Valid machine learning algorithms for multiparameter methods

Steffen Uhlig, Bertrand Colson, Karina Hettwer, Kirsten Simon (QuoData.de, Dresden)

Carsten Uhlig (Akees.com, Berlin)

Stefan Wittke (University of Applied Sciences Bremerhaven)

Manfred Stoyke, Ulrike Steinacker, Petra Gowik (BVL, Berlin)

Monday 14 May 2018

Eurachem 2018, Dublin

Background (1)

- In the light of recent food fraud cases, the issue of food authenticity is receiving increasing attention. New analytical methods and evaluation approaches are currently being proposed in order to address this issue.
- In this framework, the evaluation of mass spectral profiles constitutes a promising avenue, e.g. for the determination of food origin, but also for the identification and characterization of microbiological parameters.
  - *Analytical methods:* LC-MS, Maldi-TOF, …
  - *Data preprocessing:* peak extraction steps include soft filtering, baseline correction, peak picking …
  - *Evaluation methods:* PCA, logistic regression, random forest, artificial neural networks, and many more …

Background (2)

- Method standardization is crucial in ensuring reliable results in routine food monitoring applications. For this reason, the § 64 LFGB (German food act) working group "Mass spectrometric protein analysis" was founded. The aim of the working group is to incorporate validated liquid chromatography mass spectrometry (LC-MS) methods in the *Official Collection of Methods of Analysis and Sampling* and to develop guidelines for the validation of such methods.

- Currently, a team of bioinformaticians and mathematicians from
  - QuoData Quality and Statistics GmbH, Dresden in cooperation with
  - Akees Data Intelligence GmbH, Berlin,

  is working on evaluation methods and statistical criteria for the validation of such methods.

Aim of validation

- Comparison of evaluation methods
- The question is whether the combination of analytical method, data preprocessing steps, and data evaluation approach across different laboratories will yield reliable classification criteria – making it possible to answer questions such as
  - Is the fish species *Red Snapper* or not?
  - Is the bacterium species *Staphylococcus aureus* or not?
- How can the performance of these classifications be characterized?
- Specifics regarding the analytical method and data preprocessing steps are not discussed in this presentation.

Structure of data

- Structure of (mass spectral) data:

  Let

  $x_{k1} = (x_{k11}, \ldots, x_{k1m})$
  $x_{k2} = (x_{k21}, \ldots, x_{k2m})$
  ...
  $x_{kn} = (x_{kn1}, \ldots, x_{knm})$

  denote m features (signal intensity values, peak) of *n* samples measured in laboratory *k*. The samples are denoted *i = 1, 2, …, n*

- Each of the samples *i = 1, 2, …, n* has been characterized, e.g. with respect to species or origin, i.e. each sample *i* is assigned to a class z_i.

Types of evaluation methods

- There are many statistical and machine learning tools; the selection of a suitable tool depends on the underlying question:
  - *Explorative Analysis:* the aim is to ***find*** new structures and relationships in the data; exploratory tools look for distribution, correlation, clusters and outliers; many tools do not make use of information regarding class membership
  - *Single-class classification:* the aim is to ***identify*** samples of a specific species or origin amongst all other species or origins.
  - *Binary or multiclass classification:* the aim is to ***distinguish*** between two or more species or origins

Evaluation methods

- In this presentation we only consider two machine learning tools:
  - *Principal Component Analysis (PCA)*,
    probably the most popular method for the analysis of multivariate data, and
  - *Artificial Neural Networks (ANN)*,
    which is possibly the method with the greatest potential
- ANN has been applied in connection with the analysis of mass spectra for approximately 30 years, and PCA for nearly 50.

Principal Component Analysis

- PCA can be thought of as fitting an m-dimensional ellipsoid to the m-dimensional mass spectral data, where each axis of the ellipsoid represents a principal component. The size of each axis is proportional to the variance along that axis, and by removing a relatively small axis from the data sets representation, only a relatively small amount of information is lost.

- In many cases, only the two largest axes of the ellipsoid are retained, and the result is an ellipse in a two-dimensional scatter plot.

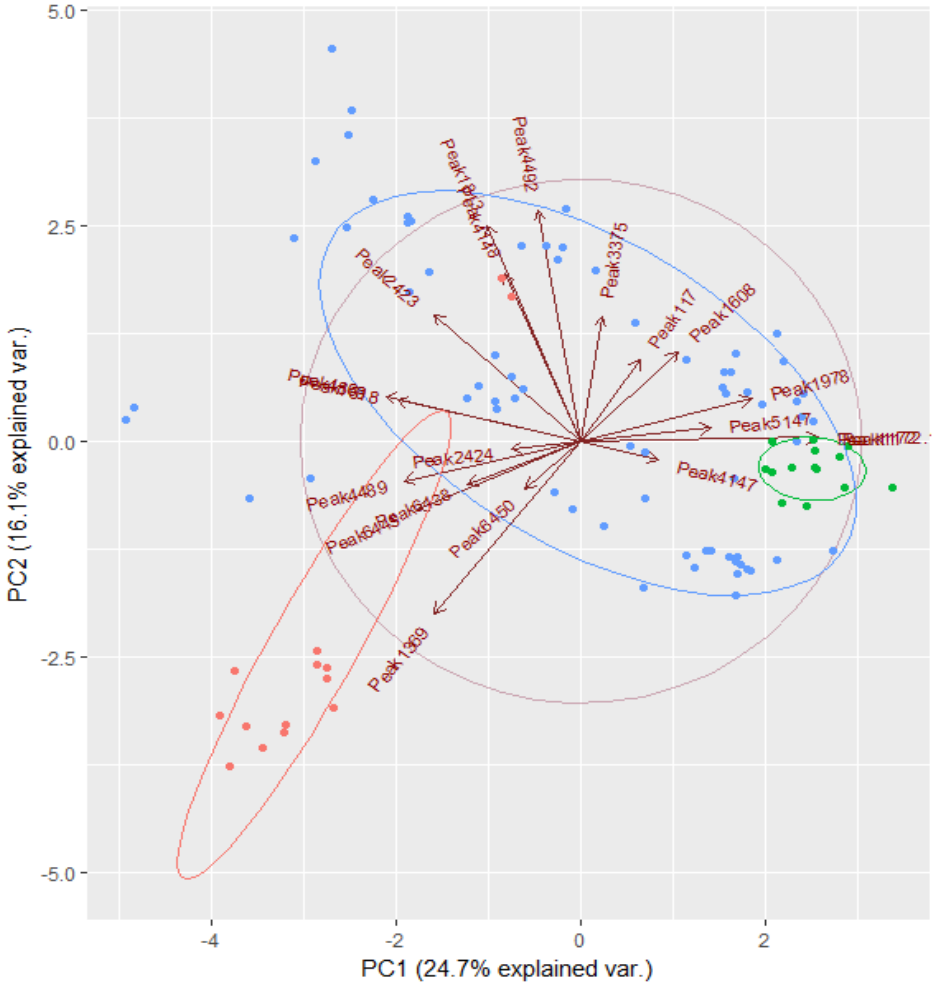
PCA – Example 1

PCA based on 20 features (peaks) selected from 10k intensity values. Two specified fish species, and many unknown (Source: Stefan Wittke).

The chart represents not only the grouping of samples but also the peaks which explain the differences between these groups.

Only 40.7% of the variance of the data is represented in the chart.

Green = Red Snapper, Red = Pangasius, Blue = Miscellaneous

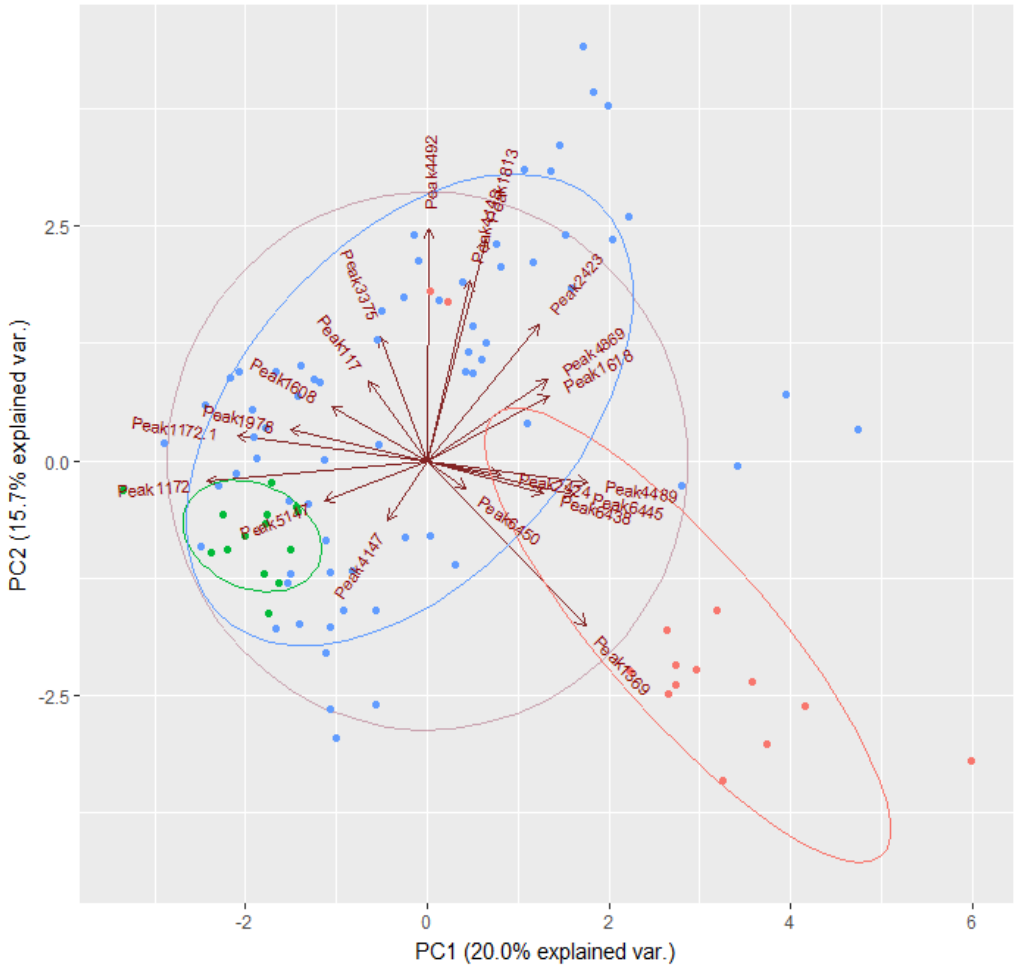


PCA – Example 2

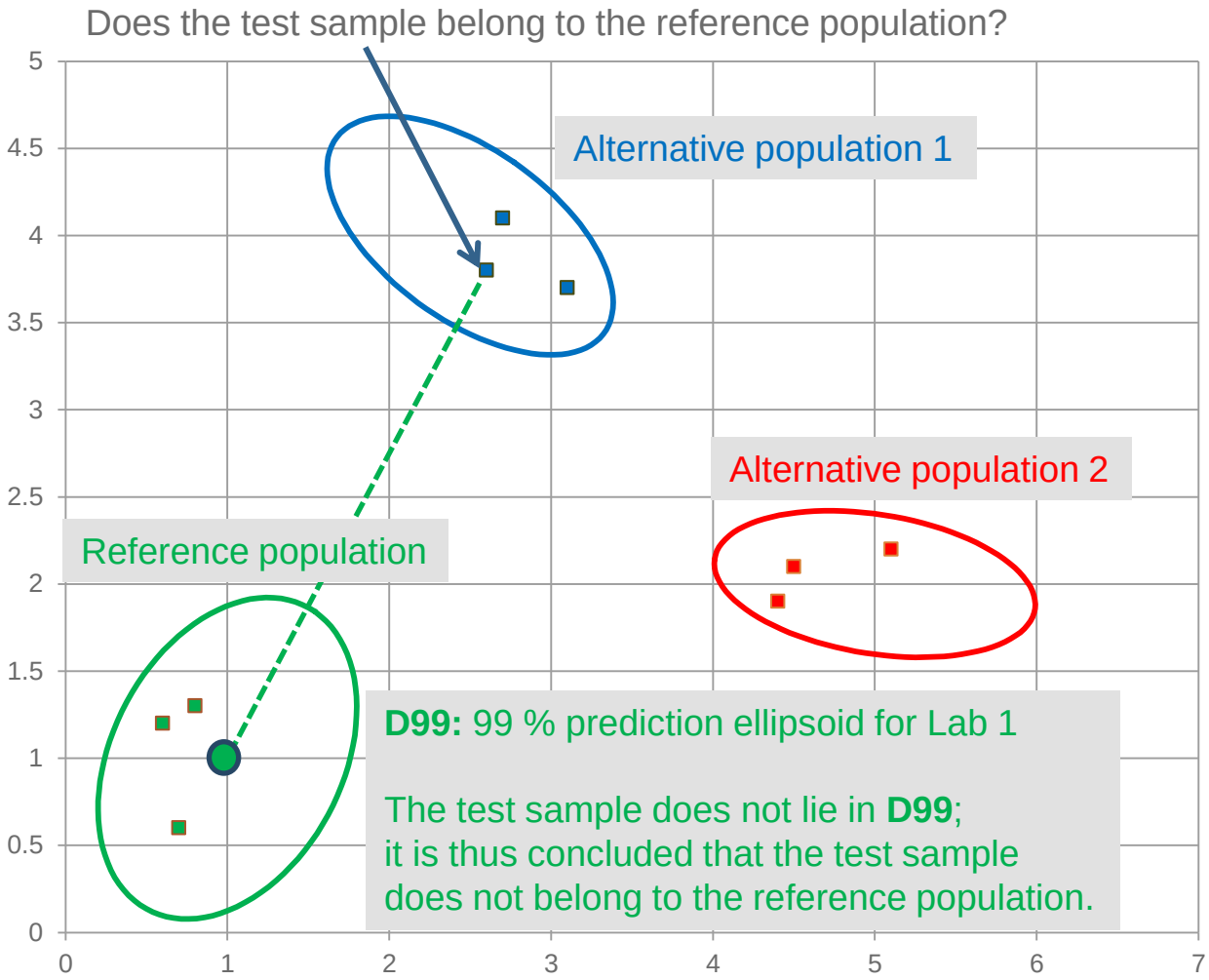
Example 2 is obtained from
Example 1 + 10 % random noise.

Grouping is similar, but it becomes clear that PCA is not very robust against measurement error.

Green = Red Snapper, Red = Pangasius, Blue = Miscellaneous



PCA: Single-class classification rule (single lab)



PCA: Single-class classification rule across labs

- Step 1: Coordinated study in at least four labs.
  - Each laboratory randomly selects at 8-20 samples; in total at least 80 samples
  - PCA analysis is conducted on the basis of the data from all laboratories

- Step 2: Interlaboratory study
  - Samples are sent to at least eight laboratories (in case of factorial design: at least four laboratories)
  - No PCA analysis: the principal components from Step 1 are taken, and the data are projected onto the corresponding 2-dimensional (or higher-dimensional) space
  - The computation of the classification rule is conducted in this two dimensional space
  - Evaluate precision data in this two dimensional space (next slide)

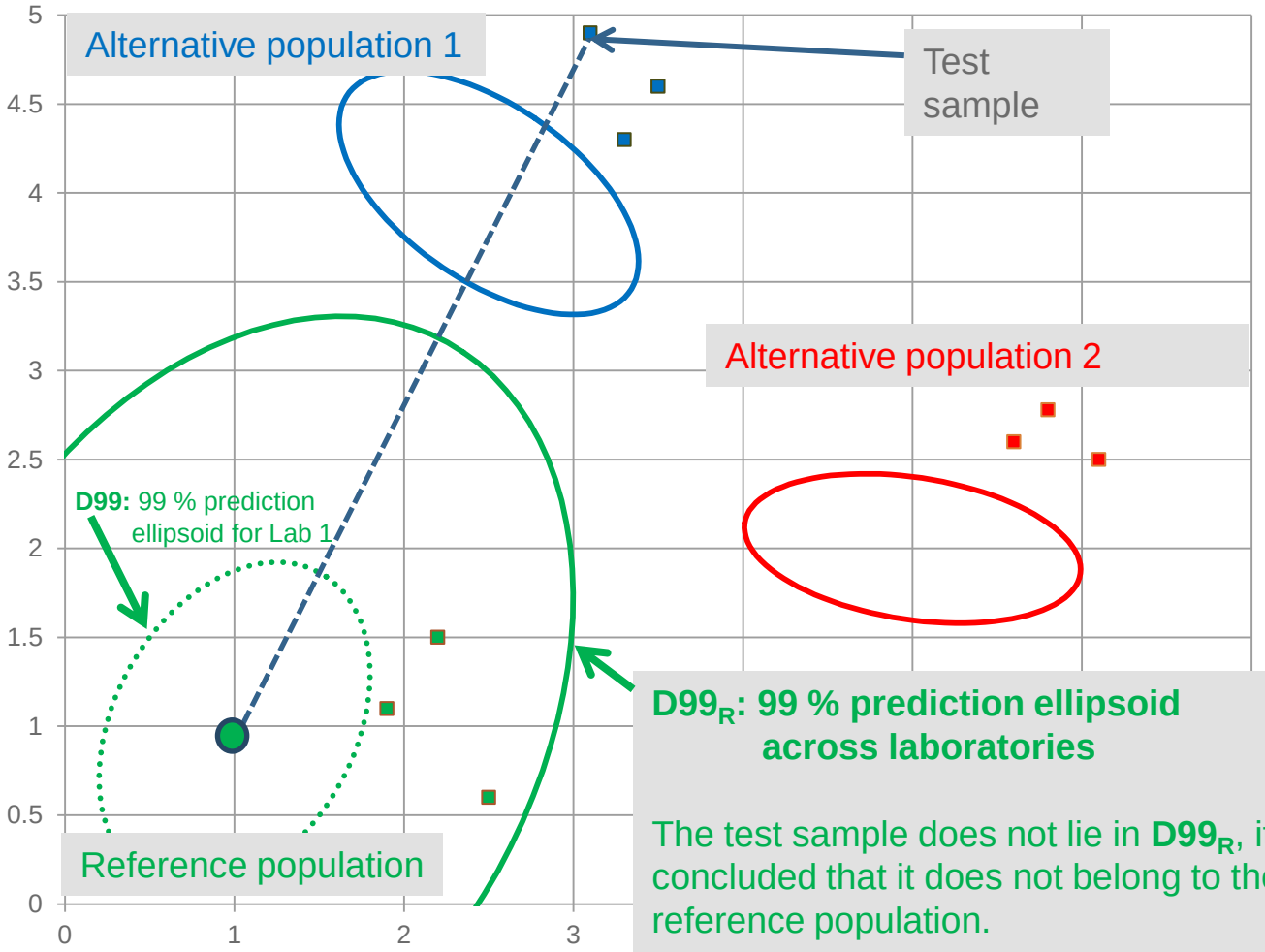
PCA: Single-class classification rule across labs

- Determination of reproducibility prediction ellipsoid $D99_R$ is based on a variance component model with the following error components:
  - Repeatability covariance matrix
    (variation in the 2-dimensional space under repeatability conditions)
  - Reproducibility covariance matrix
    (variation in the 2-dimensional space under reproducibility conditions)
  - “Sample matrix” covariance matrix
    (variation in the 2-dimensional space due to variation between samples belonging to the reference population)

- Formal decision rule:

Test sample not in $D99_R$	⇒	Sample does not belong to the reference population at a significance level of 1 %
Test sample in $D99_R$	⇒	The assumption that the sample belongs to the reference population is not rejected.

PCA: Single-class classification rule across labs



$D99_R$: 99 % prediction ellipsoid across laboratories

The test sample does not lie in **$D99_R$**, it is thus concluded that it does not belong to the reference population.

PCA – Summary and closing remarks

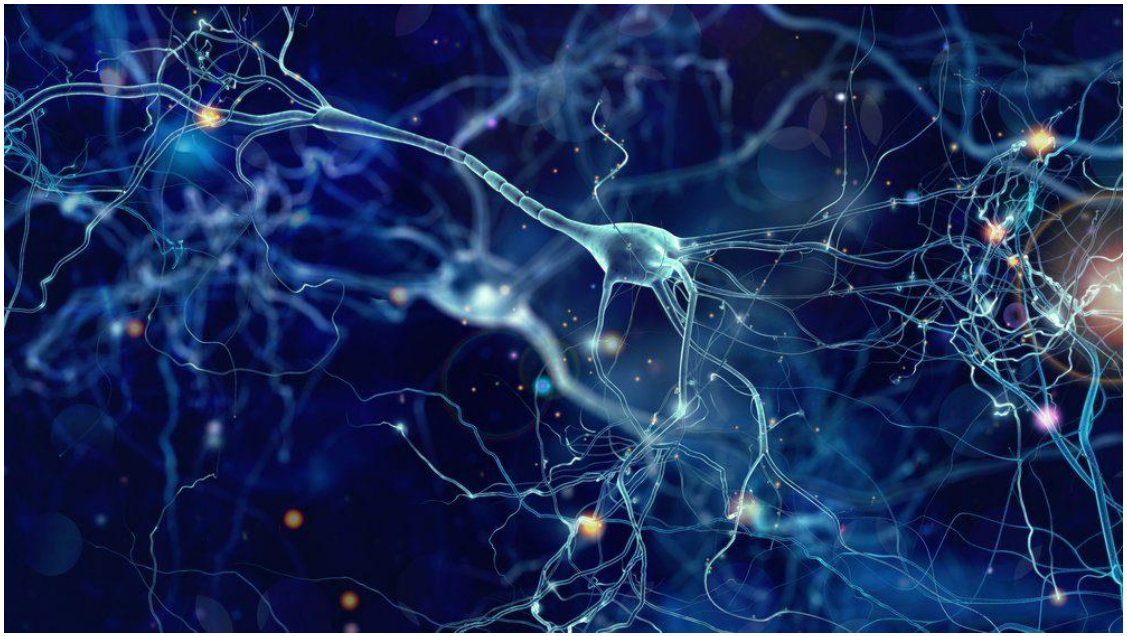
PCA

- Single-class classification across laboratories based on 99% prediction ellipsoid computed in two steps (Step 1 = PCA, Step 2 = Calculation of precision and $D99_R$).
- Is highly sensitive to measurement error, laboratory bias, and to different types of data standardization (Uhlig 2011 [1]).
- The usefulness of the approach has not yet been demonstrated: an interlaboratory study is currently being prepared.

[1]Uhlig S, Eichler S (2011), Are the results of customary methods for analyzing dioxin and dioxin-like compound congener profiles court-proof? Journal of Chromatography A, 1218, pp 5688-5693

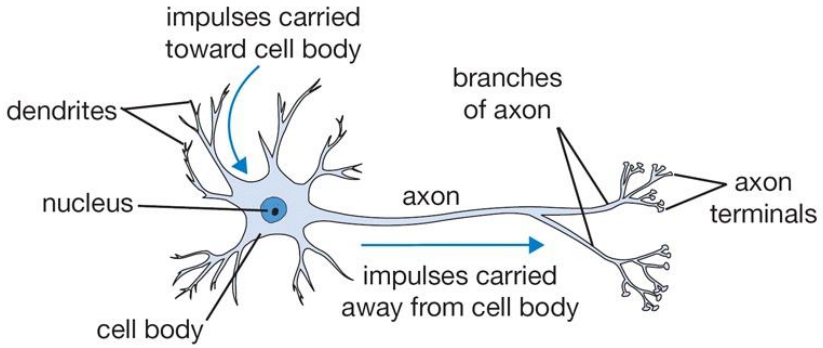
Artificial Neural Networks (ANN)

- Artificial **Neural Networks** (ANN) have been used in mass spectrometry classification for many years now, see e.g. Curry, B (1990), “MSnet: A Neural Network That Classifies Mass Spectra”
- Idea: Simulate the human brain through every single neuron → processing input and output
- Why?: How do we create **knowledge, creativity, cognitive** skills, etc.

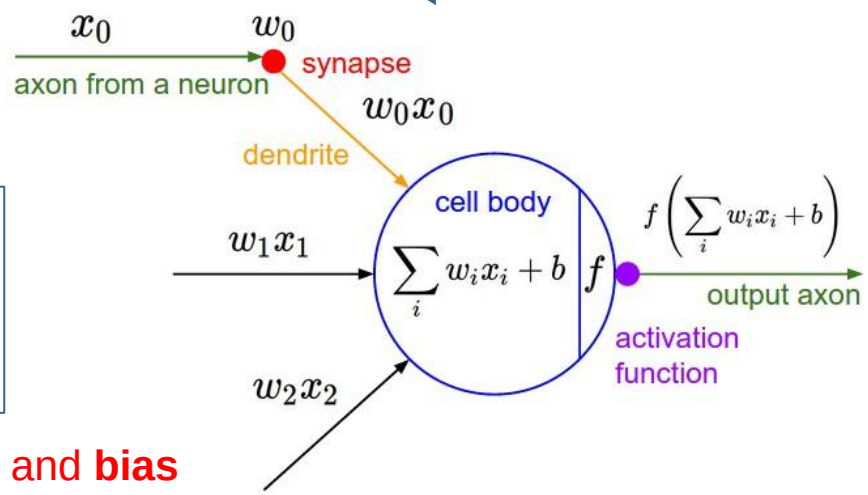


Credit: whitehoune/Shutterstock

ANN Simulation

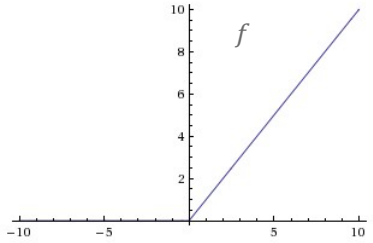


- The **input signal** for a neuron is **processed** alongside other signals from other neurons arriving
- Depending on the **learnt threshold** the neuron fires another signal to the next neuron(s)

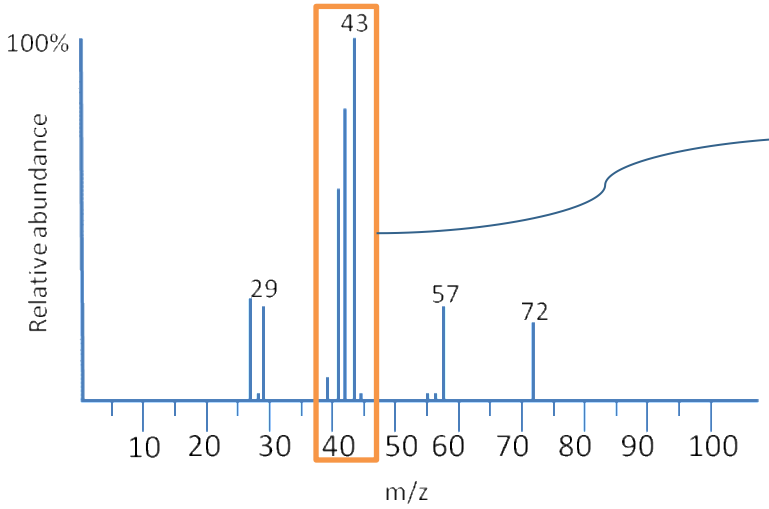


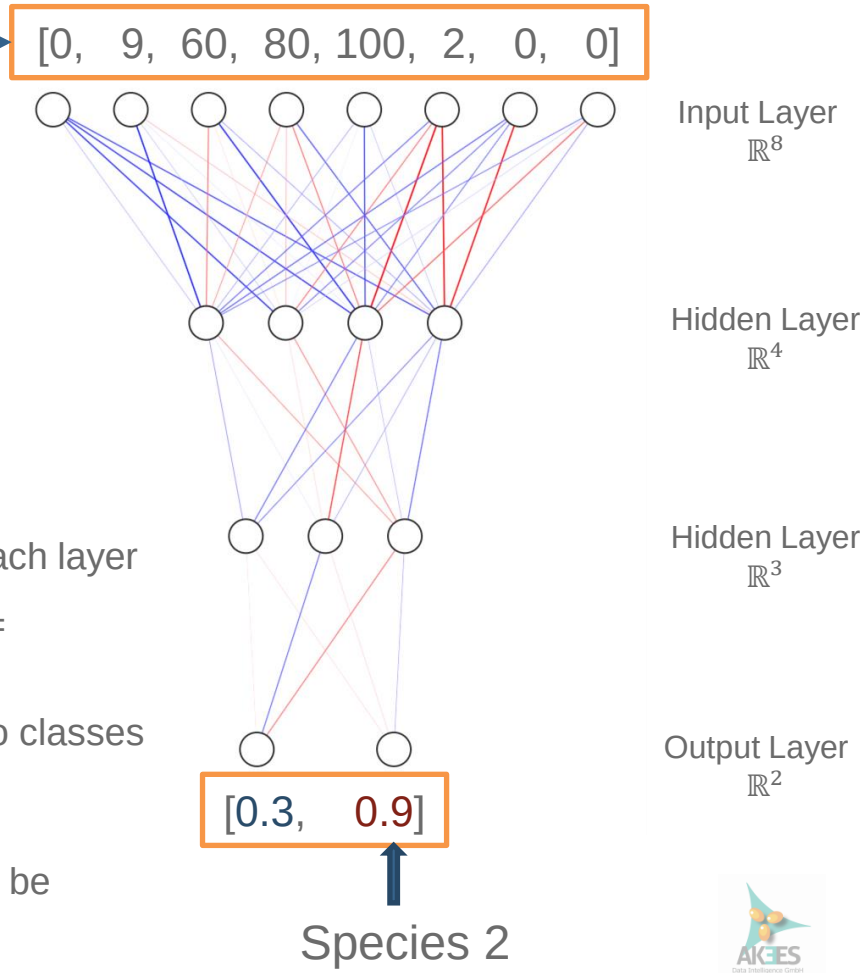
x_i	impulses
w_i	weights
b_i	bias
f	activation function

Determination of **weights** and **bias**
➔ learning or "**training**" of ANN



ANN Structure





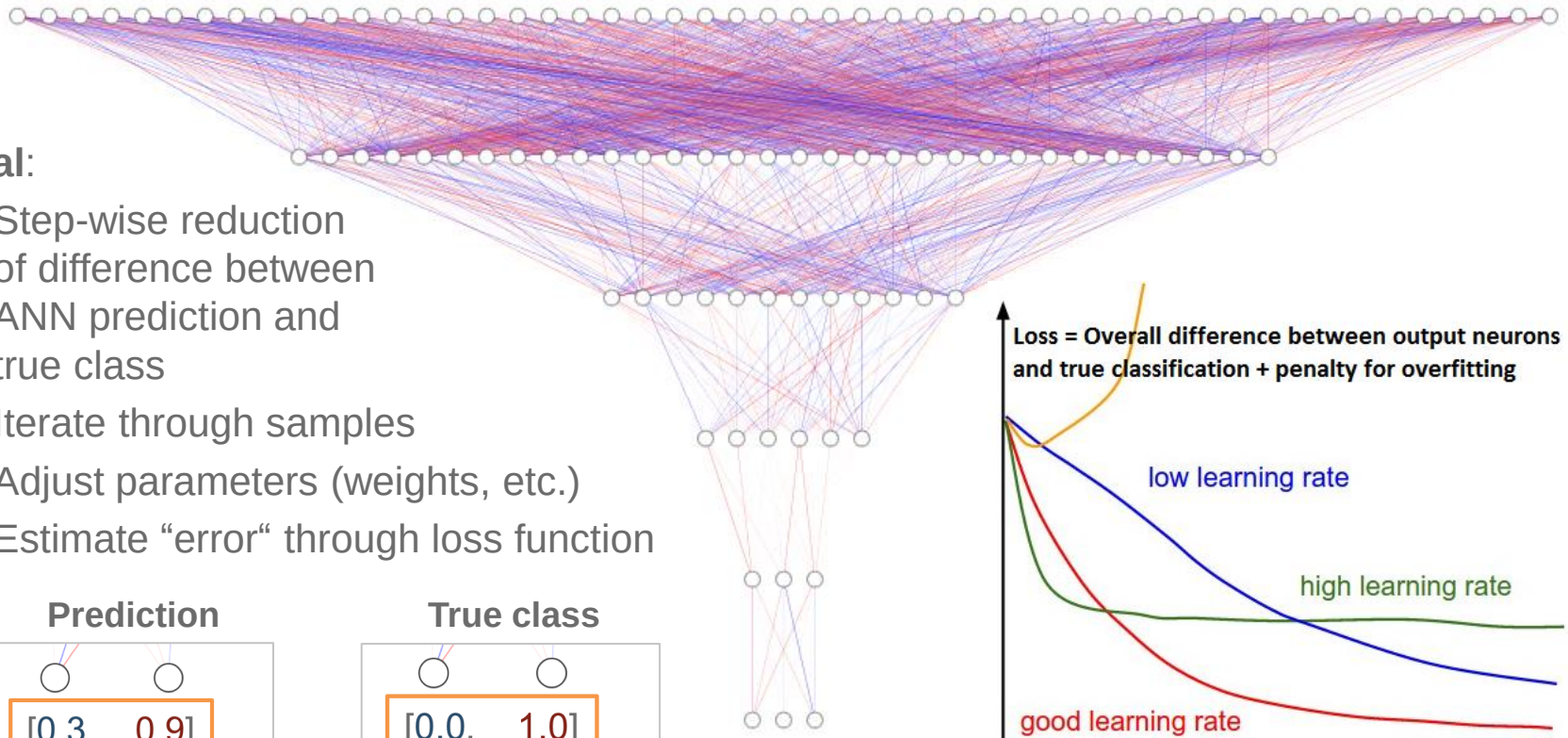
Example ANN Structure

- Two **hidden** layers with 8, 4, 3, 2 neurons for each layer
- Edges represent **weights** (blue = positive, red = negative)
- Output layer: Two neurons, representing the two classes (e.g. species 1, species 2)
- With values between 0 and 1
  Predicted class = maximum of output layer; can be interpreted as the likelihood of being species y

How to make ANN learn

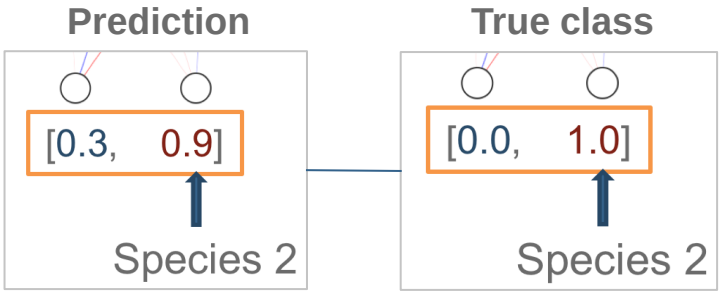
ANN Structure in Reality (this example: 50 input, 3 output classes)

- Input layer: 20 – 20 k Peaks (for mass spec data)
- More than two hidden layers (e.g. 2-20++)

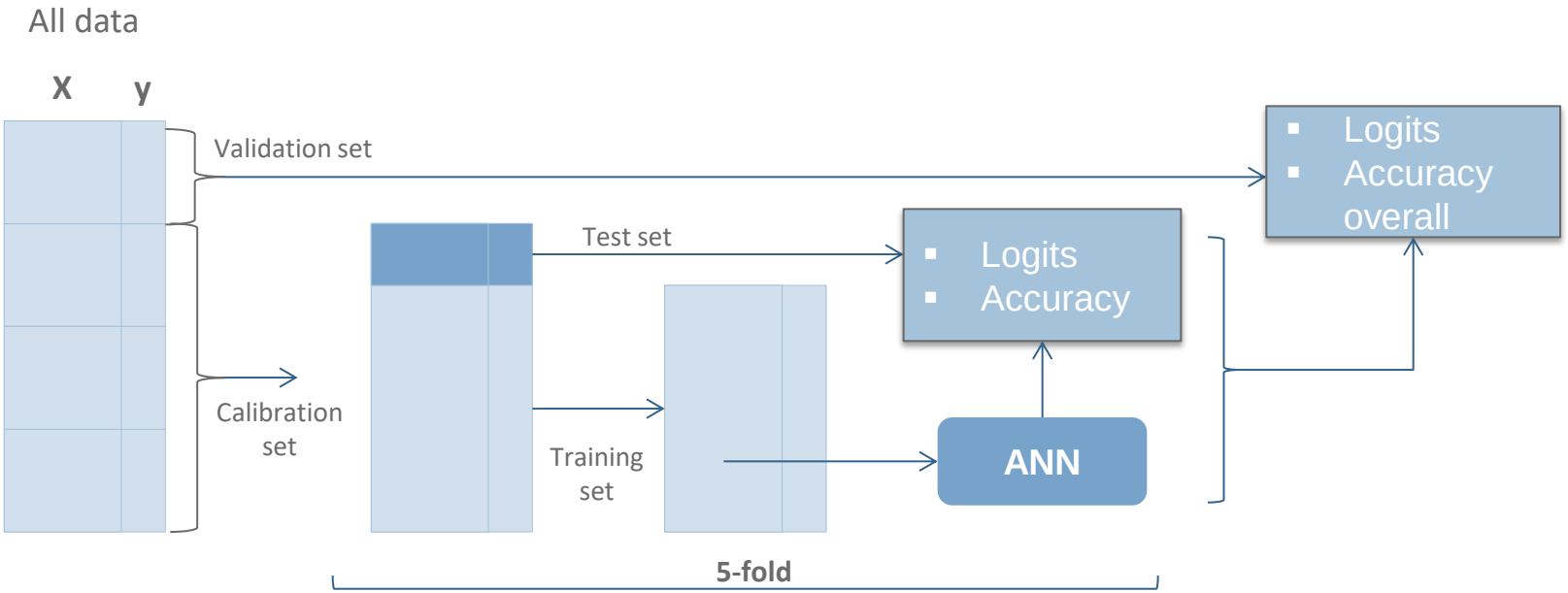


Goal:

- Step-wise reduction of difference between ANN prediction and true class
- Iterate through samples
- Adjust parameters (weights, etc.)
- Estimate “error“ through loss function



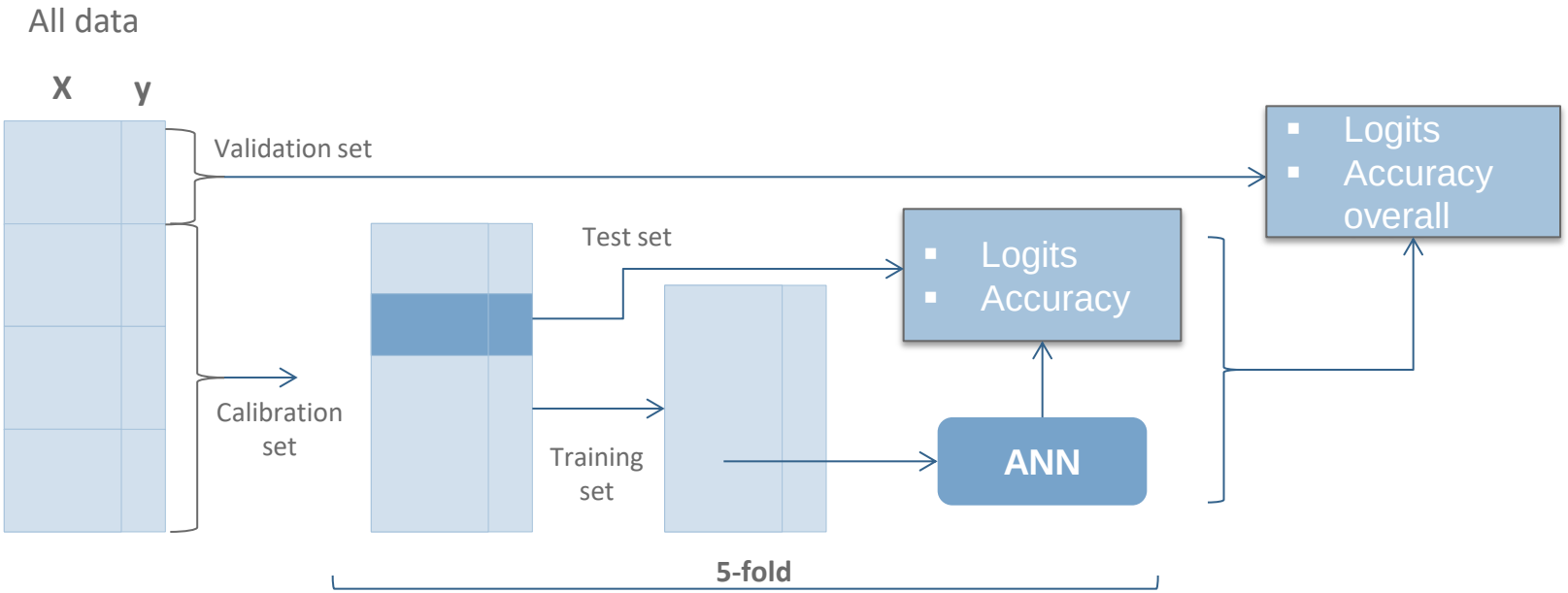
How to make efficient use of sample data
Training and testing



***Inner loop**: develop ANN with training set and validate with test set*

***Outer loop**: use ANNs to predict classes of unseen data*

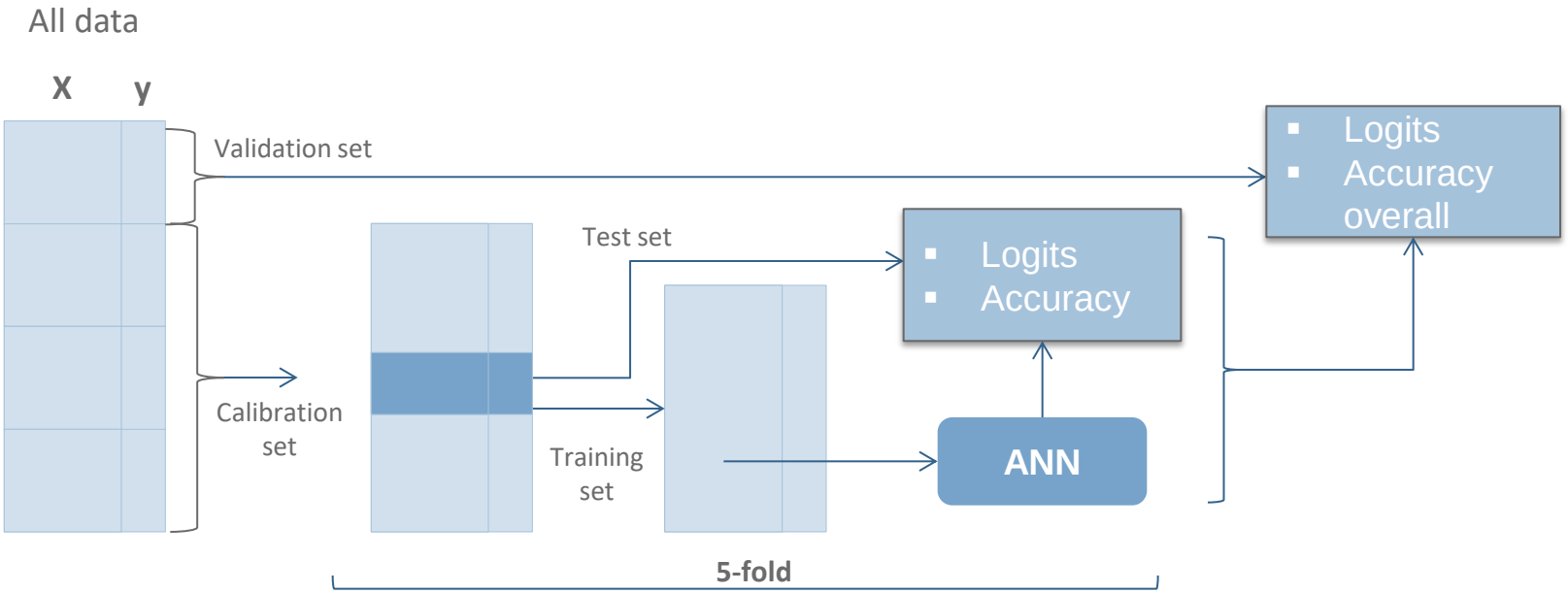
How to make efficient use of sample data
Training and testing



***Inner loop**: develop ANN with training set and validate with test set*

***Outer loop**: use ANNs to predict classes of unseen data*

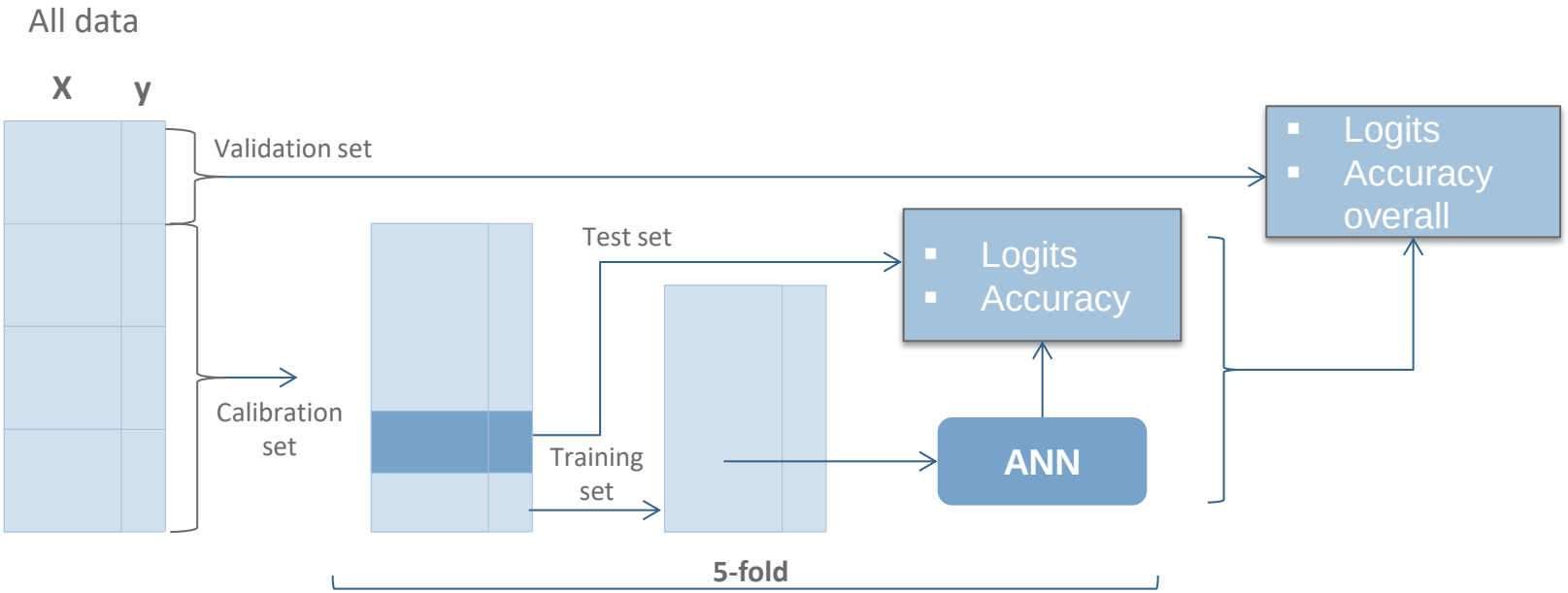
How to make efficient use of sample data
Training and testing



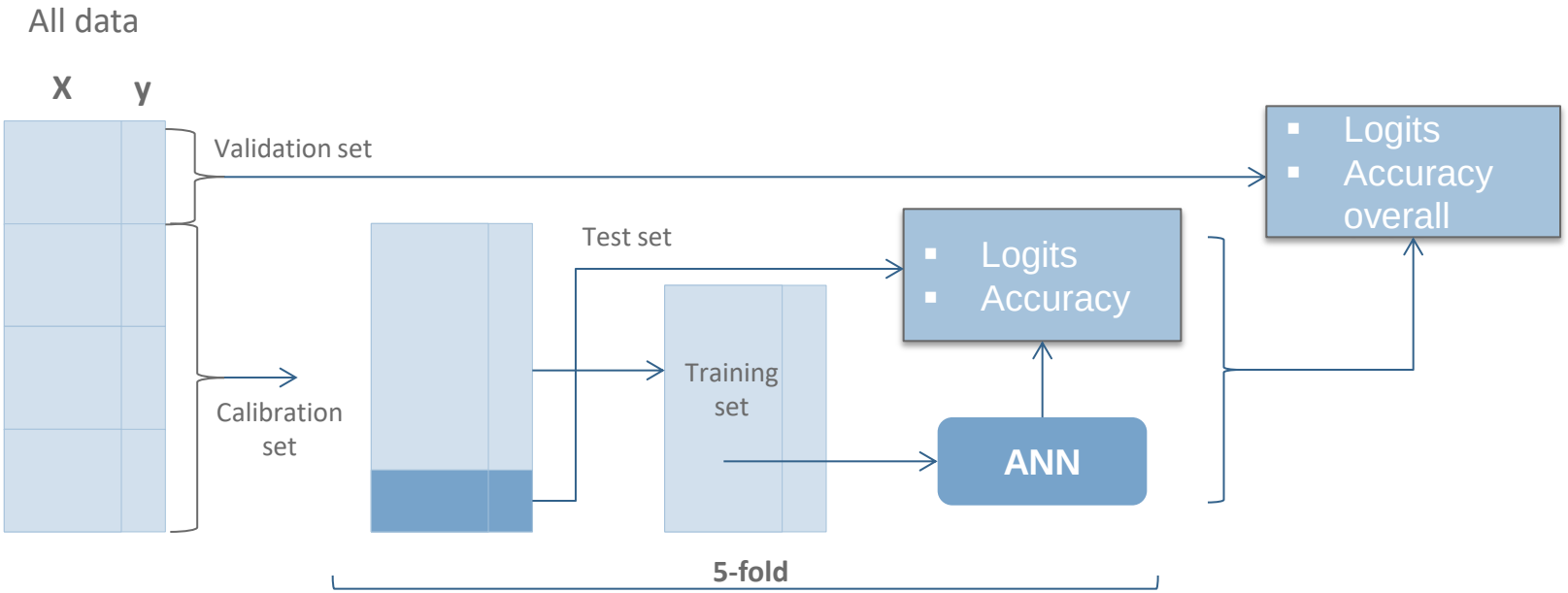
***Inner loop**: develop ANN with training set and validate with test set*

***Outer loop**: use ANNs to predict classes of unseen data*

How to make efficient use of sample data
Training and testing



How to make efficient use of sample data
Training and testing



Datasets

- LC-MS method for identification of fish species:
  Data: 96 samples with 9300 intensities from

  Red Snapper: 14 samples
  Pangasius: 14 samples
  Miscellaneous (8 species): 68 samples
  Results were obtained with Triple TOF 4600 ESI-LC-MS/MSMS ABSciex

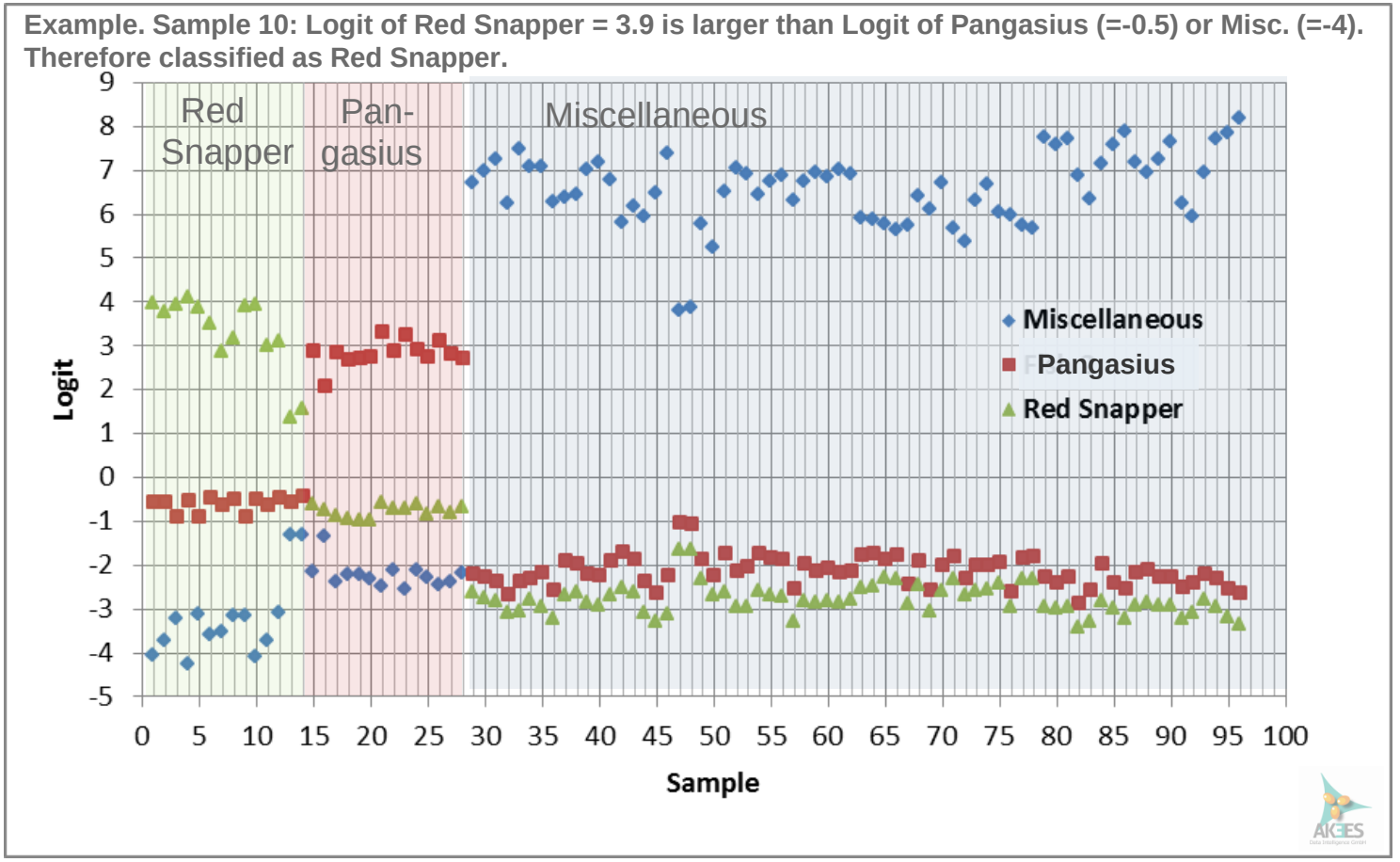
  Source: Stefan Wittke, University of Applied Sciences Bremerhaven

- MALDI-TOF MS method for identification of bacteria:
  Data: 302 samples with 21000 intensities from

  Staphylococcus aureus (110 samples)
  Staphylococcus intermedius (92 samples)
  Miscellaneous (100 samples)
  Results were obtained within one laboratory over a period of +12 months

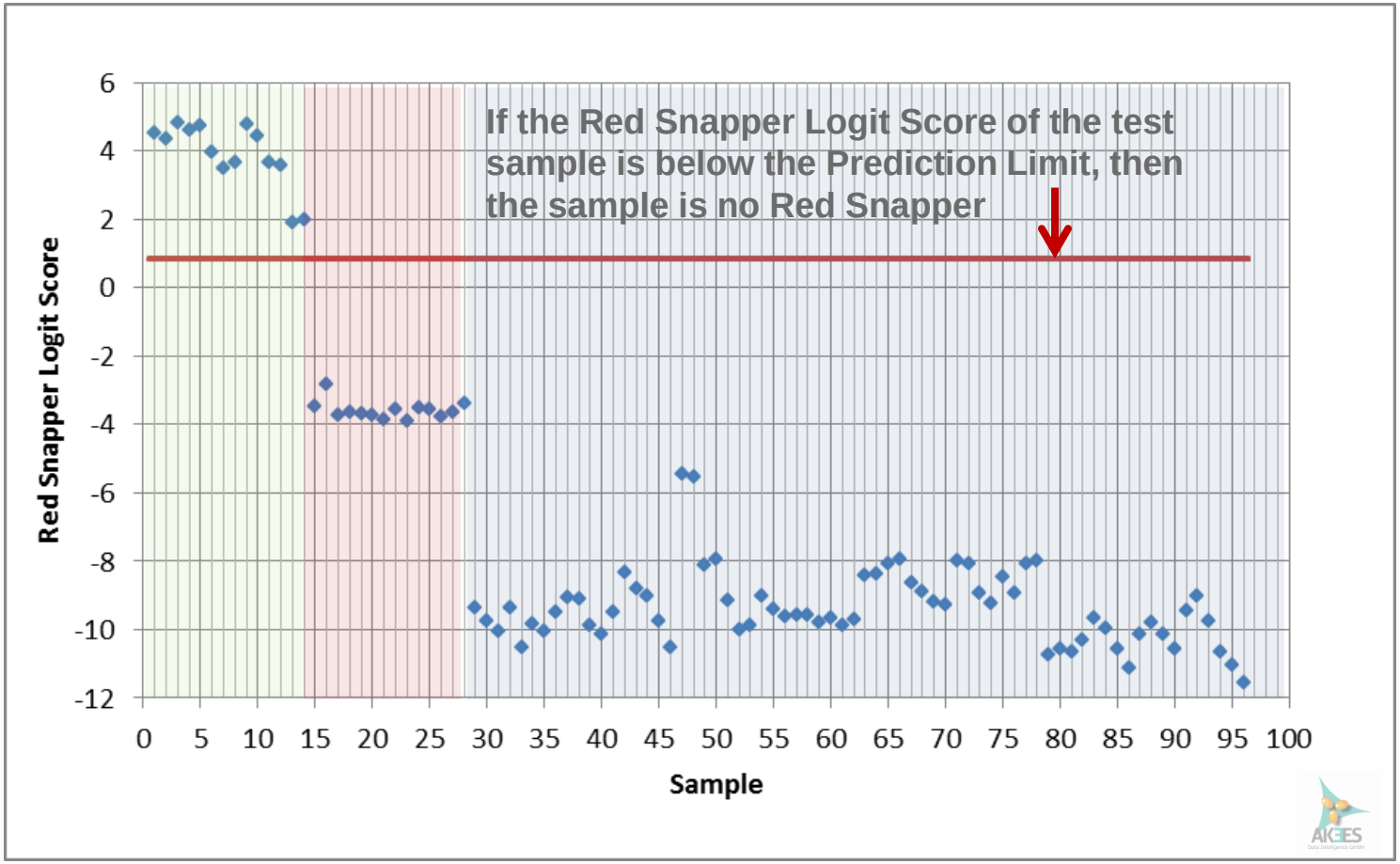
  Source: Ulrike Steinacker, BVL

Analysis of fish data
ANN logits of test samples

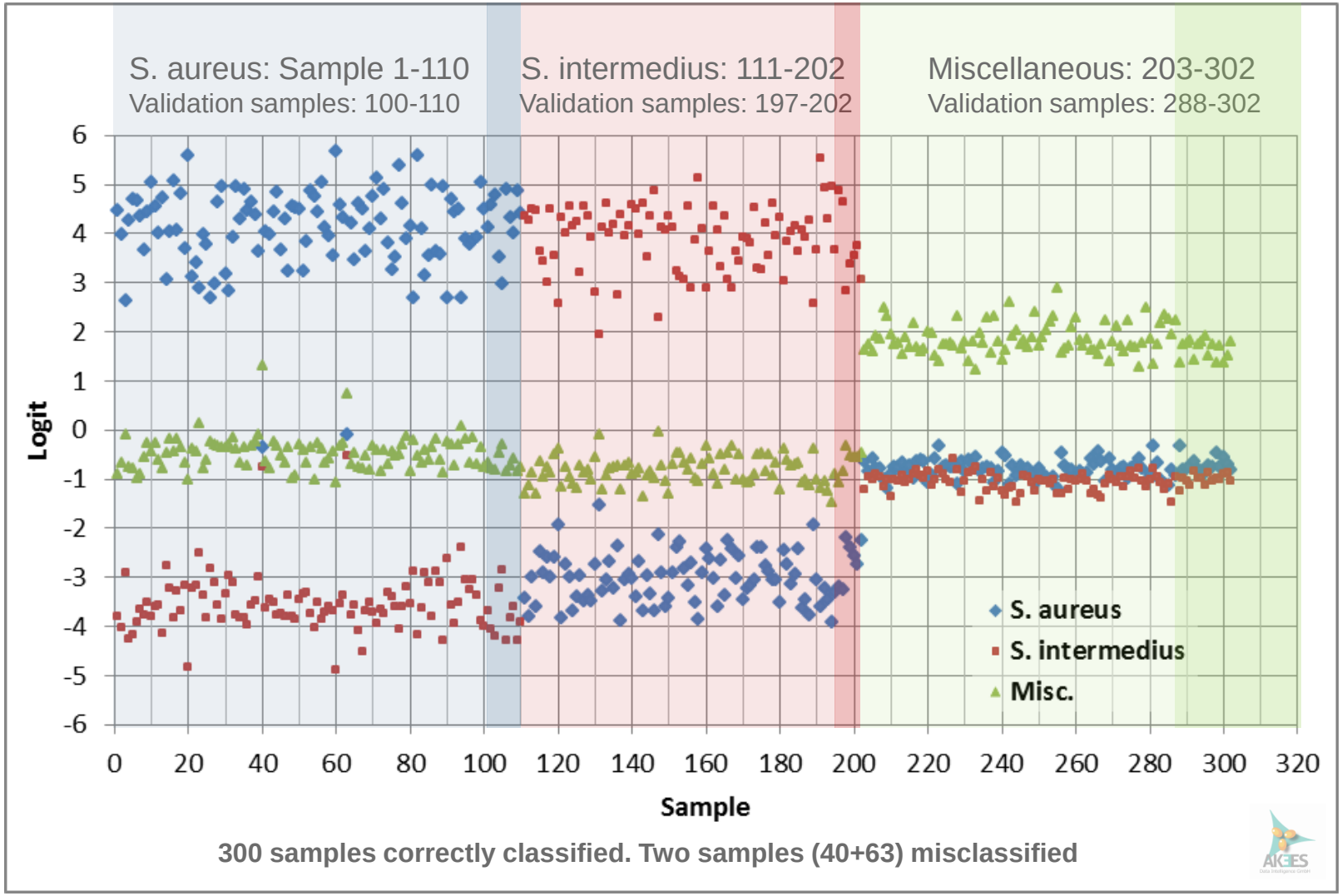


All samples are classified correctly

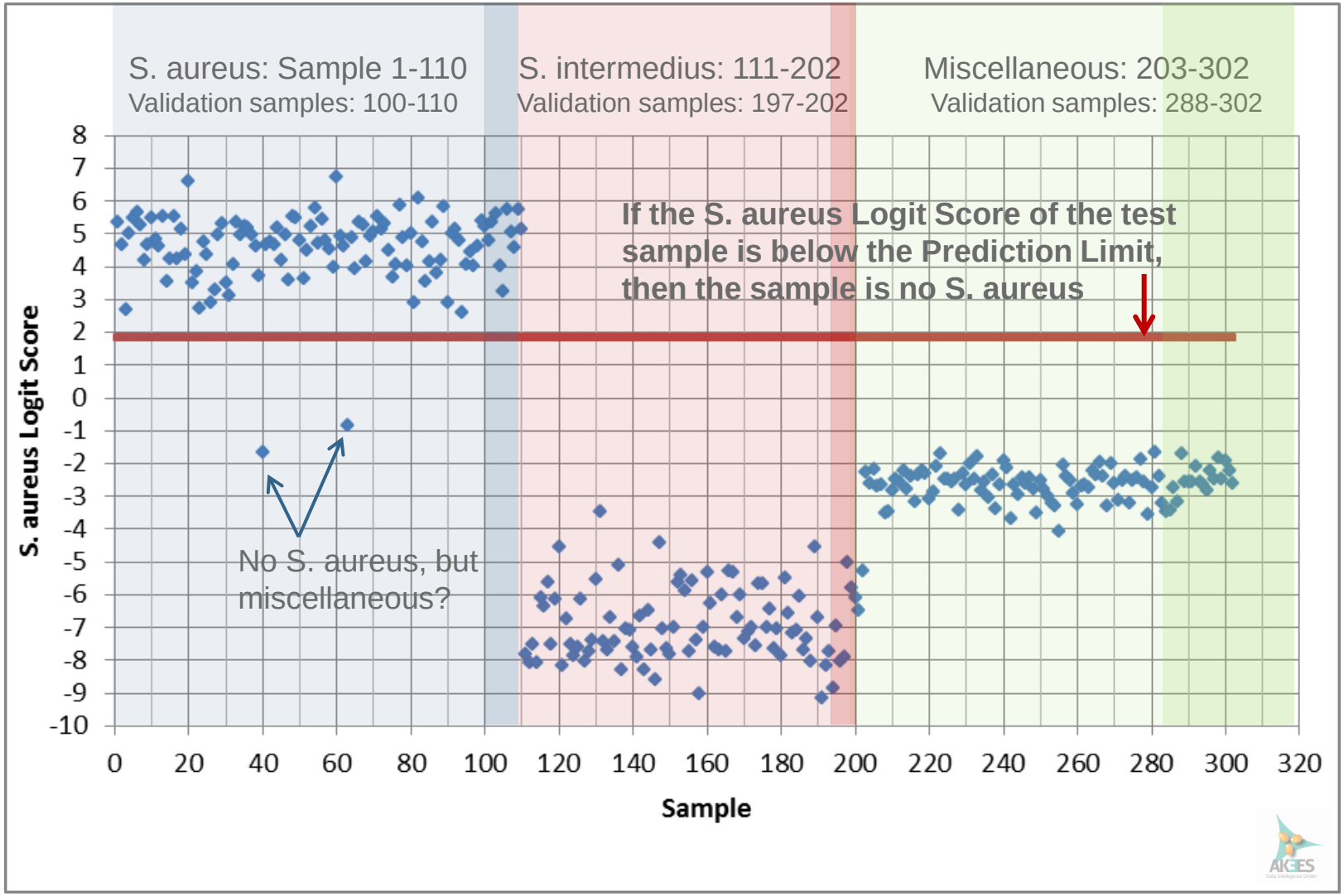
Analysis of fish data
Single-class classification: Red Snapper yes or no?



Analysis of bacteria data
ANN logits of test samples



Analysis of bacteria data
Single-class classification: S. aureus yes or no?



ANN Summary

- Analysis of 31 unseen **validation samples** of the bacteria dataset confirms the validity of the classification procedure
- For the fish data, the number of samples is **too low** for an independent validation
- Perfect classification with only **two misclassifications** in the bacteria dataset, possibly due to error in data (**outliers**)
- The logit scores **across samples** of the same species can be used to **establish** statistically valid classification rules
- Next steps: Conduct **interlaboratory studies** with **different instruments** in order to
  - train the ANN for differences due to different **equipment** and different **data** preprocessing steps
  - **evaluate reproducibility standard deviation across matrices** of the logit scores (standard deviation across samples and across laboratories)
  - analyze **additional samples** from **other species**

Thank you for your attention!

quo data
*QUALITY & STATISTICS!
uhlig@quodata.de

AKEES
Data Intelligence GmbH
carsten.uhlig@akees.com

Federal Office of Consumer Protection and Food Safety
manfred.stoyke@bvl.bund.de

Hochschule Bremerhaven
swittke@hs-bremerhaven.de