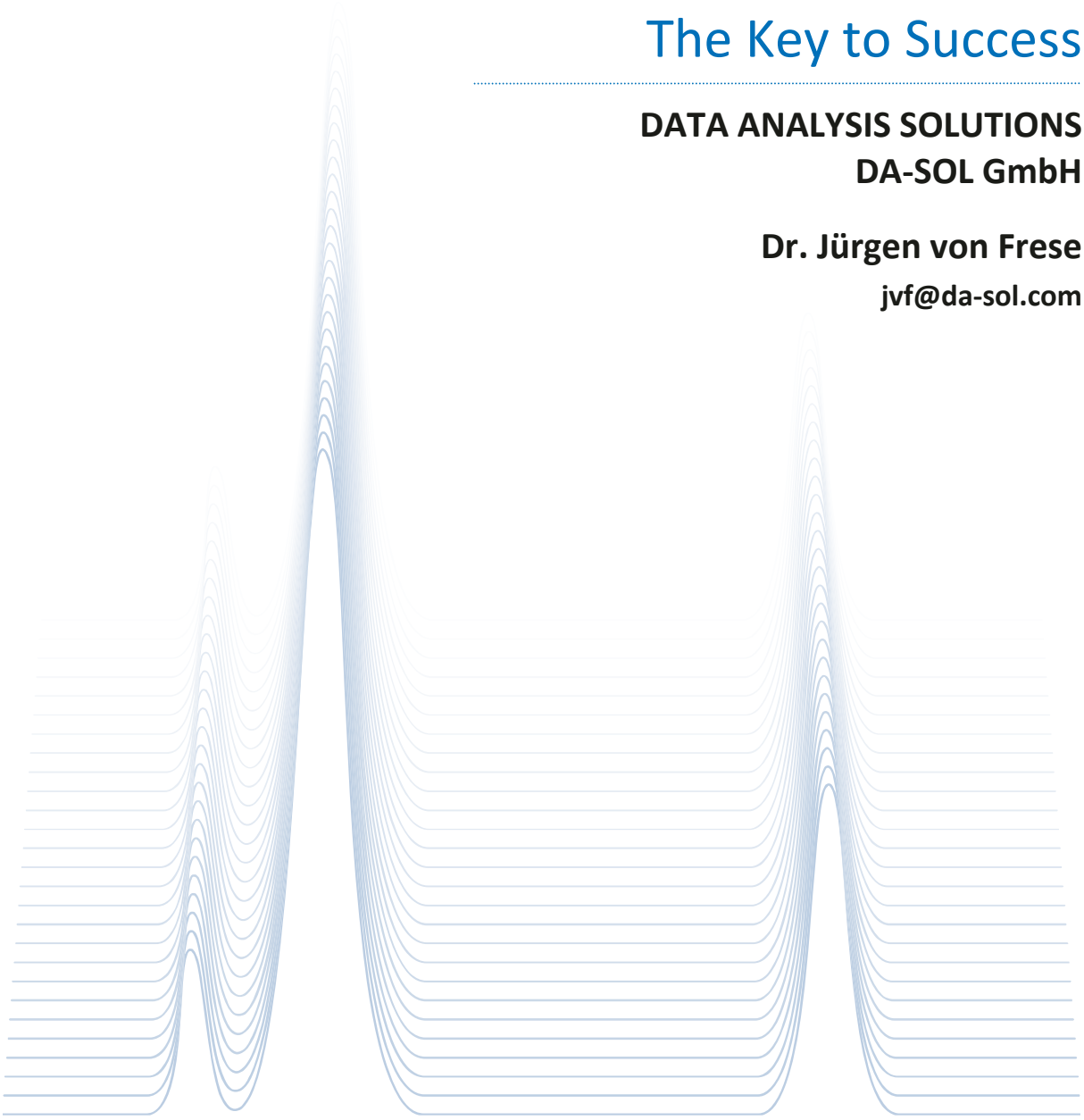
The Key to Success

DATA ANALYSIS SOLUTIONS
DA-SOL GmbH

Dr. Jürgen von Frese
jvf@da-sol.com



THE KEY TO SUCCESS

"A chain is no stronger than its weakest link"

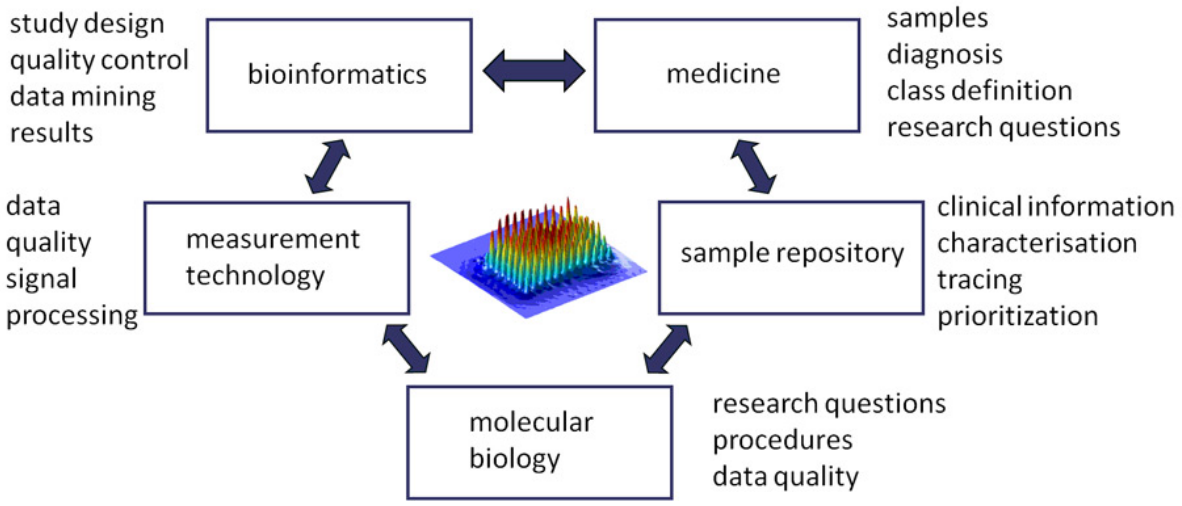
With today's complex measurement technology producing thousands or more measurement values for each sample, data analysis has often become the most critical component for harvesting the true value of these costly efforts.
Traditionally, data analysis (i.e. bioinformatics, chemometrics, pattern recognition etc.) is seen as a very final effort, something to be considered when everything else is finished and complete data is available. Naturally this is also largely the academic viewpoint where methodological research often starts based on publicly available data from some repository. But this is exactly where the large gap arises between hundreds of promising publications and the far fewer examples of successful implementations in practice.

One can consider data analysis in close analogy to photography: Performing the analysis means providing a picture of the data. But producing a good photo consists of so much more than just pressing the shutter release. For a good photo an interesting, appealing motif, proper lighting conditions and a good viewpoint are at least as important.
A good photographer will spend much more time (and resources) on that than on finally taking the picture. In fact, he will not "take a photo" - he will "make it". In contrast, with an uninteresting motif, poor lighting conditions and a bad viewpoint, even the best photographer and equipment will be of no help.

Data analysis is always part of a whole project chain - where it is generally true that:

"The chain is no stronger than its weakest link."

That is, the least performing step will critically determine the possible overall achievement. It is of no use to put in a lot of effort and apply sophisticated algorithms at a particular sub-step, when other crucial aspects are neglected and thus the actual performance is lost elsewhere. Therefore, data analysis considerations cannot wait until everything is "said and done".



Typical process chain in a biomedical or pharmaceutical research project, e.g. using gene expression, proteomics or metabolomics. Note that the bioinformatics/medicine interaction plays an important role both at the start (study design) and in the end (analysis).

Some of the most crucial contributions of "data analysis" from start to finish are:

- Study design is about "asking questions". Without the right questions, poor or misleading answers will be obtained. - In order to extract information from data, it has to be in there in the first place. Study design is where this is critically determined. Likewise, study design really means drafting any future analysis. - Making things worse - study design is usually an irreversible step, i.e. errors and omissions cannot be corrected easily after the measurements have been started.

- Varying data quality, batch effects, drift and gross outliers are often an unavoidable part of real-life measurements. Randomizing measurements, establishment of controls and standards, tailored replication and effective data quality control are all critical contributions to the generation of good data in the laboratory in the first place. Likewise, with today's complex data it is often impossible for the colleagues in the laboratory to evaluate the quality of their own data or to detect batch effects etc.

- We apply technologies which are much more powerful than anything used in the past. With tools which are able to measure whole genome expression, thousands of proteins or metabolites, we might be in for some surprises. It is often the case that these technologies reflect more biological complexity than we have anticipated or called for. Thus, it is of outmost importance to perform exploratory analyses in order to understand which additional information is expressed in the data and how it might influence the original aims of the study.

Thus, integrating data analysis from the very earliest project stages is not just one of the most critical measures for project success but also one of the most cost effective decisions for dramatically improving result quality. Using powerful algorithms is certainly part of the trade, but tackling the various project steps properly and taking an active, intervening role certainly has a much larger impact overall.