Intégration de données

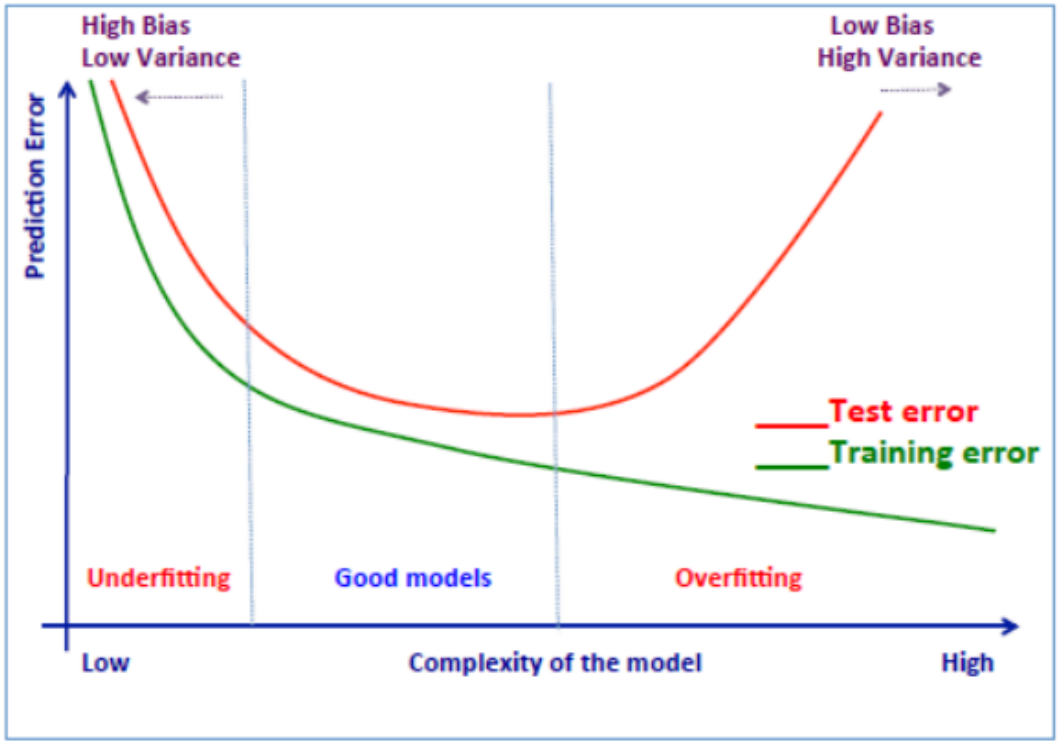
Cross validation

Ana Karina Fermin

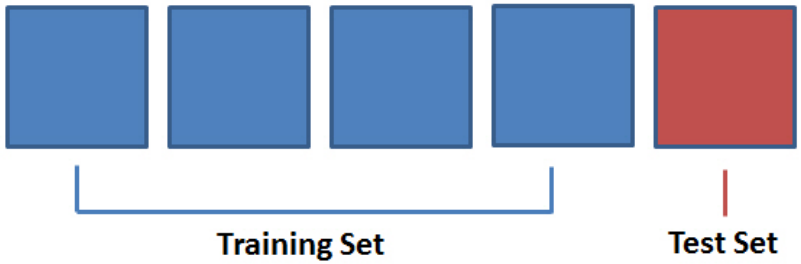
Université Paris Nanterre

M2 Miage APP

http://fermin.perso.math.cnrs.fr/



- We can determine whether a predictive model is underfitting or overfitting the training data by looking at the prediction error on the training data and the test data.
- How to estimate the test error ?



- **Very simple idea:** use a second learning/verification set to compute a verification error.
- Sufficient to avoid over-fitting!

Cross Validation

- In K-fold cross validation, the sample is randomly partitioned into K separate subsamples.
- Each time, use $\frac{K-1}{K}n$ observations to train and $\frac{1}{K}n$ to verify
- The error estimation is averaged over all K trials to get total effectiveness of our model.

- Most classical variations:
  - Leave One Out, Hold Out
  - K-fold cross validation.
- Accuracy/Speed tradeoff: $K = 5$ or $K = 10$!