Stability of Clustering
Hans Ulrich Simon (Ruhr-University)

The goal of cluster analysis is to partition observed data into groups (called "clusters") so that the pairwise dissimilarities between data assigned to the same cluster tend to be smaller than those in different clusters. Although clustering is one of the most widely used techniques for exploratory data analysis in various fields, the building of a solid theoretical foundation is still in progress. Among the methods in cluster analysis that currently attracts considerable attention from both sides, theoreticians and applied researchers, we find the "stability approach". The intuitive idea behind "stability" is as follows: if we repeatedly sample data points and apply a clustering algorithm, then a "good" algorithm should produce clusterings that do not vary much from one sample to another. In the tutorial, we present some new results that aim at a "theory of stability" by providing a rigorous mathematical analysis within a clean formal framework. In the first lecture, we derive necessary and sufficient conditions for stable behavior of a clustering algorithm. In the second lecture, we apply the general results to so-called risk-minimizing algorithms whose risk function is induced by a "dissimilarity measure". These clustering algorithms represent a rich family with the popular "k-means" algorithm as a special case.

The topic of the tutorial is "brand-new" so that we cannot provide pointers to textbooks dealing with stability of clustering. In order to get a general idea of cluster analysis, we recommend to follow (one or more of) the following suggestions:

- Apply GOOGLE to the keyword "Clustering" and see what you get.
- Have a look in Chapter 6 of the classical book "Pattern Classification and Scene Analysis" by Duda and Hart.
- Have a look in Chapter 14.3 of the book "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman.

Later (but *before* the tutorial), we will provide further material about clustering stability that will be fairly self-contained (including up-to-date pointers to the relevant literature and talk slides.)