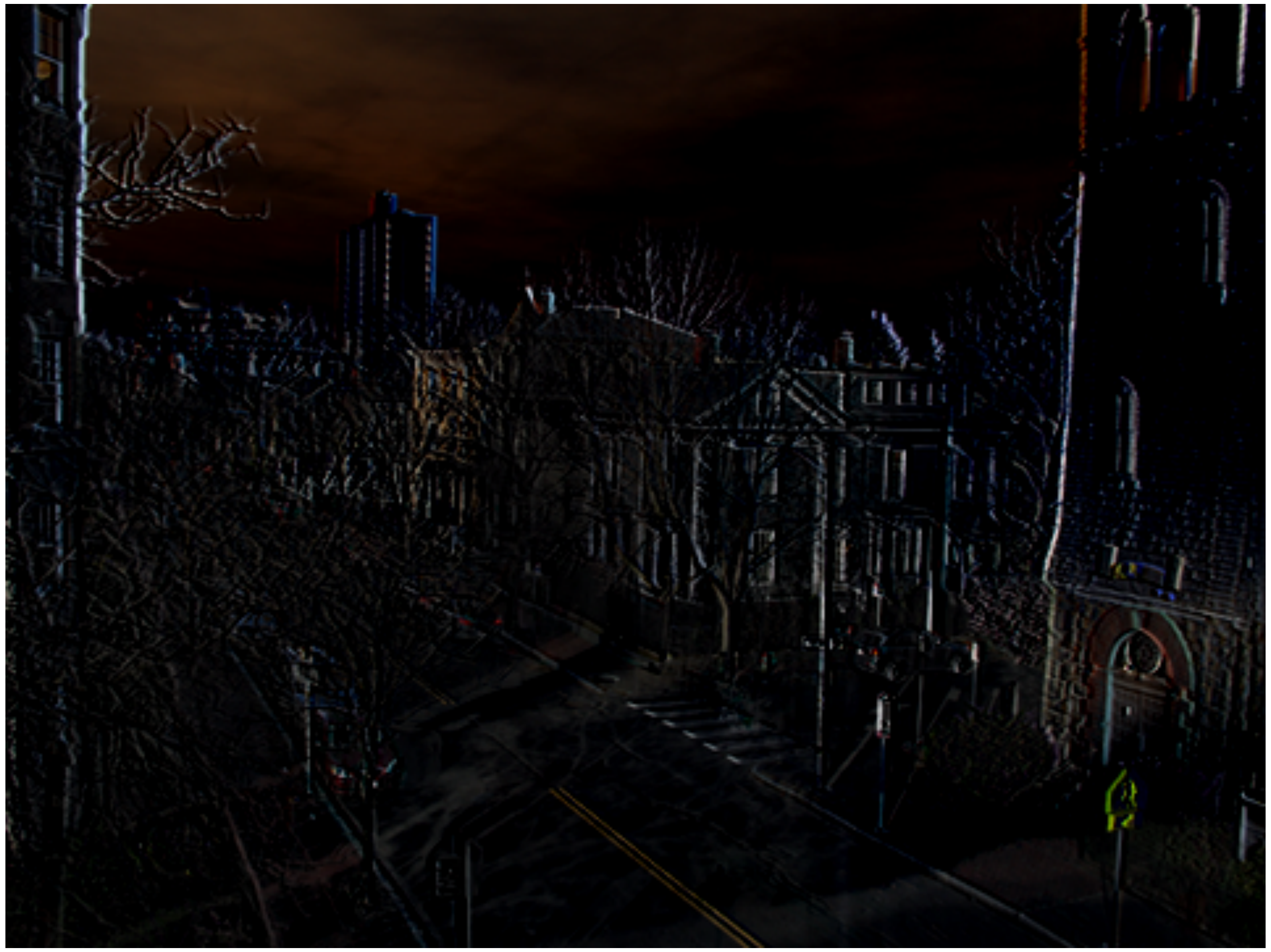
Negative contribution

Variance in this component: 1.46% of total

Fifth component



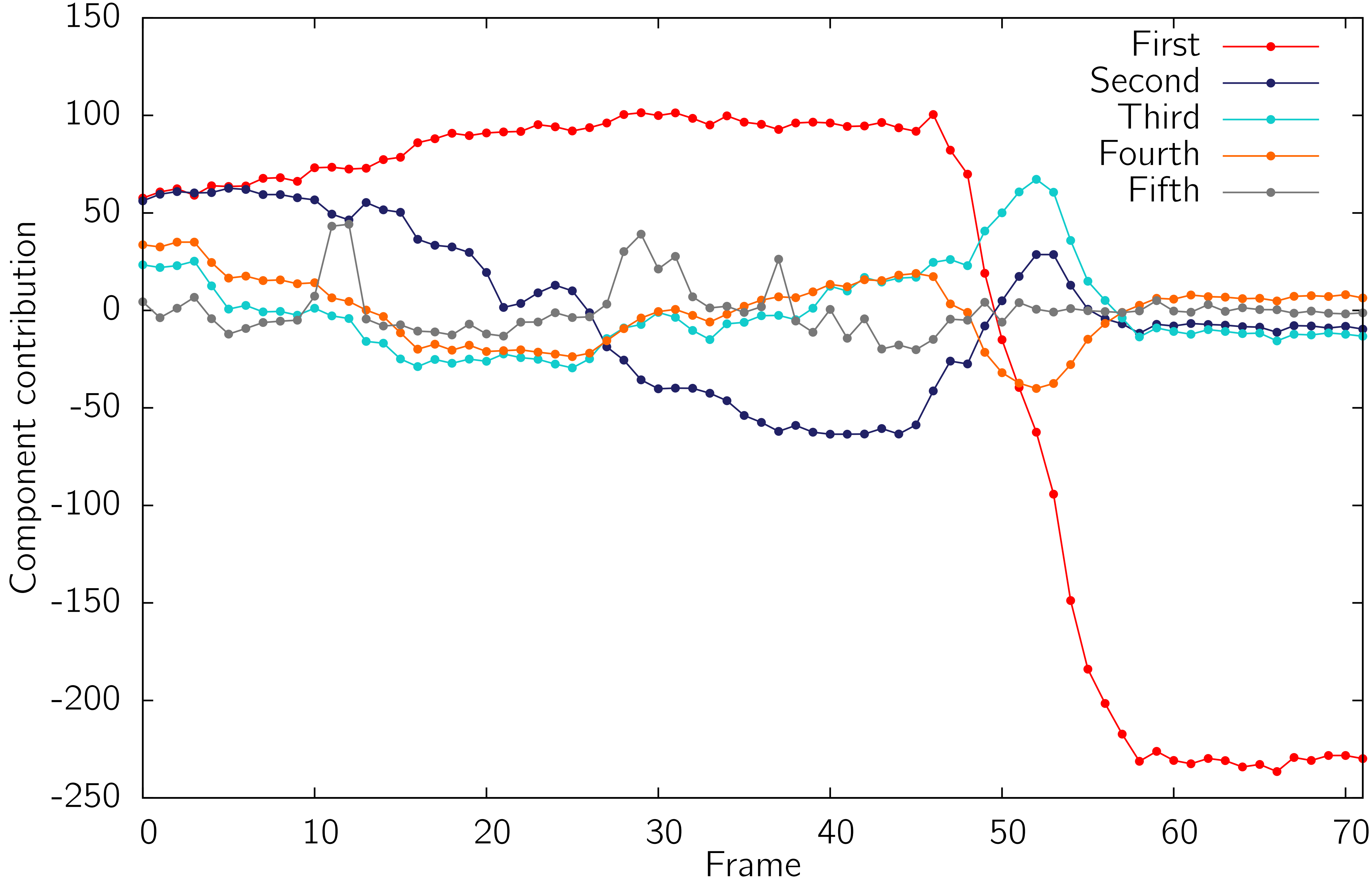
Positive contribution



Negative contribution

Variance in this component: 0.78% of total

Evolution of components



Classification

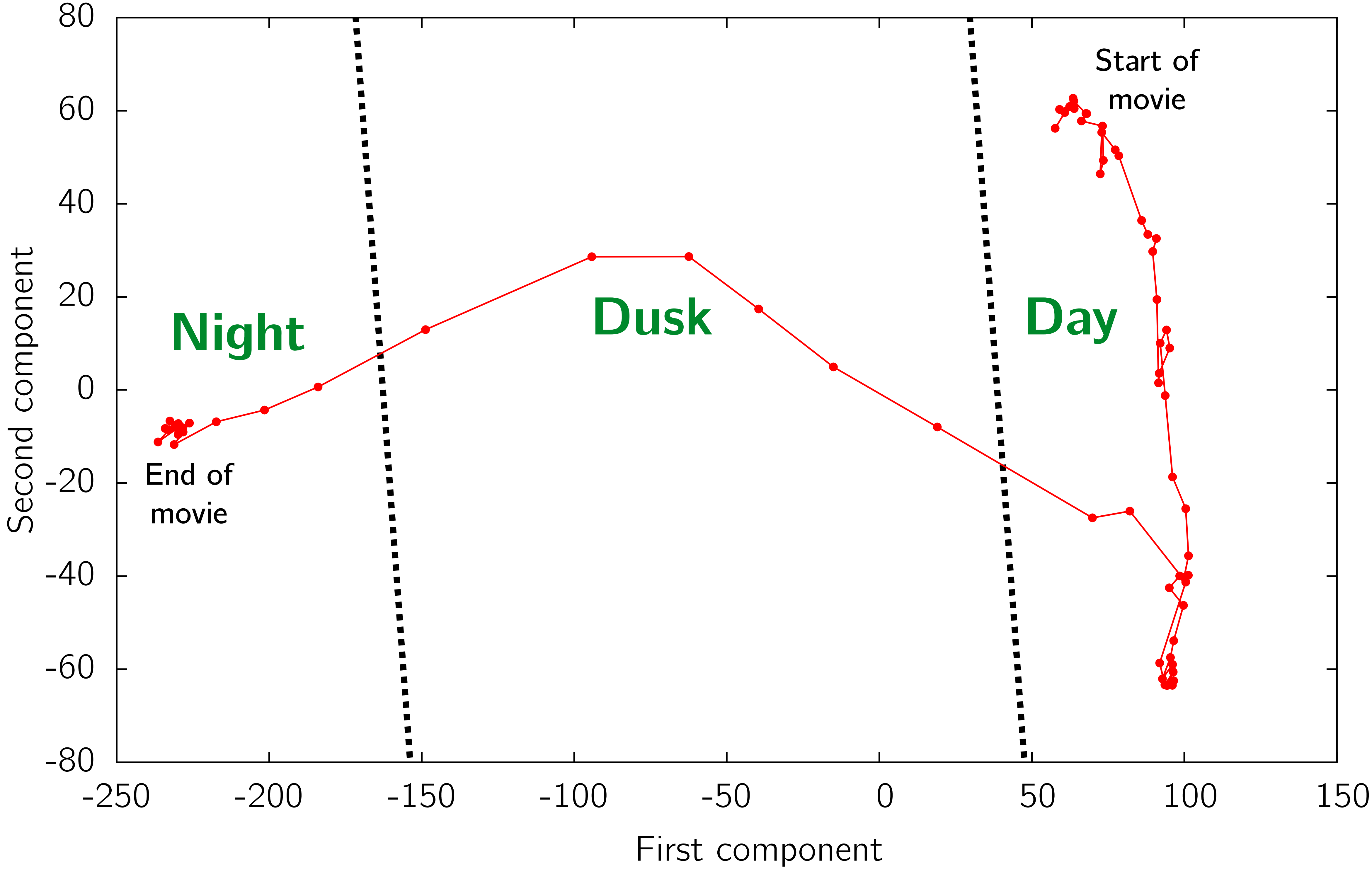
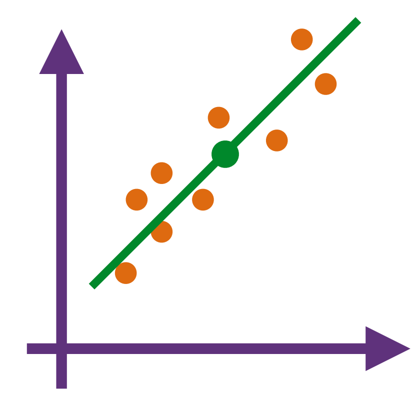


Image reconstruction

Original movie

One-mode representation



http://seas.harvard.edu/~chr/teach/pca/one_mode.mov

Image reconstruction (3 modes)

Original movie

Three-mode representation

http://seas.harvard.edu/~chr/teach/pca/three_modes.mov

Image reconstruction (10 modes)


Original movie


Ten-mode representation

http://seas.harvard.edu/~chr/teach/pca/ten_modes.mov