# 3 - Data handling

- Referencing
- Machine learning and AI

Referencing

The referencing of spectra is one of the most important steps in the pre-processing of spectra, as the aim is to measure precisely and correctly. Once spectra have been recorded in insufficient quality or incorrectly referenced, improvement is not possible even with mathematical methods. The results of chemometric and machine learning methods also become invalid if gross errors have already occurred in the spectra recording.
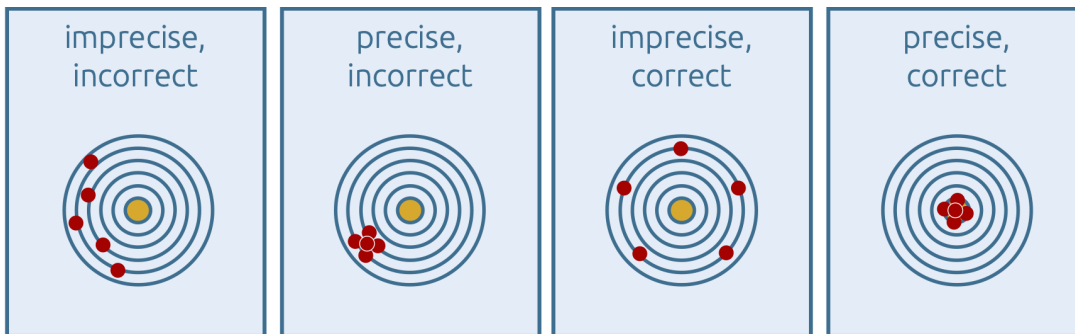
Precision and correctness

All electronic detectors have a noise that is thermally or electronically induced. This noise is often referred to in the respective software as 'background', 'dark image', 'detector reference' or 'dark reference'. This noise is subject to a drift that depends on the (mains) voltage and the ambient or detector temperature. If this background correction does not take place, the measurements will be imprecise. In some cases, the background of the device, i.e. the characteristics of the radiation source and all optical components, are also included in the background measurement.

In other cases, the measurement of the device background can also be found separately in the measurement of the 'white reference' or simply 'reference measurement'. For all reflection methods, it is advisable to measure a reflection standard with known reflection (as a rule, standards with consistently high, practically complete (100 %) reflection over the entire spectral range are used). In practice, gold, aluminum, silicon or PTFE standards are used for this purpose. There are certified white standards whose absorptions/reflections are precisely known. For transmission measurements, the substrate or solvent can be removed directly from the sample spectrum in this way. If the measurement of the device background or reference is omitted, incorrect spectra are obtained which are not comparable. In general, referencing can be calculated in the spectra data as follows (I Intensity):

$$I_\lambda = \frac{(I_{\lambda,\ measured} - I_{\lambda,\ dark\ reference})}{I_{\lambda,\ white\ reference}}$$

Wavelength referencing should also be carried out from time to time. Certified wavelength standards based on rare earths, for example, are available for this purpose, but materials with a known spectrum (stored in a database or literature) can also be used as a wavelength reference. Incorrect wavelength assignment also leads to incorrect spectra. For MCT detectors, spectral

interpolation of the faulty pixels of the detector is also necessary.

Machine learning and AI

The multivariate methods of chemometrics - often also referred to as data mining, machine learning or deep learning/AI - offer the possibility of making superimposed and hidden information visible in the spectra. When using machine learning or AI methods, precise knowledge of the input data is a fundamental prerequisite. The methods of exploratory data analysis are suitable for structuring the input data (value range, distribution and characteristic values of features) and reducing the amount of data.

Machine learning methods make it possible to differentiate between spectra which, from a visual point of view, show almost no differences. With the help of algorithms and mathematical statistics, existing differences can be clarified and data that do not contribute to the distinction can be separated. What remains is the information with the maximum variance or distinction. The individual methods will be discussed elsewhere in this guide.