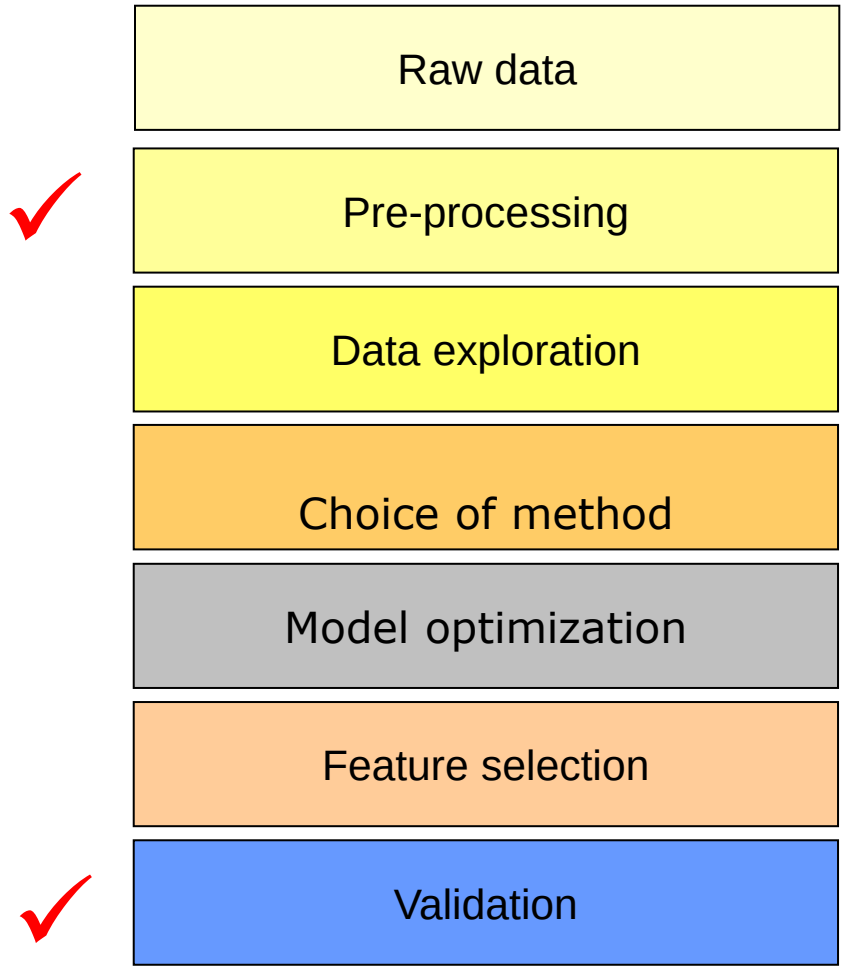
Model Validation

Hadi Parastar

Department of Chemistry, Sharif University of Technology
Tehran, Iran

Strategy of data analysis

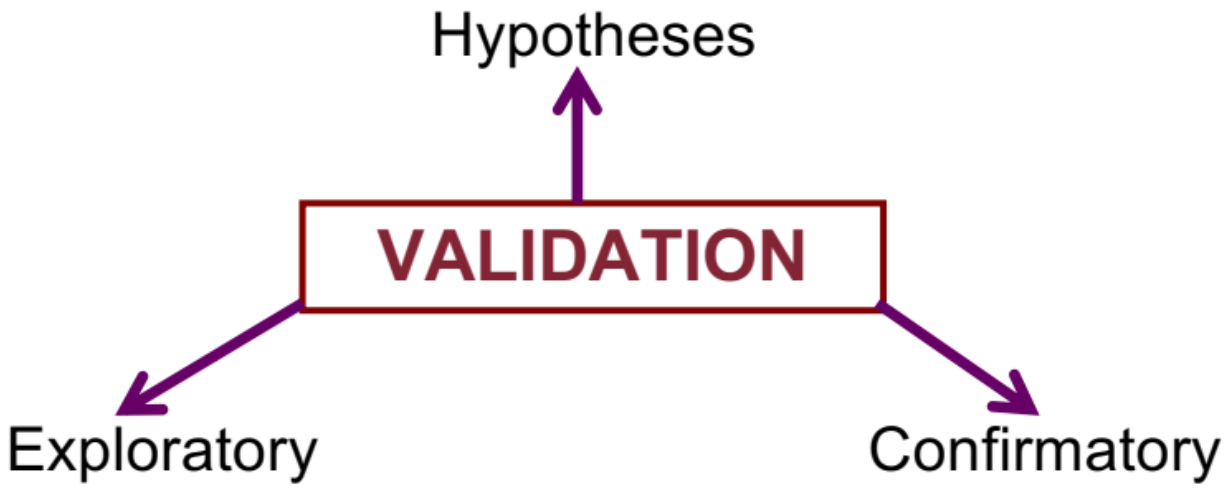


Model Validation

Far too often, solutions obtained by multivariate procedures—including factor analysis, multidimensional scaling, and cluster analysis—are interpreted, and even published, without adequate evaluation of their reliability or validity. Particularly among inexperienced users, there is an uncritical and somewhat cavalier approach to determining what parts (or which version) of an analysis to accept. Clusters or dimensions are frequently taken to be "real" whenever an interpretation can be projected onto them by the imagination of the analyst. On the other hand, dimensions that don't fit preconceptions and are hard to interpret tend to be dismissed too easily. While some users may make a feeble attempt at justifying their choice of dimensionality by examining improvements in fit values, little effort is otherwise expended in determining whether clusters or dimensions are stable or reliable, whether the model is appropriate for the data, whether the algorithm achieved correct convergence, whether serious outliers are present in the data, and so forth.

In H. G. Law, C. W. Snyder, Jr., J. Hattie, & R. P. McDonald (Eds.), Research methods for multimode data analysis (pp. 566-591). New York: Praeger.
Available from: http://psychology.uwo.ca/faculty/harshman/

Validation



BUT ALSO:

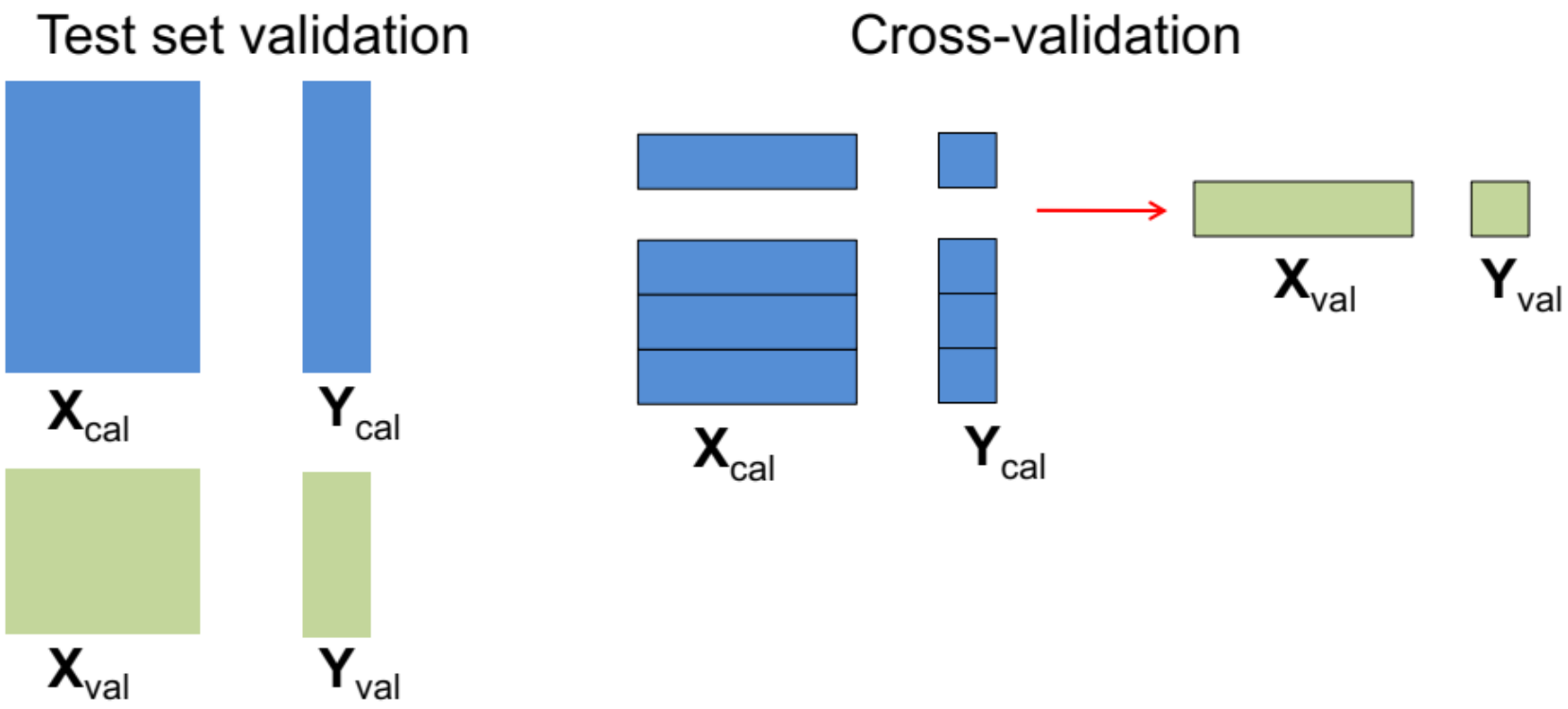
- was an appropriate model chosen?
- are outliers and/or highly influential points present?
- is the selected subspace stable?
- has the algorithm converged?

The concept of validation

- Verify if valid conclusions can be formulated from a model:
  - Able to generalize parsimoniously (with the smaller nr. of LV)
  - Able to predict accurately
- Define a proper diagnostics for characterizing the quality of the solution:
  - Calculation of some error criterion based on residuals
- Residuals can be used for:
  - Assessing which model to use;
  - Defining the model complexity in component-based methods;
  - Evaluating the predictive ability of a regression (or classification) model;
  - Checking whether overfitting is present (by comparing the results in validation and in fitting);
  - Residual analysis (model diagnostics).

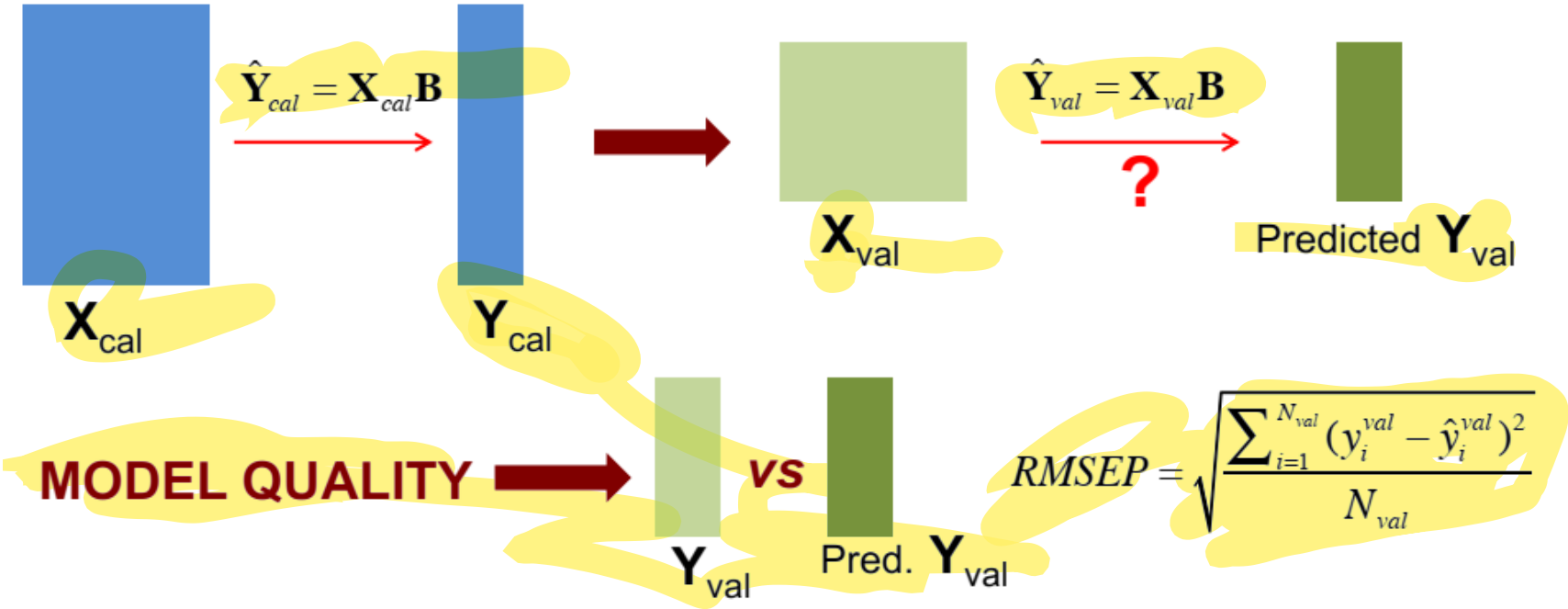
The need for "new" data

- The use of fitted residuals would lead to overoptimism:
  - Magnitude and structure not similar to the ones that would be obtained if the model were used on new data.



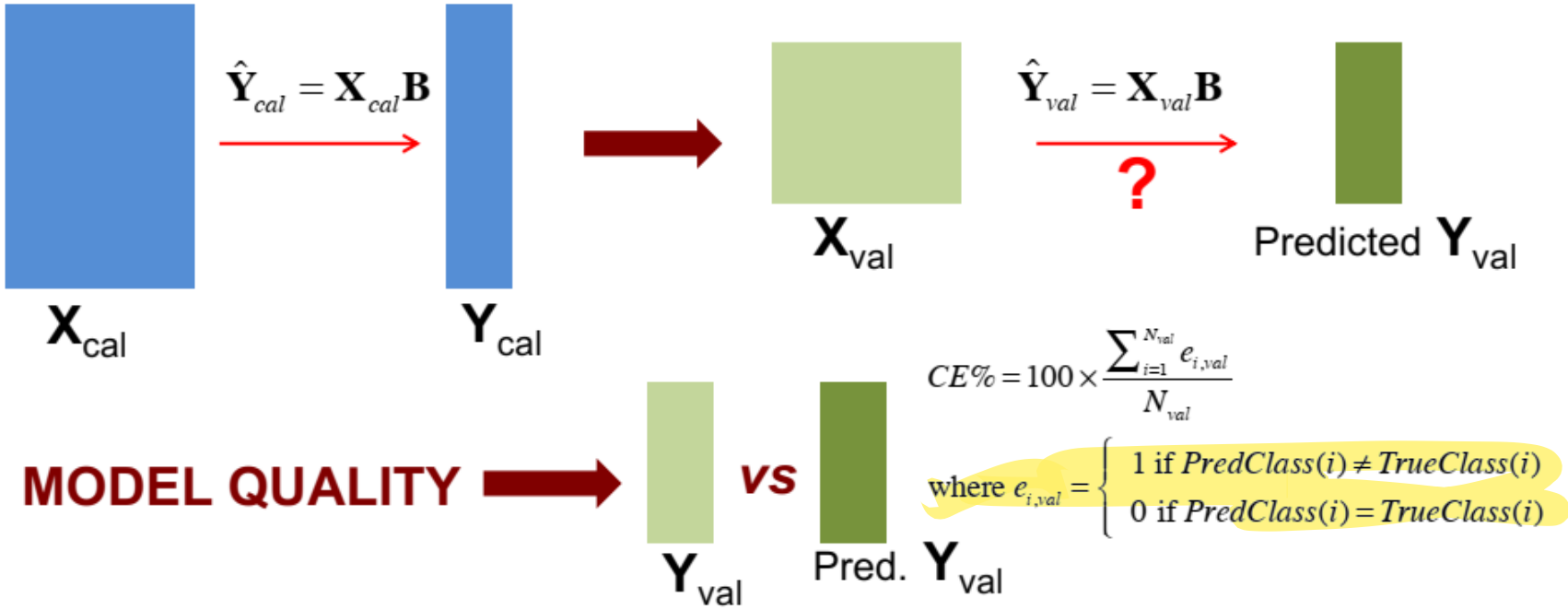
Test set validation (Regression)

- Carried out by fitting the model to new data (test set):
  - Simulates the practical use of the model on future data.
  - Test set should be as independent as possible from the calibration set (collecting new samples and analysing them in different days…)
  - A representative portion of the total data set can be left aside as test set.



Test set validation (classification)

- Carried out by fitting the model to new data (test set):
  - Simulates the practical use of the model on future data.
  - Test set should be as independent as possible from the calibration set (collecting new samples and analysing them in different days…)
  - A representative portion of the total data set can be left aside as test set.

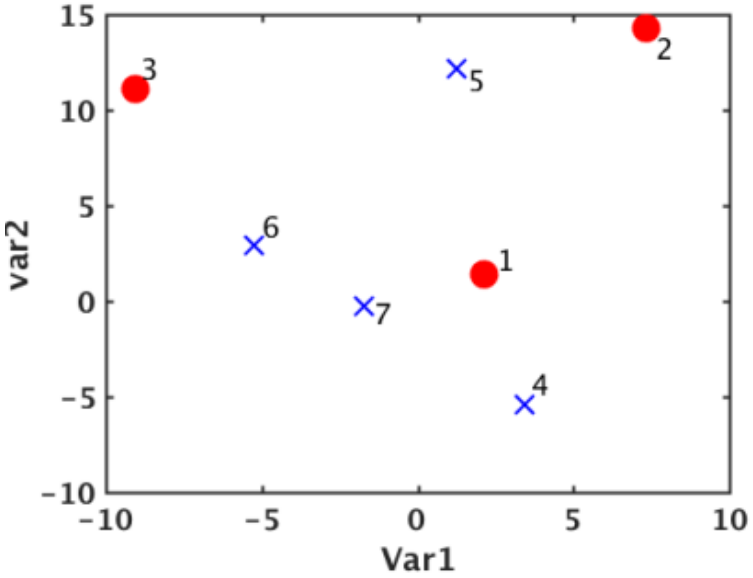


How to split data?

- Intelligent choice of the samples to be put in each set → reliable considerations based on the obtained results.
- Different criteria have been proposed in the literature to operate an intelligent splitting
- They all share the same concept:
  - try to span the sample space as uniformly as possible.
- Just to cite a few:
  - Kennard-Stone
  - Duplex
  - D-optimal criterion
  - Kohonen-based

Kennard-Stone algorithm

- The most diverse samples are placed in the training set
- All the remaining ones are left out as test set
- The "diversity" of a new samples from the ones already selected is defined by the ***maximin*** criterion:
  - The sample with the maximum value of the minimum distance to the ones already selected is added to the training set



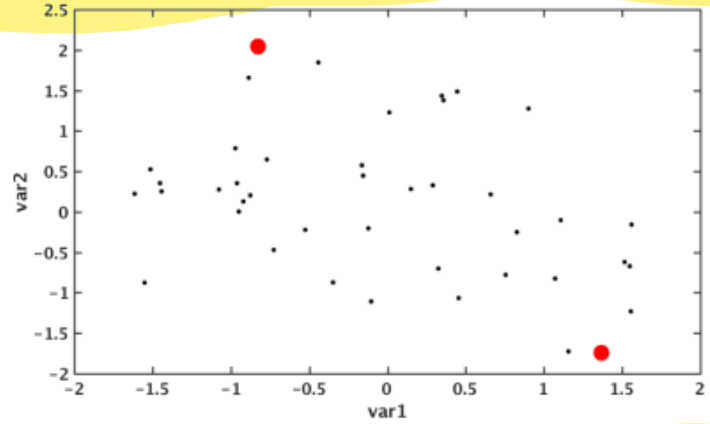
Samples	1	2	3	Min distance
4	6.9	20.1	20.7	6.9
5	10.8	6.4	10.4	6.4
6	7.5	17.0	9.0	**7.5**
7	4.2	17.1	13.5	4.2

Sample 6 would be selected as the next one to be included

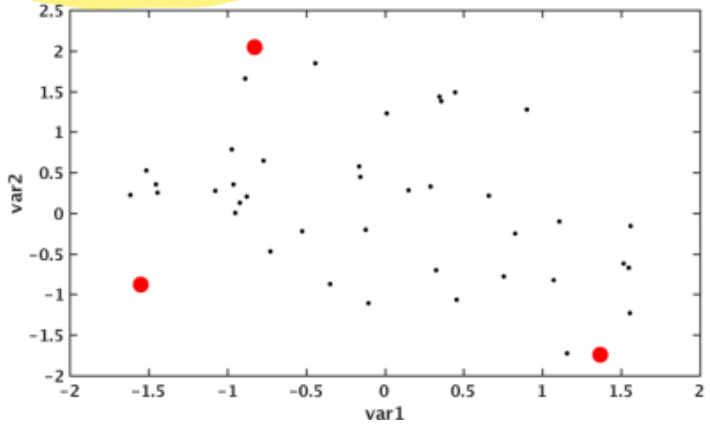
R.W. Kennard, L.A. Stone, *Technometrics*, **11** (1969), 137-148

Kennard-Stone Algorithm

- The two most distant samples are selected to be the initial training set

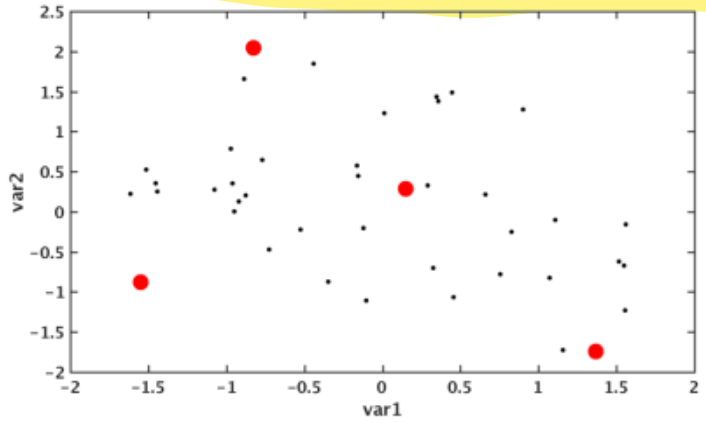


- The third training sample is selected as the one having the maximum minimum distance from the two already chosen

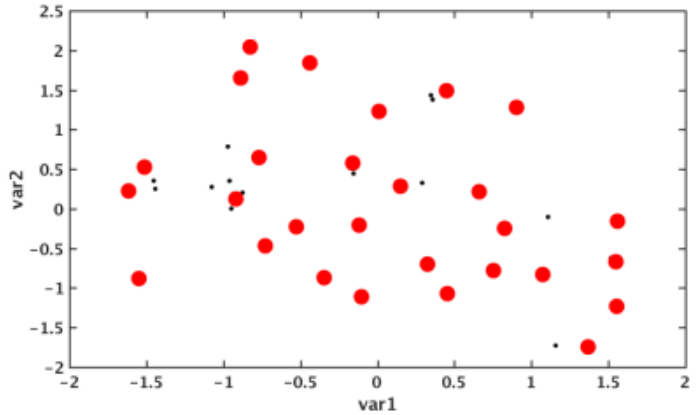


Kennard-Stone Algorithm

- The next samples are also added according to the maximin criterion



- Until the desired number of training object is reached (the left out object will form the test set)



Duplex Algorithm

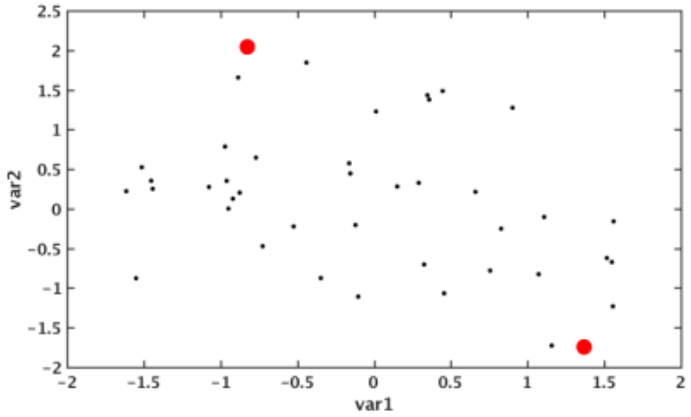
- Kennard-Stone approach tries to concentrate as much of the data diversity in the training samples
- It can lead to overoptimistic results
- A modification of the algorithm aimed at maintaining a comparable diversity between the two sets was proposed by Kennard himself (even though it was left unpublished until it was discussed by Snee).

DUPLEX

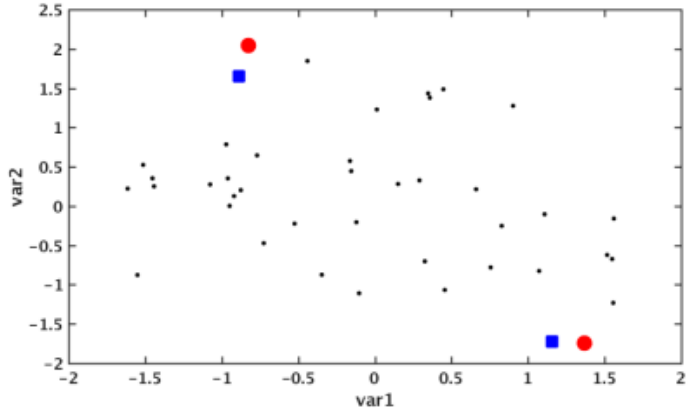
R.D. Snee, *Technometrics*, **19** (1977), 415-428

Duplex algorithm

- The two most distant samples are selected to be the initial training set

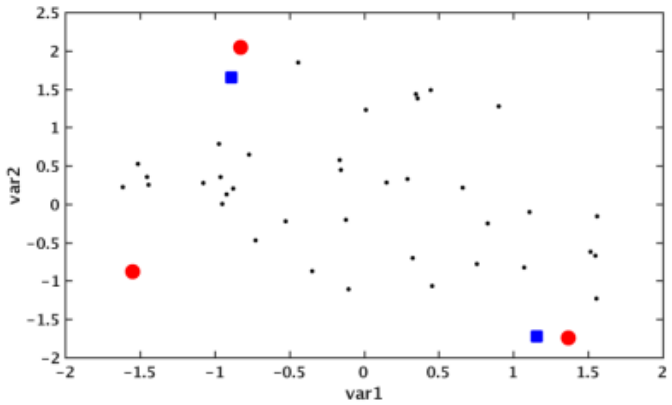


- The remaining two most distant samples are selected to be the initial test set

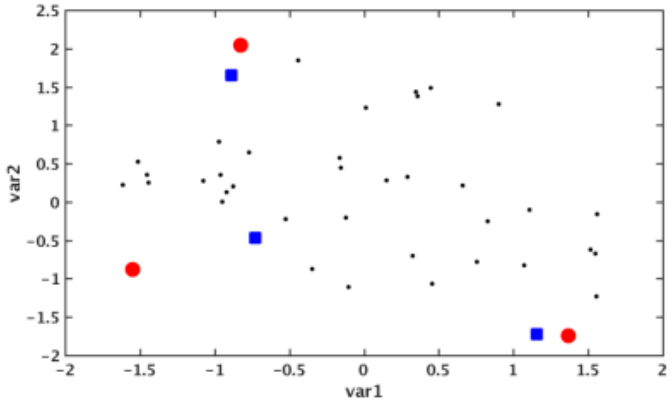


Duplex algorithm

- The third training sample is selected according to maximin from the training set

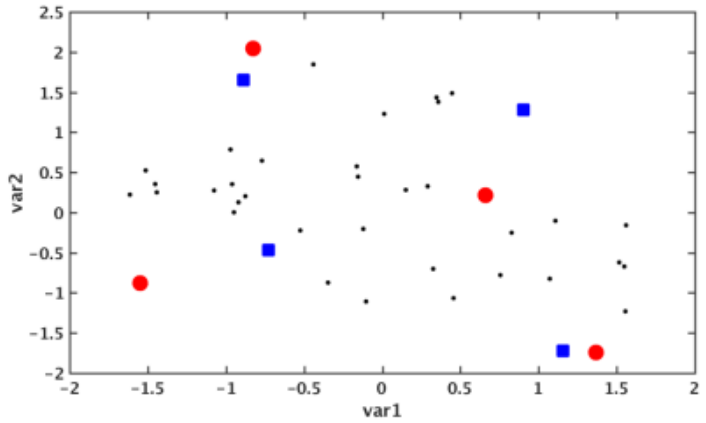


- The third test sample is selected according to maximin from the test set

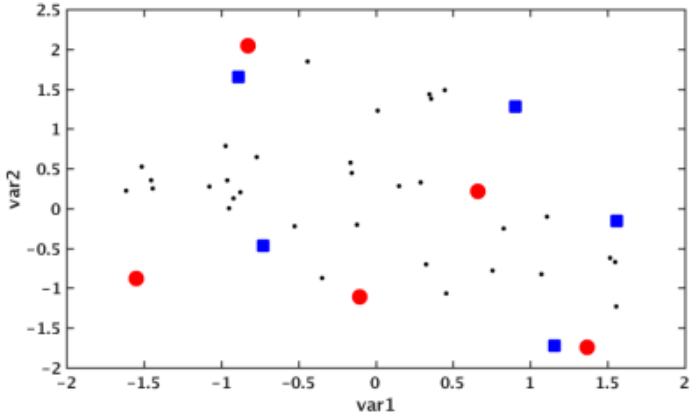


Duplex algorithm

- The same procedure is followed for the fourth training and test samples

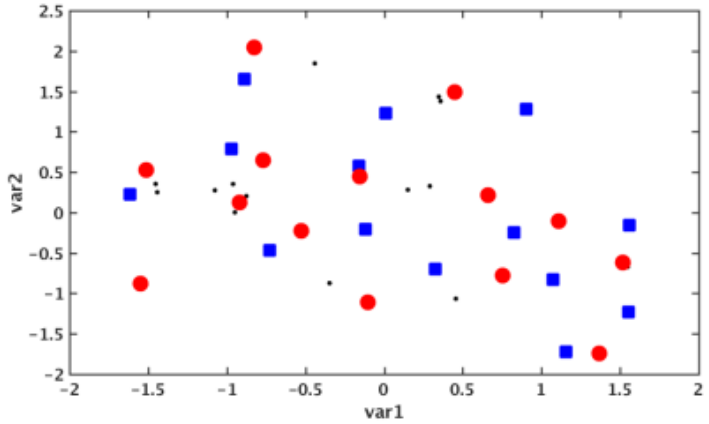


- …For the fifth...

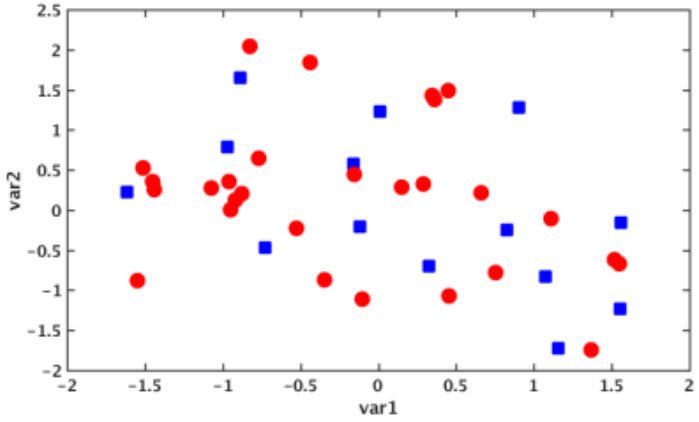


Duplex algorithm

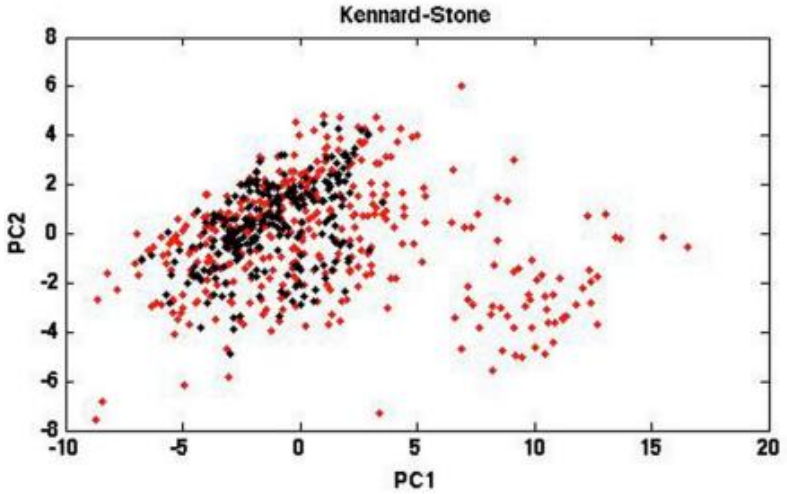
- …Until the desired number of test samples has been selected.

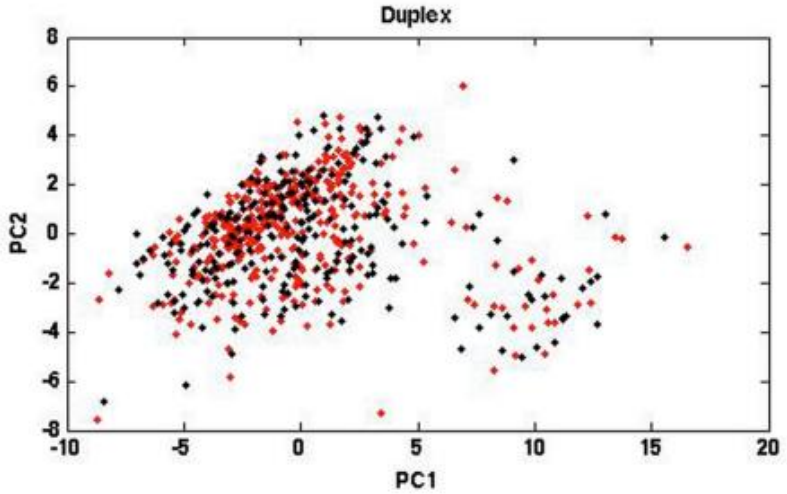


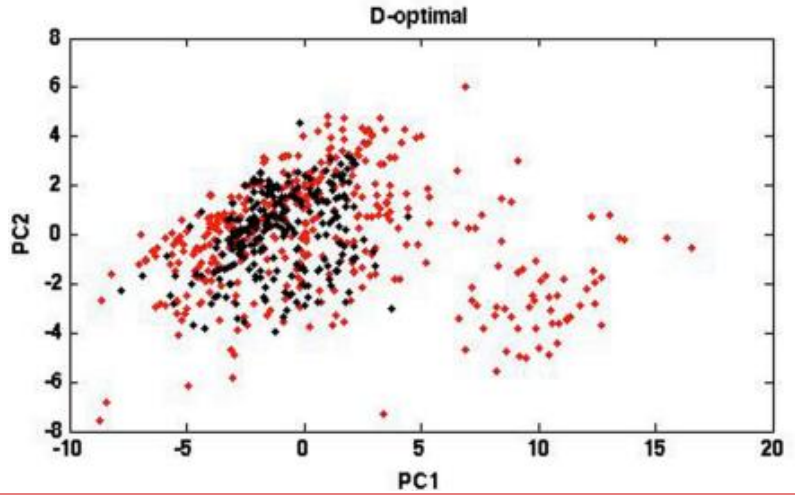
- Then, the remaining objects are added to the training set.

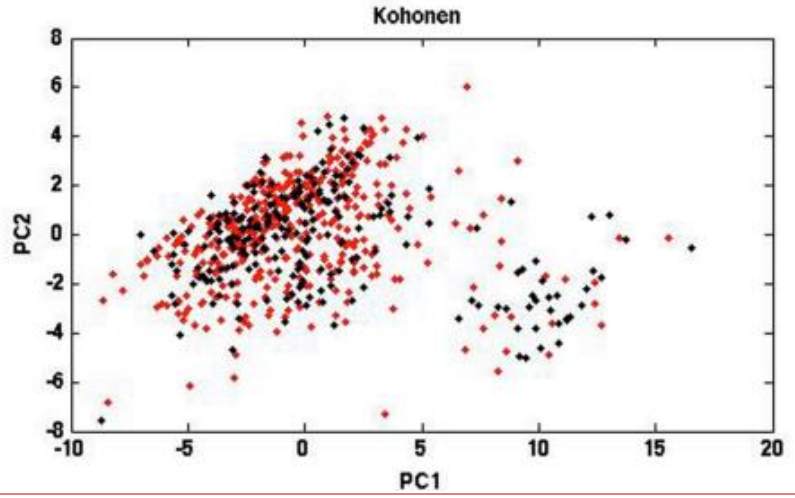


Comparing different splitting methods





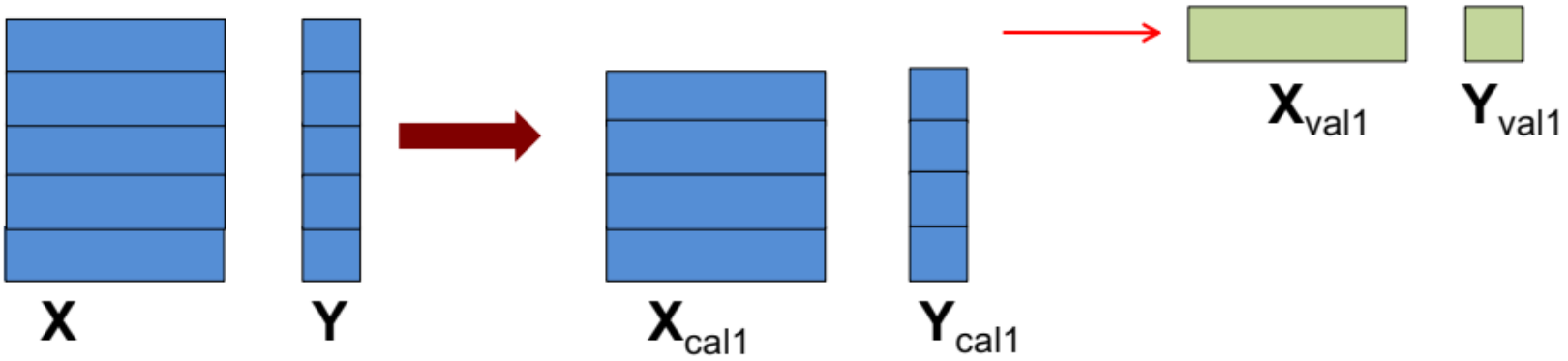




Cross-validation

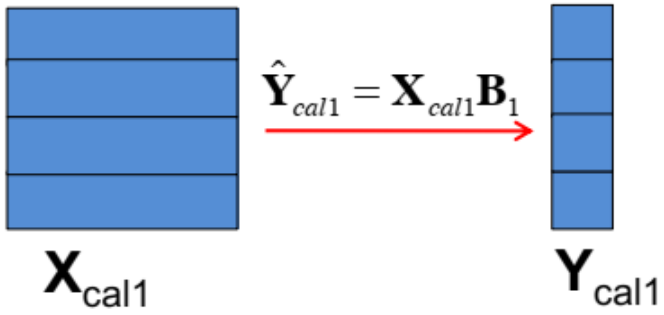
- Internal resampling method:
  - Simulates test set validation by repeating a data splitting procedure where different object are in turn placed in the validation set.
  - Particularly useful when a limited number of samples are available.
- Schematically, it consists of the following steps:

1. Leave out part of the data values

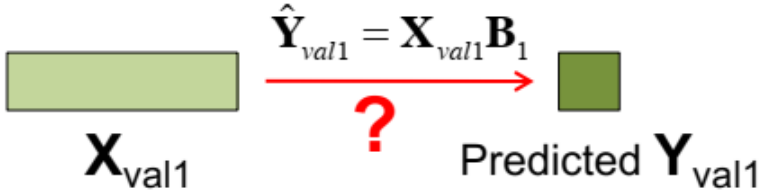


Cross-validation

2. Build the model without these data



3. Apply the model to the left out values and obtain predictions;

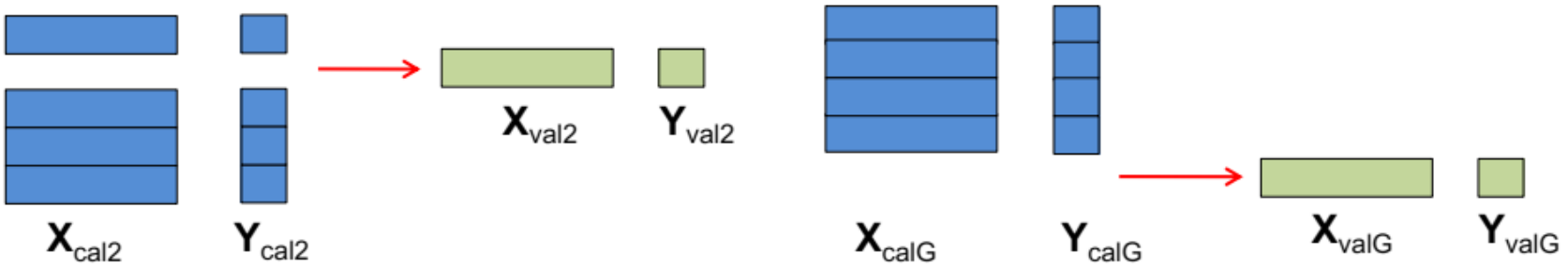


Cross-validation

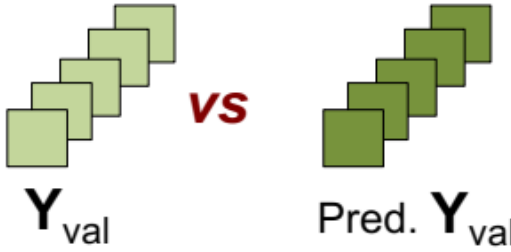
4. Calculate the corresponding residual error

$$PRESS_1 = \sum_{i=1}^{N_{val1}} (y_i^{val1} - \hat{y}_i^{val1})^2$$

5. Repeat steps 1-4 until each data value has been left out once



6. Collect all the residuals into an overall error criterion



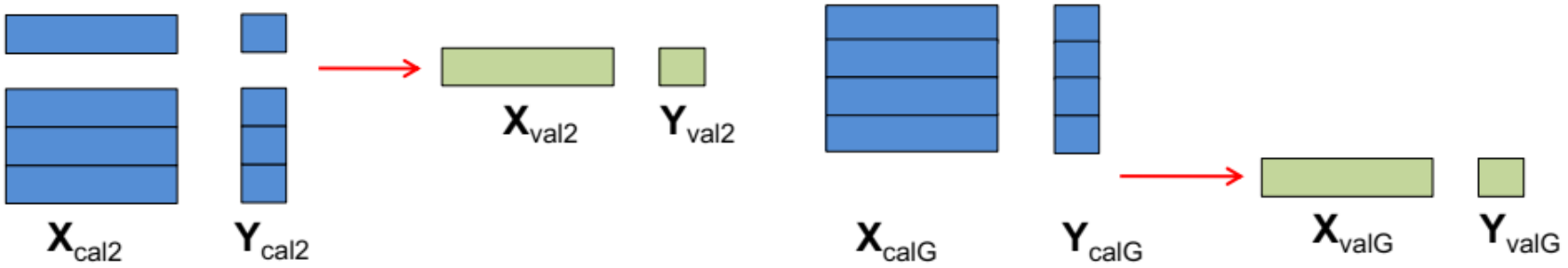
$$RMSECV = \sqrt{\frac{\sum_{j=1}^{G} PRESS_j}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_{-i})^2}{N}}$$

Cross-validation (classification)

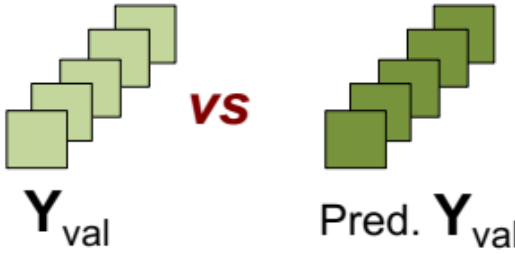
4. Calculate the corresponding residual error

$$CE_1 = \sum_{i=1}^{N_{val1}} e_i^{val1}$$

5. Repeat steps 1-4 until each data value has been left out once



6. Collect all the residuals into an overall error criterion



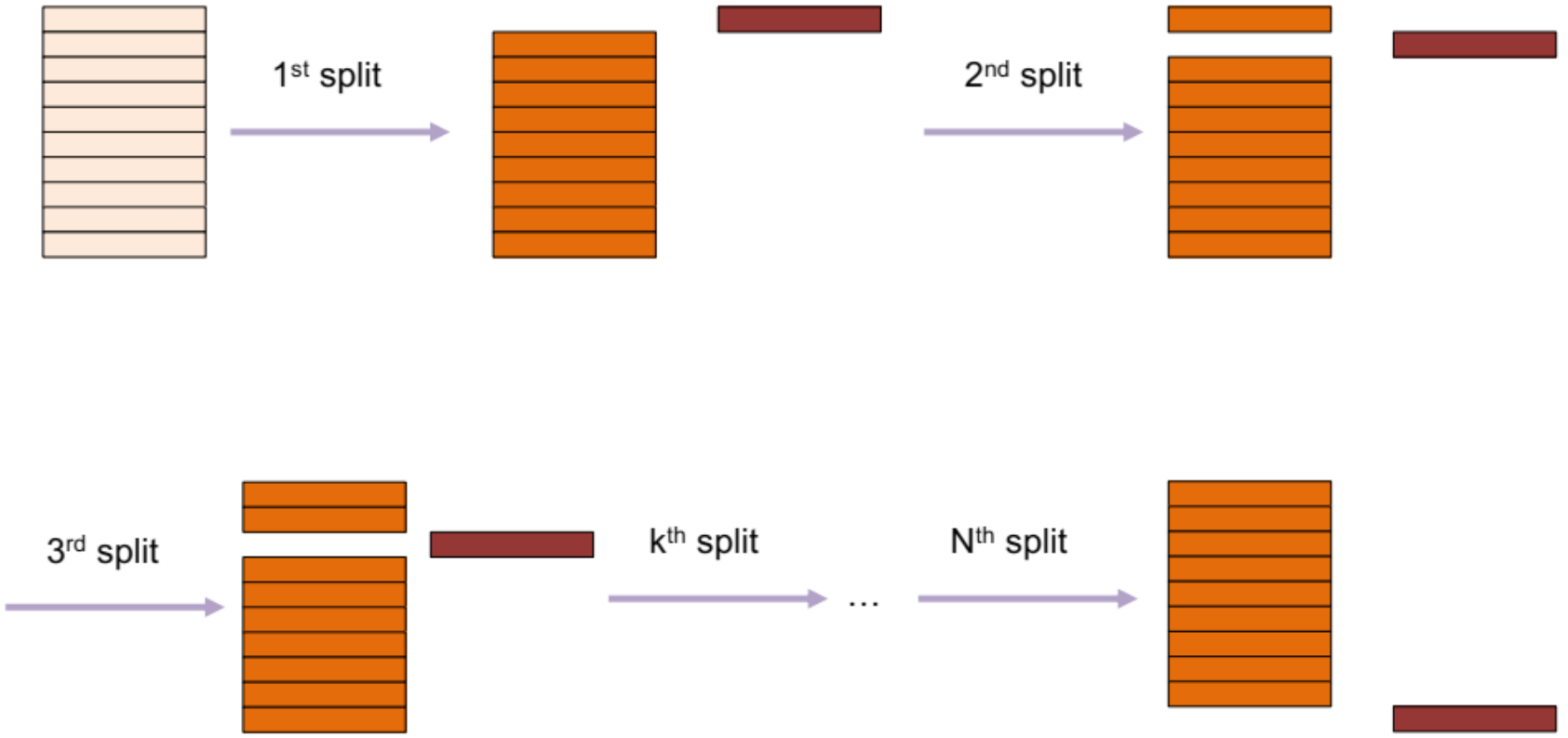
$$CE_{CV}\% = 100 \times \frac{\sum_{j=1}^{G} e_j}{N}$$

Cross-validation

- Number of objects is limited
- Understand the inherent structure of the system ←→
  Estimating model complexity
- Objects in a data table can be stratified into groups based on background information:
  - Across instrumental replicates (repeatability)
  - Reproducibility (analyst, instrument, reagent...)
  - Sampling site and time
  - Across treatment/origin (year, raw material, batch...)

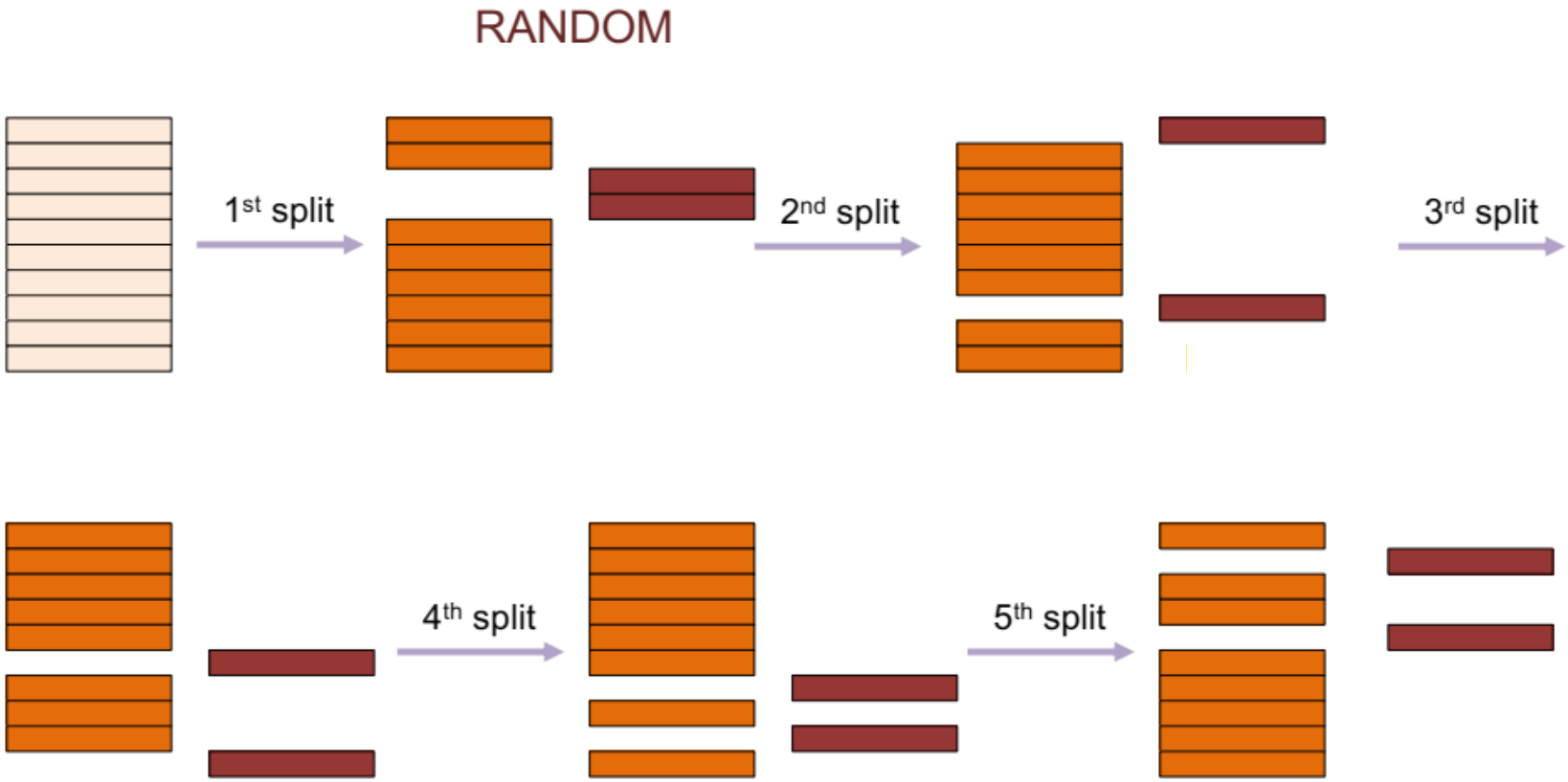
Full Cross-Validation (Leave-One-Out)

- Cancelation groups are made by a single object.
- Not advisable if the number of samples is larger than 20.



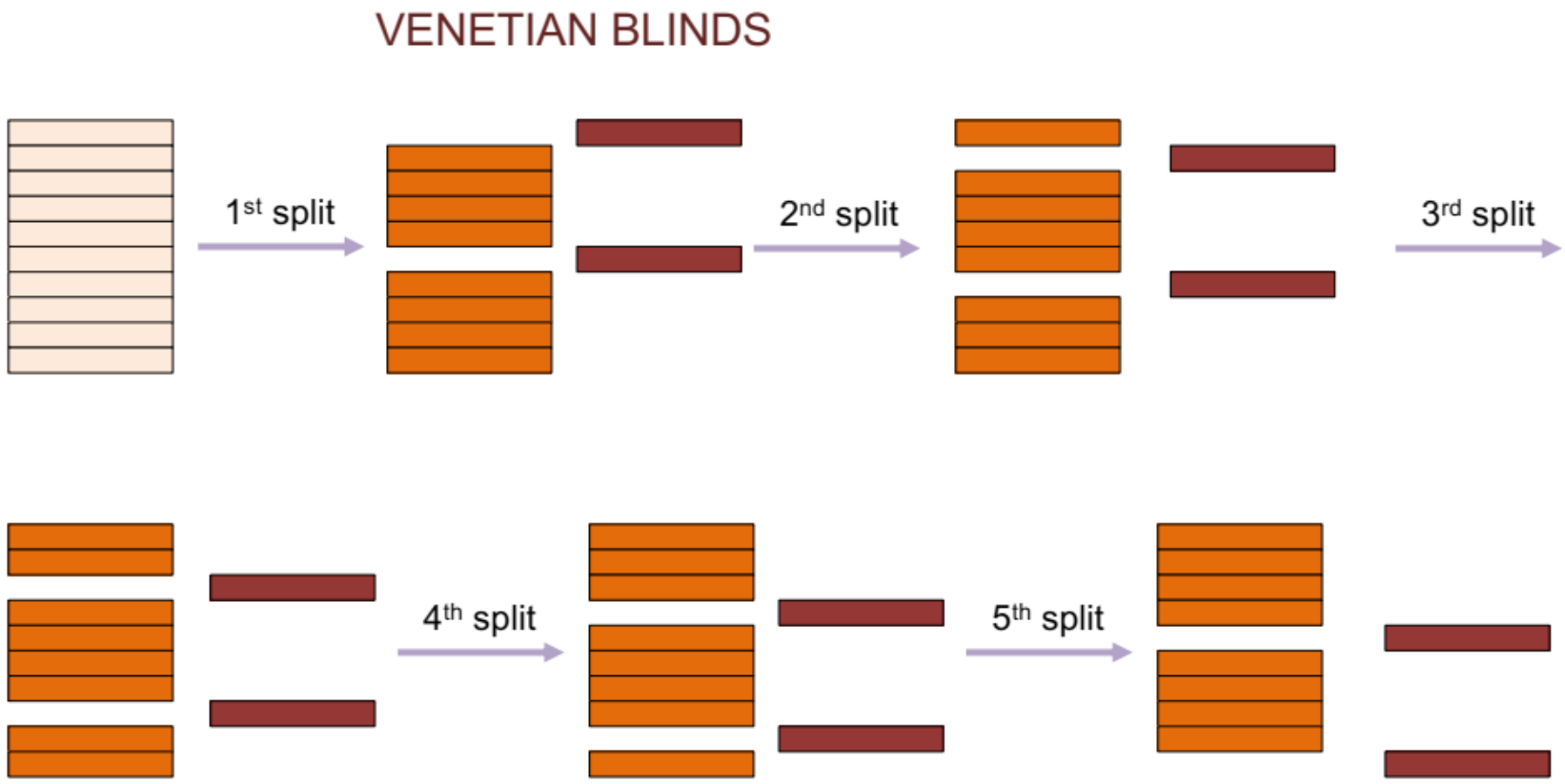
Segmented Cross-validation

- Cancelation groups are made by more than one object.
- Usually 5-10 cancelation groups allow a good estimate of the prediction error



Segmented Cross-validation

- Cancelation groups are made by more than one object.
- Usually 5-10 cancelation groups allow a good estimate of the prediction error



Segmented Cross-validation

- Cancelation groups are made by more than one object.
- Usually 5-10 cancelation groups allow a good estimate of the prediction error