Validation of Classification Models

Chemometrics for Spectroscopists

Intensive Course Kraków
2021-11-29 – 12-03

Claudia Beleites
Chemometric Consulting Claudia Beleites

Topics

1. Introduction
2. Figures of Merit
3. Excursion: Design of Experiments
4. Verification Schemes
5. Resampling Techniques
6. Model Stability
7. Excursion: Model Aggregation
8. Sample Size Planning
9. Validation
10. Data-driven Model Optimization and Hyperparameter Tuning

Validation and Verification

Verification: Making sure/measuring/showing that the model meets the specifications.

Validation: Making sure that the model meets the application needs.

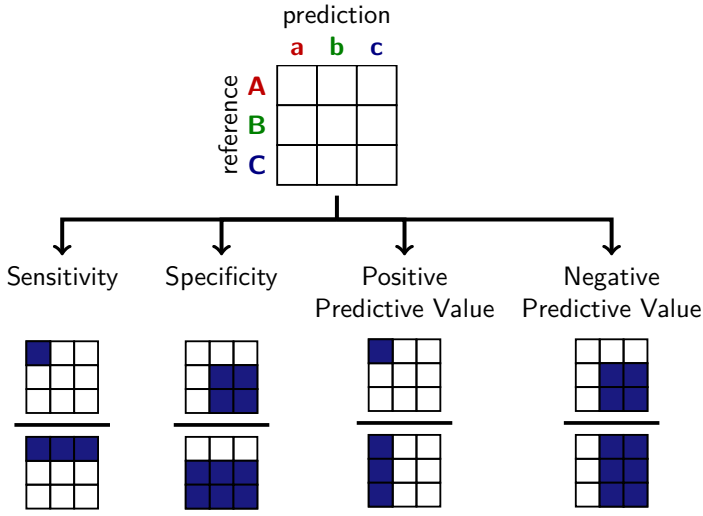
- Chemometric model validation ⇝ typically verification rather than validation is done.
- Characterize model by
- measuring its predictive performance

Model Validation Recipe

Ingredients

- Ready-to-use classifier
  treated as black box: *case* $\mapsto$ *prediction*
- Figures of merit (performance measure)
  Overall Accuracy, Sensitivity, Specificity, Predictive Values, MSE, ...
- Validation *scheme*: *How to get test cases?*
  ~~Autoprediction~~, Resampling, Test Set, Validation study

Figures of Merit: Proportions



Proportion Questions

Sensitivity: of all truly class *A* cases, which fraction is correctly recognized as class *A*?

Specificity: of all cases truly not belonging to class *A*, which fraction is correctly recognized as not belonging to class *A*?

Positive Predictive Value: of all cases predicted to belong to class *A*, which fraction does truly belong to class *A*?

Negative Predictive Value: of all cases predicted not to belong to class *A*, which fraction does truly not belong to class *A*?

accuracy: correct proportion among all predicted cases

error rate: misclassified proportion among all predicted cases

κ: chance-corrected accuracy, inter-observer agreement

Proportions: Characteristics

✔ well-known, widely used

✘ often misunderstood:

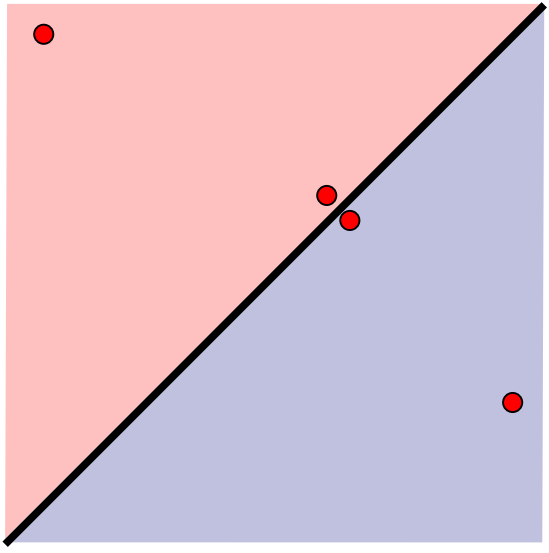
- sensitivity & specificity
  - ✔ easy to measure: test *n* cases of each class, record results
  - ✘ low relevance for application
- predictive values (positive/negative)
  - ✔ high relevance for application
  - ✘ difficult to measure: need to know relative class fequencies under application conditions
    weight rows of confusion matrix accordingly
- “single” figures of merit accuracy, error rate, κ
  - ✘ useless unless corrected for relative class fequencies under application conditions

More figures of merit

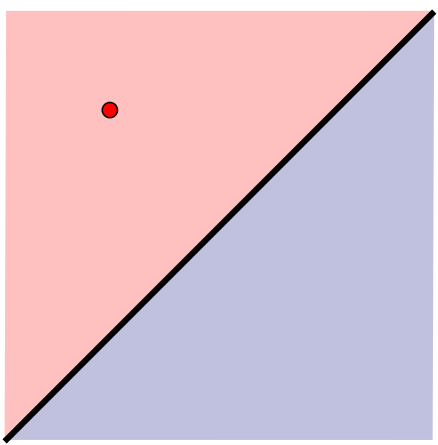
- chance-corrected: κ
  - ✔ rescaling possible for other figures of merit
  - ✔ alternative: report chance agreement (or naive model performance) together with figure of merit

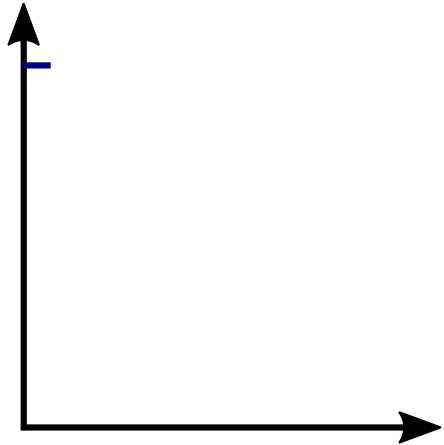
- Information gain
  - **positive likelihood ratio:** $LR_A^+ = \frac{Sens_A}{1-Spez_A}$
    How much do the odds to belong to class A increase when a case is predicted to belong to class A?
  - **negative likelihood ratio:** $LR_A^- = \frac{Spez_A}{1-Sens_A}$
    How much do the odds to belong to class A decrease when a case is predicted not to belong to class A?
  - ✔ independent of relative class fequencies under application conditions

Proportions: Behaviour

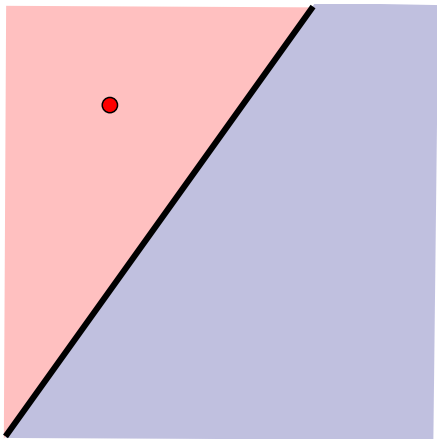


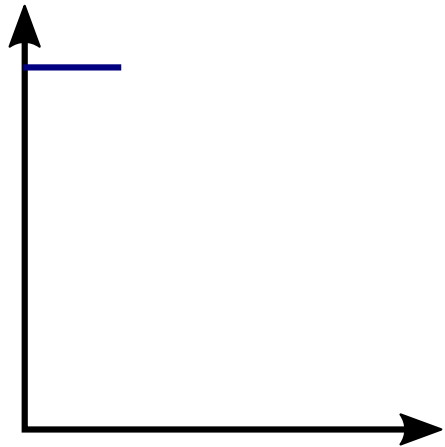
Proportions: Behaviour



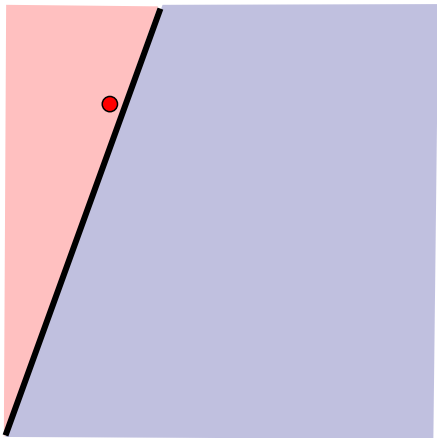


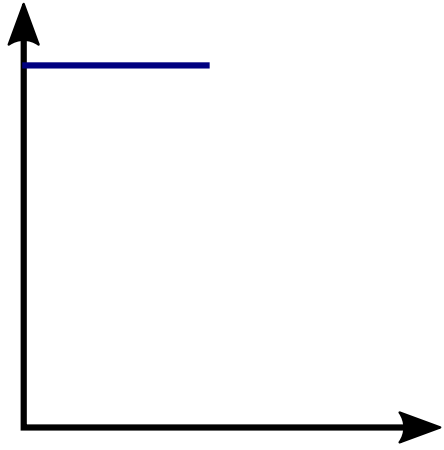
Proportions: Behaviour



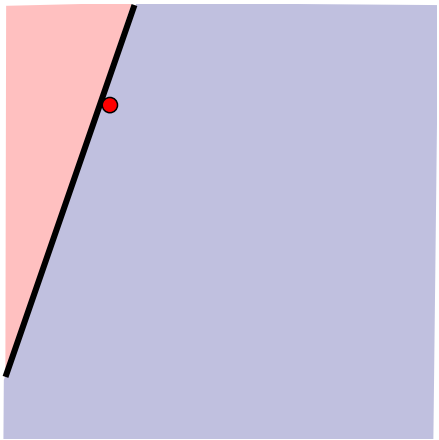


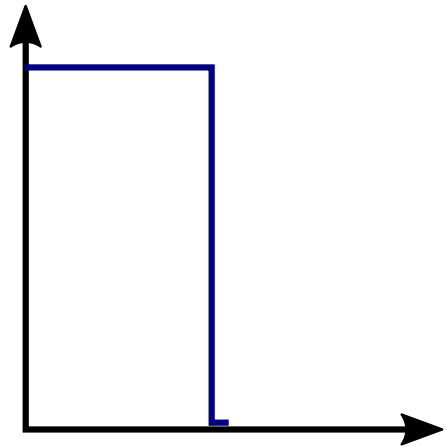
Proportions: Behaviour



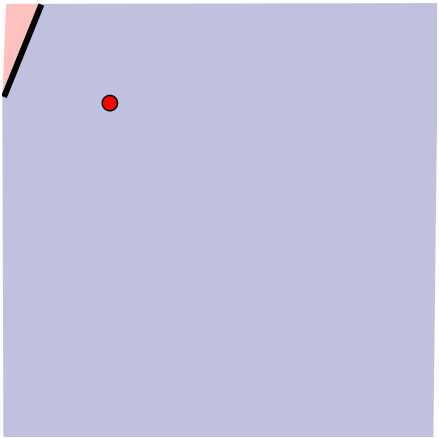


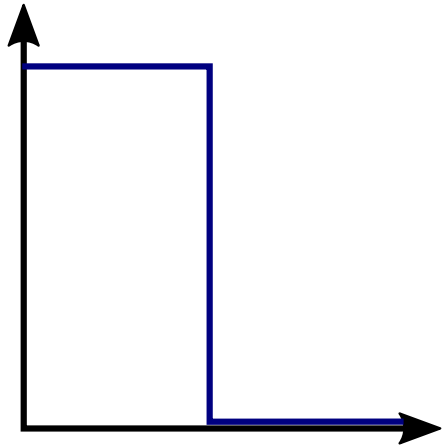
Proportions: Behaviour



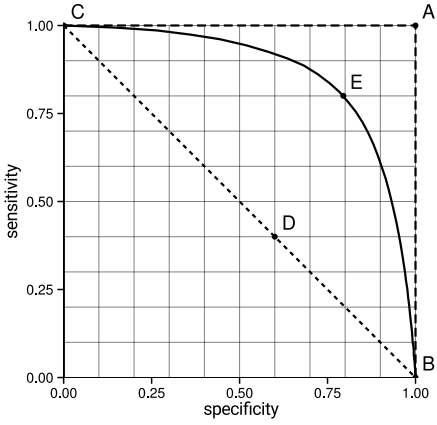


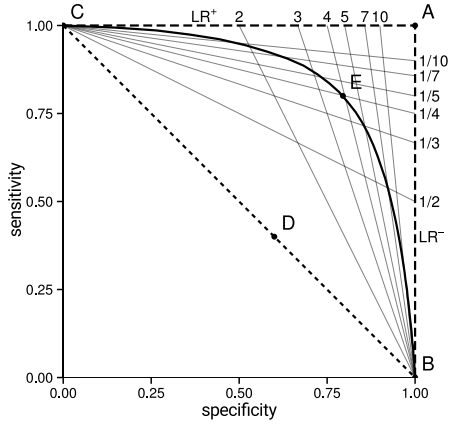
Proportions: Behaviour





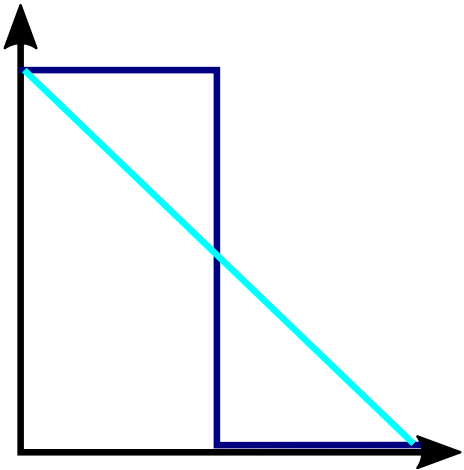
Receiver Operating Charcteristic/Specificity-Sensitivity-Diagram





CHEMO METRIX

(Strictly) Proper Scoring Rules

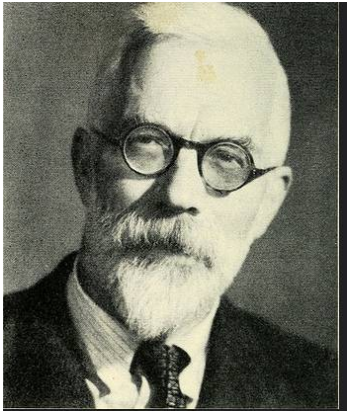


Wanted: Figure of merit that …

- …continuously penalizes closeness to class boundary
- …continuously reacts to changes in the model
- …slight deterioration ⇝ slight drop in measured performance
- …has exactly one optimum
- at the best classifier.

Brier's Score: Mean Squared Error for Classification

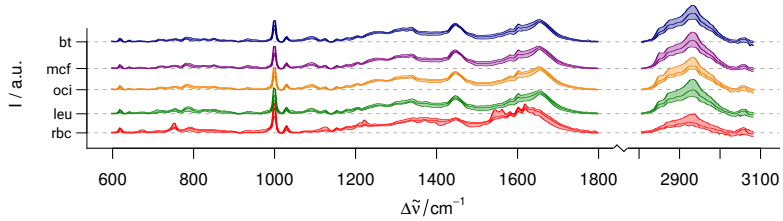
- Classifier that predicts class membership probability rather than labels
- Idea: of all cases where classifier predicts x % class membership, x % should belong to class in question
  a.k.a. *well calibrated prediction*
- Brier's score: $BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i)^2$ or
  $BS = \frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{N}(\hat{p}_{ij} - p_{ij})^2$ (multiclass version) with
  N . . . number of cases
  i . . . case in question
  R . . . number of classes
  j . . . class in question
  p . . . class membership, usually $\in \{0, 1\}$
  $\hat{p}$ . . . predicted class membership $\in [0, 1]$



To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination.
He can perhaps say what the experiment died of.

— R. Fisher, 1938

Example: Tumor Cell Identification



rbc	normal red blood cells	5 donors	372 spectra
leu	normal leukocytes	5 donors	569 spectra
oci	acute myelotic leukemia cell line OCI-AML	5 batches	518 spectra
mcf	breast cancer cell line MCF-7	5 batches	558 spectra
bt	breast cancer cell line BT-20	5 batches	532 spectra
total			2549 spectra

Neugebauer2010a
Beleites2013

CHEMO METRIX

Aims

Ensure and/or establish

- validity
- reliability
- repeatability, reproducibility

- statistical power and sensitivity
  - necessary sample size
  - efficiency: get information with lowest possible experimental effort

of the experimental data ⇝ derived model ⇝ interpretation

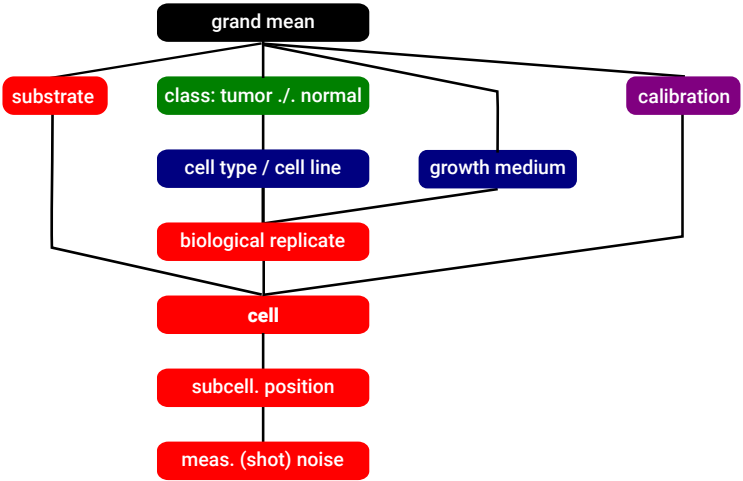
“GLP” for experiment set-up

Vocabulary: factors and effects

- A **factor** influences the system we study (the spectra)
  ⇝ **effect**
- A **confounding factor** is a hidden factor that influences/disturbs our experiment
- The same factor may be of interest or disturbing depending on the study!

- Distinguish: *fixed* vs. *random* factors
- Relations between factors: *crossed* vs. *nested* factors
- Design: *orthogonal* and *balanced* designs/data

Hasse Diagrams



Crossed Factors

. . . occur independent of each other:

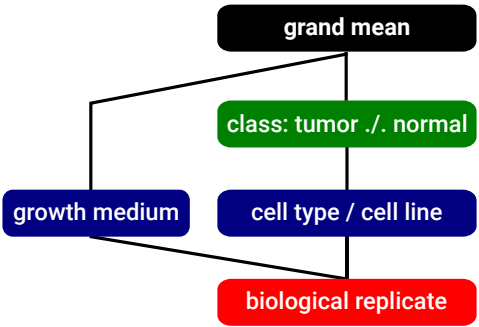
- **Fully Crossed:** measurements available for all combinations

	cell line		
medium	bt	mcf	oci
RPMI	46	51	57
DMEM	49	56	78

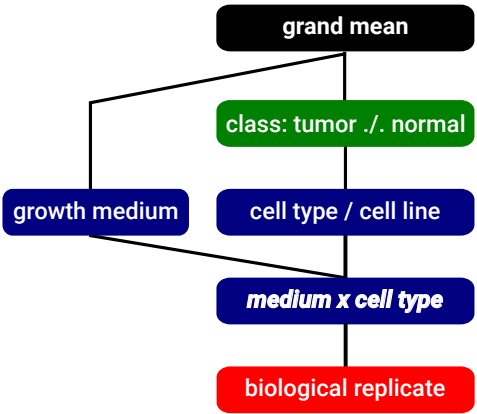
- **Partially Crossed:** measurements only of some combinations

	cell line		
medium	bt	mcf	oci
RPMI	·	51	57
DMEM	49	56	·

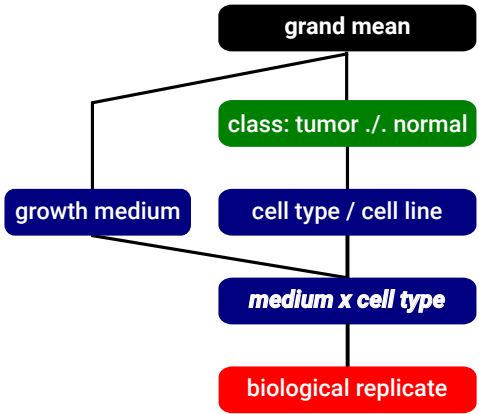
Excursion: Interactions

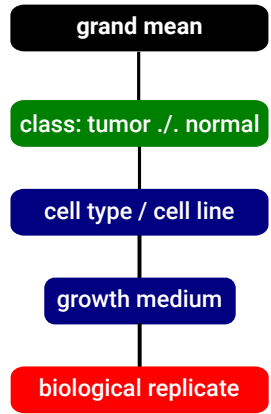


Excursion: Interactions



Excursion: Interactions





CHEMO METRIX

Nested Factors

Levels of inner factor occur only within *one* level of outer factor

e.g. ***cell within sample****: another sample never contains the same cell*

✔ Explicit Coding:

	cell									
sample	1	2	3	4	5	6	7	8	9	10
A	10	10	·	·	·	·	·	·	·	10
B	·	·	·	·	10	10	10	10	·	·
C	·	·	10	10	·	·	·	·	10	·

Nested Factors

Levels of inner factor occur only within *one* level of outer factor

e.g. ***cell within sample****: another sample never contains the same cell*

✔ Explicit Coding:

	cell									
sample	1	2	3	4	5	6	7	8	9	10
A	10	10	·	·	·	·	·	·	·	10
B	·	·	·	·	10	10	10	10	·	·
C	·	·	10	10	·	·	·	·	10	·

✘ Implicit Coding:

	cell			
sample	1	2	3	4
A	10	10	10	·
B	10	10	10	10
C	10	10	10	·

Nested Factors

Levels of inner factor occur only within *one* level of outer factor
e.g. ***cell within sample****: another sample never contains the same cell*

✔ Explicit Coding:

	cell									
sample	1	2	3	4	5	6	7	8	9	10
A	10	10	·	·	·	·	·	·	·	10
B	·	·	·	·	10	10	10	10	·	·
C	·	·	10	10	·	·	·	·	10	·

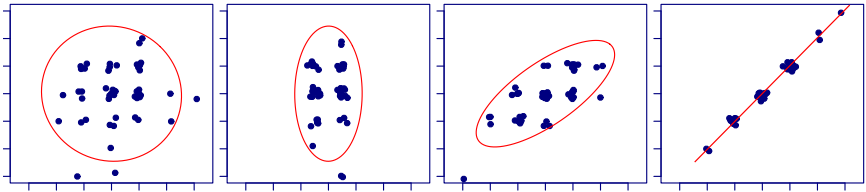
✘ Implicit Coding:

	cell			
sample	1	2	3	4
A	10	10	10	·
B	10	10	10	10
C	10	10	10	·

✔ Make sure crossed factors do not appear nested in your DoE!
Crossed factor with interaction ≠ nested factor!

Design: Orthogonality and Balance

- *Orthogonality:* no correlation between factors

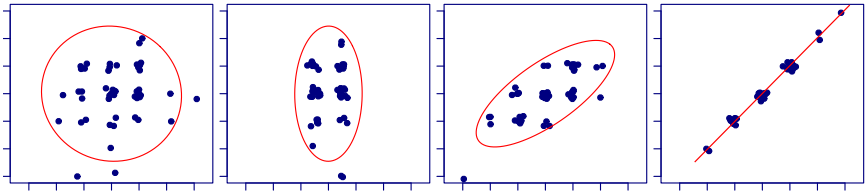


- *Balance:* for each factor, all other factor levels occur with the same frequency

	day								
cell line	a	b	c	d	e	f	g	h	i
bt	50	50	50	50	50	50	50	50	50
mcf	50	50	50	50	50	50	50	50	50
oci	50	50	50	50	50	50	50	50	50

Design: Orthogonality and Balance

- *Orthogonality:* no correlation between factors



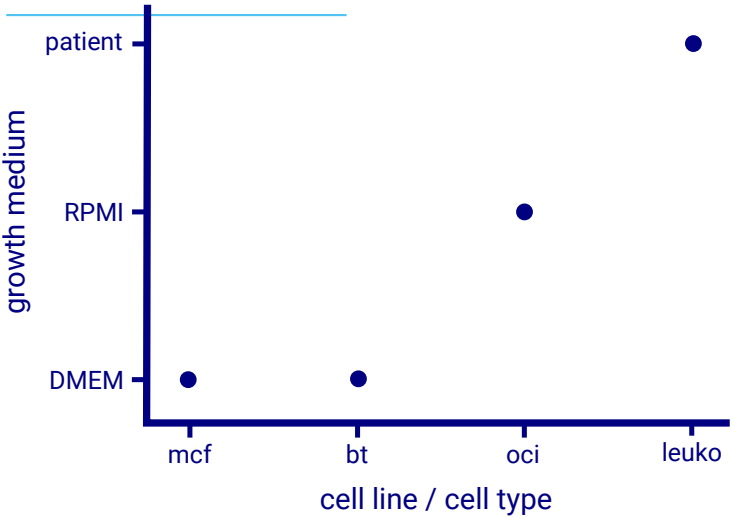
✔ Orthogonal design (data): effects of factors can be completely separated

- *Balance:* for each factor, all other factor levels occur with the same frequency

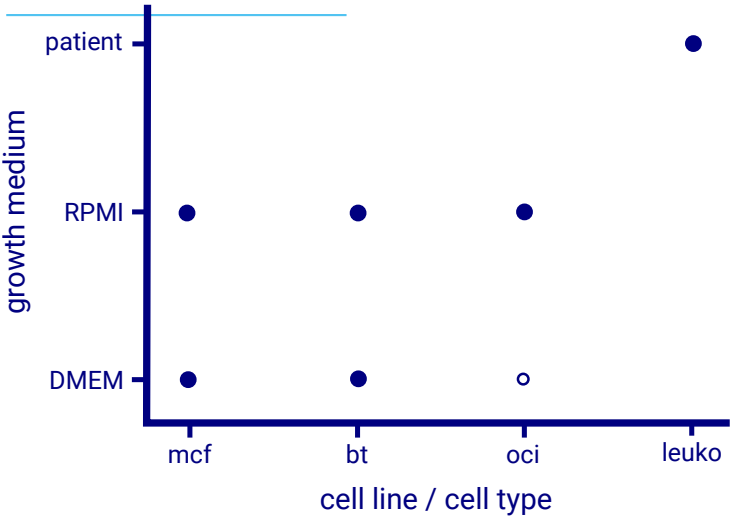
	day								
cell line	a	b	c	d	e	f	g	h	i
bt	50	50	50	50	50	50	50	50	50
mcf	50	50	50	50	50	50	50	50	50
oci	50	50	50	50	50	50	50	50	50

✔ Balanced data easier to analyze

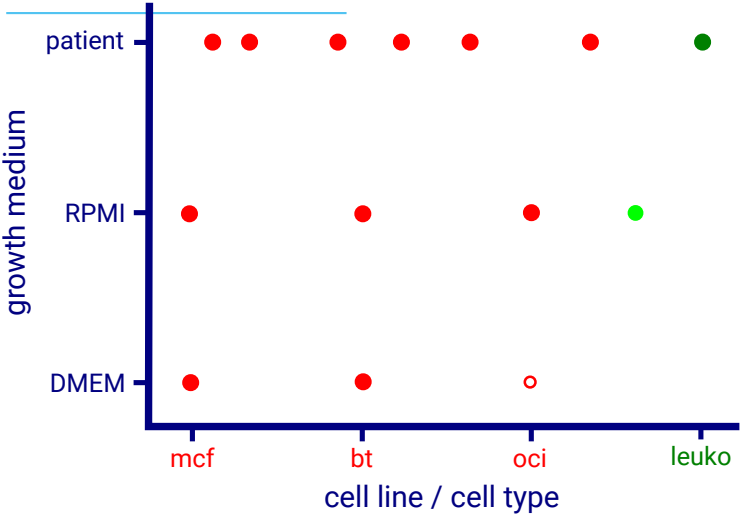
Your turn! Crossed? Partially crossed? Nested? Orthogonal design? Balanced?



Your turn! Crossed? Partially crossed? Nested? Orthogonal design? Balanced?



Your turn! Crossed? Partially crossed? Nested? Orthogonal design? Balanced?



Random and Fixed Factors

Fixed Factor

- occurs at levels which
  - ✔ We can either *know* and *reproduce*, or even
  - ✔ *set* (*fix*).

Random Factor

- occurs at levels which
  - – We can *know*, but
  - ✘ will never meet again (*reproduce*)

⚠ The same factor may be fixed in one study/application and random in another!

Random and Fixed Factors

Fixed Factor

- occurs at levels which
  - ✔ We can either *know* and *reproduce*, or even
  - ✔ *set* (*fix*).

✔ Recognize/predict the level/value for new unknown data

Random Factor

- occurs at levels which
  - – We can *know*, but
  - ✘ will never meet again (*reproduce*)

✘ New unknown data always corresponds to new level of a random factor

⚠ The same factor may be fixed in one study/application and random in another!

Random and Fixed Factors

Fixed Factor

- occurs at levels which
  - ✔ We can either *know* and *reproduce*, or even
  - ✔ *set* (*fix*).

✔ Recognize/predict the level/value for new unknown data

✔ Account for *known* level/value for new unknown data

Random Factor

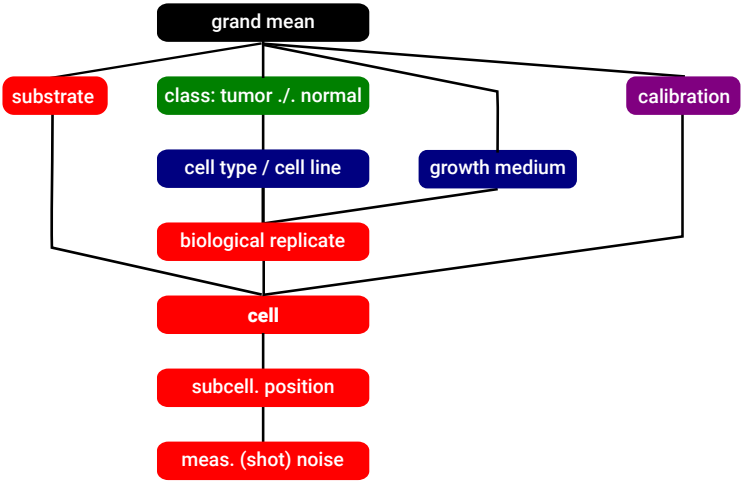
- occurs at levels which
  - – We can *know*, but
  - ✘ will never meet again (*reproduce*)

✘ New unknown data always corresponds to new level of a random factor

✘ Account for random factors only by general (population) behaviour.

⚠ The same factor may be fixed in one study/application and random in another!

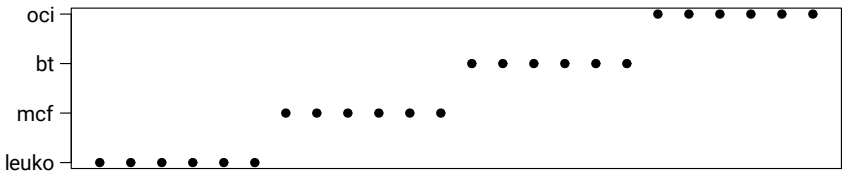
Hasse Diagrams



Dealing with Factors Disturbing our Experiment

- ✔ Include fixed factors in model
- ✔ Reduce variance as much as possible:
  - measure and reduce systematic influence: e.g. instrument calibration
  - standard operating procedures
  - automation
  - quality control, e.g. positive and negative controls
- ✔ *Representative sampling* ⇝ train model with *reduced influence*.
- ✔ Randomization
- ✔ Blocking/Stratification
- *Disregard* confounders which are *known* to be unimportant
- Possibly: reproducible (= fixed factor) subgroups
  - use *local* models

Randomization and Blocking/Stratification



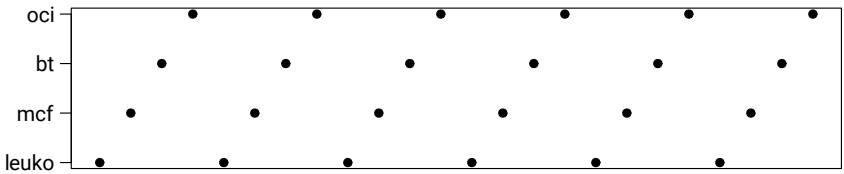
- **Blocking:** Factor causing groups ⇝ repeat smaller (sub)experiment for each group

✔ **Randomize** assignment of samples/cases to fixed factors

✔ **Randomize** measurement order for random factors

CHEMO METRIX

Randomization and Blocking/Stratification

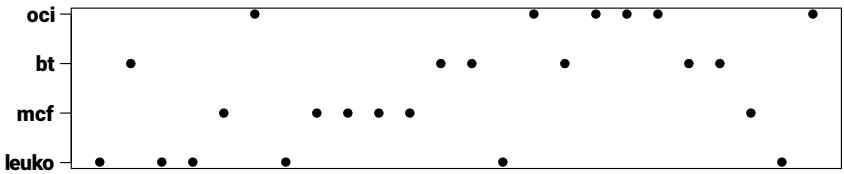


- **Blocking:** Factor causing groups ⇝ repeat smaller (sub)experiment for each group

✔ **Randomize** assignment of samples/cases to fixed factors

✔ **Randomize** measurement order for random factors

Randomization and Blocking/Stratification

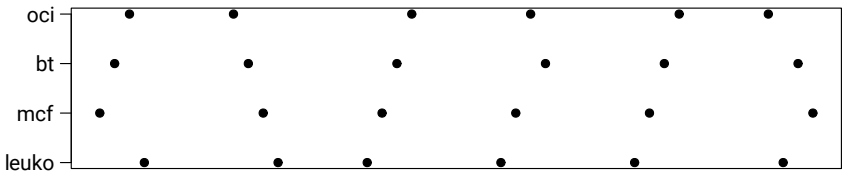


- **Blocking:** Factor causing groups ⇝ repeat smaller (sub)experiment for each group

✔ **Randomize** assignment of samples/cases to fixed factors

✔ **Randomize** measurement order for random factors

Randomization and Blocking/Stratification



- **Blocking:** Factor causing groups ⇝ repeat smaller (sub)experiment for each group

✔ **Randomize** assignment of samples/cases to fixed factors

✔ **Randomize** measurement order for random factors

Model Testing: Measure the Model's Performance

Different kinds of test samples
⇝ different performance measures

Goodness of fit: *training samples*
⇝ residuals

Generalization error: statistically independent samples
resampling,
test set measured at same time as training set

Future performance: samples measured *after* training samples
dedicated test set for detection of drift

Resampling for Classifier Validation

✘ We don't have enough samples

- Training:
  - Model quality depends on ratio $n_{train} : d.f.$
  - Linear classifier: 5 samples/(variate · class)

✔ We want to use all samples for training

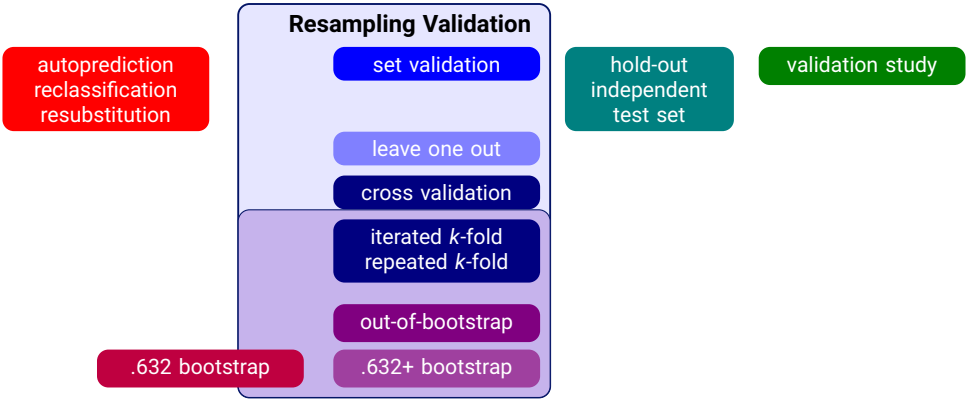
Resampling for Classifier Validation

✘ We don't have enough samples

- Training:
  - Model quality depends on ratio $n_{train} : d.f.$
  - Linear classifier: 5 samples/(variate · class)
  - ✔ We want to use all samples for training

- Testing:
  - ✔ We want to know whether the model is stable
  - Quality of the performance measure depends on n_{test}
  - Width of 95 % confidence interval $\overset{!}{\leq}$ 10 % for $p = 90\,\%$: $n_{test} \geq 140$
  - ✔ We want to use all samples for testing

Validation Schemes: Overview

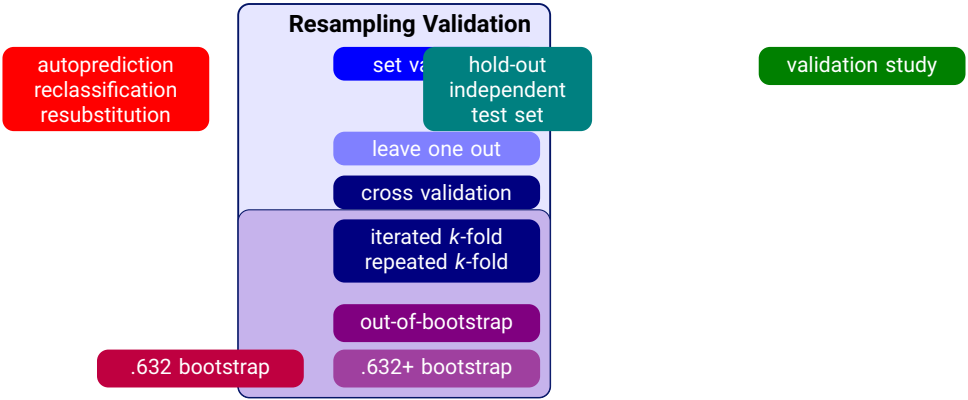


R Kohavi: **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In: *Proc. 14th International Joint Conference on Artificial Intelligence*, 1995, Morgan Kaufmann, USA, 1995, 1137–1145.

Beleites2005

CHEMO METRIX

Validation Schemes: Overview

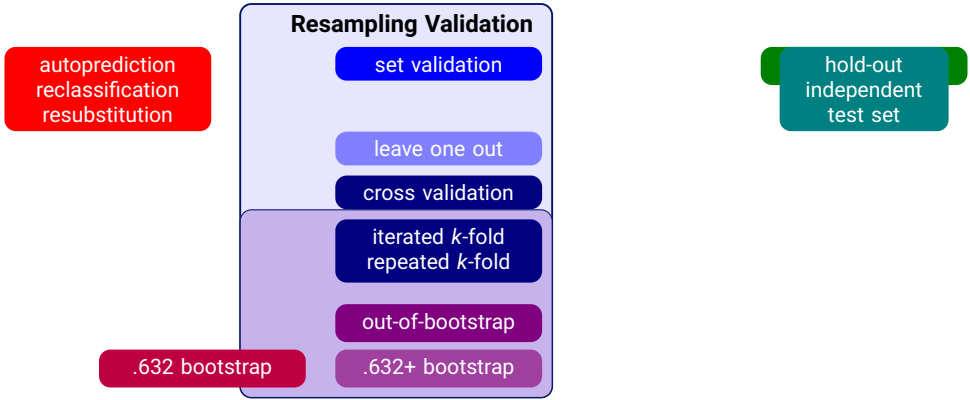


R Kohavi: **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In: *Proc. 14th International Joint Conference on Artificial Intelligence*, 1995, Morgan Kaufmann, USA, 1995, 1137–1145.

Beleites2005

CHEMO METRIX

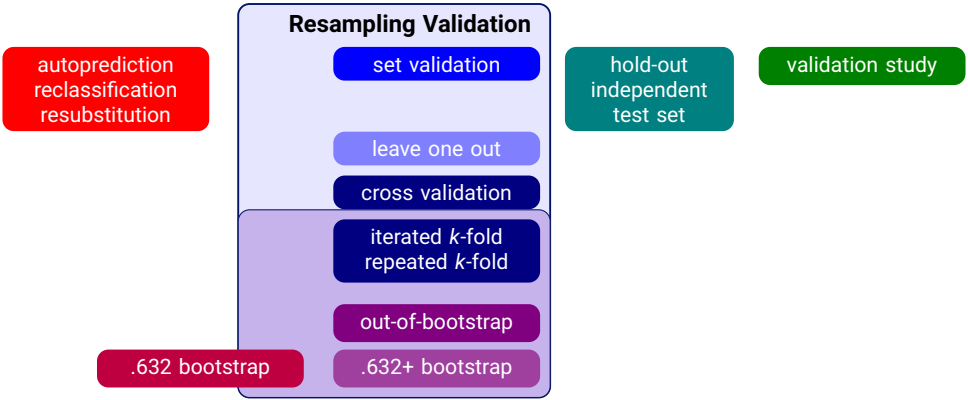
Validation Schemes: Overview



R Kohavi: **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In: *Proc. 14th International Joint Conference on Artificial Intelligence*, 1995, Morgan Kaufmann, USA, 1995, 1137–1145.

Beleites2005

Validation Schemes: Overview

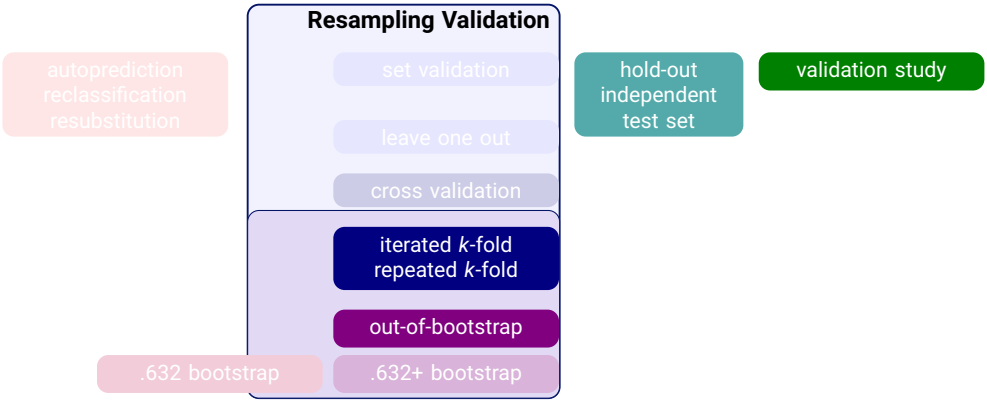


R Kohavi: **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In: *Proc. 14th International Joint Conference on Artificial Intelligence*, 1995, Morgan Kaufmann, USA, 1995, 1137–1145.

Beleites2005

CHEMO METRIX

Validation Schemes: Recommendations



Kohavi1995

Beleites2005

CHEMO METRIX

Resampling vs. Validation Study

	Resampling	Validation Study
statistical properties		
bias	✔ pessimistic (low)	✔ unbiased
variance	$f(n)$	$f(n_{test})$
efficient use of cases	✔	✔
measure model stability	✔ iterated	✔/✘
measure drift	✘	✔ DoE
future case performance	✘	✔ DoE
out-of-spec cases	✘	✔ DoE
practical properties		
independence	⚠ splitting error prone	✔
effort	✔ computational	✘ experimental

Beleites2008
Esbensen2010

CHEMO METRIX

Resampling vs. Hold Out

	Resampling	Hold Out (Set) Validation
statistical properties		
bias	✔ pessimistic (low)	✔ unbiased
variance	✔ $f(n)$ lower	✘ $f(n_{test})$ HUGE
efficient use of cases	✔	✘
measure model stability	✔ iterated	✘
measure drift	✘	✘
future case performance	✘	✘
out-of-spec cases	✘	✘(✔)
practical properties		
independence	⚠ splitting error prone	⚠ same as resampling
effort	✔ computational	✔ low

Beleites2008
Esbensen2010

CHEMO METRIX

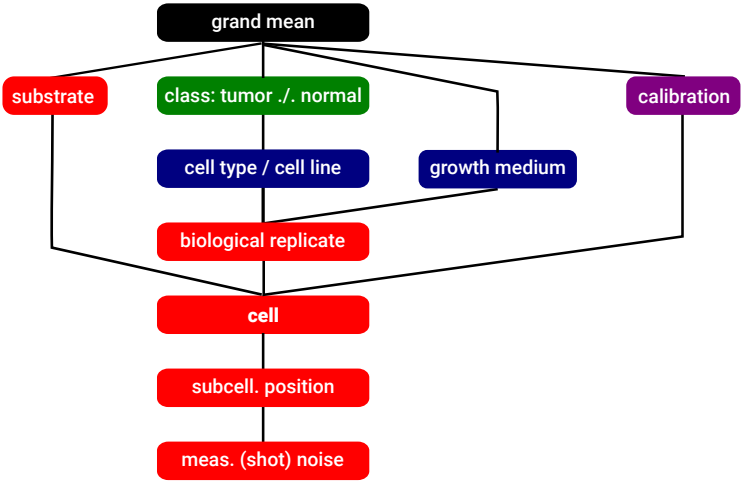
Resampling vs. Hold Out

	Resampling	Hold Out (Set) Validation
statistical properties		
bias	✔ pessimistic (low)	✔ unbiased
variance	✔ $f(n)$ lower	✘ $f(n_{test})$ HUGE
efficient use of cases	✔	✘
measure model stability	✔ iterated	✘
measure drift	✘	✘
future case performance	✘	✘
out-of-spec cases	✘	✘(✔)
practical properties		
independence	⚠ splitting error prone	✔ by organization
effort	✔ computational	✔ organizational

Beleites2008
Esbensen2010

CHEMO METRIX

Hasse-Diagram revisited



CHEMO METRIX

Validation Check List

- Randomize order of measurements
- Split at highest level in sample hierarchy and independently for known confounders use Hasse diagram to determine
- Split before 1st step that involves multiple cases
- Additional independent validation for data-driven optimization/tuning/model selection
- Test cases: reference labels must be independent of cases (measurements, spectra, chromatograms, . . .)
- Ensure correctness of code

CHEMO METRIX

Validation Check List

- Randomize order of measurements
- Split at highest level in sample hierarchy and independently for known confounders
  use Hasse diagram to determine patients, strains, cell lines,
- Split before 1st step that involves multiple cases
- Additional independent validation for data-driven optimization/tuning/model selection
- Test cases: reference labels must be independent of cases (measurements, spectra, chromatograms, . . .)
- Ensure correctness of code

Validation Check List

- Randomize order of measurements
- Split at highest level in sample hierarchy and independently for known confounders
  use Hasse diagram to determine patients, strains, cell lines,
  day of measurement, before/after new calibration, . . .
- Split before 1st step that involves multiple cases
- Additional independent validation for data-driven optimization/tuning/model selection
- Test cases: reference labels must be independent of cases (measurements, spectra, chromatograms, . . .)
- Ensure correctness of code

Validation Check List

- Randomize order of measurements
- Split at highest level in sample hierarchy and independently for known confounders
  use Hasse diagram to determine patients, strains, cell lines,
  day of measurement, before/after new calibration, . . .
- Split before 1st step that involves multiple cases
  centering, PCA preprocessing, . . .
- Additional independent validation for data-driven optimization/tuning/model selection
- Test cases: reference labels must be independent of cases (measurements, spectra, chromatograms, . . .)
- Ensure correctness of code

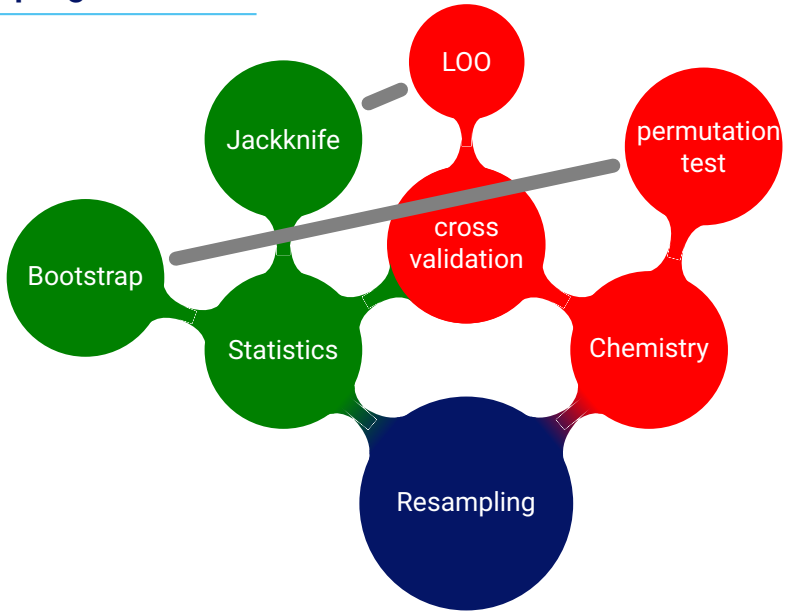
Validation Check List

- Randomize order of measurements
- Split at highest level in sample hierarchy and independently for known confounders
  use Hasse diagram to determine patients, strains, cell lines,
  day of measurement, before/after new calibration, . . .
- Split before 1st step that involves multiple cases
  centering, PCA preprocessing, . . .
- Additional independent validation for data-driven optimization/tuning/model selection
  nested/double cross validation or train-validate-test ⇝ necessary patient numbers
  HUGE
- Test cases: reference labels must be independent of cases (measurements, spectra, chromatograms, . . .)
- Ensure correctness of code

Validation Check List

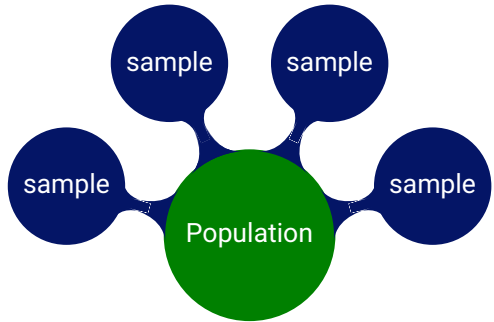
- Randomize order of measurements
- Split at highest level in sample hierarchy and independently for known confounders
  use Hasse diagram to determine patients, strains, cell lines,
  day of measurement, before/after new calibration, . . .
- Split before 1st step that involves multiple cases
  centering, PCA preprocessing, . . .
- Additional independent validation for data-driven optimization/tuning/model selection
  nested/double cross validation or train-validate-test ⇝ necessary patient numbers
  HUGE
- Test cases: reference labels must be independent of cases (measurements, spectra, chromatograms, . . .)
  cluster analysis to assign labels → OK for training cases
- Ensure correctness of code

Resampling: Fields of Use



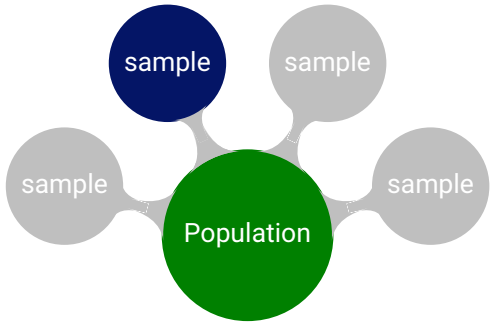
CHEMO METRIX

The Concept behind Resampling

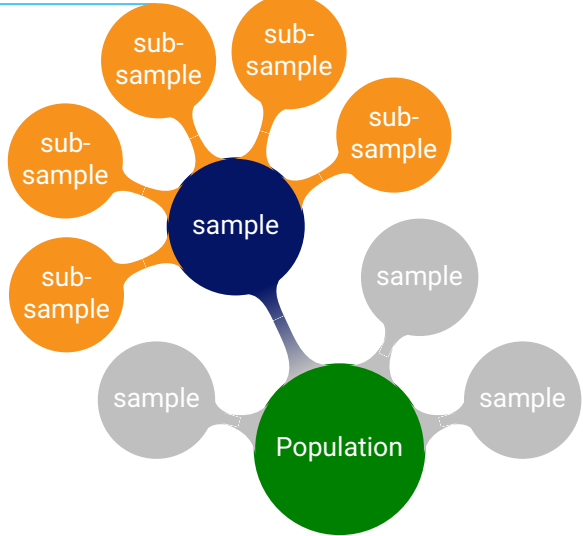


CHEMO METRIX

The Concept behind Resampling



The Concept behind Resampling



- Subsamples are approximations of (more) real samples
- Subsample is perturbed version of the real sample

CHEMO METRIX

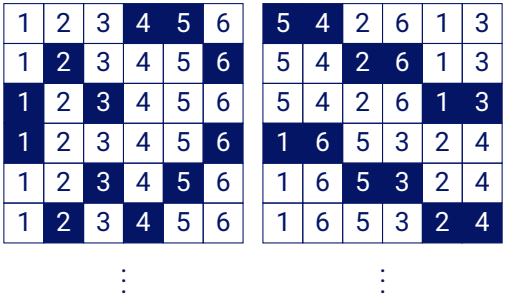
Cross Validation: Drawing without Replacement

1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6

5	4	2	6	1	3
5	4	2	6	1	3
5	4	2	6	1	3

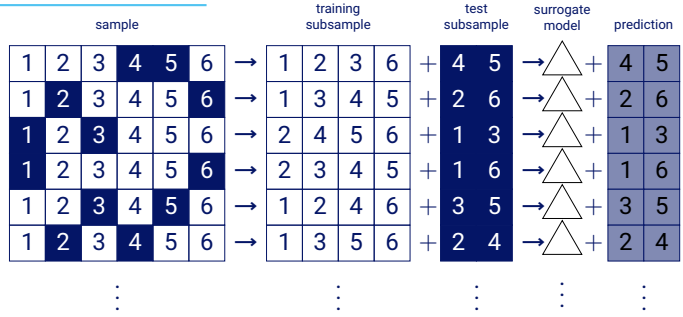
- ✔ Each sample is left out exactly once
- ✔ No copies

Cross Validation: Drawing without Replacement



- ✔ Each sample is left out exactly once per iteration
- ✔ No copies
- • Iterations possible also with *k*-fold or leave-*n*-out cross validation
- ✘ Leave-one-out cannot be iterated

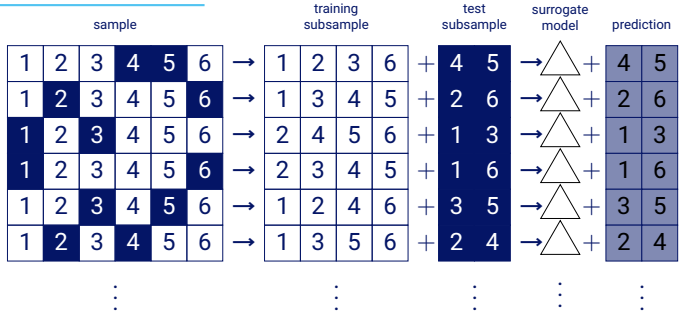
Resampling for Model Validation: Assumptions



- Surrogate model equals model of real sample
- Surrogate models equal to each other
- All samples are equal (come from the same distribution)

CHEMO METRIX

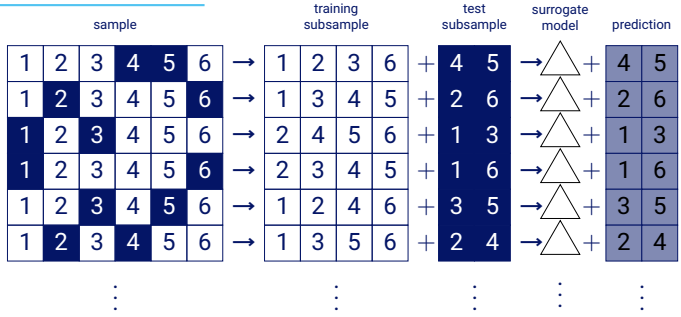
Resampling for Model Validation: Assumptions



- Surrogate model equals model of real sample
- ✘ Violation ⇝ pessimistic bias
- Surrogate models equal to each other
- All samples are equal (come from the same distribution)

CHEMO METRIX

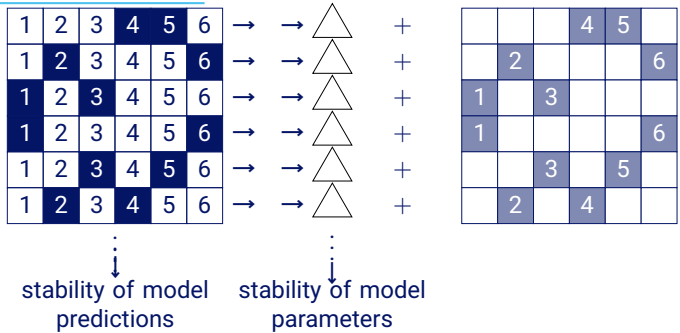
Resampling for Model Validation: Assumptions



- Surrogate model equals model of real sample
- ✘ Violation ⇝ pessimistic bias
- Surrogate models equal to each other
- ✘ Violation (instability) ⇝ higher variance
- All samples are equal (come from the same distribution)

CHEMO METRIX

Model Stability

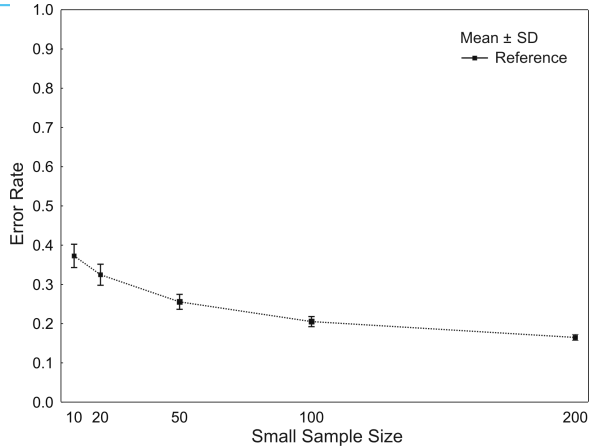


- Subsamples are perturbed versions of real sample
- ✔ Measure stability of model
  - Stability of model parameters
  - Stability of predictions

- Iterated cross validation reduces variance due to instability of surrogate models.

CHEMO METRIX

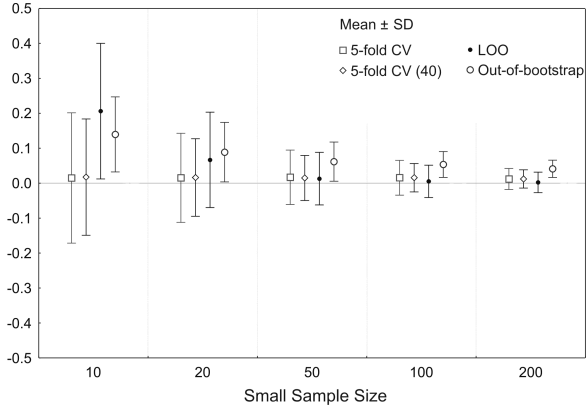
Bias and Variance of Classifier Performance



- PLS-DA with 2 latent variables
- simulated data: 200 variate normal distribution
- overlap of classes ca. 10 %

CHEMO METRIX

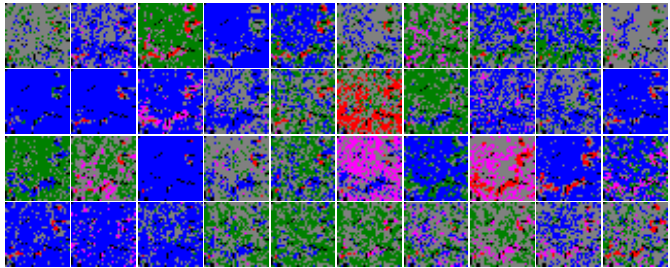
Bias and Variance of Classifier Performance



- PLS-DA with 2 latent variables
- simulated data: 200 variate normal distribution
- overlap of classes ca. 10 %

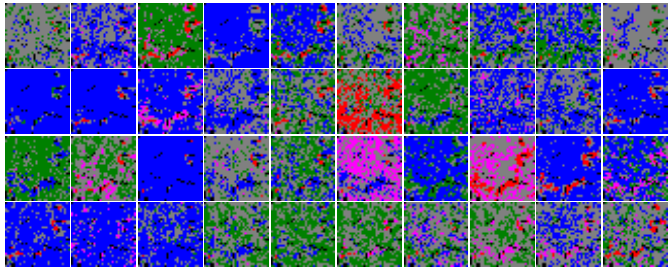
CHEMO METRIX

Model Stability: 40× 8-fold cross validation



- FTIR images of tumour sections (normal, °II, °III, °IV)
- total: 150 images of 58 patients: 133 000 spectra
  smallest class: °II, 4 800 spectra (3 patients, 5 images)
- LDA after automatic selection of 8 spectral regions
- reject spectra with posterior probability <0.85

Model Stability: 40× 8-fold cross validation

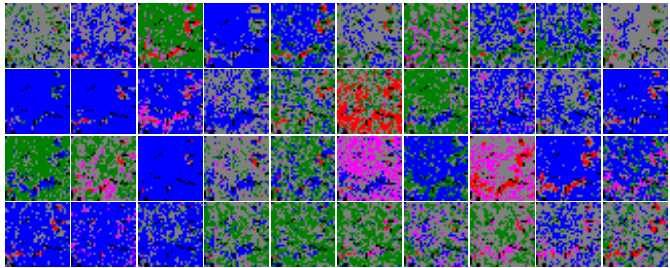


✘ Classification: unstable = bad

- Deviation is always in direction “wrong”

CHEMO METRIX

Model Stability: 40× 8-fold cross validation

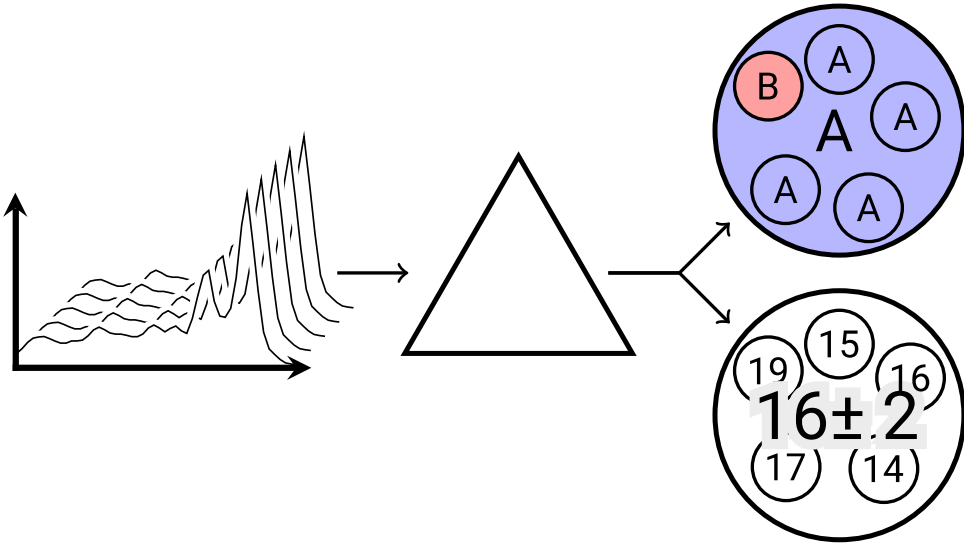


✘ Classification: unstable = bad

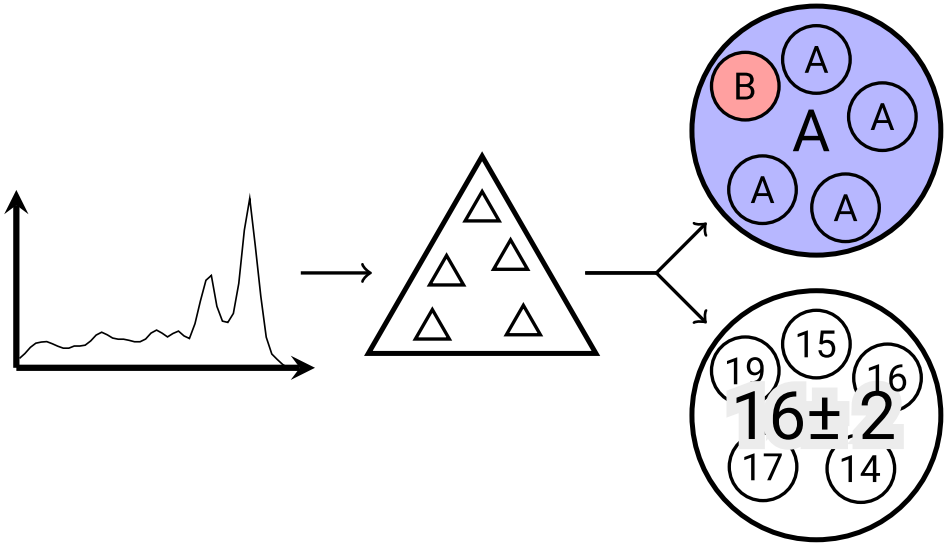
- Deviation is always in direction “wrong”
- Stabilization of the model ⇝ improved performance

CHEMO METRIX

Model Aggregation

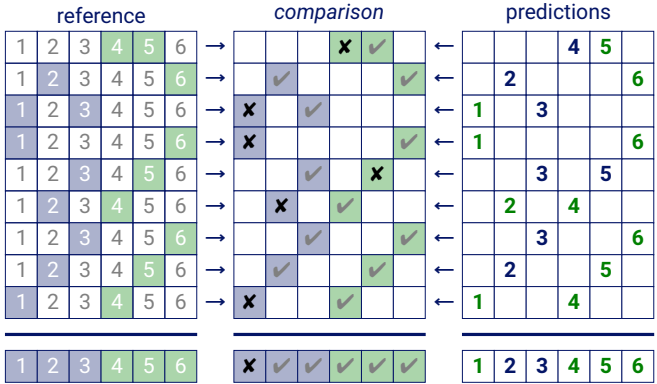


Model Aggregation



CHEMO METRIX

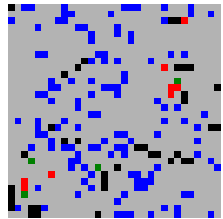
Model Aggregation: Testing

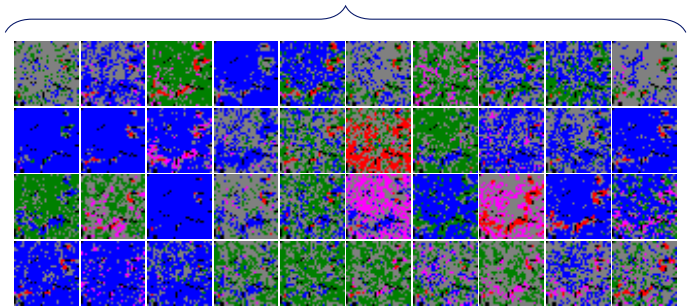


- 3× 3-fold cross validation: $\frac{6}{18} = 33\,\%$ errors
- aggregated model (majority vote): $\frac{1}{6} = 17\,\%$ errors

CHEMO METRIX

Again the Tumour Sample





CHEMO METRIX

How many cases do we need?

... **to train a good classifier?**

- rules of thumb
  linear classifier: $\frac{n}{p} \geq 3 - 5$ in each class

... **to measure the classifier's performance?**

How many cases do we need?

... **to train a good classifier?**

- rules of thumb
  linear classifier: $\frac{n}{p} \geq 3 - 5$ in each class

$\Rightarrow$ learning curve

... **to measure the classifier's performance?**

How many cases do we need?

. . . to train a good classifier?

- rules of thumb
  linear classifier: $\frac{n}{p} \geq 3 - 5$ in each class

⇒ learning curve

. . . to measure the classifier's performance?

⇝ confidence intervals for test results

How many cases do we need?

... **to train a good classifier?**

- rules of thumb
  linear classifier: $\frac{n}{p} \geq 3 - 5$ in each class

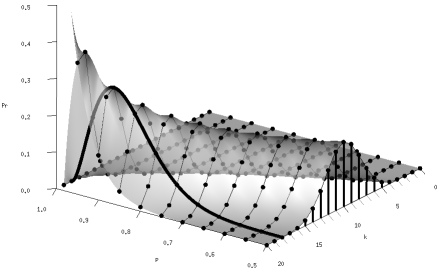
⇒ learning curve

... **to measure the classifier's performance?**

⇝ confidence intervals for test results

- rules of thumb
  100 test cases to estimate a proportion

Confidence Intervals for Proportions



- Statistical description: Bernoulli trial
- Uncertainty on proportion: $var(\hat{p}) = \frac{p(1-p)}{n_{test}}$

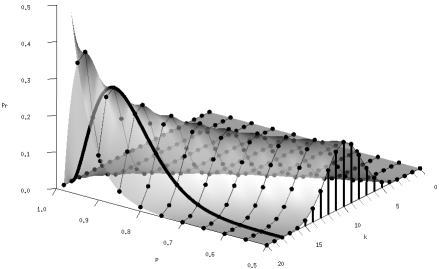
✘ normal approximation appropriate only with $np \geq 5$ and $n(1-p) \geq 5$

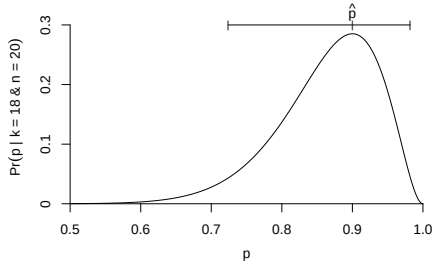
✔ ⇝ use binomial distribution

⇝ Estimate necessary n_{test}

CHEMO METRIX

Confidence Intervals for Proportions





- Statistical description: Bernoulli trial
- Uncertainty on proportion: $var(\hat{p}) = \frac{p(1-p)}{n_{test}}$

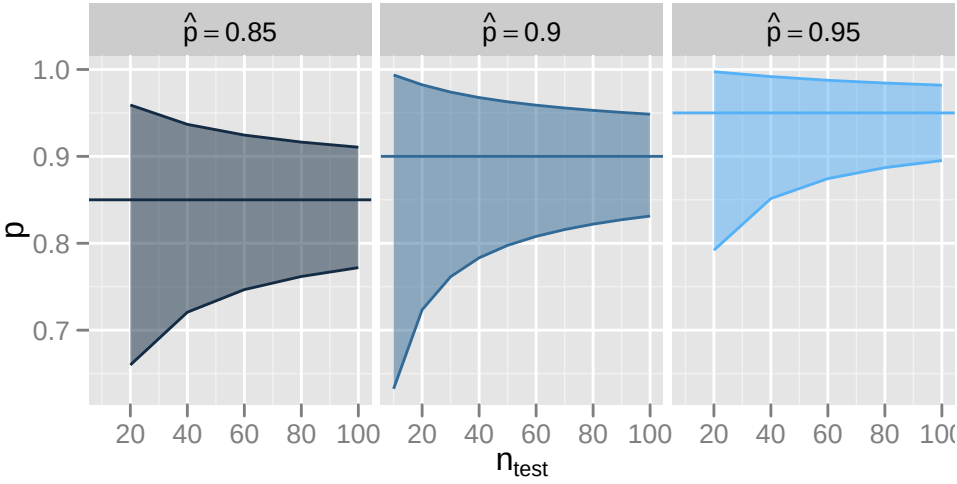
✘ normal approximation appropriate only with $np \geq 5$ and $n(1-p) \geq 5$

✔ ⇝ use binomial distribution

⇝ Estimate necessary n_{test}

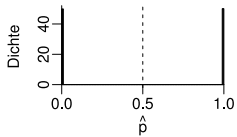
CHEMO METRIX

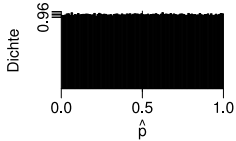
Sample size from Confidence Interval

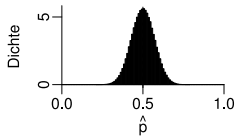


CHEMO METRIX

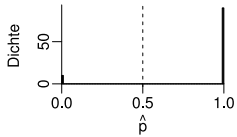
Variance Uncertainty: Brier's Score

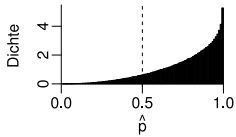


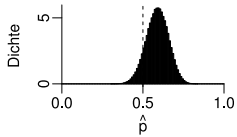




	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{hart})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{hart})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{hart})}$
$Sens^{(hart)}$	0,50 ± 0,050		0,50 ± 0,050		0,50 ± 0,050	
$Sens^{MSE}$	0,50 ± 0,050	1,00	0,67 ± 0,030	0,36	0,75 ± 0,007	0,02



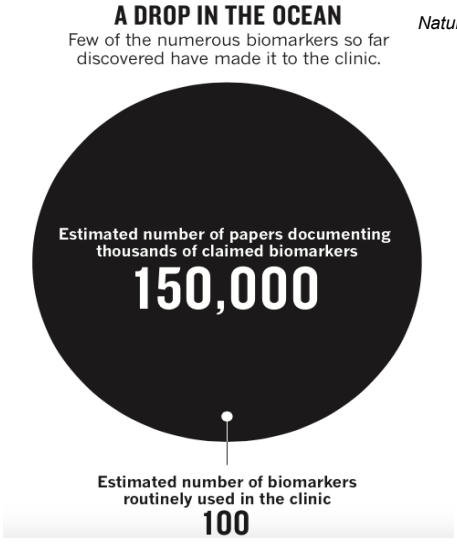




	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{hart})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{hart})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{hart})}$
$Sens^{(hart)}$	0,90 ± 0,030		0,90 ± 0,030		0,90 ± 0,030	
$Sens^{MSE}$	0,90 ± 0,030	1,00	0,92 ± 0,013	0,18	0,83 ± 0,006	0,04

CHEMO METRIX

Reproducibility!?



Nature **469**, 156-157

CHEMO METRIX

Validation: Questions

- Did you ask the right question?
- Or did you use a surrogate?
  - Is that surrogate appropriate?
  - What are the limits?
- Is your classifier set up correctly?
  - Is it really a classification problem?
  - one-class vs. discriminative?
  - open-world vs. closed-world?
- Do you use the correct controls/base class?
- What happens with out-of-spec cases (unknown class? bad measurements?)

Validation: Questions

- Bias introduced by data acquisition procedure?
  - Labeling procedure with self-fulfilling prophecies (e.g. cluster analysis as basis for labeling, semi-supervised label generation)?
- What about borderline cases?
  - Do your *labeled* cases correctly represent them?
  - No exclusion of “difficult” cases in the reference labeling step?
- What other confounders could exist?
- What are the limits of your method?
- reading:
  - **Buchen2011**
  - **Begley2012**
  - **Ioannidis2005**
  - …

Ruggedness: Perturb Data

- How robust are the predictions?
- Which factors (confounders) have most influence?
- Perturb Data
  - Repeated cross validation:
    How do predictions vary if *a few* training cases are exchanged?
    ⇝ stability of predictions
  - Simulate instrument related distortions:
    Measure respective drop in performance

Beleites2008
Sattlecker2010

CHEMO METRIX

Drift

- production use a model is almost always extrapolation in time!
  Training cases were collected *before* the prediction cases

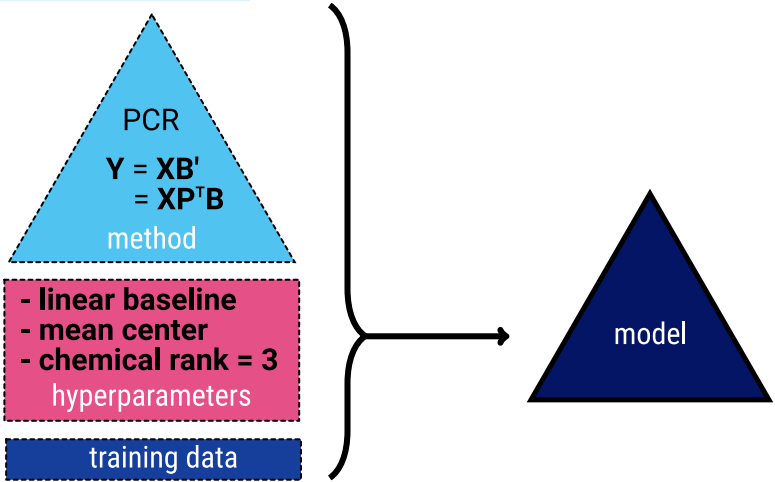
✘ Instrument drift

- Question: low long into the future does the model work well?
  ⇝ “recalibrate daily”

✔ dedicated experiments to check long-term stability

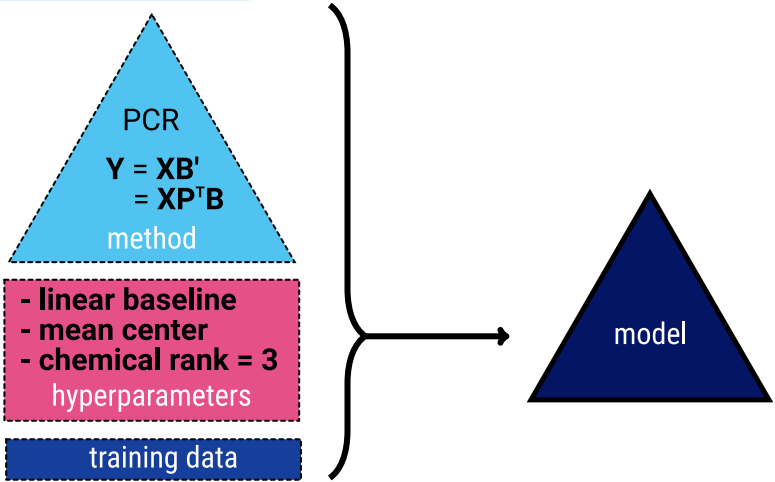
✘ Influence of drift/aging *cannot* be estimated by resampling
  glimpse at drift: look at splits where model is trained on old cases, tested on new cases

Hyperparameters



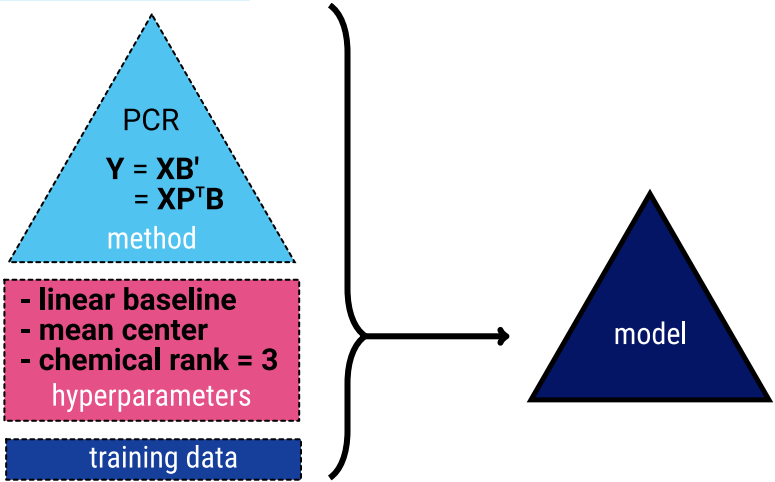
CHEMO METRIX

Hyperparameters



- available: PCR (**$X_{train,preprocessed}$**, `m, center = TRUE`)

Hyperparameters



- available: PCR ($\mathbf{X}_{\mathtt{train,preprocessed}}$, m, center = TRUE)
- wanted: tuned.PCR ($\mathbf{X}_{\mathtt{train}}$)

CHEMO METRIX

Data-driven Model Optimization

training data

Idea:

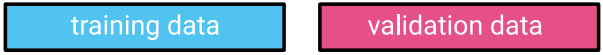
1. sensible range of hyperparameters
2. build covering this search space
3. validate them ⇝ figure of merit (performance)
4. take the best

⇒ **Optimize predictive performance**

✔ Large variety of numerical optimizers available

exhaustive grid search, genetic optimizers, simulated annealing, …

Data-driven Model Optimization



Idea:

1. sensible range of hyperparameters
2. build covering this search space
3. validate them ⇝ figure of merit (performance)
4. take the best

⇒ **Optimize predictive performance**

✔ Large variety of numerical optimizers available

exhaustive grid search, genetic optimizers, simulated annealing, …

CHEMO METRIX

Data-driven Model Optimization

Idea:

1. sensible range of hyperparameters
2. build covering this search space
3. validate them ⇝ figure of merit (performance)
4. take the best

⇒ **Optimize predictive performance**

✔ Large variety of numerical optimizers available
exhaustive grid search, genetic optimizers, simulated annealing, …

✘ Careful: valdiation data enters model building process
⇒ need another independent set to validate the *final* model

Data-driven Model Optimization

Idea:

1. sensible range of hyperparameters
2. build covering this search space
3. validate them ⇝ figure of merit (performance)
4. take the best

⇒ **Optimize predictive performance**

✔ Large variety of numerical optimizers available

exhaustive grid search, genetic optimizers, simulated annealing, …

CHEMO METRIX

Data-driven Model Optimization

- fit normal parameters (coefficients) with *training* set
- fit hyperparameters with validation set
- validate chosen model with test set

Data-driven Model Optimization

- fit normal parameters (coefficients) with *training* set
- fit hyperparameters with validation set
- validate chosen model with test set

Data-driven Model Optimization

- fit normal parameters (coefficients) with *training* set
- fit hyperparameters with ~~validation set~~ optimization set
- validate chosen model with ~~test set~~ final verification set

Data-driven Model Optimization

- fit normal parameters (coefficients) with *training* set
- fit hyperparameters with ~~validation set~~ optimization set
- validate chosen model with ~~test set~~ final verification set

✔ resampling version: nested/double cross validation

Data-driven Model Optimization

- fit normal parameters (coefficients) with *training* set
- fit hyperparameters with ~~validation set~~ optimization set
- validate chosen model with ~~test set~~ final verification set

✔ resampling version: nested/double cross validation

✔ `train (X, hyperparameters)` vs. `tuned.train (X)`
  - tuned training function: additional internal split for tuning

CHEMO METRIX

Data-driven Model Optimization

- fit normal parameters (coefficients) with *training* set
- fit hyperparameters with ~~validation set~~ optimization set
- validate chosen model with ~~test set~~ final verification set

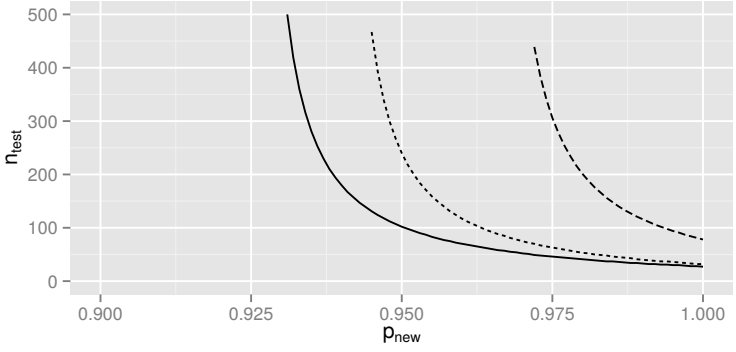
✔ resampling version: nested/double cross validation

✔ `train (X, hyperparameters)` vs. `tuned.train (X)`
  - tuned training function: additional internal split for tuning
  - ✔ treat `tuned.train (X)` like any other training function

CHEMO METRIX

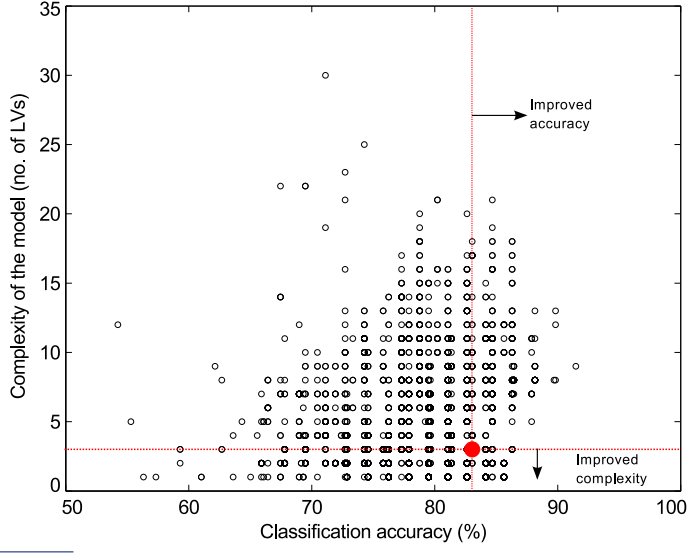
Sample Number Planning: Model Comparison as Statistical Test

		accepted risk		
Type I error	α =	5 %	10 %	5 % wrongly reject non-superiority hypothesis
Type II error	β =	20 %	10 %	5 % wrongly accept non-superiority hypothesis



Assume: old model $p = \frac{19}{25} \approx 75\,\%$

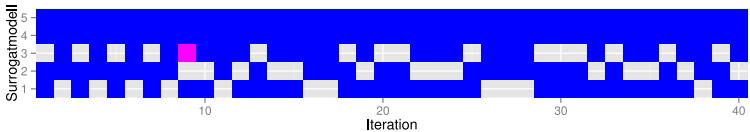
Autotuning: Preprocessing, method selection



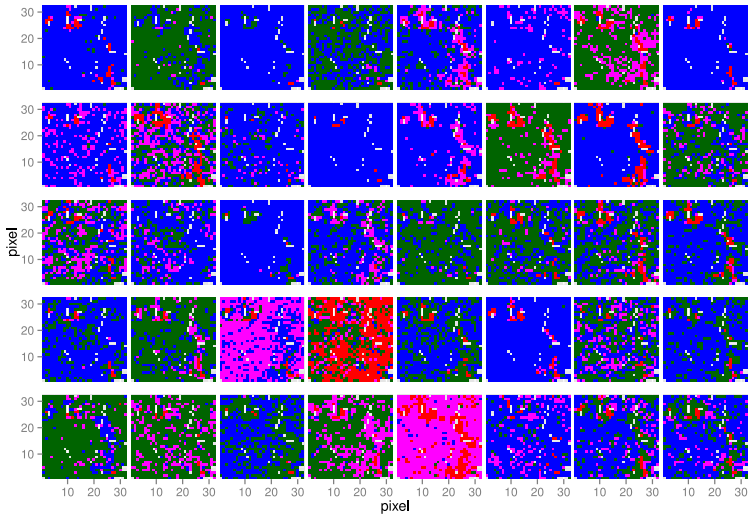
J. Engel et al., Trends in Analytical chemistry 50 (2013) 96 – 106.

CHEMO METRIX

Internal vs. External Performance Estimate

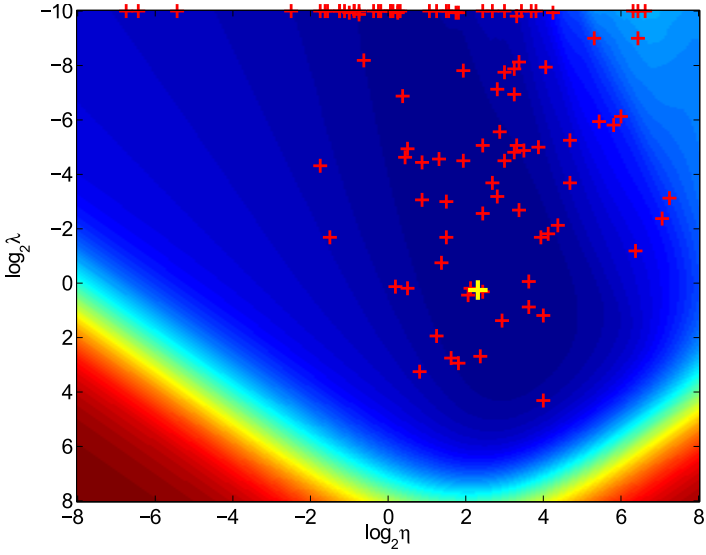


Internal vs. External Performance Estimate



CHEMO METRIX

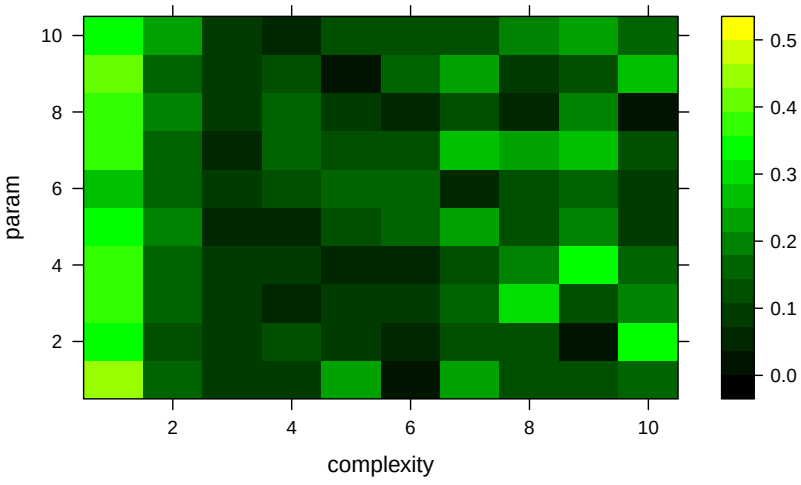
Grid Search



Cawley2010

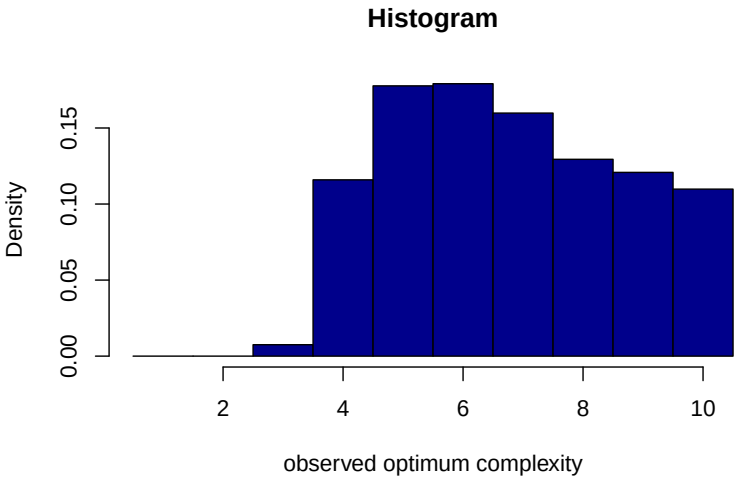
CHEMO METRIX

Grid search



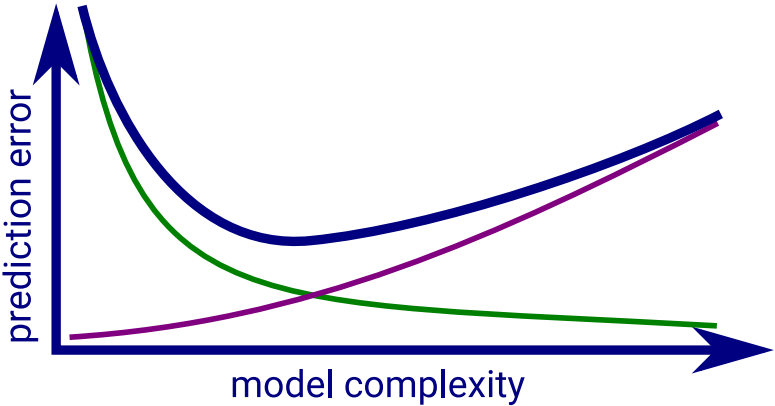
CHEMO METRIX

Grid search: Stability of Solution



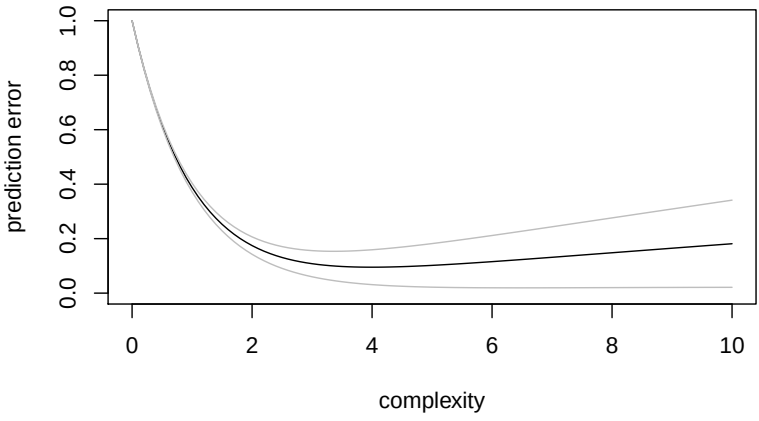
CHEMO METRIX

Grid search: Stability of Solution

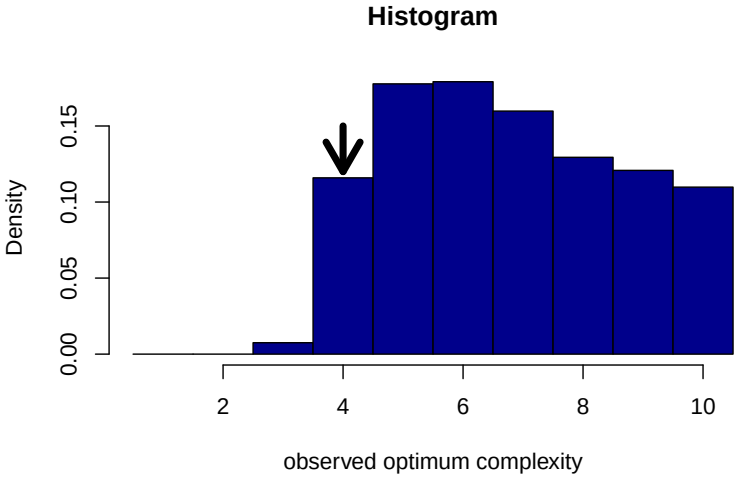


CHEMO METRIX

Grid search: Stability of Solution

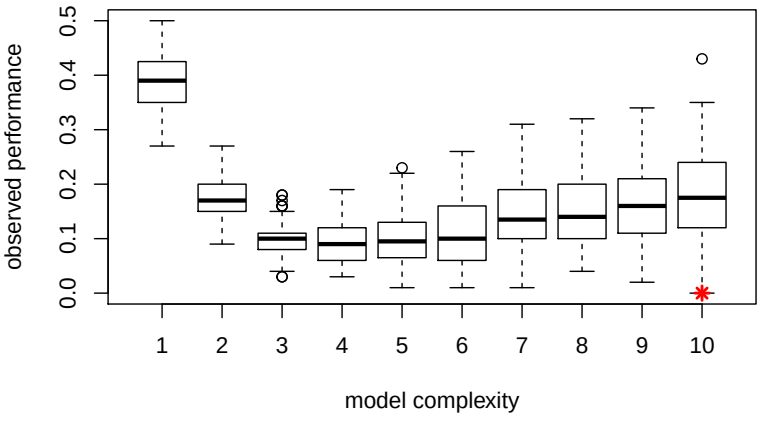


Grid search: Stability of Solution

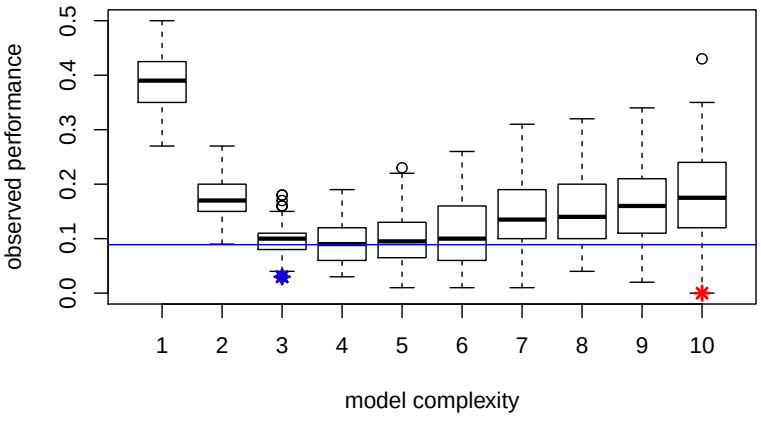


CHEMO METRIX

Grid search: Stability of Solution



Grid search: Stability of Solution



CHEMO METRIX

Summary: Validation

- ✔ Think hard about your data, model, and application!
- ✔ Randomize (and block) your measurements
- ✔ Sample size planning: calculate from required precision of validation results
- At some point, validation studies are needed.
  Before that, use repeated cross validation or out-of-bootstrap.
- ✔ Use DoE (Hasse diagram) to determine independent splitting
- ✔ Check stability of predictions and – if possible – model parameters
- ✔ Aggregation improves *unstable* models
- ✘ Resampling cannot detect drift
- ✘ Hold-out is inefficient and prone to the same errors as resampling!

CHEMO METRIX

Summary: Data-driven model Optimization

- ✔ Needs internal performance estimate plus outer independent validation
- ✘ ⇝ large sample size required
- ✔ wrap optimization in `autotuned.model` function
- ✔ validate output of `autotuned.model` like any other model training function
- ✔ Check stability of optimization
- ✔ Use 1-sd-rule to guard against overfitting
- ✔ Class membership probability predicted: MSE (Brier's Score) has low variance and is proper scoring rule
  ⇝ suitable for optimization