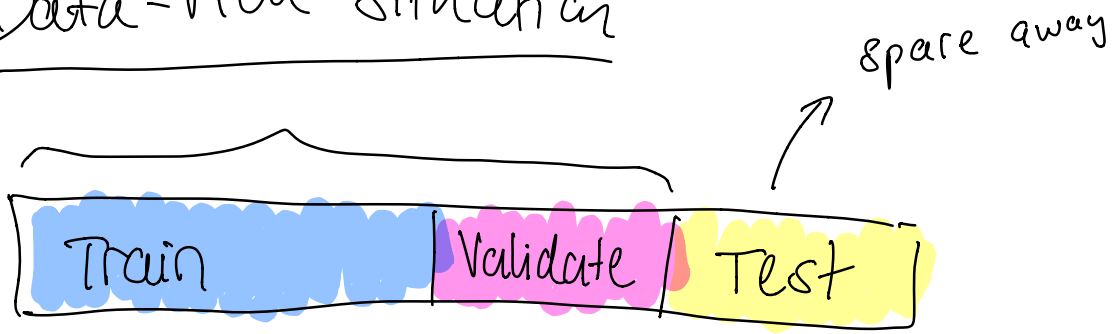
Data-rich situation

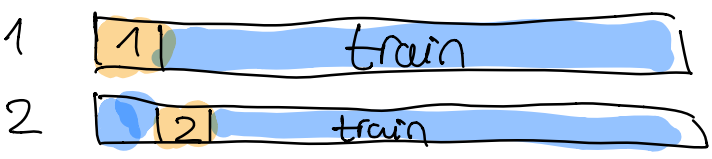


e.g.: 50% 25% 25%

Model selection — Model assessment

Validation set approach

2) LOOCV



$MSE_1 = (y_1 - \hat{y}_1)^2$

$MSE_2 = (y_2 - \hat{y}_2)^2$

$$MSE_n = (y_n - \hat{y}_n)^2$$

$$CV_n = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

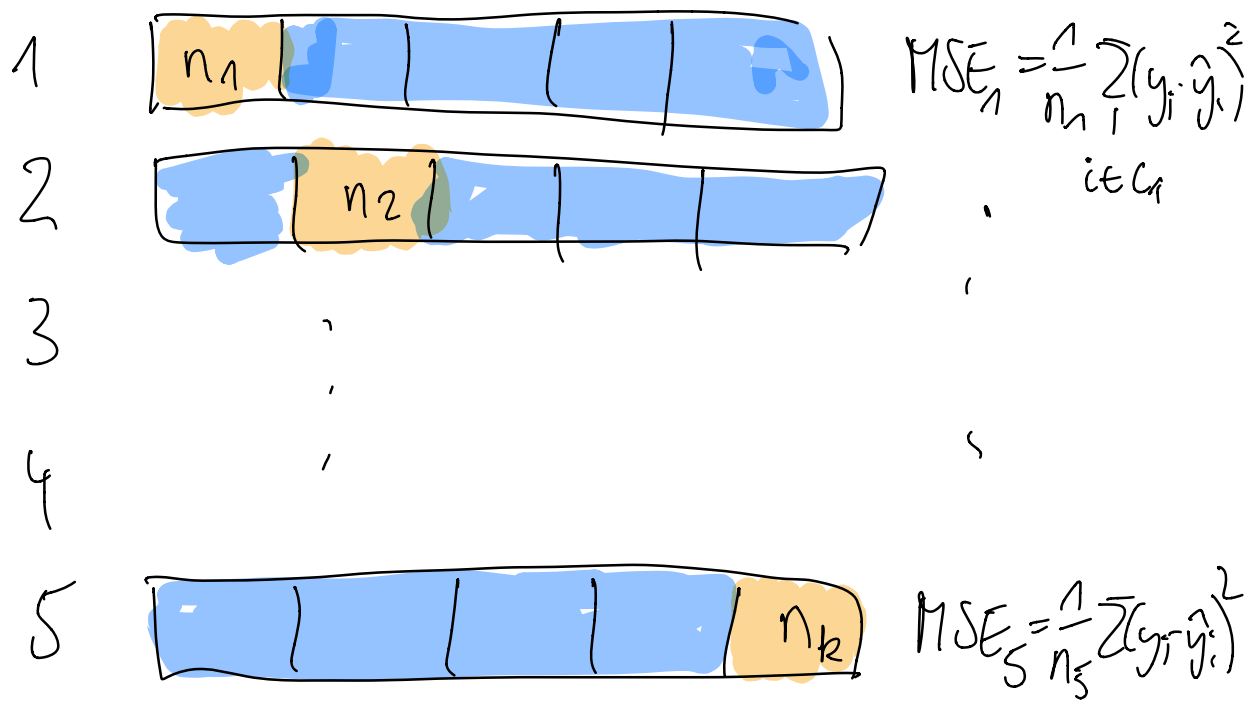
Cons: • The models are trained on almost the same dataset
• Computationally expensive

Pros:
• No randomness

---

3) k-fold CV

for example $k = 5$

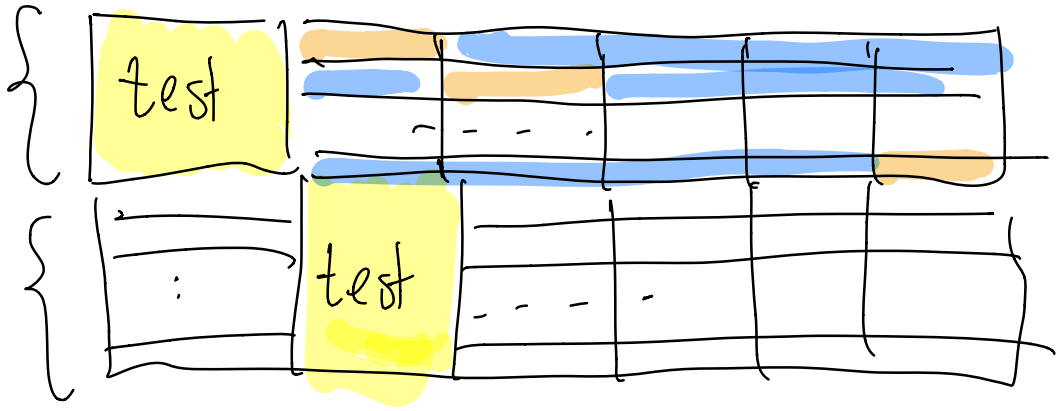


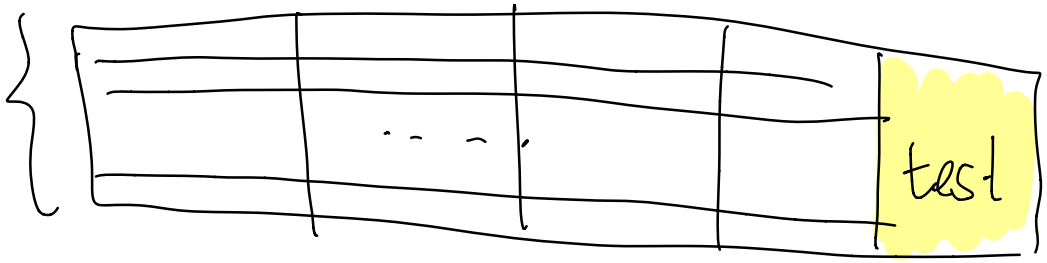
C_1 = indices in fold 1

$$CV_5 = \frac{1}{n} \sum_{j=1}^{5} n_j MSE_j$$

$$= \frac{1}{5} \sum_{j=1}^{5} MSE_j$$

if $n_1 = \ldots = n_5$





Nested
cross-validation