Screen Shot Chronicle for Initial Rhesus Macaque Data Analysis

To illustrate the utility of multi-dimensional visualization methods in unstructured data analysis – i.e., that where an objective such as increased yield or decreased switching time is not explicitly known – Dr. Stephan Bour of the National Institute of Health kindly provided clinical data from HIV testing in primates. This was a relatively small body of information, having fewer than 100 variables and observations of interest spanning only the period when two rhesus macaques, both infected with identical strains of HIV at almost the same juncture, remained in viable condition. Interestingly, however, one of those macaques remains alive and relatively healthy to this writing while the other perished within seven months of infection. Hence any clues on what events during the disease process allowed one individual to survive while the other eventually required euthanasia would be of potential scientific merit. Before engaging in this, however, it is worth mentioning the basic transform from N-Dimensional orthogonal space into parallel coordinates portrayed in Figure 1, below. This will be used with all data discussed in

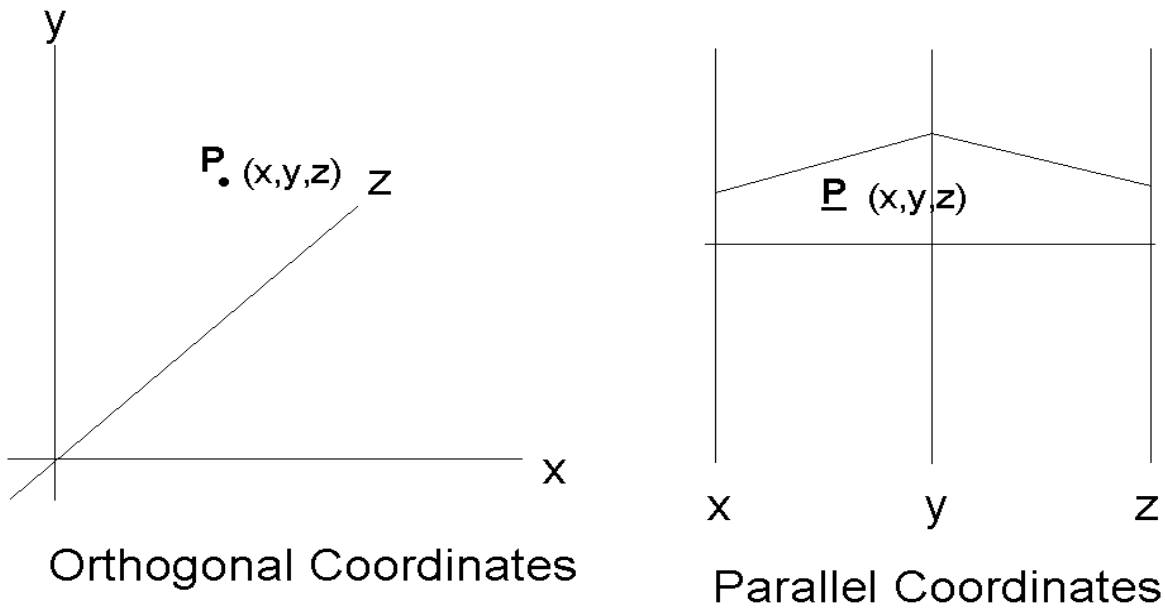

Figure 1: Transform from N-Dimensional orthogonal space (left) to Parallel Coordinates (right).

the analysis to follow, and allows us to map a high dimensionality data space – i.e., one where there are many variables, all conventionally laid out on axes (only three of which are simultaneously visible) at right angles to each other and intersecting at a common origin – onto a two-dimensional surface. The latter property allows us to visualize many more dimensions (in fact, there is no technical limit) at once, and hence many more variables (and their interactions) at once, than conventionally feasible. Figure 2 illustrates this further, with a single observation – the black polygonal line – isolated from a dataset of over 400 observations, where each of the original observations had 16 dimensions (X1 through X16) capturing the interplay among 16 variables in a situation, as it turned out, of non-trivial complexity.

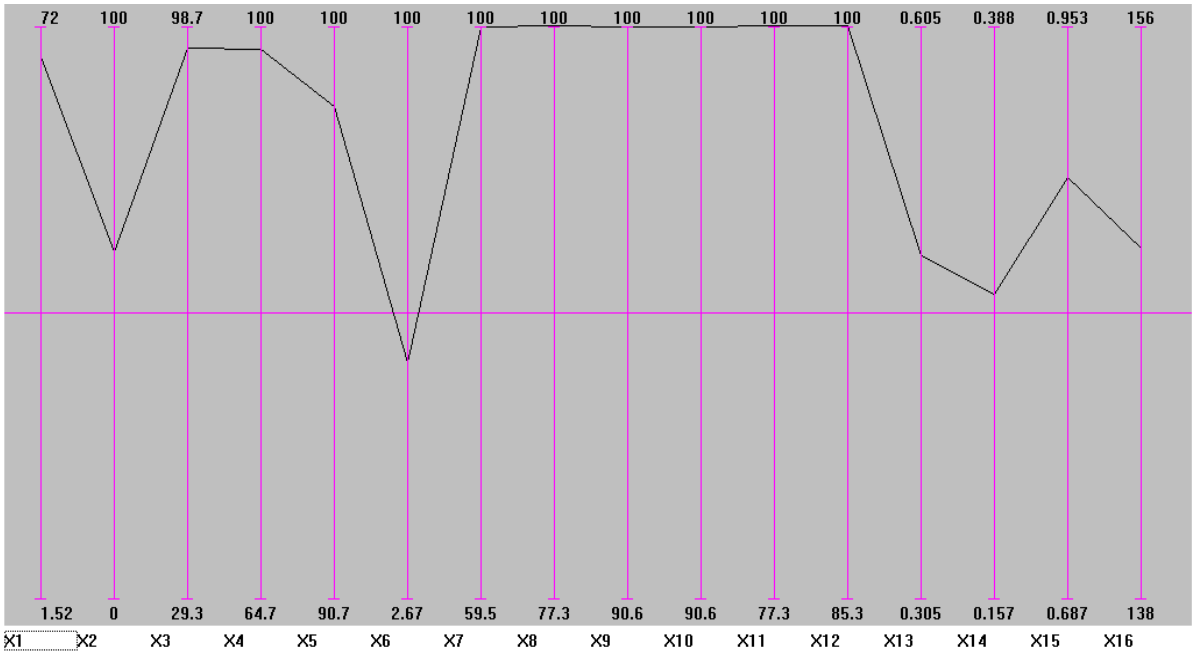

Figure 2: A single, 16-dimensional observation, not visible by conventional means, portrayed in parallel coordinates.

Turning now to the Life Sciences data provided by Dr. Bour, we begin by laying out all the information available on parallel coordinates and drawing queries to start isolating patterns of behavior in the data. Since the most useful of these to understand would be why one macaque lived while the other was rapidly overcome by HIV, our initial query will be to isolate the individual which lived (Subject 1) in yellow and its less fortunate counterpart (Subject 2) with fuchsia colored lines on the parallel coordinate plot captured by Figures 3a through 3c.

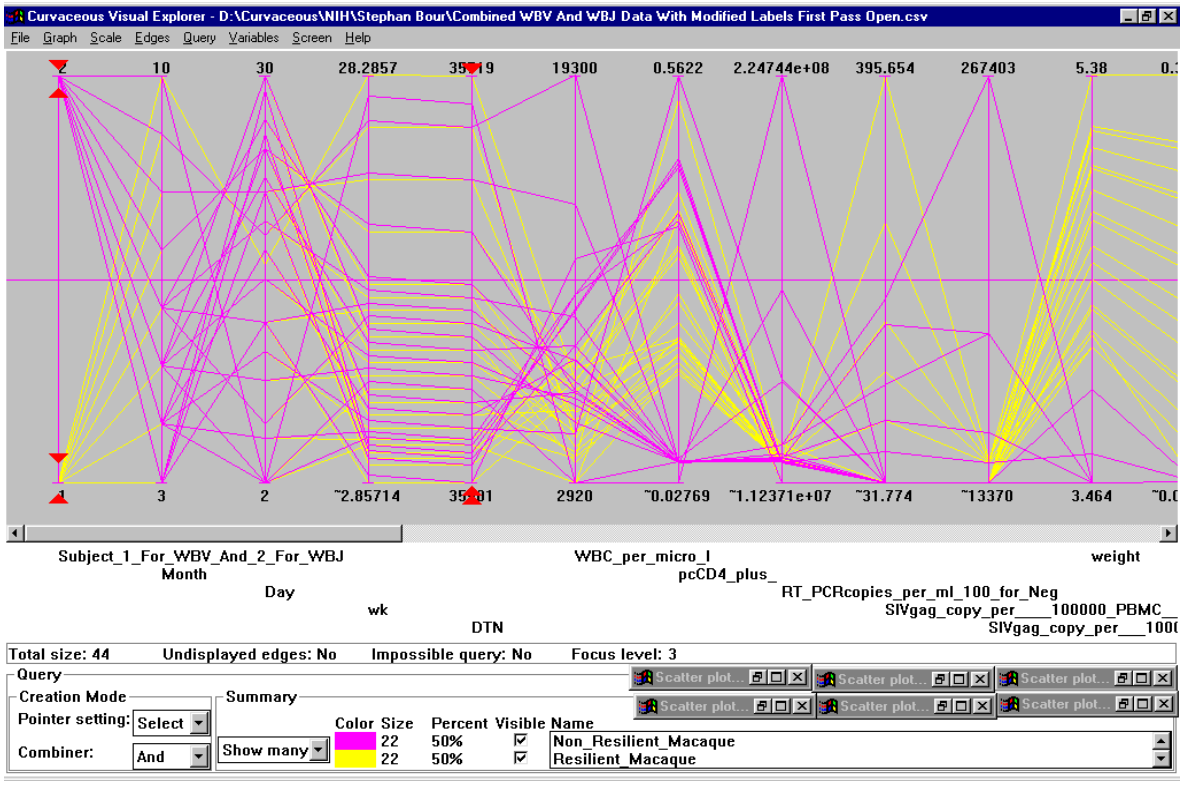

Figure 3a: Parallel coordinate display for leftmost 11 variable axes in NIH primate HIV data.

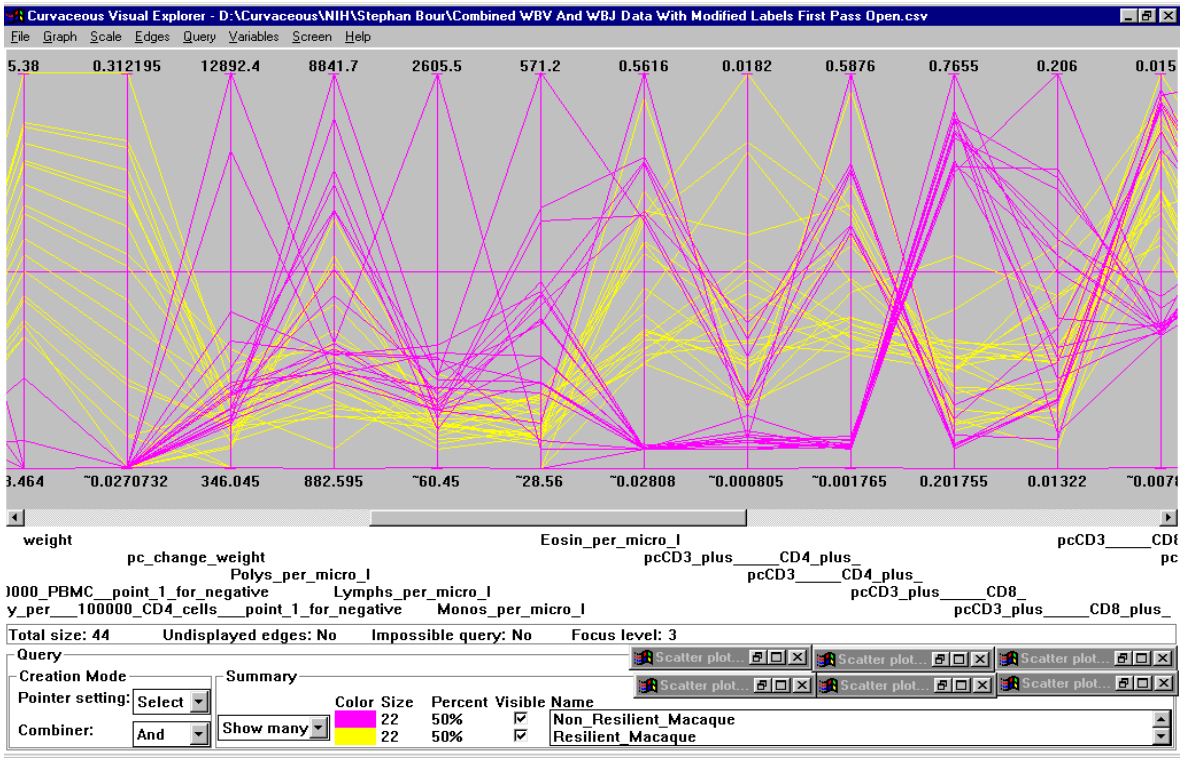

Figure 3b: Next 11 variable axes with all raw clinical data displayed from NIH primate study.

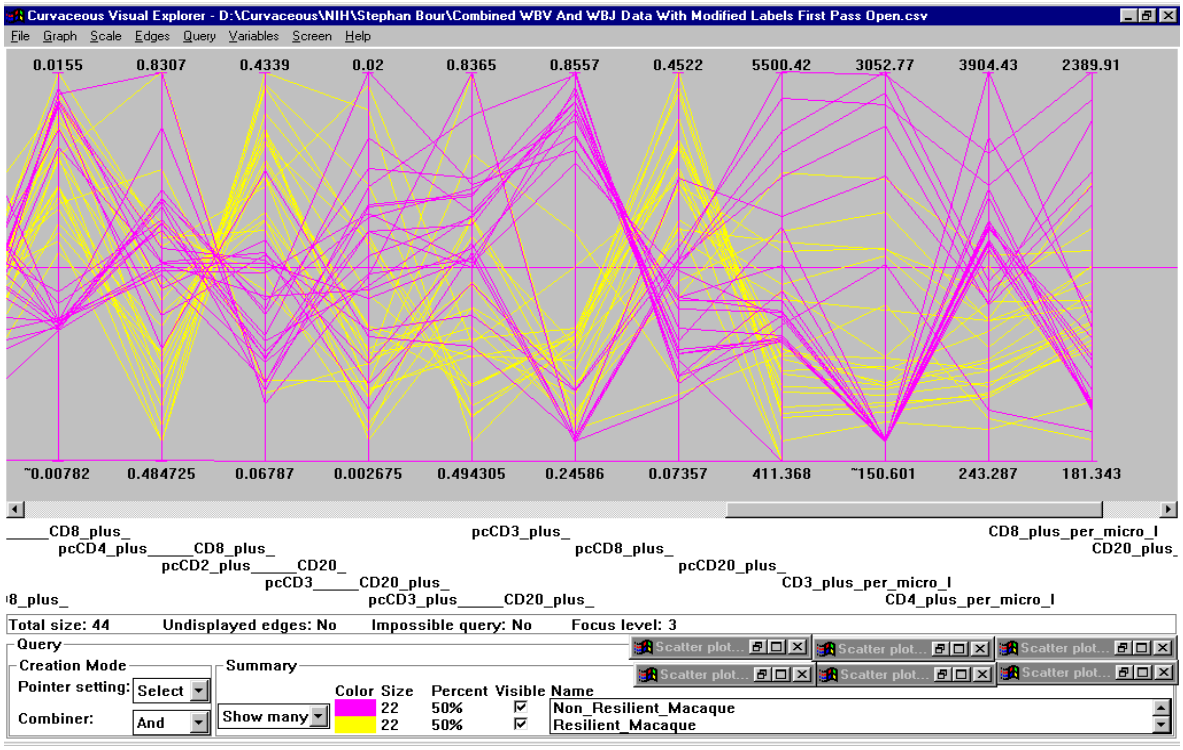

Figure 3c: Final 10 variables from original NIH study.

Looking at the plots of Figure 3 it is clear that although there is a stark difference in weight – the healthy macaque was always heavier – and a few noteworthy zones where one sees yellow but not fuchsia or fuchsia but not yellow, in general there is a great deal of mixing along the ranges of measurements for both subjects. Classically, given that we know there is a distinct difference

among the two data sources – one lived and the other died – this suggests that interaction among variables may be substantially more important than raw levels alone. Hence, although patterns such as the yellow patches along central regions of "pcCD4_plus" and "pcCD3plus__CD4_plus" (among others) as well as domination in the upper reaches of Polys, Lymphs, Monos, and Eosin per micro liter by fuchsia are undoubtedly part of the puzzle, a few interaction plots are probably the next step. A pair of these are captured in Figure 4, below, where an unusual (and potentially off-pattern) leg of the relationship between pcCD8_plus_and_pcCD20_plus has been decorated in blue and the time (weeks since infection) variable axis placed on the far left end of the screen. This does tell a

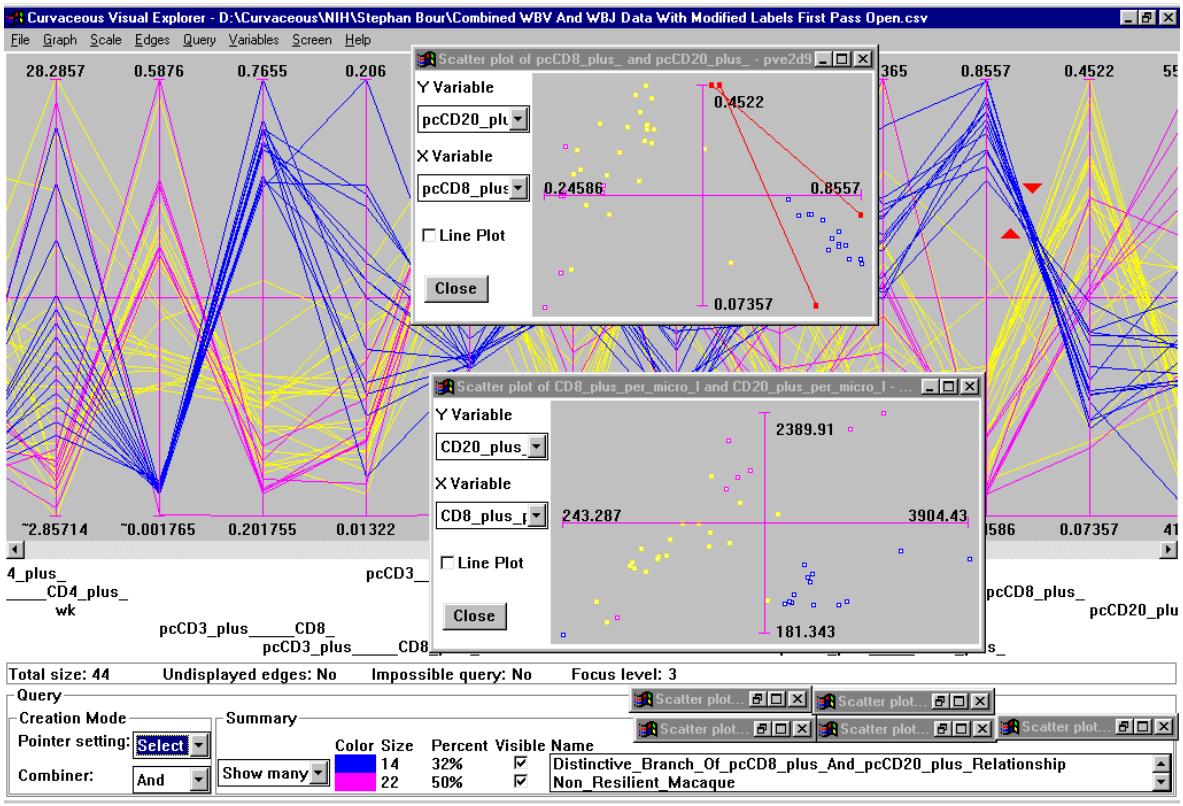

Figure 4: Scatter plot probes of variable interactions with time since infection (far left) kept in view and unusual branch of pcCD8_plus_and_pcCD20_plus relationship highlighted in blue.

tale, as the scatter plot relationships between points painted blue – which, conspicuously, start abruptly after a given juncture in time (wk) – and those remaining in fuchsia or yellow are clearly different. Moreover, since the points colored blue in Figure 4 were all fuchsia (the non-survivor's color) in Figure 3, this strongly suggests that an event in disease progression has been detected with a distinct temporal key. Thus, before a given time both macaques react to the HIV infection similarly, but after that juncture clinically measurable behavior of the doomed subject, which will continue fighting the infection for some months, diverges substantially. This leads to further scatter plot probing, where we can see both N-dimensional and Two-dimensional views simultaneously, with a query on time also incorporated to allow temporal pinpointing of whatever event disabled the fuchsia macaque's reaction to infection in Figures 5 and 6. Studying Figure 5, it becomes immediately evident that points decorated blue not only all belong to the fuchsia macaque but also correspond to a situation where certain reactions of the doomed subject's system went essentially dead. Figure 6 pursues this further, bringing in a time marker (in green) for the span immediately before certain immune system reactions of the fuchsia macaque became all but negligible. Hence, if the event circled in red on Figure 6 can be interdicted – or its cause at least determined – one must hypothesize that further understanding of HIV progression within primates can be readily obtained by multi-dimensional visual analysis of existing data.

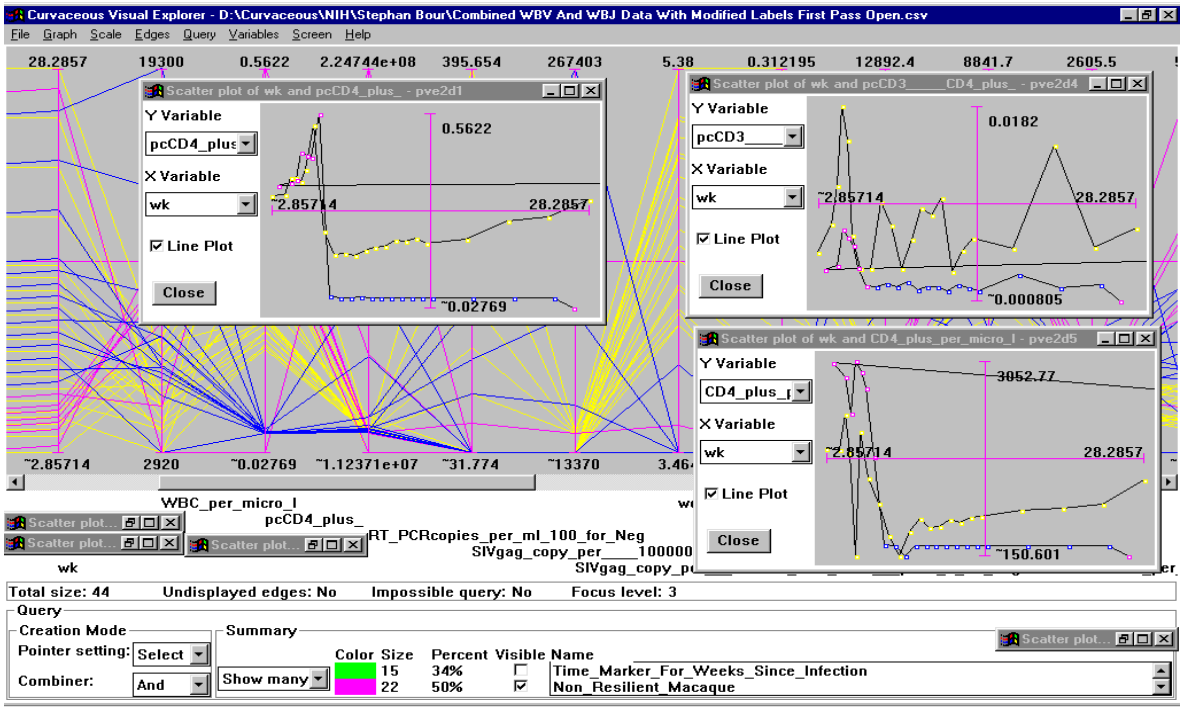

Figure 5: Further scatter plot probes (2-D views) as a function of weeks since infection superimposed on the N-Dimensional view.

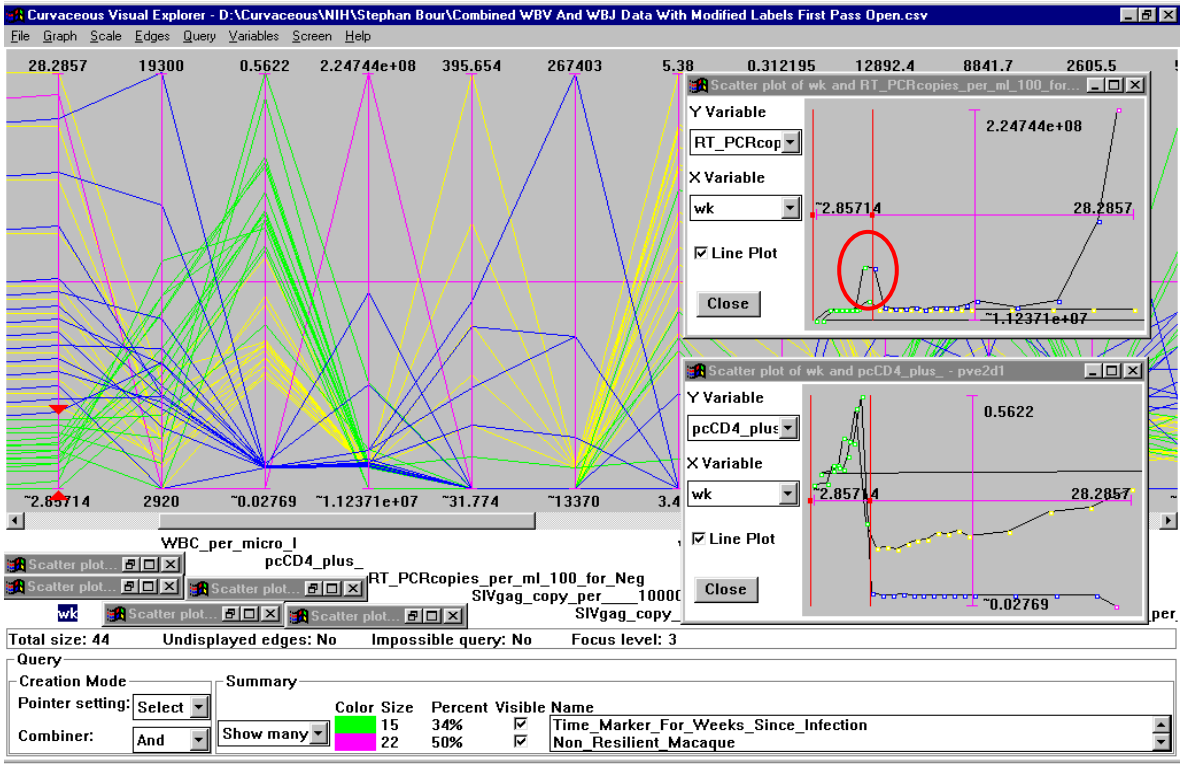

Figure 6: A smoking gun…probable damage event for the non-surviving subject in the spike evident on the upper right scatter plot for fuchsia points (overlaid by green and blue) but not yellow.

Continuing this line of inquiry, one is led to the question of whether it is possible to determine a macaque's fate before the 16th day "smoking gun" event. The latter is presumably irreversible, as many of the fuchsia subject's clinical measurements (e.g., pcCD4_plus) essentially flatline after that juncture, but earlier detection might yield sufficient warning for intervention. To devise such a mechanism when, from the 2D plots of preceding figures clinical measurements of the fuchsia and yellow macaques were ostensibly very similar before the 16th day of infection, one must first look for commonalities among macaques that lived. Fortunately, this is simple, as displaying a few key clinical parameters for macaques that survived SHIV infection from data provided by Dr. Bernard Lafont as a function of time (weeks since infection) on conventional 2D axes yielded striking similarity. This is reproduced as Figures 7(a) and 7(b), with parallel coordinate display of all variables for macaques that lived in Figure 7(c). Hence, even if parallel

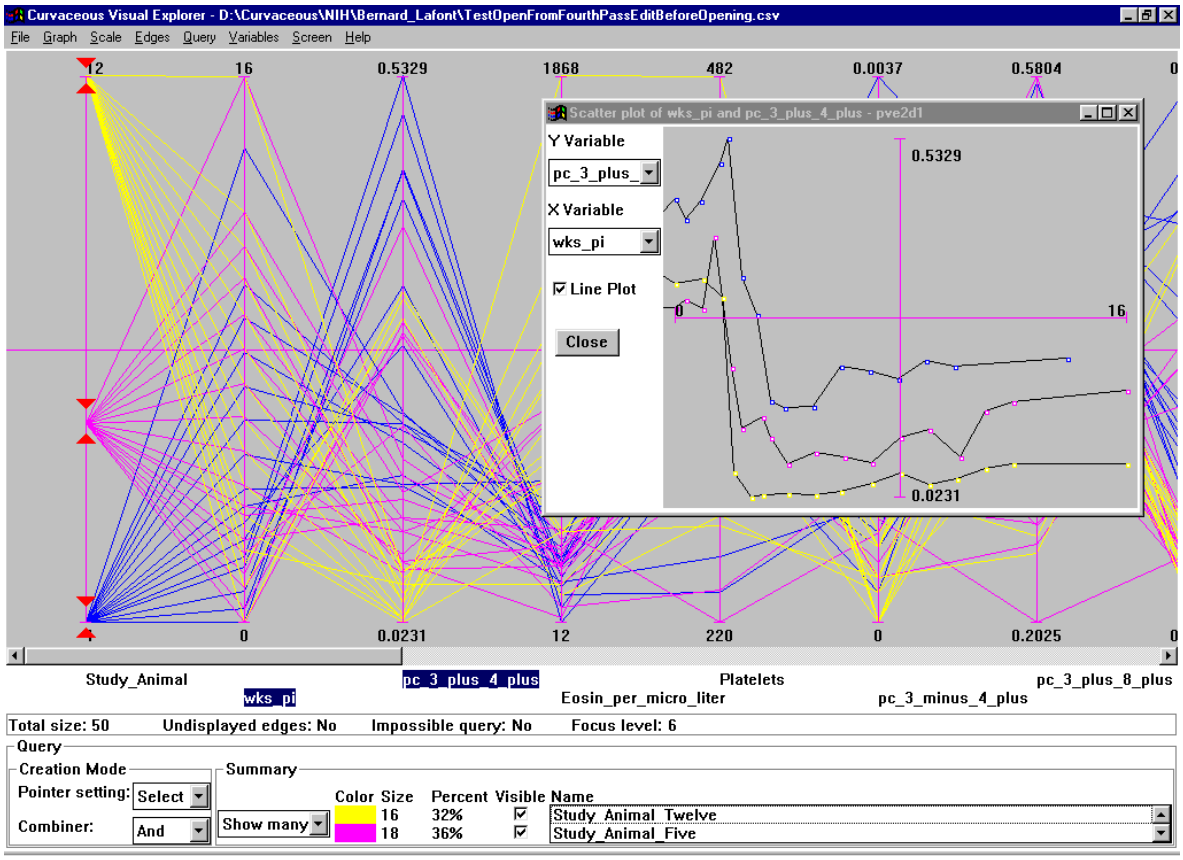

Figure 7a: Two dimensional plot of pc_3_plus_4_plus as a function of time for surviving macaques.

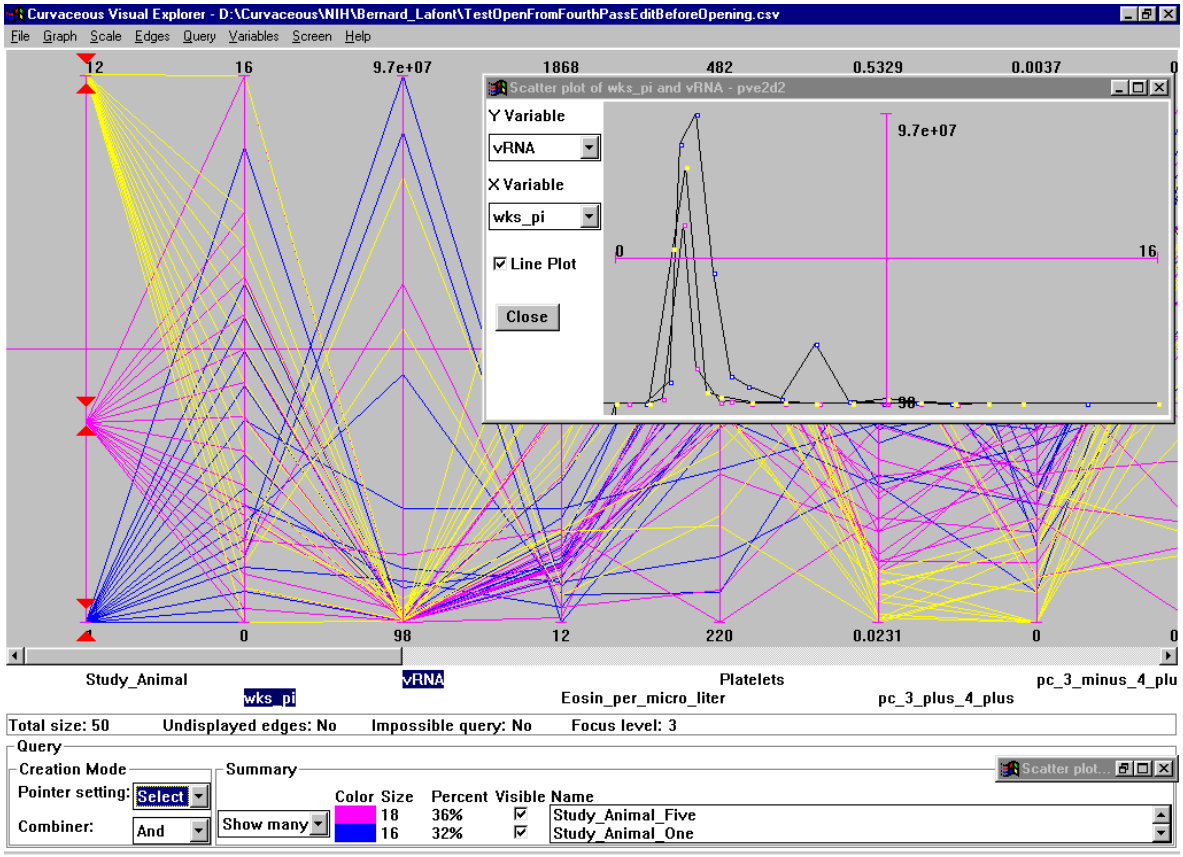

Figure 7b: Two dimensional plot of vRNA as a function of time for surviving macaques.

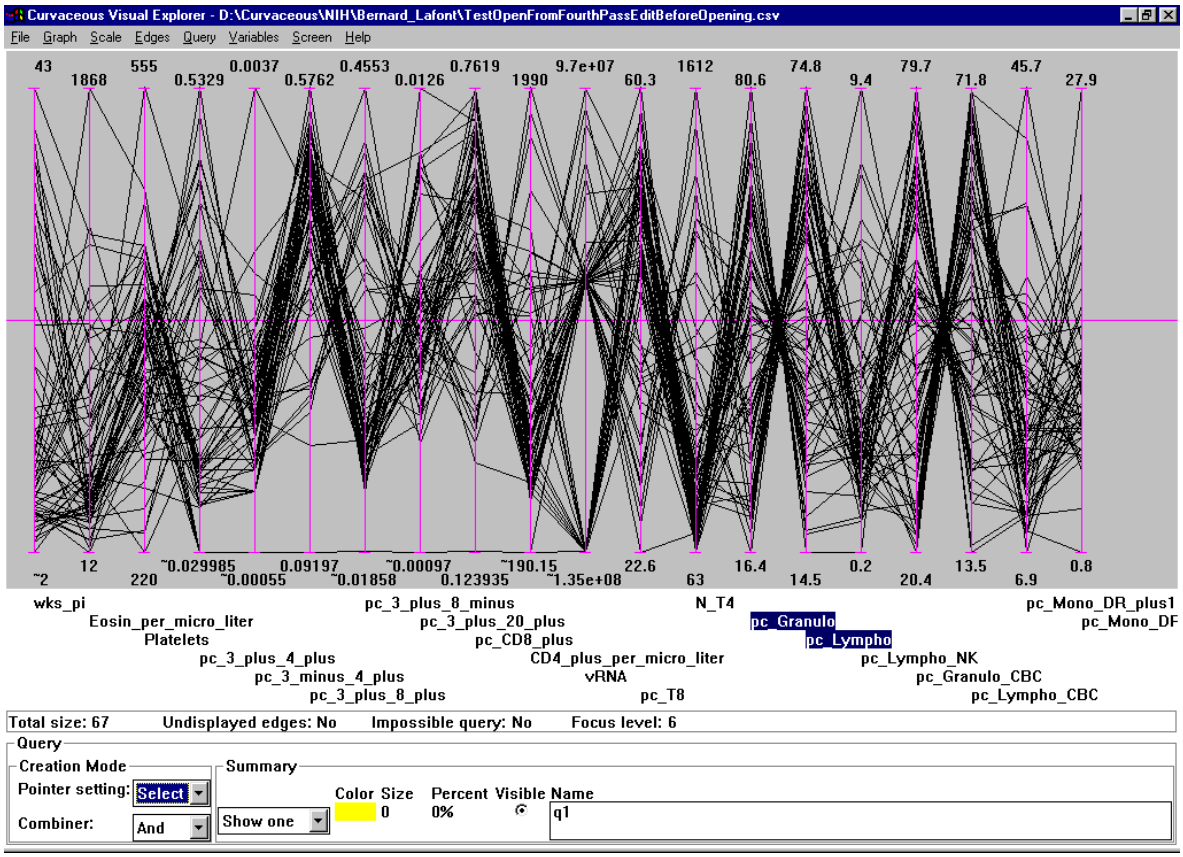

Figure 7c: Parallel coordinate display of all clinical data variables for surviving macaques.

coordinate display of data from macaques that lived is a tangled mass of black lines, embedded in it is a striking degree of temporally based similarity. That similarity, moreover, effectively comprises a time based trajectory which is the pattern of recovery for animals who survive SHIV infection. Consequently, the question becomes whether nuances and interdependencies among a wide variety of clinical measurements for surviving test subjects along with time based ranges of those parameters can be faithfully encapsulated into a model. More importantly, if so, when would the clinical data of a dying macaque would fail to fit that pattern?

Looking again at Figure 7c, note the crossing pattern apparent between pc_Granulo and pc_Lympho. This is indicative of a linear relationship, as one would expect, between pc_Granulo and pc_Lympho clinical measurements for test subjects which live as illustrated in Figure 8 and captured, interestingly enough, in the telltale "X" shape between those two variables. Hence,

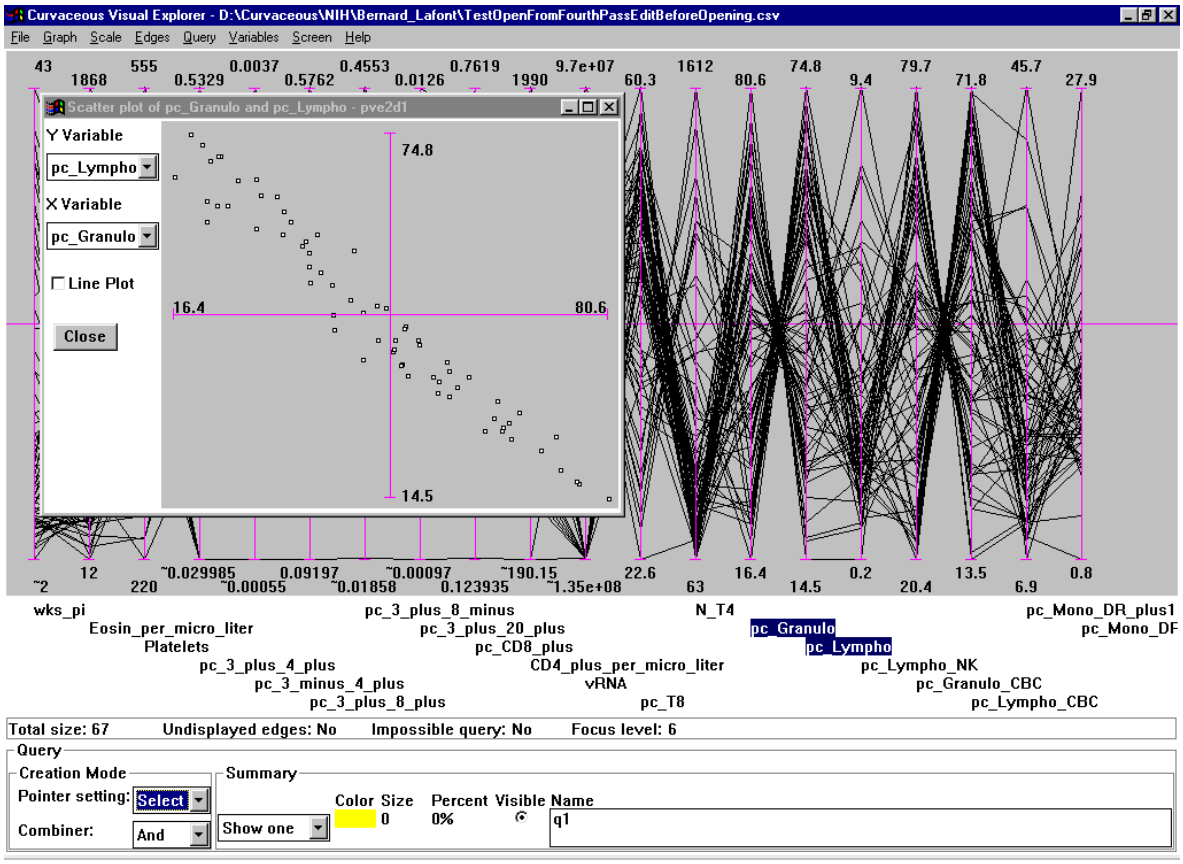

Figure 8: Expected linear relationship between pc_Granulo and pc_Lympho indicated by crossing pattern on parallel coordinate plot.

in parallel coordinates, where all variables can be seen simultaneously the shape or outline between adjacent clinical parameters keynotes the relationship between them. Further, by permuting axes on a parallel coordinate plot with respect to one another other shapes, and hence relationships, will become apparent. Consequently, we can proceed by capturing the interrelationships in clinical data between measurements for macaques which lived in the outline of their parallel coordinate display. Then, permuting the axes of that display to bring all dependencies and interrelationships to the surface – i.e., render them visible – creates a highly sensitive test for whether time based clinical data from any SHIV infected macaque does or does not fit the pattern and range of clinical data for animals which live. This is done in Figure 9, where red lines capture the pattern of clinical variable ranges and interrelationships for macaques which lived, green lines represent restrictions imposed by those interrelationships for any week since infection (the leftmost axis) when all axes are permuted with respect to one another, and violations of those available ranges are annotated. In this case, we have tested clinical data for a macaque which died against the multi-dimensional pattern –

also known as a geometric model – for those which lived. Note that all measurements (the blue balls) fall within ranges (extreme values for axes) exhibited by surviving test subjects at the upper

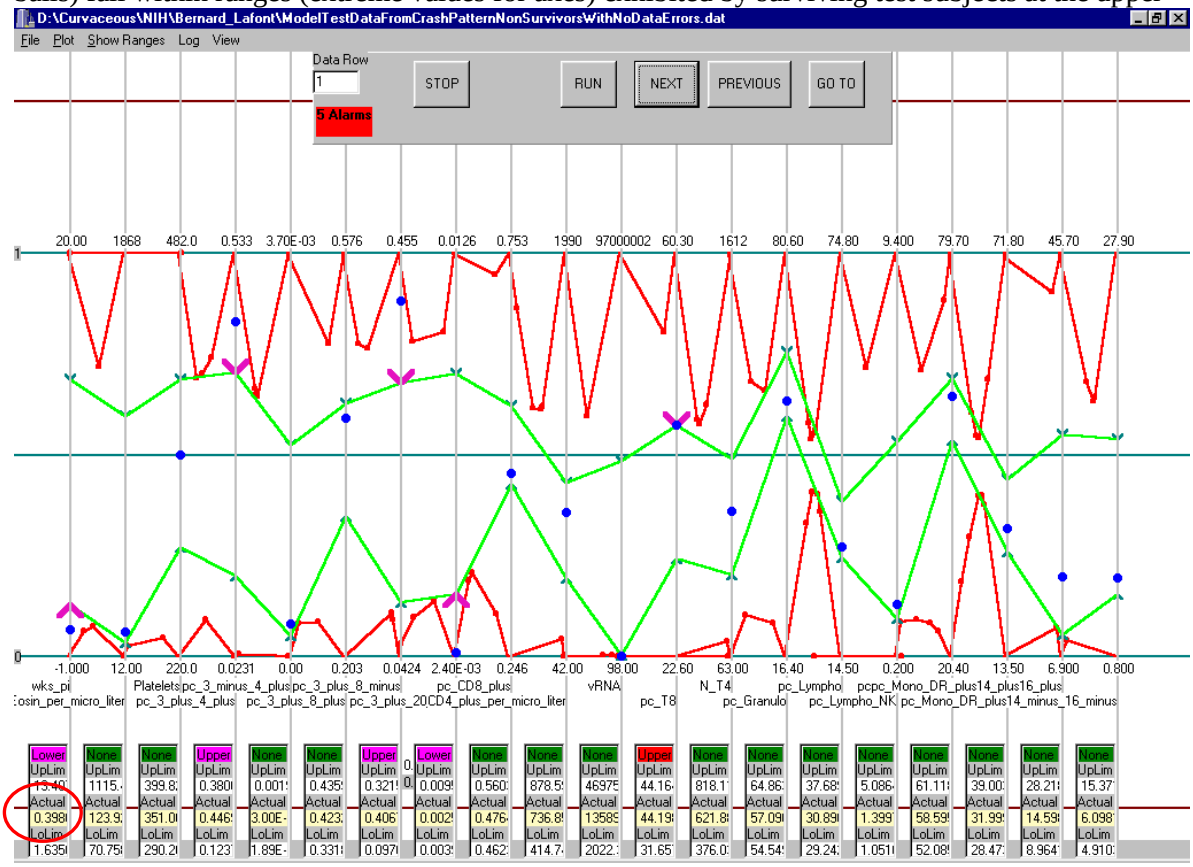

Figure 9: Geometric model testing doomed macaque against recovery pattern from surviving animals.

and lower most extremes of their geometric model (the red lines). At less than 0.4 weeks past infection, however, they are transgressing the green lines which enumerate constraints imposed by relationships among clinical variables for animals on a trajectory to recover… and hence signaling, far in advance of the 16th day past infection, that this animal will eventually be found dead or euthanized. Consequently, by taking advantage of multi-dimensional relationships latent in clinical data and brought to the surface by a geometric model, we can devise a very sensitive detection and recognition mechanism for use with data throughout the Life Sciences.