Common Mistakes in Building QSAR Models

Damjan Krstajić
Research Centre for Cheminformatics
Belgrade, Serbia
www.rcc.org.rs

- The fourth Sheffield Chemoinformatics Conference (2007)
- Lessons from Bioinformatics
  - Microarray data similar to QSAR data?
- *"There is nothing more practical than a good theory*." (Kurt Lewin)

Early 1970s

- Problem of overfitting

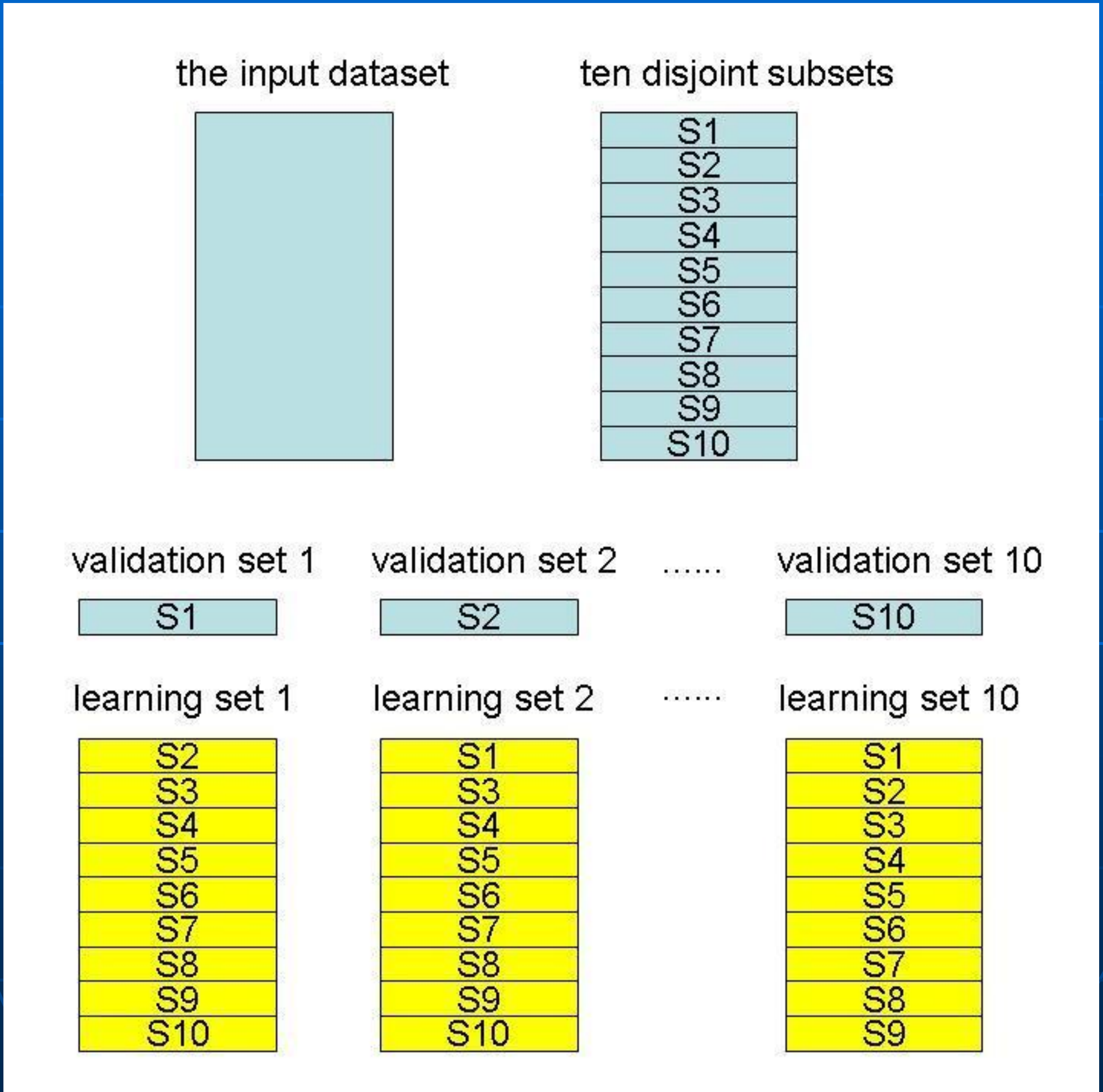
$$\sum_i (y_i - x_i^T \beta)^2$$

- Ridge regression

$$\sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- How to choose λ? (similar in PLS – how to choose number of components?)
- Cross-validation was introduced by
  - D. Allen (1974)
  - M. Stone (1974)
  - S. Geisser (1975)

10-fold cross-validation



Marvyn Stone's paper

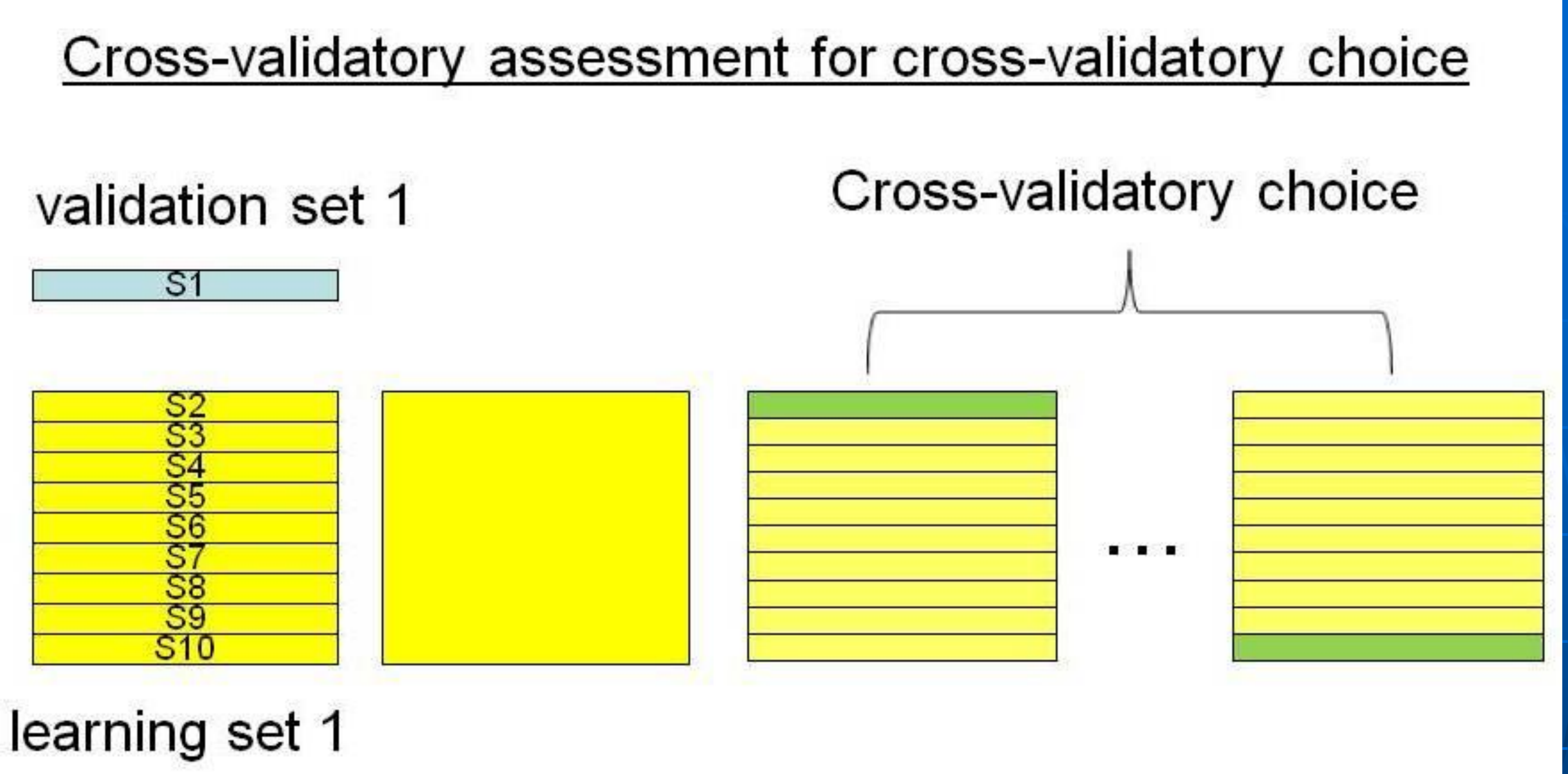
- M. Stone, Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society*, Vol. 36 No. 2 (1974) 111 – 147
- Clear difference between *Model Selection* and *Model Assessment*.
- My mistake in the past has been not to differentiate between the two processes.

Marvyn Stone's terminology 1

- In linear regression we want to find an optimal β and estimate its error.
- *Naive choice* of β
- *Naive assessment* of this naive choice
- *Cross-validatory assessment* of the naive choice
  - Build a model with naive choice on each of 10 learning sets and predict the corresponding validation sets.
  - Sum up the errors from the validation sets.
- One model selection procedure (naive choice) and two model assessments (naive and cross-validatory).

Marvyn Stone's terminology 2

- Ridge regression $$\sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$
- Lets say we have 76 different λ values.
- *Cross-validatory choice*
  - Build 76 models (one for each λ) on each of 10 learning sets and predict corresponding validation sets.
  - For each λ sum up the errors from validation sets.
  - Optimal λ is the one for which the sum is minimal.



- "two-deep" cross-validation
- *nested cross-validation* by Varma S, Simon R, Bias in error estimation when using cross-validation for model selection, *BMC Bioinformatics* 2006, 7:91

Why do we need to perform nested cross-validation?

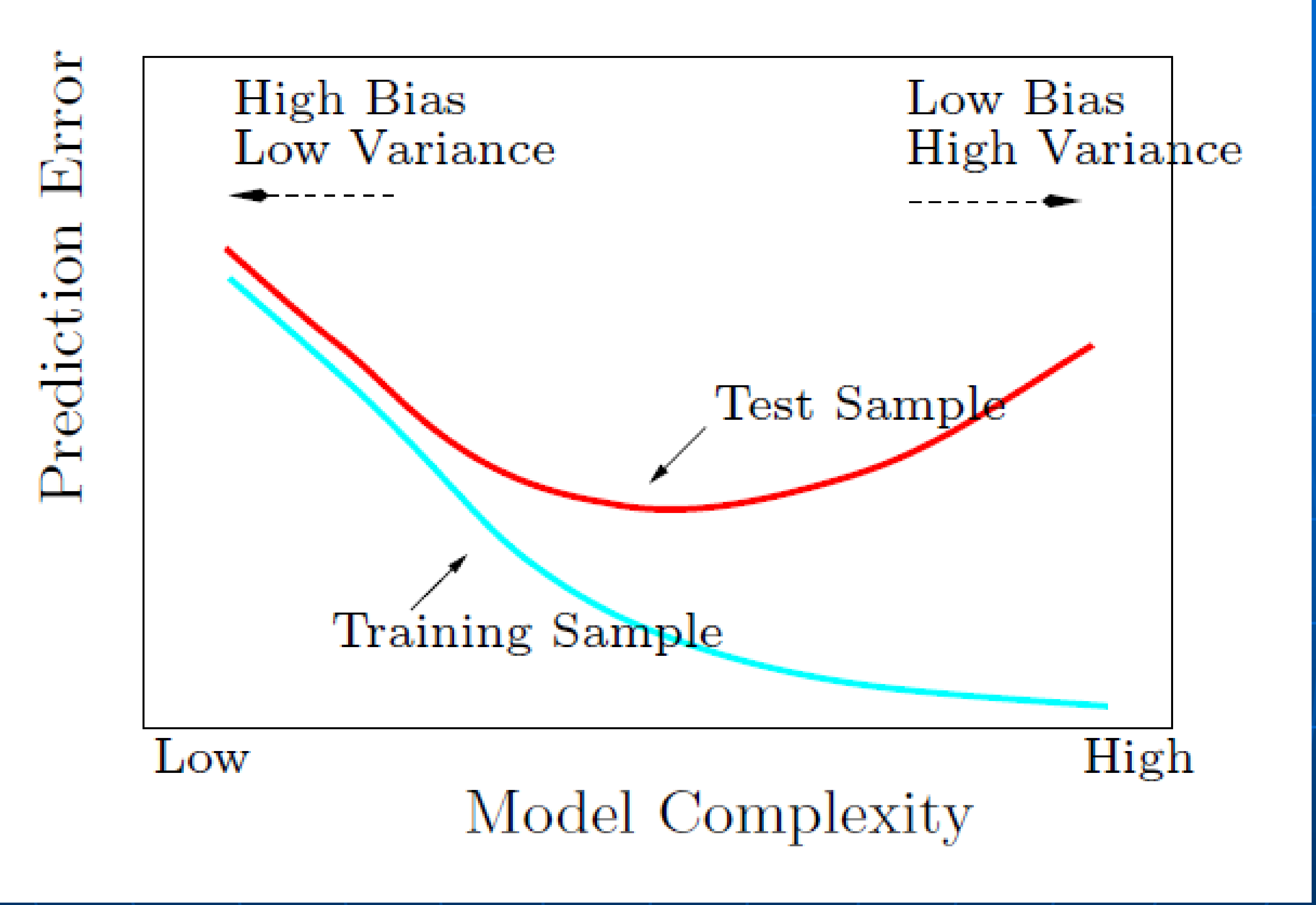
- (Varma & Simon) The difference between the CV error estimate and the true error can be greater than 20% more than one-fifth of the time.
- (Varma & Simon) Almost unbiased estimate of the true error.
- Not the same as repeated CV.

Issue of variable (descriptor) selection

- Ambroise C, McLachlan GJ, Selection bias in gene extraction on the basis of microarray gene-expression data, (2002) *PNAS*
- Variable selection should be executed within and not prior to cross-validation.
- Cartmell J., Enoch S., Krstajic D, Leahy D., Automated QSPR through Competitive Workflow, *J Comput Aided Mol Des* (2005)
- *Mea culpa*

How should we implement descriptor selection in our QSAR model selection process?

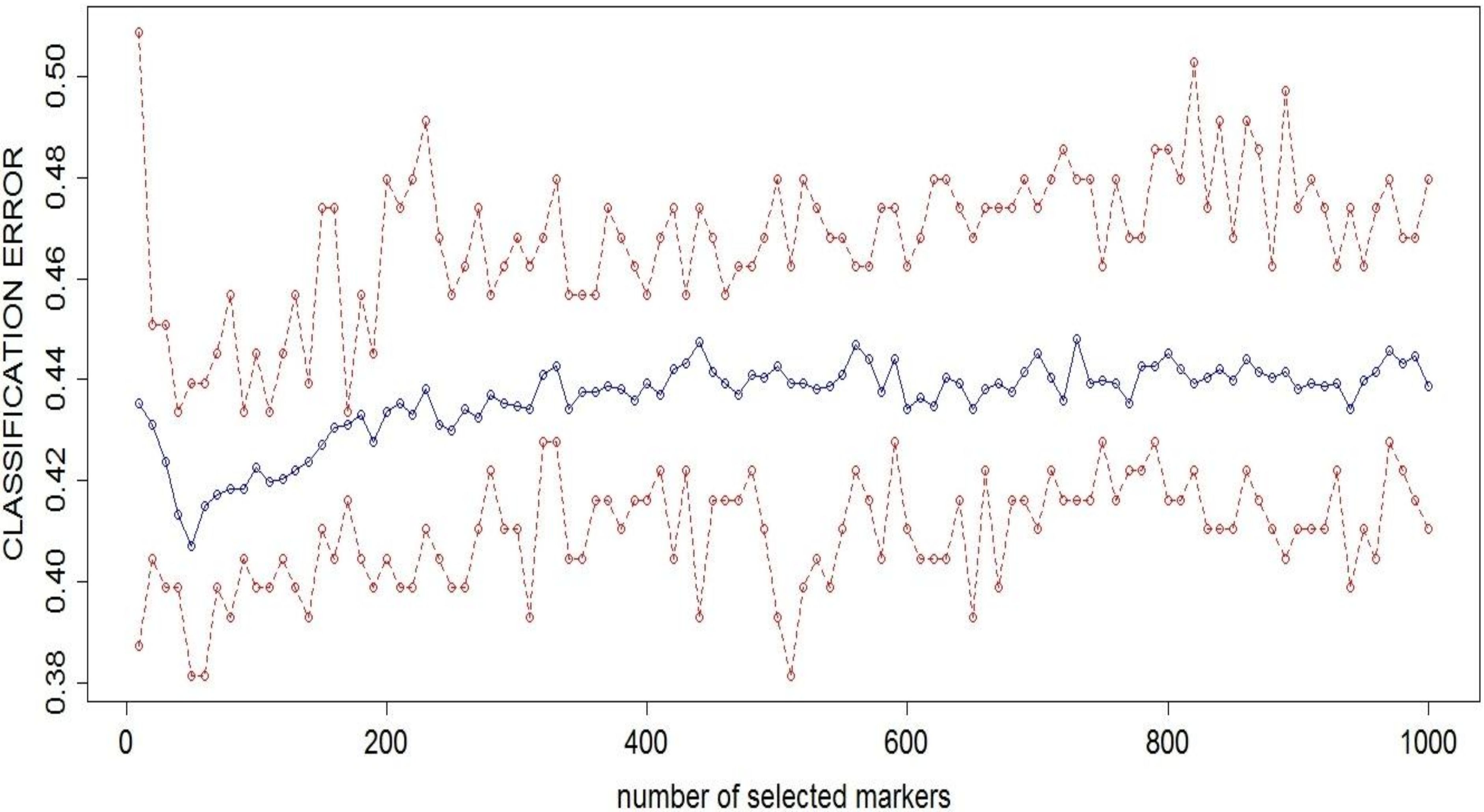
- In 10-fold CV very likely to select different descriptors on each learning set.
- Back to theory (Kurt Lewin)
- Prediction Error = $Bias^2$ + Variance



- Goal is to find optimal model complexity, i.e. optimal number of descriptors.

Descriptor selection in our QSAR model selection process?

- Define a set of numbers of descriptors to select (e.g. 10, 20, 30, …, 1000)
- For each number N (10, 20, 30, …, 1000) repeat 50 times following cross-validation procedure.
- Select N descriptors on each of 10 learning sets and predict corresponding validation sets.
- Sum up the errors from validation sets.
- Calculate min, mean, max error from 50 experiments for each N


CLASSIFICATION ERROR
0.38
0.40
0.42
0.44
0.46
0.48
0.50
0
200
400
600
800
1000
number of selected markers

QSAR Error statistics

- q2
- Y-randomisation
- Latest developments in the statistics methodology make a case for revision of current QSAR *model selection* and especially QSAR *model assessment* procedures.

Acknowledgments

- Prof. David E Leahy
- Dr Ljubomir J Buturovic
- Dr Simon Thomas

- Comments and criticisms are welcome!