This Week's Citation Classic®

CC/NUMBER 25
JUNE 24, 1985

Johnson S C. Hierarchical clustering schemes. *Psychometrika* 32:241-54, 1967.
[Bell Telephone Laboratories, Murray Hill, NJ]

This paper developed a useful correspondence between any system of clustering based on empirical data and a certain kind of distance measure [The *Science Citation Index*® (*SCI*®) and the *Social Sciences Citation Index*® (*SSCI*®) indicate that this paper has been cited in over 770 publications since 1967]

Stephen C. Johnson
AT&T
190 River Road
Summit, NJ 07901

May 24, 1985

Placing similar objects together into groups is a natural human tendency, one even engaged in by social scientists. A *cluster* is such a group formed by someone who wishes to be thought a mathematician.

Repeating this clustering process by clustering the clusters, and so on, one obtains hierarchies such as naturally arise in biology and bureaucracy. A *clustering method* attempts to form clusters and hierarchies based on some empirical data about the objects being grouped, for example, a matrix of perceived distances between the objects. Clustering methods are appealing because clusters are easier to understand than are large tables of numbers. Clustering is dangerous because a great deal of information is lost.

My paper studied this loss of information. The distance measures described perfectly by cluster hierarchies correspond to the *ultrametric* distances of mathematics. A clustering method fits an empirical distance measure by an ultrametric distance measure. Looked at from this point of view, a couple of extreme "approximations," maximum and minimum, suggested themselves; these have the advantage that they are independent of the scaling of the data. A computer program was written to perform these analyses and was advertised in the paper.

This was my first published paper; it was written during the summer of 1965 while I was working at Bell Labs, taking a vacation from my mathematics PhD (in *category* theory, not even distantly related to *clustering* theory). The paper went through many drafts and owes much to the encouragement and criticisms of my colleagues at Bell, especially Roger Shepard and Doug Carroll. The paper represents a borrowing from mathematics of ideas whose time had clearly come; similar work was done independently by at least half a dozen others at nearly the same time.

I suspect that a major reason for the paper's popularity was the computer program. It was written in a highly portable dialect of FORTRAN, made pretty output pictures, was cheap to run, widely distributed (it was free!), and sinfully uncritical of its input data. Thus, many researchers could, and did, try the methods; those that liked the results could, and did, publish. Moreover, there were no burdensome significance tests to suggest that the data might not fit the clustering model. Thus, it was a fine way to obtain a computer's blessing without confronting the data's deficiencies.

Most empirical data is probably not well modeled by clusters. Such clusters as are truly present will be found by almost any clustering method. Thus, in the absence of significance tests, cheap methods seem preferable to expensive ones. Significance tests imply a null hypothesis; the many null hypotheses that suggested themselves to me were all unnatural and nearly intractable mathematically; in frustration, I became a computer scientist. Without significance tests, it is uncertain to what extent clustering results can be replicated, and replication is the cornerstone of science.

A good source of recent information on clustering is the *Journal of Classification*, a representative paper from which is cited in reference 1.

1 **Day W H E & Edelsbrunner H.** Efficient algorithms for agglomerative hierarchical clustering methods *J Classification* 1 7-24, 1984

