

# Characterizing and Measuring Data to Discover Relationships in that Data

October 31, 2005

## Overview

Here is a new and very different way to learn from data independent of semantic meanings associated with that data. The technique is accurate and simple once its principles are learned. The measures used work for individual patients and for sets of new patients or cases and work from any given data.

Discovering relationships in data is typically done using statistical analyses [1]. Supervised Learning is the discipline of finding relationships involving training data and Unsupervised Learning is discovering whatever relationships exist in the data itself. This new analytical technique accomplishes these same Learning objectives by first *characterizing* the data to create:

1. *Matched* groups of values in each variable,
2. The *probability* that each of those groups of values occurs (in that dataset),
3. Statistically *standardized* values of the raw data,
4. Numeric names of the *patterns of the values* in that data,
5. Numeric names of the *patterns of the probabilities* in that data.

The patterns are named and summarized for each dimension of its database. The dimensions of a relational database are its horizontal, vertical, and relational dimensions connecting all subsets of data in that database. In tensor notation, these dimensions are the dimensions of the tensor. When working with 2D spreadsheets, these two dimensions are the dimension of columns and the dimension of rows.

Scatter plots reveal internal relationships in the characterized multi-variate data. There are no concerns about minimum or maximum number of variables and cases. Analytical accuracy is achieved for the data as submitted. The number of variables in a 2D spreadsheet is the number of columns and the number of rows is the number of cases, or patients, or instances of an experiment. Data ranging from 2 columns to 55,000 columns and from 6 rows to 64,000 rows has been analyzed using just a PC. Larger sized data can also be handled with the same programs.

Once the data is characterized and visualized, the third step is to measure it. Classical mathematical *measures* applied to the characterized data are:

1. The vector-sum of each characterization of the data in each of the dimensions of its relational database [2],
2. The Shannon Information measure [3] of each characterization,
3. Algebraic proximities of each characterization in relevant subsets of the data,

These measures summarize each characterization. The vector-sums (vector-fused resultants) of multi-variate data, when scatter-plotted, display the relationships in the characterized data. Linear, circular, cardioidal, or hyperbolic relationships are displayed as those geometric relationships when such relationships exist in the data. If non-linear (and 1:1) relationships exist in the multi-variate data, the shape of the non-linear relationship is displayed as a line of that shape. If 1:1 relationships do not exist in the multi-variate data, clouds of vector-fused resultants are displayed much as clouds generally appear in statistically analyzed data. In these situations, proximity measures applied to the vector-fused resultants produce point-by-resultant-point conditional probability measures for the likely associations of each point in the cloud. Shannon Information produces even more precise measures of the relationships of individual cases with classes of training data outcomes.

A specific application of Shannon Information to colorectal cancer gene-expression data [5] has shown 99% accuracy in predicting individual survival beyond 36 months for 116 patients each characterized by the pattern of the data giving the amount of each of 54,676 genes in their tumor.

Further analyses using both the horizontal and vertical dimensions of this data that has 54,676 variable and 116 patients is expected to improve the accuracy and the understanding of how and why this Shannon Information measure works so effectively.

Inherent in this work is the step of being able to see the raw, the characterized, and the measured data. Seeing can pinpoint what and where significant structure exists in each of these classes of analysis.

Further analyses are expected to improve accuracy, knowledge, and appreciation of where and how to apply which characterization and measure to specific kinds of data.

## Background: seeing data

Multi-variate data can be seen or visualized using a Manhattan Diagram such as is provided in Excel. Figure 1 uses one perspective angle putting each column (or variable) of data values “in perspective” and thereby displays all 6 variables in one 3D perspective drawing. The first figure in the Appendix is another way to see all of one’s data using Excel. Vector-fusion [2] then shows how to see the vector-sum measure of multi-variate data in one 3D perspective drawing by using a different perspective angle for the values of each variable. The challenge to see large data sets is merely size: deciding what other artifacts to use to see huge data in one view.

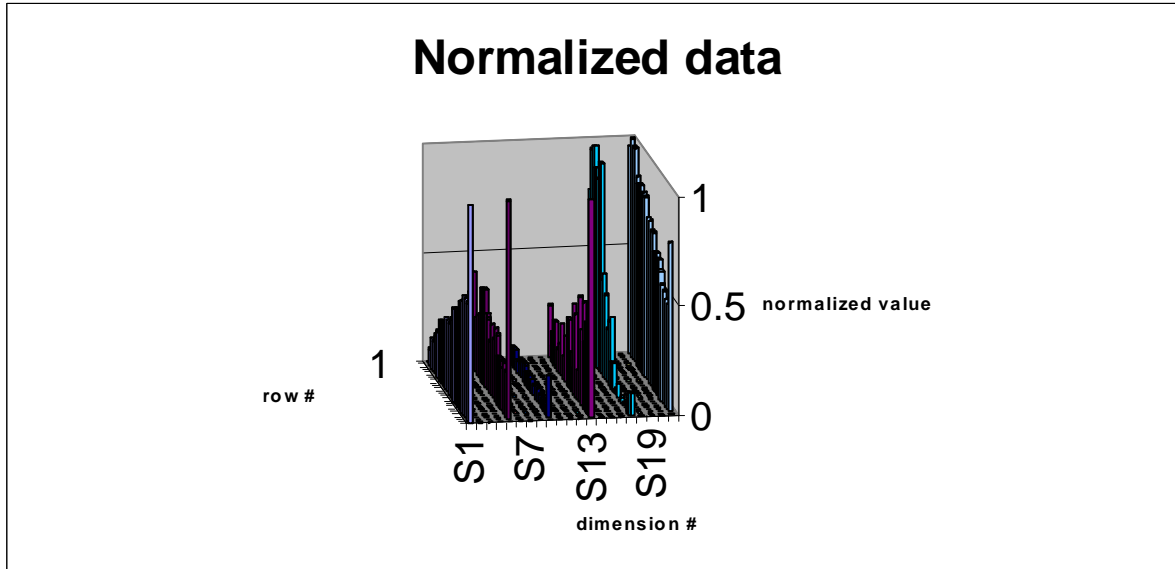


Figure 1. Excel Chart of 20 values each in six dimensions

In principle, seeing multi-variate data is easy if one does not insist on seeing all variables plotted perpendicular to each other. Being able to see multi-variate data of any number of variables, and of any size, is the first step in understanding that data.

## Background: discovering relationships in data

The second step in understanding data is to see relationships in that data. The reason for plotting each variable of data perpendicular to the other variables is to discover the (orthogonal)(or independent) relationship among all the variables. A scatter plot is a 2D orthogonal plot of two variables of data; if there is a linear relationship between those variables, one gets a straight line. If there is a circular relationship, one gets a circle. The names of classic geometries describe other familiar shapes. The shape of the locus in a scatter plot describes the (independent) relationship between the values of its 2 variables of data.

To see the relationship between 3 variables of data, one vectorizes each variable at  $90^\circ$  and plots the vector sum of all three variables in 3D orthogonal display space (a 3D scatter plot). Figure 1 however is a plot of 6 data-space variables seen in 2 dimensions of display space using a single perspective angle to represent the 3<sup>rd</sup> variable in 2D display space. Figure 1 does not show any relationships among the 6 data variables however. To see relationships in data, one has to plot the vector-sum of each variable, not just the values of each variable in the data.

Vector-sum-resultants for vectors at angles other than  $90^\circ$  also capture relationships between the plotted variables. To see the relationships between 4 variables of raw data one could vectorize each of the first two variables perpendicular to each other and plot the next two variables each at a different perspective angle, all in 2D of display space. Straight lines in 4 variables are revealed as straight lines in 2D perspective drawings of 4 variable data; circular relationships in 4 variables appear as elliptical relationships in 2D perspective drawings with 2 perspective angles. Ellipses occur because while the last two vectors are not independent, there is a geometric relationship between them. Vector-fusion extends the multiple perspective angle idea and vectorizes all variables, beyond the first two, each at their own different perspective angle. The 2D scatter plot of the vector-sum of all n

vectorized variables in vector-fused data thus displays a shape or form of whatever relationship exists in that n-variable data.

### Background: properties of patterns in data

The vector-sum of vectorized data is both a name and a measure of the pattern of the values in the n variables of the data [2]. Vector-fusion is a useful way to capture, in one 2D number, the vector-sum property as a measure of the pattern of all n variables of data in a row of data. The measure generated by vector-fusion can also be considered a form of hashing where this hash total captures properties of the “not-randomized” values. The properties captured by the vector-sum include the curvature among the sampled data points in each column. The vector-sum is a numerical “name” and a measure for the pattern in the row of data in its many columns that is functionally derived from the discrete values of data in each row.

### Background: Statistical Measures.

Essentially the only data mining tools generally available today have been statistical tools. Statistical analyses of multi-variate data are complicated and difficult to understand and assess [1] Statistical tools are largely predicated on probability distributions and measures on those distributions. Most statistical tools are derived assuming the data has a binomial or other mathematical distribution. Most experimental data, at best, only approximates a mathematical distribution, and the significance of statistical error analyses is often hard to appreciate. Drawing specific conclusions for individual patients or cases is typically not attempted in statistics; at best, a patient or individual case can be said to be in “this” percentile.

Regression in statistics [1] estimates a vector  $\mathbf{w}$  of parameters for an unknown function  $y = f(x, \mathbf{w}) + \epsilon$  where  $\epsilon$  has a standard distribution of zero mean and fixed standard deviation, and  $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \dots \mathbf{w}_{M-1})$  where each  $\mathbf{w}_i$  is a component vector orthogonal to all other component vectors  $\mathbf{w}_i$ . Training data has a single predictor variable  $x$  where  $\mathbb{R}^1 \ni x$ . Polynomial regression approximates a continuous function  $y$  with a polynomial:

$$y = \mathbf{w}_1 x^{M-1} + \mathbf{w}_2 x^{M-2} + \mathbf{w}_3 x^{M-3} + \dots + \mathbf{w}_{M-1} x^1 + \epsilon \quad (1)$$

This yields a probability that the vector parameters  $\mathbf{w}$  fit  $f(x, \mathbf{w})$  with minimum error  $\epsilon$ .

For comparison, vector-fusion [2] maps a multi-variate vector  $y = f(\mathbf{w})$  where  $\mathbf{w}$  is the 2D vector:

$$\begin{aligned} \mathbf{w} &= w_1 e^{i\theta_1} + w_2 e^{i\theta_2} + \dots + w_M e^{i\theta_M} \\ &= \sum_M w_i \cos \theta_i + i \sum_M w_i \sin \theta_i \\ &= (w_{\text{sum } x}, w_{\text{sum } y}) \\ &= (\text{SBP}_x, \text{SBP}_y) \text{ using the notation from [2].} \end{aligned} \quad (2)$$

and each  $w_i$  is the value in each cell of M columns for each row of raw data. Each variable (column) of raw data is assigned its own unique phase angle  $\theta_i$  and the vector sum of all values  $w_i$  is computed as the vector-fused resultant of all M component vectors. The vector sum is precise; there is no error  $\epsilon$  in this mapping. The vector-sum here corresponds to the linear “approximating function” of statistical analyses. Other values of  $w_i$  may duplicate this vector sum, but there is no error with the vector sum “approximating function” itself. In vector-fusion, the approximating error  $\epsilon$  is zero.

### Mathematical Measures

There are many mathematical measures that have been developed to capture properties of data [4]. Familiar statistical measures are the average, the median, variance, and standard deviation. These kinds of measures are called intrinsic measures because the measure of the selected sample varies with, and is intrinsic to, sample size.

A different kind of measure called an extrinsic measure is used in thermodynamics and Information Theory. This measure uses the arithmetic product of the probability of an event multiplied by the logarithm of

that probability, and sums this product over all events. In thermodynamics this measure is called entropy; in Information Theory [3] this measure is called (Shannon) Information. Shannon Information  $H$  is computed:

$$H = - \sum_n p_i * \log p_i \quad (3)$$

where  $\sum_n p_i = 1$  and  $p_i$  is the probability of an event, and  $n$  is the total number of events  $i$ .

The unit of measure defined by Shannon for  $H$  in 1948 was the (now familiar) name “bit”. For comparison with statistics where the logarithm of  $p_i$  is a frequent measure, in Information Theory, the Information *Capacity* of a system (in bits), such as the capacity of a computer memory to store Shannon Information as computed by Equation (1), is the simple logarithm measure of that Information.

Entropy and Information differ by a constant called Boltzman’s constant which was chosen by Boltzman to give entropy the units of temperature and energy and normalize the measure for the number of events (molecules) in a standard volume. Shannon entered no constant in his Information measure and assigned the name “bit” to the measured value of this sum. Otherwise, entropy and Information use the same extrinsic measure ( $p \log p$ ).

One significant property of this ( $p \log p$ ) measure is that it gives the same value for a sample as for the value of the entire set by normalizing the  $p_i$  for the number of values in each variate. Thus sample size does not matter with this extrinsic measure. This extrinsic property is in important contrast to the situation encountered when analyzing or measuring data as is done typically using statistical measures where are all measures are based on averages and medians that are sensitive to sample size. The probability distribution generated by the  $p_i$  for Shannon Information is the exact distribution contained in the data, not an approximation to a binomial (or other) distribution as in statistics.

Another property of the ( $p \log p$ ) measure is that it measures disorder or uncertainty, not order or “knowledge”. Maximum (Shannon) Information occurs when nothing is known about the data; for a given set of  $n$  events  $i$ , maximum Information  $H$  occurs when each and all  $p_i = 1/n$ . Thus it is critical to retain all “tiny” or single occurrences of values since there are normally so many of them summed in Equation (3). In most statistical analyses, outliers and small probability occurrences tend to be ignored.

In statistics, data with the most variance is said to be of most interest because it contains the most “information” [1]. “information” in this usage probably means knowledge rather than Shannon Information, but the meaning of “information” needs to be carefully understood when using statistical measures. Statistical “information” sometimes measures the % of the variance attributable to a variable. Shannon Information does not measure or involve variance. In any case, using data with the most variance, as is done in Principal Component Analysis, such data when used in a signal processing application would mean using the data with the most noise. In signal processing, one wants to use data with the least noise as being more reliable than data with the most noise.

In discussing noise, variance, and probability, one should start with basic definitions again. Variance involves differences measured with respect to the median (an intrinsic measure); (discrete) probability usually is interpreted to be measured by the ratio of the number of occurrences of a given type to the total number of all occurrences. Probabilities in statistics are usually considered in terms of their mathematically continuous distributions, such as a binomial or standard distribution. Thus data in statistical analyses is characterized by its named or approximated probability distribution.

To measure the (Shannon) Information in each variable (column) of a dataset, the probability of occurrences of identical values in that variable is computed and then normalized such that the probability of all values in that variable sums to one. This normalized value of probability for each value in each variable is the  $p_i$  used in equation (3) above. Then the product ( $p_i * \log p_i$ ) is computed replacing each value of data in each row and in each column of the data. The sum by column, for all rows in the dataset of  $m * n$  cells, is the (Shannon) Information (measure)  $H$  of each column or variable of data. Significantly, blanks (where no data is available) are treated naturally and appropriately as: no data available; missing data does not affect the probability  $p_i$  of values that do occur.

### **Application of Shannon Information measure to colon cancer data.**

Data characterizing the colon cancer tumors of 116 patients using 54676 genes with gene-expression values of about 9 decimal digits each was supplied by Dr. Tim Yeatman of Moffitt Cancer Institute of Tampa

Florida [5]. The analysis he desired was to predict survival of these patients beyond 36 months, as characterized by their separately known survival life in months. Moffitt's statistical analyses of this data produced correct predictions for about 65% of the patients.

The Shannon Information measure for each patient was computed using equation (3). The normalized probability  $\pi_i$  of each exact gene-expression value occurring among all 54676 genes in each patient's tumor was used to compute the incremental Information value for each of the 54676 genes. The sum of all 54676 incremental Information values per equation (1) is the Shannon Information for that patient. Figure 2 plots the Information measure  $H$  for all patients that survived and all those that did not survive the 36 month threshold.

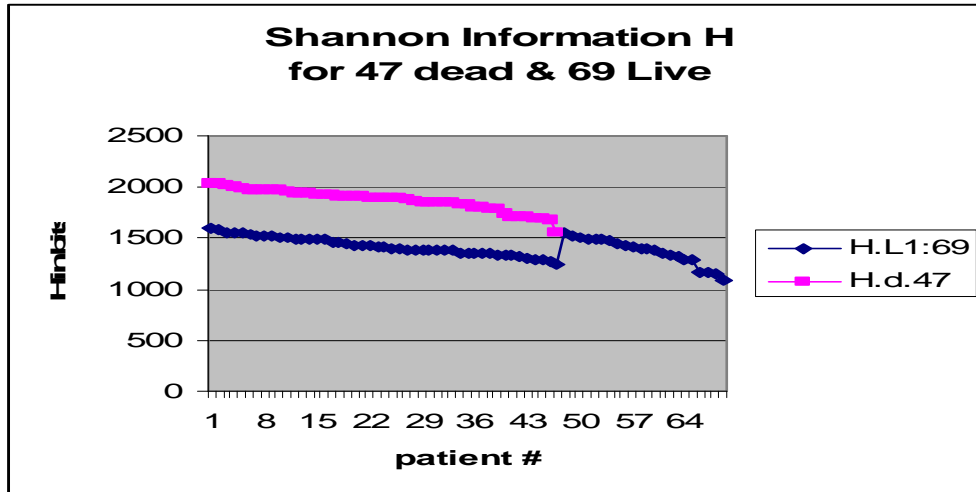


Figure 2. Shannon Information  $H$  for 116 colon cancer patients

One patient that died is incorrectly predicted by the Information measure to live beyond the 36 month threshold, an accuracy of 115/116 or better than 99%.

A second measure using incremental Shannon Information for each gene for each patient was computed as the Euclidean proximity of (each patient's gene) to the (sum of incremental Information for all 47 patients), for each group of 47 patients that survived and for each of the 47 patients who did not survive beyond 36 months. Equal sized 47 patient groups were used for summing incremental Information per group by gene to provide a comparable basis for proximity calculations since 47 patients died and 68 patients lived. Information-proximity was defined to be the square root of the sum of the squared differences above, summed over all 54676 genes. This Information-proximity measure made no errors in predicting survival beyond 36 months, a 100% accurate prediction.

### Line and Area Relationships

A scatter plot can be points filling an area (as in a cluster) or a line of points showing a 1:1 relationship between the values expressed by the axes of the plot. The shape of a line relationship defines the shape of the 1:1 relationship between the 2 variables being plotted.

An area or cluster relationship can be visually ineffective in that no 1:1 shaped relationship between the 2 variables is visually obvious. In statistical analyses, various mappings or rotations of the variables are attempted to separate the area covered by each cluster to enable associations with the groups of rows (i.e. subsets of patients) generated by subsets of rows. With vector-fusion, line relationships within clusters can be determined that allow proximity measures to determine nearest proximity to loci of subsets of grouped and identifiable vector-summed points that are associated with each group of rows. This allows both proximity-association and proximity-measured-probability ratings for nearest neighbor-to-locus calculations.

Relationships revealed as lines of any shape or profile are more useful and accurate than are area or cluster relationships. Line loci such as in Figure 2 (and Figure 5 below) where each line represents a visually separated grouping of patients or rows, are desired.

## Normalizations

In order to equalize the effect of significantly different value sizes in different variables, various techniques are used to normalize or standardize data prior to analysis. Other issues are usually managed in the standardization process such as how to treat missing data, how to treat alphanumeric data, and how to analyze binary (0 / 1 valued) data. Each kind of normalization affects some of the basic measures used in statistical analyses where varying meanings can apply to missing data for example. Normalizing the data values in each variable so that the range of all values in each variable fits between 0 and 1 for example changes the variance or standard deviation measure of each raw variable. Different kinds of normalization can change the shape of the relationships between the variables. When normalizing data, knowledge of the meanings of the data and the effect of normalization is important because the choices can have significant effects on results.

Different normalizations also can affect the shape of visually detected relationships discovered using vector-fusion. For both statistical analyses and vector-fusion analyses, considerable care and effort is appropriate when normalizing data. In vector-fusion the point of normalization is first, to equalize the impact or influence of each variable in the vector-summation process, and second, to allow the use of vector-fused resultants from experiment to experiment (or oil well to oil well in different basins) for example. Statistical differences based on medians measured in different experiments are tricky to apply, especially to data among unsynchronized experiments.

Shannon Information analyses however are independent of issues such as missing data, the impact of relative range or values among the variables, and alphanumeric or Boolean representations, because the probability calculation works exactly with whatever name (and value) appears in each cell and the probability is normalized to sum to 1, for each variable. If data is missing, whatever the missing data or its value, it is missing and not counted. The probabilities  $p_i$  are normalized to sum to one for only the values that are present in each variable. The actual name of the data and its value is used in the probability calculation.

## Significance of Synchronized data.

Data synchronized in time or phase is often **not** captured in data mining and learning applications. To understand the significance of synchronized data, two cardioids as shown in Figure 3 of different diameters and orientations are analyzed using vector-fusion with synchronized and then unsynchronized data

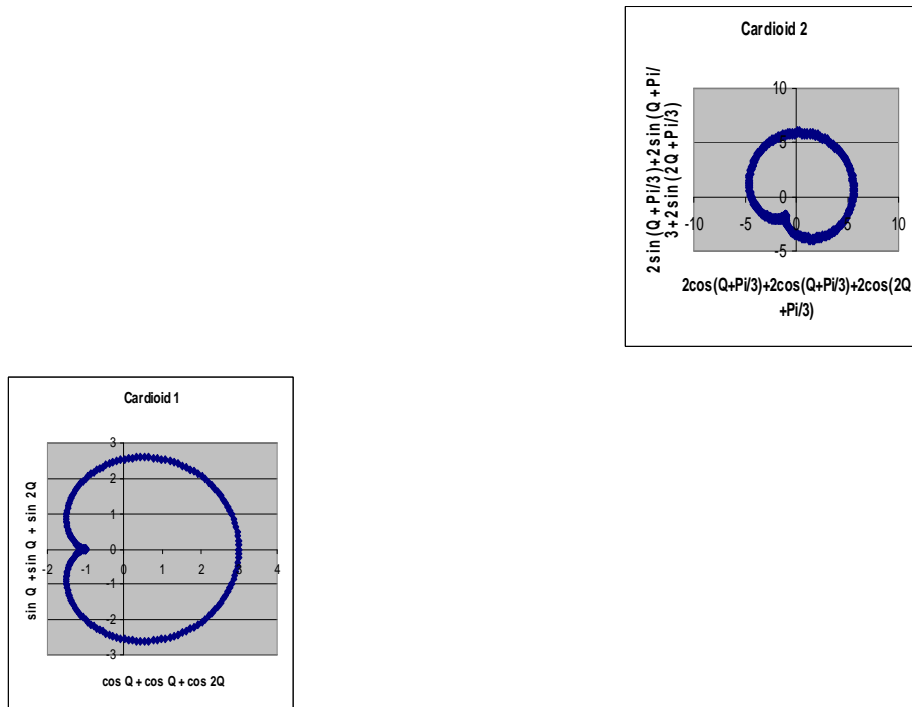


Figure 3. Two different cardioids generated with synchronized data.

The vector-fused resultant of the 4D synchronized data used to generate these 2 cardioids is given in Figure 4.

Because it is a linear transformation, vector-fusion (the vector-sum) preserves the curvature of component geometrical relationships via the unique phase-angle assignments  $\Theta_1$  (Equation [3] ) when vectorizing each of the multivariate values. The single cardioid of Figure 4 is the vector-sum resultant of the 4 component vectors used in the vector-fusion process when the data is synchronized.

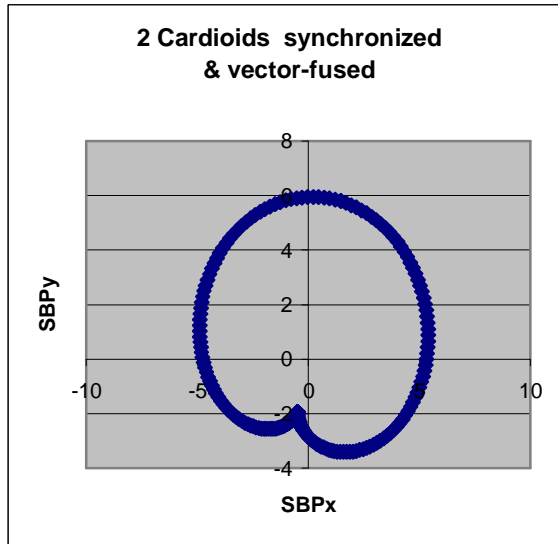


Figure 4. Vector-fused resultant of 2 cardioids with synchronized data

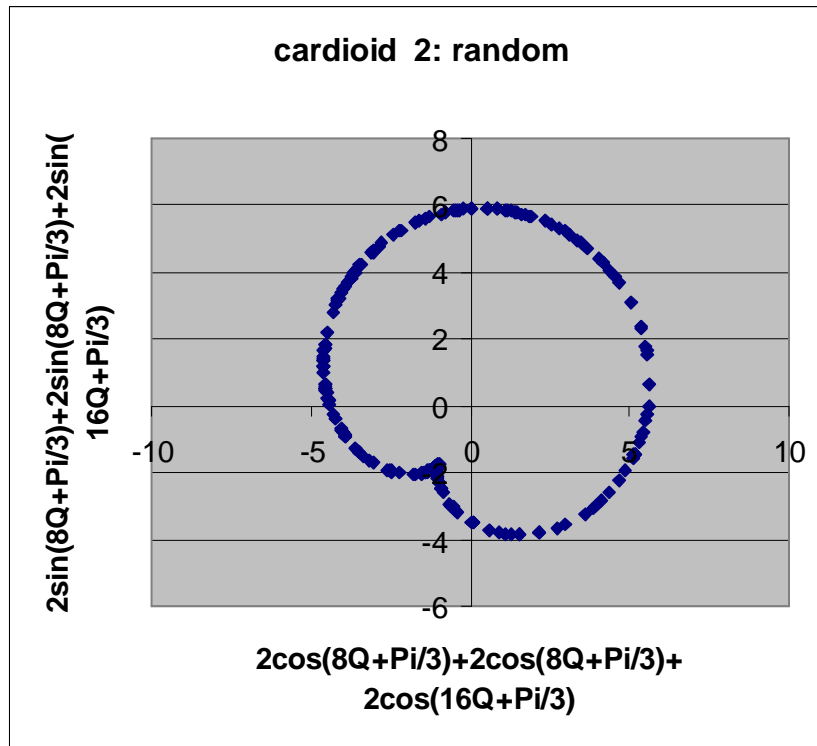


Figure 6. Cardioid 2 generated with unsynchronized (randomly assigned) data

When otherwise identical Cardioid 2 is generated with *unsynchronized* data and vector-fused with the original Cardioid 1, the composite vector-fused resultant is shown in Figure 6.

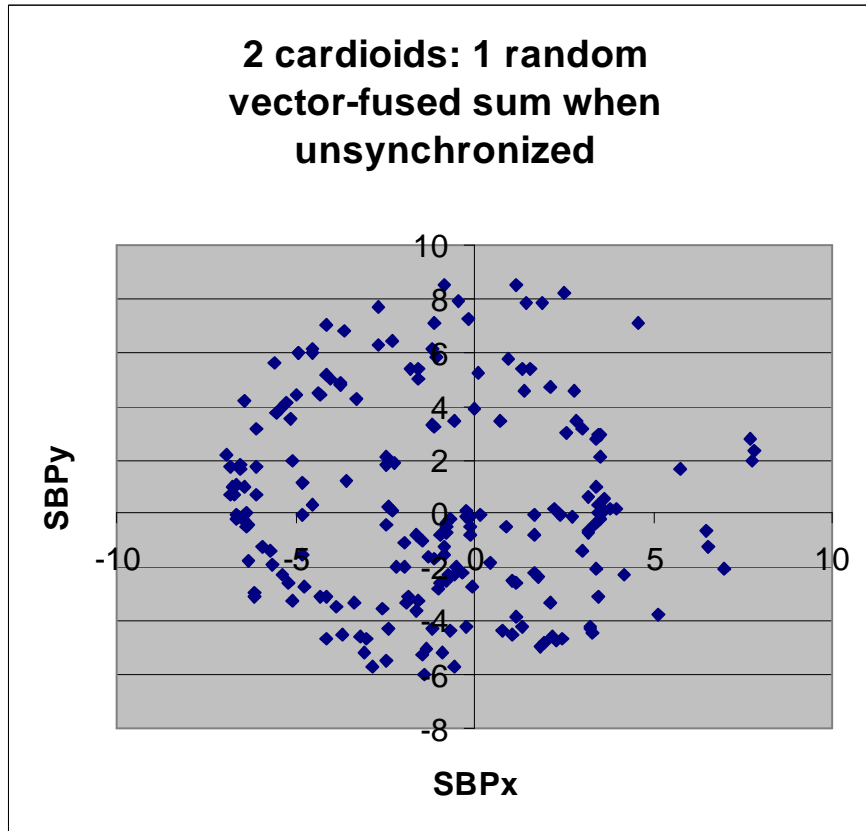


Figure 7. Vector-fused resultant of same 2 Cardioids using unsynchronized data.

Clearly, even with vector-fusion, unsynchronized data for the two known geometric shapes generates a cloud of vector-sum resultants that do not have a particularly useful shape. The cloud of points in Figure 7 is typical of the clouds generated and manipulated with statistical tools which are insensitive to synchronized or unsynchronized data. In statistical analyses, the center-of-mass of a cloud such as in Figure 7 is located; proximity of other cloud centers, or other points, to the cloud-centers is the basis for the approximations computed with statistical measures. In contrast, the locus in Figure 4 is the precise, mathematically interpolatable, extrapolatable locus describing the complete 1:1 relationship of the vector-sum resultant of the 4-variable points of synchronized data in that vector-fused (display) space.

Vector-fusion applied to sample biological data, Figure 5.

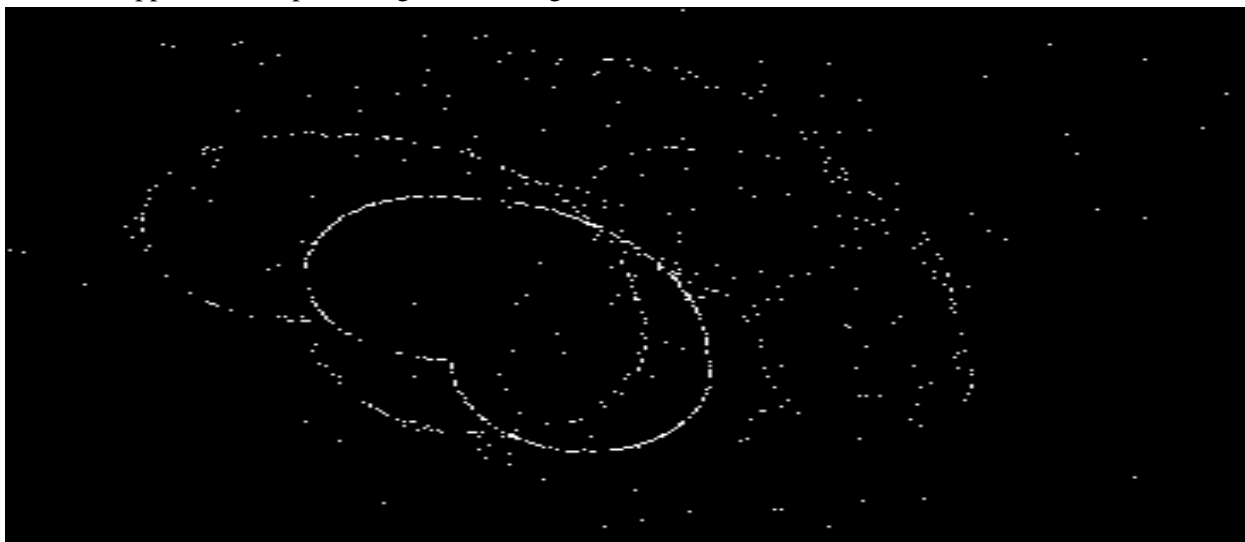


Figure 5. Families of cardioid loci. Each point is the vector-sum resultant of 746 Dimensional data.

There are 746 proteins (variables) in the pharynx of *c.elegans* [6]. The matrix of binary data that describes which of those proteins interact at least once with which other protein is a matrix 746 columns x 746 rows of ones and zeros. This data is synchronized mathematically with the protein interaction relationship. Each protein  $i$  is represented at its specific unique phase angle  $\theta_i$  assigned that protein. Using a 3D display space mapping of the 746 vector-fused resultants of this interconnection matrix, families of specifically interacting proteins appear as the cardioids in Figure 5.

Cardioids have an interesting relationship to helices. A helix (as in DNA which generates proteins) is the locus of a point on a circle as the circle rotates and moves orthogonally to the plane of the circle. A cardioid is the locus of a point on a circle that rotates around a second equal circle. If those 2 circles move orthogonally to their plane, an offset helix around the central circle is generated. Cardioids are potentially interesting in biological studies of proteins because of this offset helical property. Thus the geometry of the cardioid is used in Figures 3:7 to illustrate the importance of vector-fusion on synchronized data to capture the curvature properties of the component variables and identify the cardioid relationship among very many composite cardioid relationships.

Unique phase angle assignments with vector-fusion capture the extrinsic properties of each dimension, such as the curvature of individual geometries in multi-variable geometric studies [7]. Again, vector-fusion is a linear, precise analytic technique. Deeper understanding of vector-fusion is available [7] for background details of how this analytical technique works, and for additional illustrations of relationships discovered using vector-fusion and Shannon Information in biomedical and other applications. Appendix A is a further illustration of cardioid data to illustrate the data itself and the relationships among the component cardioids and their vector-sums.

## Inferences

The small margin in Shannon Information separating the closest survivor in Figure 2 could imply that the contribution of the very many genes that appear only once (typically at least 53,000 of them) have at least as significant an impact on survival as the genes whose values appear repeatedly. The idea that there is one (or maybe a few) genes that determine an outcome (survival, or disease sensitivity, or recovery, etc.) is subject to review in that the numeric effect of genes with large probabilities of occurrence might not sum to as large an Information measure as do the very many genes with values that occur only once. Might it just be that the pattern of all genes is as important to an outcome as the contribution of just a few genes?

Alternately, the common belief that many environmental factors (variables) other than an individual's genetic makeup, such as diet and stress and exercise, influence outcomes implies that genetic data may not itself be sufficient to accurately predict outcomes. A means to determine if genetic data is stationary or mathematically ergodic would seem to be an important new tool as data is generated and progress is made in haplotyping and in genetic understanding.

What appears relevant is collecting and analyzing all kinds of clinical, environmental, and genetic data to determine the relationships among all factors that influence outcomes of interest in biomedicine.

Tools for characterizing and measuring data are of course applicable to data derived in all disciplines of interest to mankind.

Characterizations and Measures of each of class of database, and then of the set of databases of clinical, environmental, and genetic data, would appear important in developing an overall model sufficient to make accurate predictions of outcomes. The analytical tools described here would seem to be an important addition to the arsenal for understanding the relationships and implications of the myriad datasets that appear significant in biomedical and in all other disciplines of interest.

## Contemplated business relationships by discipline

The power of analytical data mining for, and the limitations on the inferences appropriate to understanding the relationships derived from data in, each discipline that generates data to be analyzed imply that applications be developed for each discipline. The intention is to publish the techniques that have been generated for Characterizing and Measuring data once sufficient evidence and experience has shown these techniques are meaningful. The further intention then is to develop proprietary applications of specific tools for services and

analyses deemed important in applications of interest as joint ventures with specialists in each application.

Data of interest to people with specific applicational interests and skills is solicited as exploratory analyses are performed to establish credibility. Once credibility is established, mutual understanding needs to be developed to define the strategy and tactics to be applied to developing the joint venture activities appropriate to using and distributing specific knowledge, tools, and services for each application.

### **Conclusions: Relationships discovered by Characterizing and Measuring multi-variate data**

The Shannon Information measure appears to be a powerful fresh way to discover relationships in data. The particular relationship discovered in the large colon cancer data using Information in a Supervised Learning application, identifies patients by their gene-expression data and predicts a much more accurate survival rate for individual patients than was done by statistical methods.

Vector-fusion is a powerful way to discover relationships otherwise hidden in data of any size and complexity especially in Unsupervised Learning applications such as that of the 746 interacting proteins in a worm of interest.

In all Learning applications, where one is interested in discovering and measuring relationships in data, generating and using synchronized data is important.

Other interesting relationships have been and are expected to be discovered using vector-fusion and Shannon Information on data in biology and science in general. The Information measure is a powerful new way to characterize data for both Supervised and Unsupervised Learning applications in discovering relationships in data of any number of variables, and size.

Applications with data and an interest in understanding that data are solicited. Contact can be via web site [7] or directly with Prof. Johnson at [john97john@aol.com](mailto:john97john@aol.com)

Accuracy on data is at least a function of the ergodicity of that data, i.e. is the noise in that data stationary? The Shannon Information measure may also turn out to be a gauge for testing ergodicity of data using reproduceability as the criterion for ergodicity.

Being able to use and see all of the data and its incremental Information, in full (or limited) variability and size, and being able to use and see the vector-fused data resultants of all of the n-variable data is a good start to understanding "the data". Then using all of the characterizations and measures of the data, the  $p_i$  of each datum and the incremental-Information values ( $p_i * \log p_i$ ), the vector-fused patterns of the  $p_i$  and of the incremental-Information in all of its tensor dimensionalities, one can expect to provide useful insights on data in all disciplines. After seeing all of the data and its incremental Information, one should be able to select structure or relationships of interest from the visualized data for detailed visualization and analysis. Selecting subsets of data by  $p_i$ , incremental information, and Shannon Information for further visualization and analysis by patterns in all tensor-dimensions, via vector-fusion, is expected to be a deep resource for future data mining and analysis.

RR Johnson  
nDV

### **References**

- [1] Cherkassky, V., *Learning from Data*, Wiley, 1998.
- [2] Johnson, R.R., Visualization of Multi-dimensional Data with Vector Fusion, *IEEE Proceedings Visualization 2000*, Oct.2000, p. 297-302 & 570
- [3] Shannon, C.E., A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol. XXVII, No. 3., 379-423, 1948
- [4]. Papaioannon, P.C., Kempthorne, D., On Statistical Information Theory and Related Measures of Information, *Aerospace Research Laboratories*, 71-0059, March 1971
- [5] Yeatman, T., Eschrich, S., Colorectal Cancer data, Moffitt Cancer Research Institute, Tampa, FL
- [6]. Mancos, S., Interaction Data of 746 proteins in *c.elegans*, Huntsman Cancer Institute, Univ. Utah
- [7] web site for nDV: [www.n-dv.com](http://www.n-dv.com) Research papers

### **Appendix A. Significance of Synchronized Data**

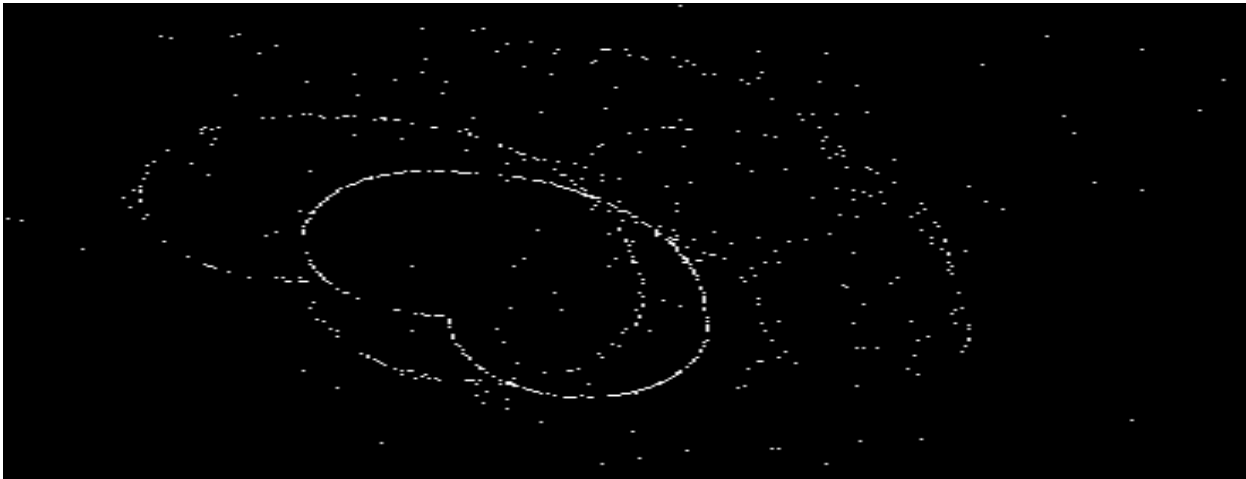
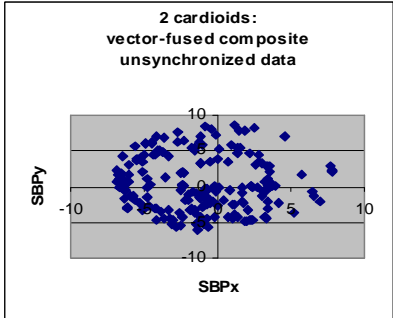
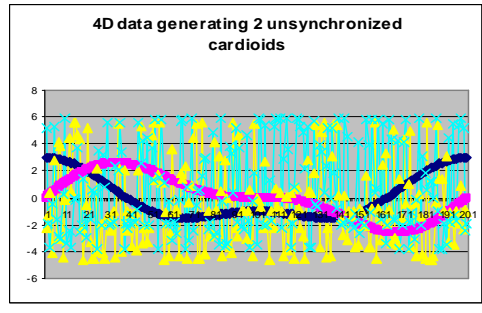
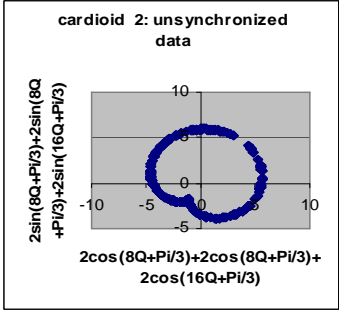
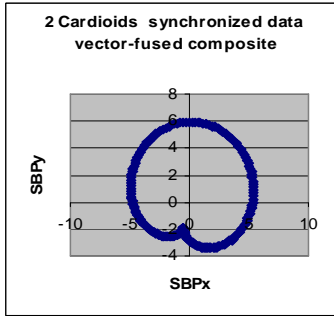
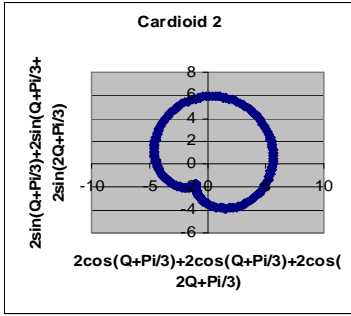
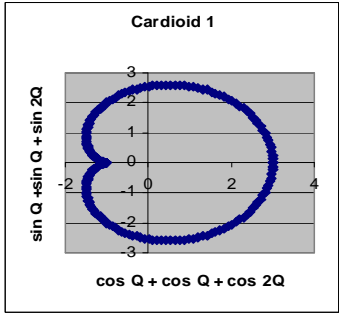
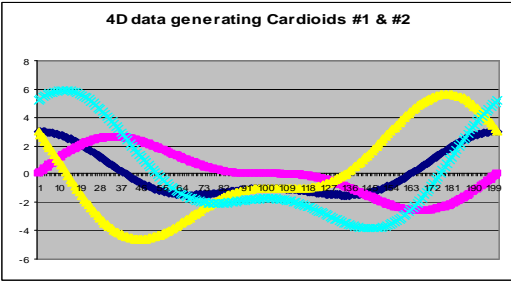
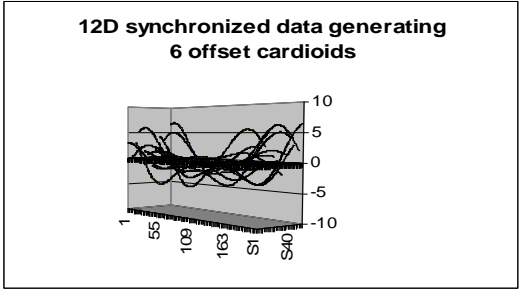


Figure 5. Families of cardioid loci. Each point is the vector-sum result of 746 Dimensional data.