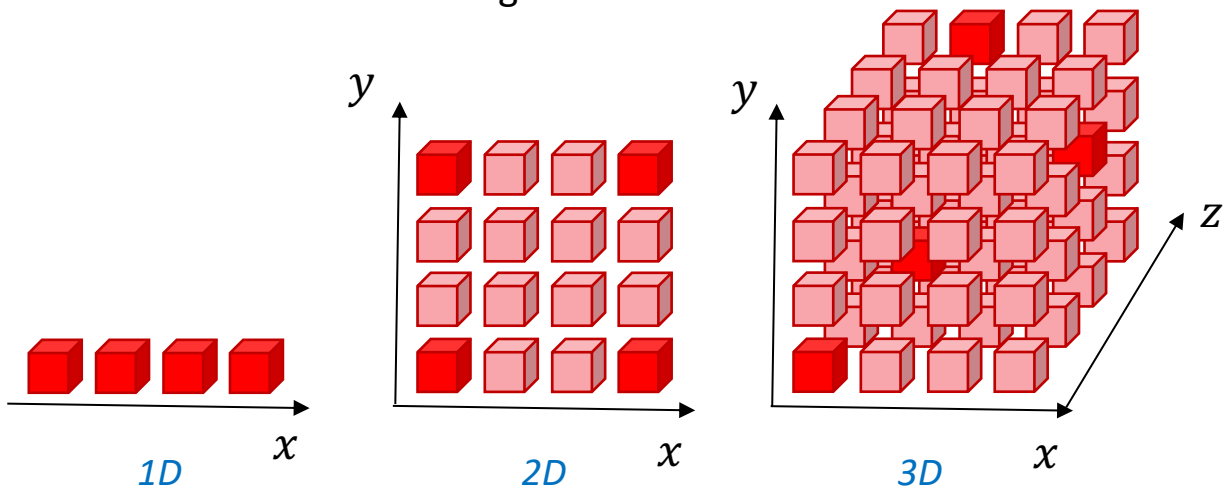
17. Dimensionality Reduction

Why Dimensionality Reduction?

- Dimensionality reduction techniques aim to reduce or transform high dimensional data points onto lower-dimensional ones, while keeping (or improving) model performance

Curse of dimensionality

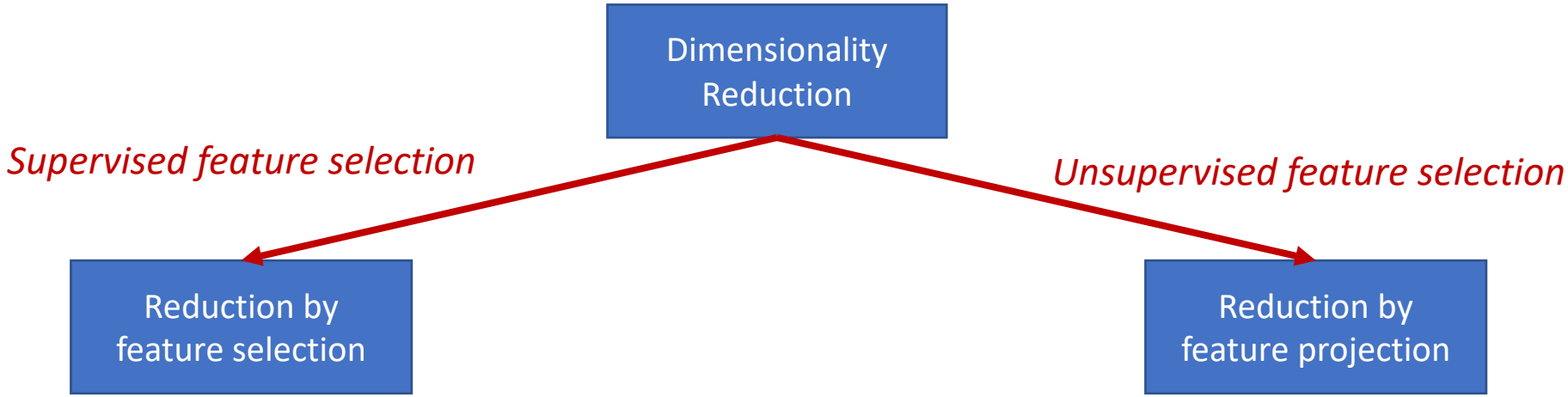
- The larger the number of dimensions, or feature space, the larger the number of data points required to maintain the same distance relationship between data points – this increases exponentially with the number of dimensions
- Therefore, a much a larger number of data points is required in order to avoid overfitting



Why feature reduction?

- Some features may be irrelevant, i.e. the intrinsic dimensionality of the dataset may be smaller than the number of dimensions
- Some features may be highly correlated, consequently redundant
- Data storage and computational requirements are reduced as the number of dimensions are reduced
- Helps visualizing high-dimensional data for analysis

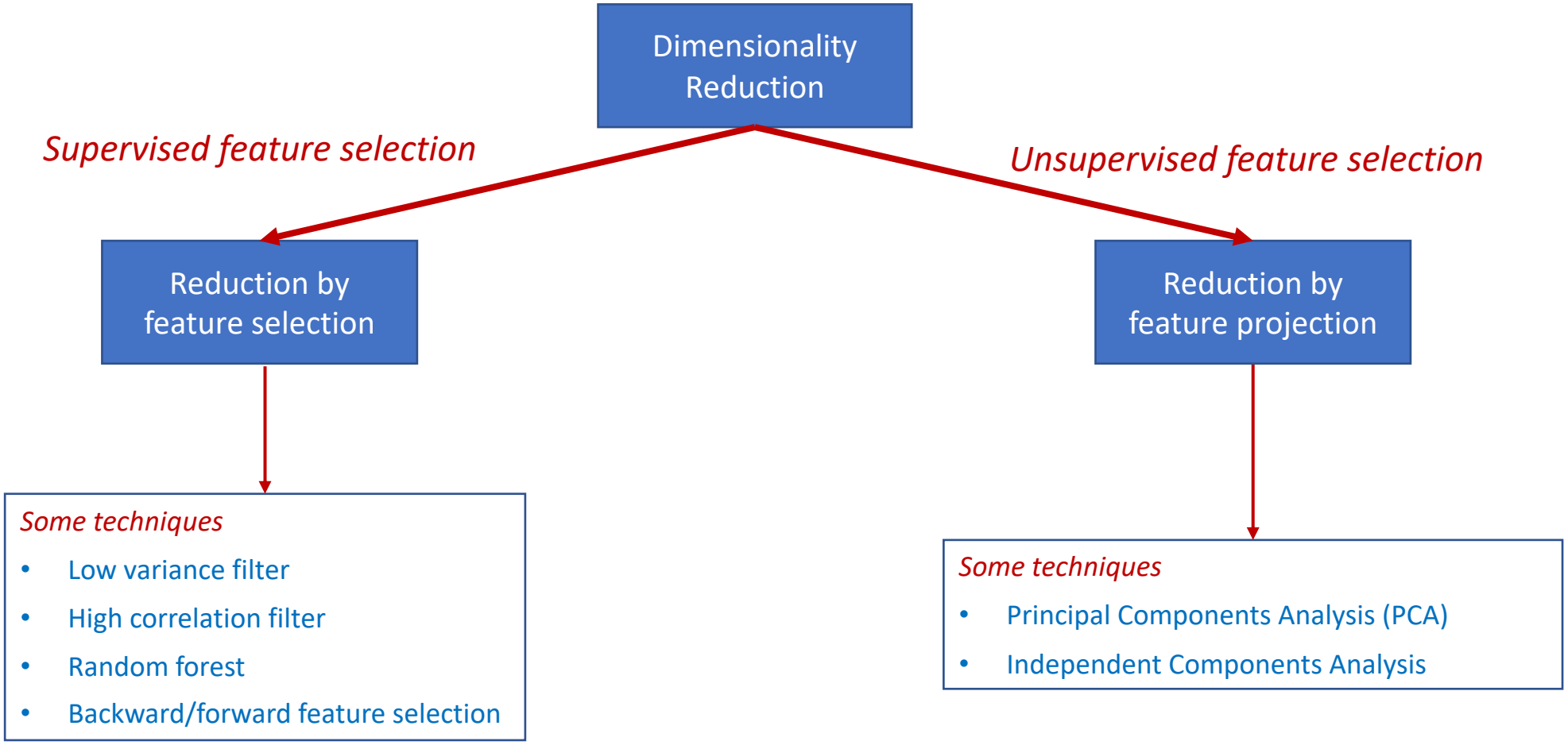
Dimensionality Reduction Techniques



- Domain-specific criteria (e.g. stop-word removal in word vector spaces)
- Score each feature based on some defined criteria such as : (1) mutual information between feature and class (2) feature impact on classification accuracy
- Perform backward/forward feature elimination:
  - eliminate feature with lowest score
  - re-score remaining features and repeat

- Create a new reduced set of features defined as functions over all features in the higher dimensional space
- Project data into lower dimensional space while preserving as much intrinsic information as possible
- The projection could minimize the squared error in reconstructing the original data

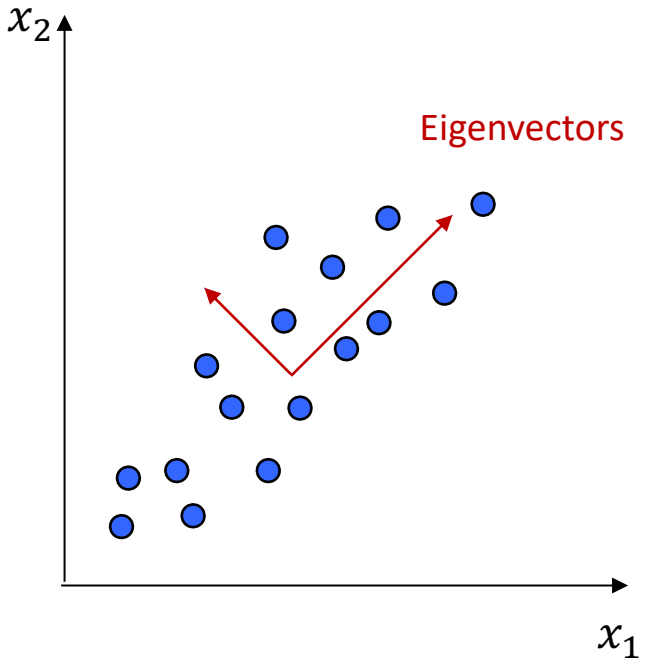
Dimensionality Reduction Techniques



Principal Component Analysis (PCA)

PCA is one of the most widely used techniques for dimensionality reduction

It reduces the original set of data features onto another one, whose components capture as much variation in the data as possible



PCA algorithm

1. Consider a set of d-dimensional data points $x_i, i = 1, \dots, n$

2. Determine the sample data mean vector given by

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

3. Determine the sample data covariance matrix given by

$$\hat{\Sigma} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (x_i - \bar{x})^t$$

4. Find eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_d$ and eigenvectors $e_1, e_2, \dots, e_d$ of the sample covariance matrix

5. Select $k < d$. The dimension reduced data is given by

$$z_i = \bar{x} + \sum_{i=1}^{k} e_i^t \cdot (x_i - \bar{x}) \cdot e_i$$