This Week's Citation Classic

CC/NUMBER 1
JANUARY 7, 1980

Lance G N & Williams W T. A general theory of classificatory sorting strategies. 1. Hierarchical systems. *Computer J.* **9**:373-80, 1967.
[Computing Res. Section, CSIRO, Canberra, ACT, Australia]

The five conventional sorting strategies are shown to be special cases of a single linear system containing four parameters. A new strategy is defined which provides a different intensity of grouping by varying a single parameter. [The *Science Citation Index*® (*SCI*®) **and the** *Social Sciences Citation Index*™ (*SSCI*™) **indicate that this paper has been cited over 195 times since 1967.]**

Godfrey N. Lance
Avon Universities Computer Centre
University of Bristol
Bristol BS8 1TW
England

August 8, 1979

"During the late 1950s and early 1960s, numerous methods for grouping individuals had been devised. The major difficulty was that each one had special properties, and thus attractions for the user, but each had to be specially programmed for a computer. (All classificatory strategies require the use of a digital computer because of the very large amount of arithmetic which has to be done.) My coauthor, W.T. Williams, and I were responsible for many of these programs and found that it was becoming impossible to see the wood for the trees.

"It was clear that if we confined ourselves to the classification techniques which were agglomerative and hierarchical then we could see a pattern. The key to success was a simple formula. We have two groups of individuals (i) and (j) with n_i and n_j members respectively and we call the inter-group distance measure d_{ij} The known sorting strategies all involved combining certain d's to determine the new distance matrix, after it had been decided which two groups were to be combined. The two groups in question were those with the smallest d_{ij} in the entire matrix —they are joined to form a hew group called k. For example, the best known strategy was nearest-neighbour and for this we define the distance between two groups as the distance between their closest elements, one in each group. This is expressed mathematically by the formula $d_{hk} = \frac{1}{2}(d_{hi} + d_{hj}) - \frac{1}{2}|d_{hi} - d_{hj}|$.

"Now we 'played' with this formula and after some experimenting arrived at a general linear combination of the distance measures containing four parameters. It is written $d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \gamma|d_{hi} - d_{hj}| + \beta d_{ij}$. (The nearest neighbour case is when $\alpha_i = \alpha_j = +\frac{1}{2}$, $\gamma = -\frac{1}{2}$ and $\beta = 0$.) We noticed that four other well-known methods were also special cases of this general expression. The squared euclid-ean distance method appears when α_i, $= n_i/n_k$ $\alpha_j = n_j/n_k$, $\beta = - \alpha_i \alpha_j$and $\gamma = 0$. So it was clear that the parameters need not be constants.

"We the-n created something new. For various reasons it is desirable to impose certain constraints, namely $\alpha_i = \alpha_j$, $\gamma = 0$, $\alpha_i + \alpha_j + \beta = 1$ and $|\beta|<1$ but even when this has been done we have a single infinity of β values, from which the a's are uniquely determined. When β is close to unity the method is space contracting and 'chaining' occurs but as the value of β decreases through zero the space is more and more dilated. We recommended that if β = —0.25 a useful amount of space dilation is obtained. This strategy has been called 'flexible' for obvious reasons. We now have one computer program which incorporates all these options.

"We believe the paper has been widely cited because it raised several questions about space-conservation and because subsequent workers have investigated the properties of the flexible strategy. Also, classification is a valuable tool in many different scientific disciplines and thus it is used by numerous different groups of researchers."