among three. But generally when more than a two-way relationship is involved, it is impossible to tabulate necessary totals in a way which will define the desired relationship without disclosure of proprietary data. Such relationships among many variables can be extended, however, at a much greater level of detail, if it is possible to apply standard statistical analysis techniques to the observations for individual respondent units over the whole range of the relevant variables. It is possible in this way to extract much more useful structural information and still insure that no disclosure of individual respondent data is contained in the results of such analyses.

In many cases, the data necessary to such an analysis require the matching of items from two or more statistical sources. One important class of analyses involves the matching or reports by the same respondent for different time periods. Some data files are so organized as to make this possible, but many are not. An even more complex problem arises when it is necessary to match data from the same respondent collected as parts of different statistical programs, by different agencies. This can be extremely difficult or impossible, though substantial progress has been made in some areas, as, for example among Census, Bureau of Old Age and Survivors Insurance (BOAST), and the Internal Revenue Service. Problems of disclosure are most difficult in this context.

There are several fundamental problems dealing with the coding and classification of original source data. Most serious is the need for uniform identification, definition, and coding of the respondent unit. Unless this is done, matching of data from diverse sources is generally impracticable if not impossible. A uniform system of classification and coding for geographic area is another major deficiency.

In general, the classification and grouping of data are dictated by the problem environment, the basic logic of the analytical model, and the kind and degree of detail in which the results must be expressed and interpreted. In practice, there is frequently need to compromise the ideal classifications and aggregations of data for several reasons; the basis and criteria of classification in the collection agency being inconsistent with the ideal requirements of the model; the lack of sufficent detail (industry, process, product, geographic location, etc.); the withholding of detail under proprietary confidentiality or security restrictions; the noncompatibility of the definitions of the respondent units in the several collection systems which could otherwise provide the information specified by the model, which can be reconciled only by coarse aggregation but with accompanying loss of information and structural detail; the noncompatibility of classification of the data by several collection agencies and information systems also capable of specious resolution by aggregation; the difficulty and cost of identifying and matching the reporting units from two or more reporting systems, so that the information about the reporting unit can be pooled; the absence of technique, staff, funds, and machine time to use large-scale data processing equipment to recode, recompile, reconcile, reclassify, and aggregate data and to perform all manner of statistical procedures upon the data.

Since there are very large numbers of ways in which most economic variables might reasonably be classified and aggregated, it is not practical to prepare the data in all of these formats in anticiption of possible requests. Nor is this necessary. The same results can be achieved with favorable logistics and great flexibility by providing for the basic records to be maintained in machine-receivable form and in as fine detail of classification as is practical. The low unit cost and high speed of modern computers can then be exploited to meet requests for data with little loss of the available information inherent in the combined resources of the participating agencies.

The availability of modern computers can meet two important requirements in this context. The first is discussed above: the conversion from finely disaggregated classes to all manner of special purpose classifications and aggregations (and, indeed, conversion to publishable forms). The second requirement is to avoid unnecessary loss of information because of proprietary and confidentiality restrictions. The fundamental rule in this case is to perform all edits and checks relating to unwanted disclosure upon the fully processed data (aggregations, summaries, averages, correlation coefficients, regressions, fitted curves, etc.) rather than upon the detailed raw data. This will assure full use of information consistent with disclosure rules. The logical capability of the computer also provides the key for the necessarily elaborate systems of rules essential to the prescribed protection.

Another major class of problems arises from the fact that errors and inconsistencies in the data as reported, transcribed, and coded are always discovered