

# 10 golden rules for applied data analysis (extended abstract)

Mats Carlin  
University of Oslo  
c/o SINTEF Electronics & Cybernetics

b/ot yaysU ot olcslCsC cr /sSy c/s Caca alaSntct  
cr CsdsSry tr4lCsU NrCsSSolT yUafcotst, -0 TrSCsl  
U4Sst aUs yUryrtsC O/of/ t/r4SC aSOant 3s vrSSrOsC,  
Baf/ nsaU a OoCs UalTs rv yaysUt ot y43Sot/sC O/of/  
3UsaRt trNs rv Nn yUryrtsC U4Sst,

**The rule of problem knowledge.** To become successful within the data analysis business, one must invest time and money in understanding the actual problem itself. Physical understanding, expert knowledge and explorative data analysis must be combined and every assumption must be checked to hold.

**The rule of method knowledge.** It is of vital importance that the data analyst knows the proper usage, benefits, assumptions and limitations of the method applied. Even the novice must have a basic understanding of standard statistical methods to succeed.

**The rule of normalising data.** Each variable should be scaled independently to the same range of values to avoid skewness in the estimation process and to assure robust and stable results.

**The rule of operation domain.** Data analysis models are only valid in those domains where the training data are dense enough for smooth model building. Whenever possible confidence intervals should be computed.

**The rule of independent testing and benchmarking.** Benchmarking should be done according to standard procedures and on several publicly available benchmark data sets. To make benchmark

tests reproducible the test data must be thoroughly described.

**The rule of validation.** Whenever possible the data used for training, test and validation should be different. The test set may be used during model adaptation. The validation should be used to ensure objectivity.

**The rule of fair comparison.** The amount of time and effort should be equal. It is easy to favour your own method by spending more time and effort in it.

**The rule of exposing weaknesses.** Whenever possible your data should be made available to ensure reproducibility. FTP or WWW are very suitable for data.

**The rule of interpretability.** The goal of data analysis is interpretability. In many ways we may gain new general insights about the actual problem. When choosing models which are easy to interpret.

**The rule of scepticism.** Always check others methods and results. Be sceptical whenever possible. The field may be wrong some times.

I wish to acknowledge Y. University of Taiwan for his suggestions of improvement. This abstract would never had emerged without the support by Hydro Aluminium Research Council (NFR) and