

## STATISTICIANS WOULD DO WELL TO USE DATA FLOW DIAGRAMS

Mark A. Martin

Bayer Diagnostics, 333 Coney Street, East Walpole MA 02032

**KEY WORDS:** Data Flow Diagram, Consulting, Data management

### Data Flow Diagrams

The *data flow diagram* (DFD) is a tool that can help streamline the flow of data. Most statisticians would welcome a reduction in the time spent collecting and cleaning data, with more time available for analysis and reporting. The purpose of this paper is to demonstrate the construction and use of data flow diagrams, so that statisticians can employ them to their benefit.

I begin by reviewing common data management issues where a DFD can be helpful. The construction of a DFD is illustrated. Examples then show how data flow diagrams can be used to streamline the flow of data, communicate the process and progress in a consulting project, and map a SAS program.

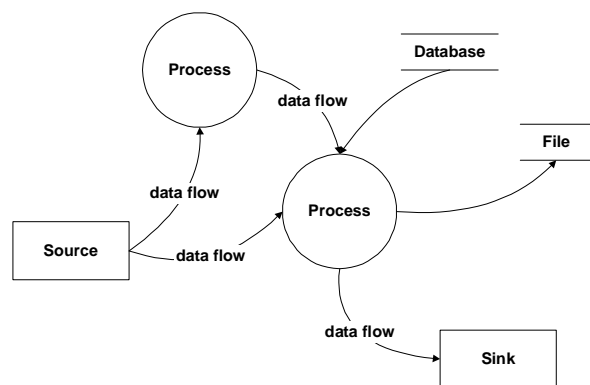
### Common Data Management Issues

There are many reasons that data processing frequently consumes more time than is desirable. After initial data collection, substantial “clean up” might be necessary. The data may need to be re-formatted before analysis can proceed. In project work, processes of data collection and formatting sometimes evolve in ways that include unnecessary or redundant steps. How can time be saved in these steps? Certainly better planning can help, but what tools can facilitate such planning? While a flow chart might help, data flow diagrams are especially useful. In fact, a DFD has some advantages over flow charts in planning for improved data processing.

### Construction of a Data Flow Diagram

A data flow diagram (DFD) is a representation of a system. It portrays the system in terms of its component pieces, and identifies all interfaces among the components. The building blocks of a DFD are illustrated in Figure 1. There are four basic elements.

- *Data Flow:* a named vector portrays a data path
- *Process:* a bubble portrays transformation of data
- *File or Database:* horizontal straight lines
- *Source or Sink:* a box portrays an originator or receiver of data



**Figure 1.** Elements of a Data Flow Diagram

The data flow diagram is best explained by example. Consider the flow chart in Figure 2, representing the process of making pizza. In Figure 3, a data flow diagram represents the same process. These contrasting types of charts illustrate some key differences in what is emphasized.

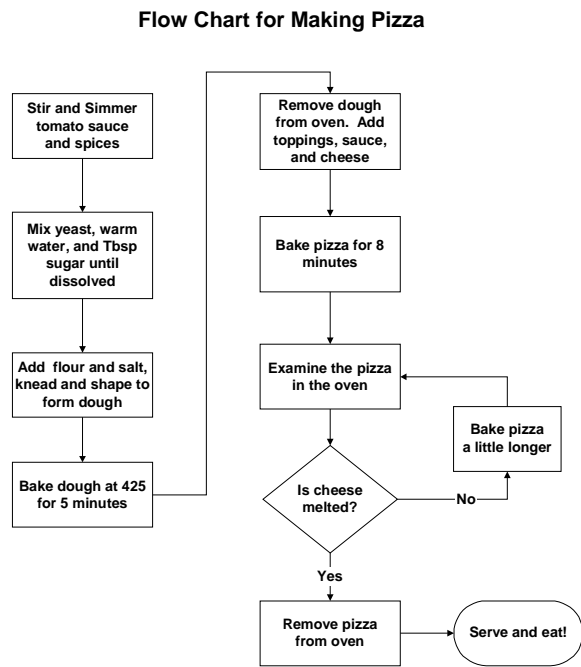
#### A Flowchart

- is drawn from the viewpoint of *those who act upon the data*
- shows the flow of *control*

#### A Data Flow Diagram (DFD)

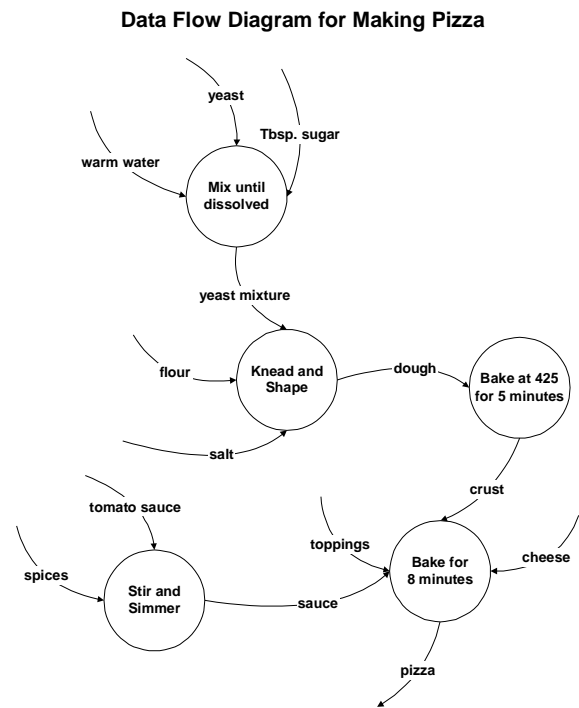
- is drawn from the viewpoint of *the data*
- shows the flow of *the data*

A DFD de-emphasizes the flow of control, and emphasizes the flow of data. This singular aspect makes a DFD very useful to the statistician who is concerned with how the data is to be obtained and processed. In the pizza example, the “ingredients” are the raw data. The ingredients are not readily



**Figure 2.** Flow Charts are drawn from the point of view of those acting upon the data (i.e. “chef acting upon the ingredients”). Flow Charts emphasize the flow of *control*.

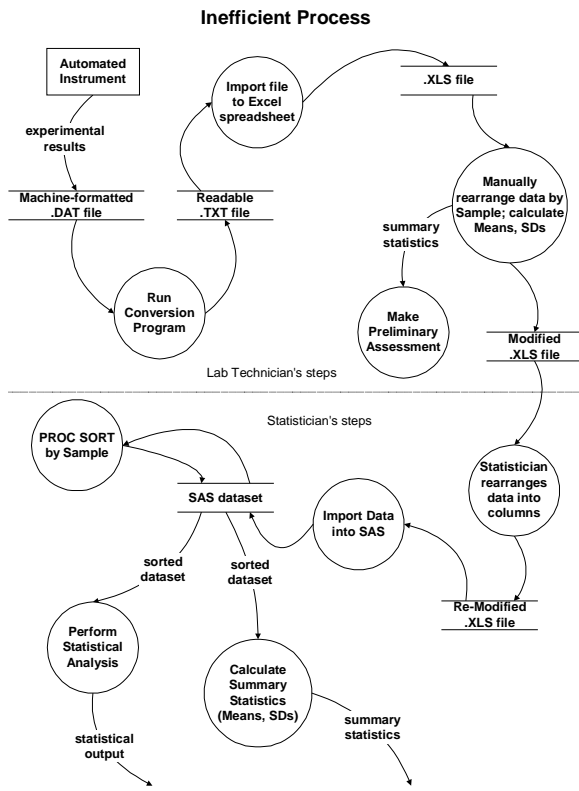
apparent in the flow chart of Figure 2, without reading the text in all of the boxes. In contrast, the ingredients are immediately apparent in Figure 3. Every ingredient begins as a labeled vector, and the arrow shows at exactly which process the ingredient is needed. The vector(s) leaving a process shows the output of that process. All interfaces among the ingredients and processes are shown in the DFD, whereas the flow chart does not reveal the interfaces. This is a key advantage of a DFD.



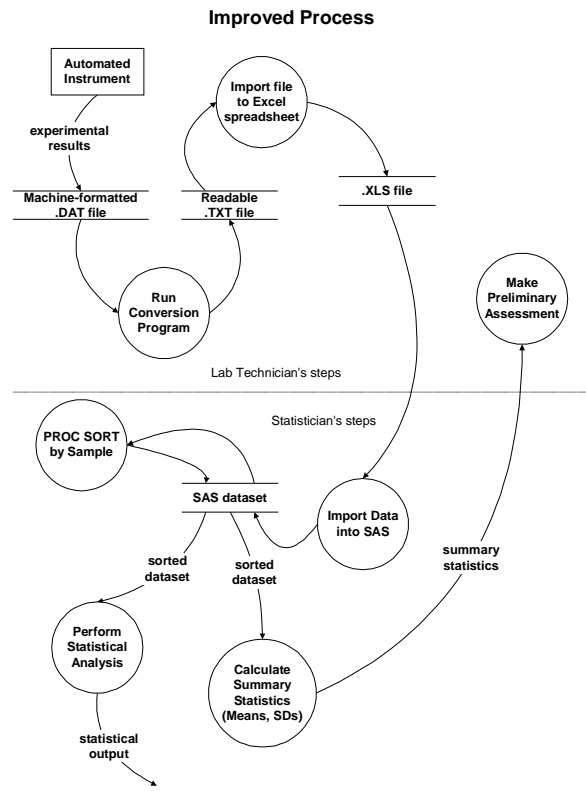
**Figure 3.** Data Flow Diagrams are drawn from the point of view of the data (i.e. the “ingredients” see the big picture). Data Flow Diagrams emphasize the flow of *data*.

The process of drawing a DFD to represent a process helps one learn whether or not the process is truly understood. Tom Demarco (1979) notes this useful property of a DFD: “When a Data Flow Diagram is wrong, it is glaringly, demonstrably, indefensibly wrong. This seems to me to be an enormous advantage of using Data Flow Diagrams.” One who knows a process well can quickly identify whether the DFD is accurately constructed.

As a side note, the data flow diagrams in this paper were made using Visio software (Visio Corporation), which provides “Data Flow Diagram Shapes” and dynamic linking for easy modifications. Software is not necessary, however. More often than not, I have drawn DFD’s by hand, and the handwritten version alone provided the desired benefit.



**Figure 4.** Diagramming a process “as is” can reveal inefficiencies. There is a redundancy in calculating summary statistics in both Excel and in SAS. In this scenario, both the statistician and the lab technician are required to manipulate data within Excel.



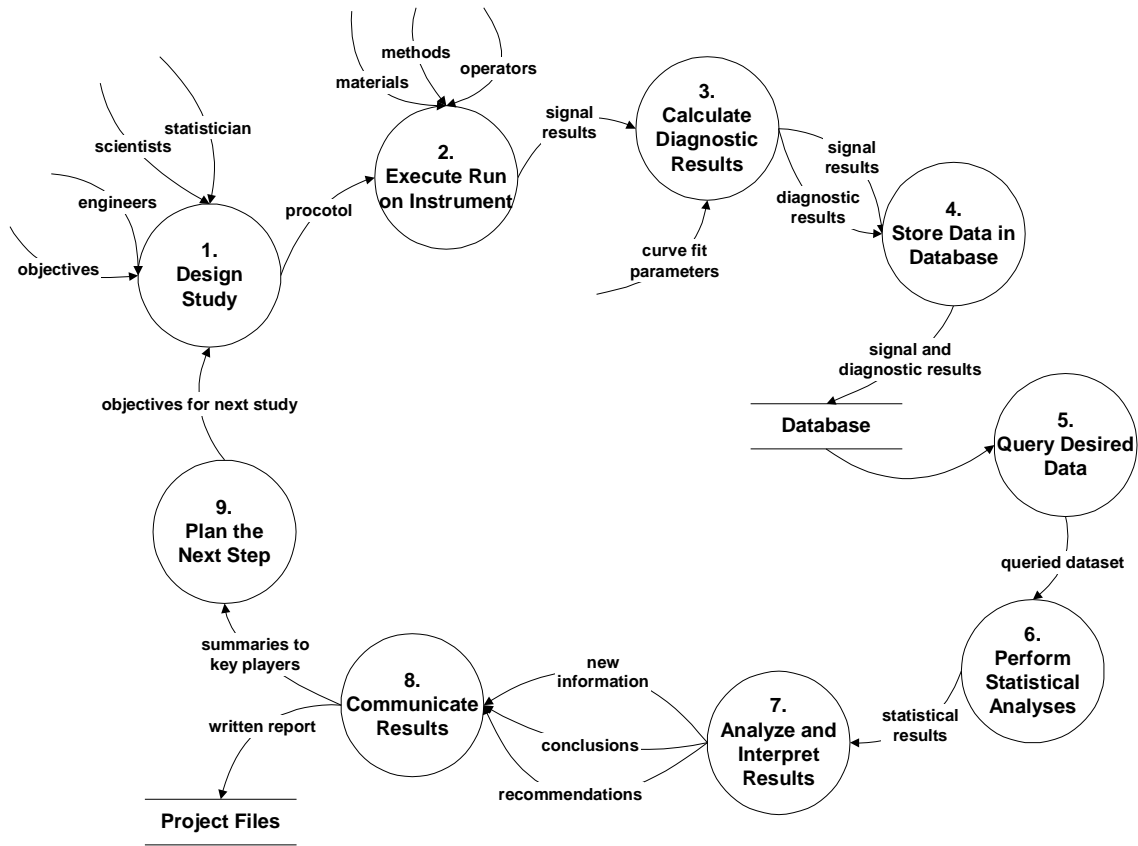
**Figure 5.** The process is re-drawn to eliminate the redundancy. The statistician can provide the summary statistics to the lab technician, and neither is required to manually rearrange the data within Excel.

### Streamlining the Flow of Data

Processes that evolve over time are often less efficient than they could be. A good way to streamline a process is to first make a DFD of the existing process. The DFD might reveal inefficiencies or redundancies in the current process. Such was the case in the process shown in Figure 4. A lab technician would dutifully perform several spreadsheet “cut and paste” operations to move data into different sections according to groups, and to calculate the summary statistics needed. Then the technician would send the modified Excel spreadsheet file to the statistician. Unfortunately for the statistician, the file had to be manipulated again to place the data into columns.

Figure 4 shows that summary statistics were being calculated in both Excel and SAS. The process was re-drawn in Figure 5 to remove the Excel file manipulations and calculations. The statistician was able to use the Excel file in its original format without the special groupings, because the SAS program would sort the data and calculate the summary statistics by group. Communication between the statistician and technician resulted in a process change and time savings for both. The data flow diagram helped pinpoint the inefficiency that could be corrected.

## Development Process for Testing and Improving Instrument Performance



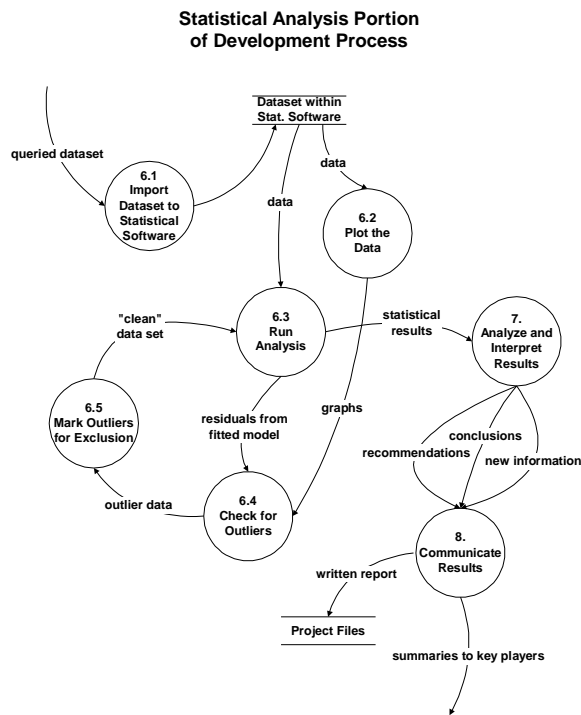
**Figure 6.** The process “bubbles” are numbered for reference. See Figure 7 for a detailed view of Process #6.

### DFD as a Statistical Consulting Tool

Data flow diagrams can be a useful communication tool in statistical consulting. In a fast-paced project, statisticians were working with a team of engineers and scientists in a series of many experiments to develop an instrumentation system. Rapid turnaround was required on each statistical analysis so that the next experiment could be planned. One of the statisticians constructed a diagram like the one in Figure 6 and distributed it to team members.

The diagram was useful in explaining which steps had to be performed prior to, during, and after statistical analysis. When there were bottlenecks, the delayed process step was pinpointed on the diagram. Decisions could be made to allocate additional resources with the right skills, if available. At other times, the diagram simply helped the team understand when and why there would be a delay.

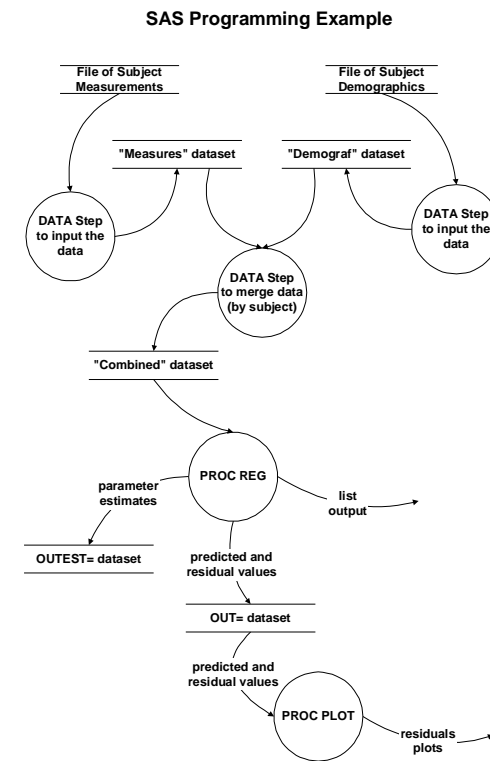
Notice that the process bubbles have been numbered for quick reference in communications. Additional DFD’s can elaborate on details within a process step.



**Figure 7.** This is a more detailed look at Process #6 (“Perform Statistical Analyses”) from Figure 6. Processes 6.1, 6.2, ..., 6.5 pertain to Process #6. Notice that the inputs to and outputs from Process #6 are consistent between the two diagrams.

### “Drilling Down” for Detail

Figure 7 includes process steps 6 through 8 from Figure 6. This DFD provides much more detail for Process #6. Process bubbles labeled 6.1, 6.2, 6.3, and so on, are components of Process #6. A statistician can use the details of this figure to show, for example, that outliers might be identified in two ways—either from graphs of the raw data, or by checking residuals from models used in the analysis. This level of detail was not provided in Figure 6, to keep that DFD from becoming too cumbersome. “Drill down” DFD’s can be constructed for any process step where it is desirable to provide more detail. The numbering system makes the relationship easily traceable.



**Figure 8.** Data Flow Diagrams for SAS Programs

- Use bubbles for Data Steps or Procedures
- Use straight lines to delineate SAS Datasets

Proc Contents can list the details of each SAS dataset

When DFD’s are constructed at multiple levels, as with Figure 6 and Figure 7, it is important to maintain consistent inputs and outputs at each process step. Notice that Process #6 in Figure 6 has one input (“queried dataset”) and one output (“statistical results”). In the more detailed Figure 7, there are 5 process bubbles associated with Process #6. While each individual step has one or more inputs and outputs, there is just one input from outside of Process #6—“queried dataset” is an input into 6.1. Likewise, just one output leaves Process #6—“statistical results” goes from 6.3 out to 7. Internal consistency between Figure 6 and Figure 7 is thus maintained. If inputs and outputs are not consistent, it is a sure sign that something needs to be corrected.

## Mapping a SAS Program with a DFD

Another useful DFD application for statisticians is mapping the flow of data resulting from software code. When writing or revising SAS code (SAS Institute Inc.), I have sometimes drawn a DFD to keep track of datasets and the procedures that use them. To illustrate, Figure 8 maps a SAS program which reads data from two separate files, merges them, performs a regression, and plots the residuals. There is a bubble for every DATA step and PROC step. Straight lines delineate SAS datasets. An arrow to a dataset indicates it is the output of a DATA or PROC step. An arrow from a dataset to a bubble indicates that the dataset is used by that DATA or PROC step.

Figure 8 shows that 3 DATA steps are employed. One reads measurement data from a file, and another reads demographic data from a second file. The third DATA step merges the two datasets by subject, so that the data needed for regression analysis is now in a single dataset.

When the PROC REG is executed, “list output” is created that includes the regression statistics. Figure 8 indicates that options were also used to create two new datasets—one for parameter estimates and another for predicted and residual values. The latter dataset is needed to create residual plots using PROC PLOT. (Incidentally, this DFD shows that dataset of parameter estimates was not used, so its creation is unnecessary in this instance.)

This application of a data flow diagram identifies the datasets created, and the procedures that use them. PROC CONTENTS can provide a detailed listing for each SAS dataset, which nicely supplements the DFD of a SAS program.

## Summary

A Data Flow Diagram (DFD) is a visual aid for understanding a system. Unlike a flow chart, a DFD focuses on the flow of *data*, rather than on *control* of process steps. For this reason, the DFD is a useful tool for statistical consulting. It can help identify opportunities to streamline data collection and processing. Statisticians who use data flow diagrams effectively may find that they buy themselves more time to devote to other activities, like analysis and reporting.

## Reference

Demarco, Tom (1979). *Structured Analysis and System Specification*. Englewood Cliffs, NJ: Prentice-Hall, Yourdon Press.