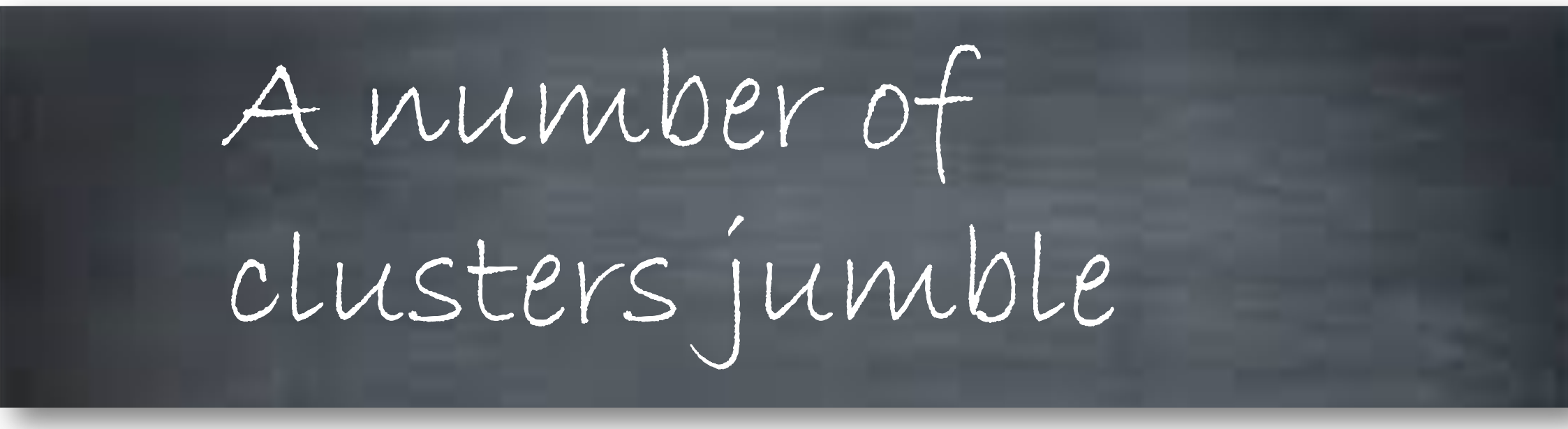
DSS Statistics Seminar
October 16, 2020, 12:00 a.m.
https://meet.google.com/rqr-ffmo-egq



C.M. Hennig
University of Bologna

The problem to decide the number of clusters in a cluster analysis doesn't have a generally accepted solution. There are good reasons for this. Most approaches in the literature do not require user input, but I will argue that any sensible approach will have to be adapted to the specific data and research question, and that any "automatic" decision of the number of clusters is deceptive.
I will present three rather different approaches to decide the number of clusters, and will try to convince you (and myself) that they all deserve a place in the cluster analysis toolbox. The first approach is based on comparing a cluster validity statistic such as the Average Silhouette Width to its distributions over different numbers of clusters under a tailor-made null model. The second approach selects the lowest number of clusters such that the observed degree of unimodality within clusters is not significantly worse than what is expected from a mixture model with the same number of clusters. The third approach is based on a weighted average of judiciously calibrated statistics that measure all cluster validity aspects that are required in the application of interest.
The first approach requires a null model but no model-based clustering; general cluster analysis methods can be used, as is the case for the third approach. The second approach is for model-based clustering. The third approach requires the simulation of random clusterings on fixed data. How do they all fit into the same presentation? We'll see...

SAPIENZA
UNIVERSITÀ DI ROMA