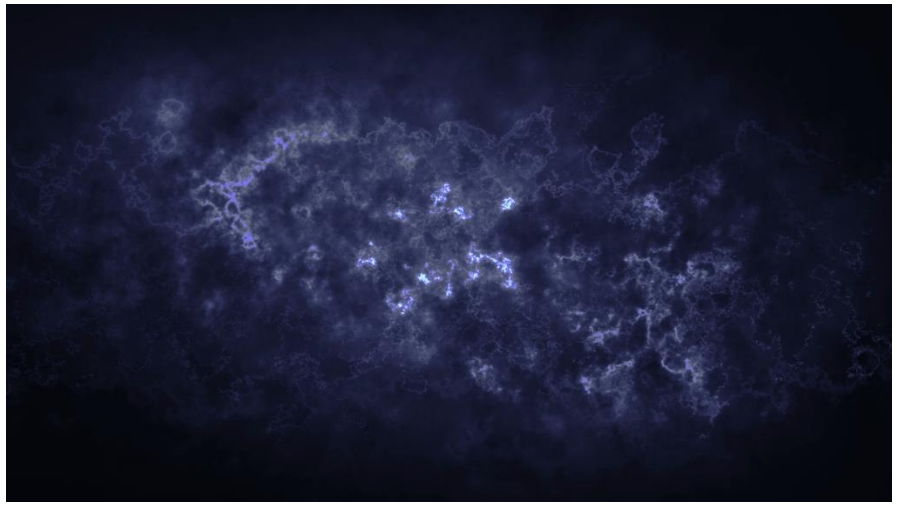
Principal Component Analysis (PCA)

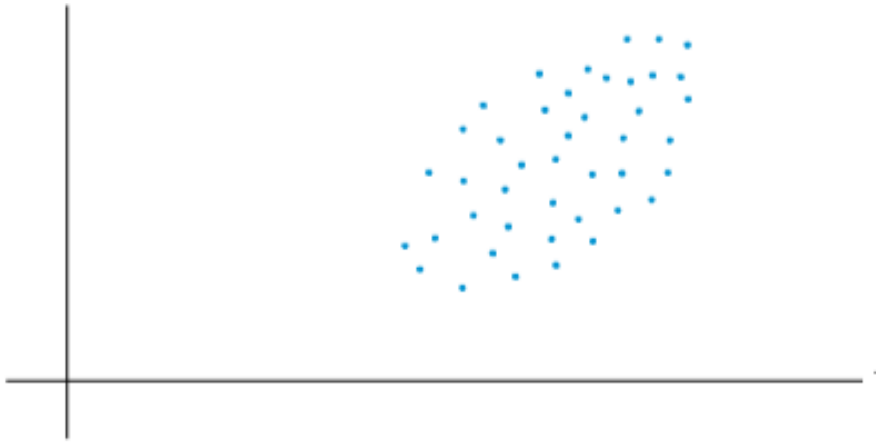
How to Reduce Dimensionality?

- Goal:
  - Try to preserve as much structure in the data as possible
  - Try to select/generate features that are discriminative
- Methods:
  - Use expert knowledge
    - E.g. An expert tells you that one of the variable can count for some other attributes
  - Feature selection
    - Simplest reduction method.
    - Select a subset of d available attributes that contribute the most in information gain.
    - Throw away rest of the attributes.
  - Feature extraction
    - Use all of the original data and combine them in some way to form a smaller set of new attributes.
    - The new attributes don’t maintain the same physical meaning as in the original data set.

Attributes and Coordinates

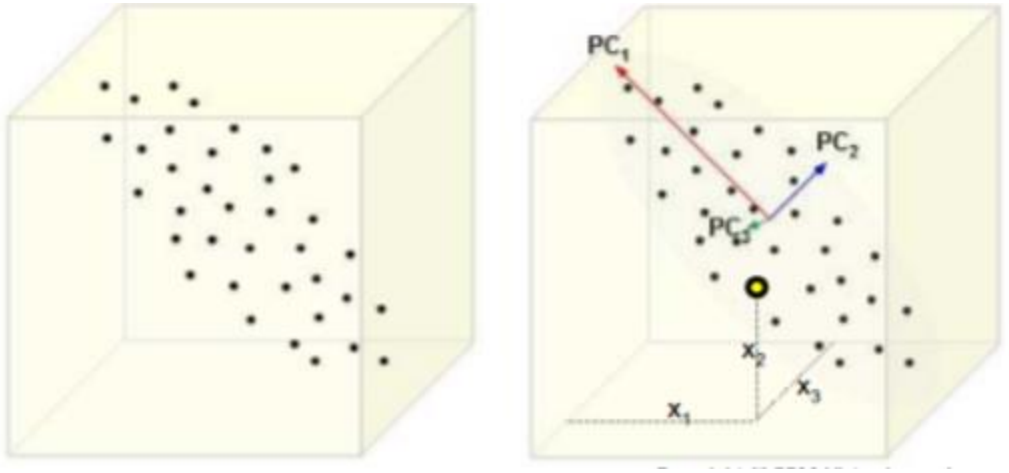
- In a dataset with k numeric attributes, you can visualize the data as a cloud of points in k-dimensional space:
  - the stars in the galaxy,
  - a swarm of flies frozen in time,
  - a two-dimensional scatter plot on paper.
- The attributes represent the coordinates of the space.
- But the axes you use, the coordinate system itself, is arbitrary.





Idea of PCA

- In machine learning there often is a preferred coordinate system, defined by the very data itself.
- The idea of principal components analysis is to use a special coordinate system that depends on the cloud of points as follows:
  - Place the first axis in the direction of greatest variance of the points to **maximize the variance** along that axis.
  - Choose the second axis in **perpendicular** to the first one, the way that maximizes the variance along it.
  - Continue, choosing each axis to maximize its share of the remaining variance.
  - Find the new coordinate by projecting the data points onto new set of axes.



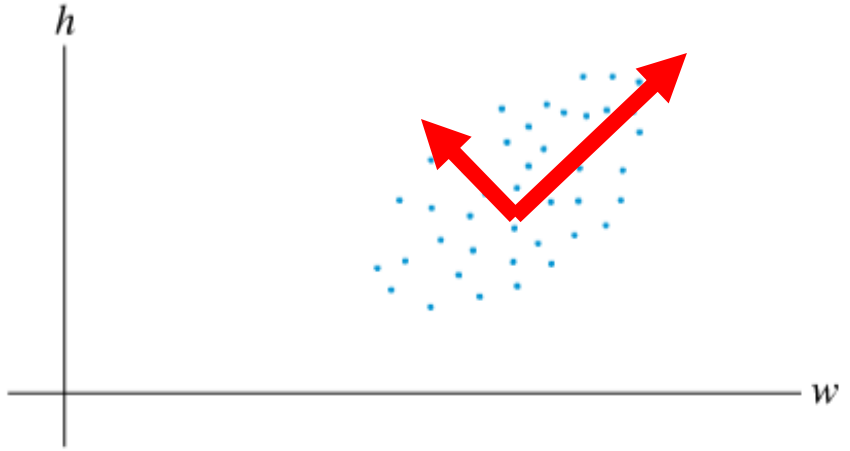
Example

An example of two-dimensional data is given by a set of weights and heights of N high school students. Let X_j denote the observation vector in R^2 that lists the weight and height of the j^{th} student. If w denotes weight and h height, then the matrix of observations has the form

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_N \\ h_1 & h_2 & \cdots & h_N \end{bmatrix}$$

$$\uparrow \quad \uparrow \qquad \uparrow$$

$$\mathbf{X}_1 \quad \mathbf{X}_2 \qquad \mathbf{X}_N$$



- Note that in two dimensions you only have one choice for the second axis. Its direction is determined by the first axis
- However, in three dimensions it can lie anywhere in the plane perpendicular to the first axis, and in higher dimensions there are even more choices, although it is always constrained to be perpendicular to the first axis

Why Greatest Variability?

- The dimensions with the greatest variability preserve the distances.
- Distance between data points is a manifestation of the data structure. Why?
- Because we assume nearby things are similar. Similarity is very important for learning algorithms.
- Therefore, the high variance dimensions preserve the structure.
- Note that while the relative distance between some points changes, the line with the largest variability preserve the most distances as accurately as possible, overall.