# Curse of dimensionality

In theory, according to the Curse of Dimensionality, the amount of data required to build a classifier increases exponentially with the number of features. There are two theoretical issues.

The first issue is that the number of samples required to build a model may increase exponentially with the number of features (dimensions). As the dimension increases the data gets more and more sparse. A consequence of this is that the number of samples required to cover a phenomenon increases exponentially with dimension.

The second issue is that the variation in distance between arbitrary points decreases as more dimensions are added. The more features used to describe the data the more similar everything appears.

# Purpose of feature selection

1. To create simpler models which are easy to explain,
2. To have shorter training times,
3. For variance reduction which increases the precision of the estimates for a given simulation,
4. To avoid the curse of high dimensionality

# Classes of feature selection

The feature selection methods are typically presented in three classes based on how they combine the selection algorithm and the model building.

## Filler method

Filter type methods select variables regardless of the model. They are based only on general features like the correlation with the variable to predict. Filter methods suppress the least interesting variables. The other variables will be part of a classification, or a regression model used to classify or to predict data. These methods are particularly effective in computation time and robust to overfitting.

Filter methods tend to select redundant variables when they do not consider the relationships between variables. However, more elaborate features try to minimize this problem by removing variables highly correlated to each other, such as the Fast Correlation Based Filter (FCBF) algorithm.

## Wrapper method

Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions amongst variables. The two main disadvantages of these methods are:

* The increasing overfitting risk when the number of observations is insufficient.
* The significant computation time when the number of variables is large

## Embedded method

Embedded methods have been recently proposed that try to combine the advantages of both previous methods. A learning algorithm takes advantage of its own variable selection process and performs feature selection and classification simultaneously, such as the FRMT algorithm.