A New Perspective of Geometry and Space as an Evolutionary Organizer of Data.

Gideon Samid
Department of Electrical Engineering and Computer Science
Case Western Reserve University, Cleveland, OH
BitMint, LLC
Gideon@BitMint.com

Abstract: Space and geometry are such a profound 'given' for our Darwinian brain that the developers of non-Euclidean geometry were reluctant to expose their insight, afraid of shame and ridicule. Scientists are still reeling from the implications of Einstein's General Relativity where the Newtonian firm space is reshaped by the substance it carries. Quantum mechanical issues of dimension-less particles, wave-corpuscular duality, and entanglement also suggest that the well-known, intimately familiar geometry is reaching its limit. In general we live in the information age: man and machine face avalanche of data to be inferred upon. This requires an effective way to organize data, which is what a good geometry will do. So we call for a fresh look upon geometry as a utility function to allow a myriad of items to be placed in good order, on proper scaffolding (geometry) to be sorted out so that they create a humanly manageable whole. We present a proposed principle with which to explore building a more useful geometry to hang our substance of interest on. Our proposal is based on the notion of *distance* as the basis of geometry (as opposed to point, line, etc.). We define distances between objects. Then we propose an estimating principle that says that the value of a property of one entity, is the best estimate for the value of the same property for another entity, and the validity of such an estimate is inversely related to the distance between the entities. We then use the validity of such estimates as a metric to sort out distance options (namely: geometries) and in particular we seek a geometry, which maximizes the validity of cross-estimates among the entities, which are fitted into this geometry. It is a broad-brush display, more to come.

1.0 A Utility View of Geometry

Motion, action, dynamics are defined over a space, a geometry. Physics was for centuries locked into Euclidean geometry, until Einstein explained the mystery of gravity through a non-Euclidean space. Quantum entanglement, the body-wave duality, and various singularity paradoxes may be reflective of our present geometry coming under terminal stress. Let us then revisit the very notion of the framework of physical science. A geometry may be likened to scaffoldings where objects are fitted and placed. We are used to treating it as a 'given'. This is because our evolutionary brain has developed a spatial conception to interpret biological sensory input. And this conception is so deeply ingrained that it comes across as what is really there, not just what we perceive to be there.

Realizing how our perception of space emerges from the happenstance of Darwinian evolution, one is tempted to consider the mechanism that established our sense of space and geometry. It may be that the sense of space is the evolutionary answer to the challenge of the sensory information inflow from the senses. The brain needed a framework to "hang the data on" and so emerged the familiar 3D world around us. If so, then we may abstract this process. We may view space, geometry as means to handle the avalanche of information with which we are bombarded. And if so then we may ask ourselves what is the most useful geometry to place our objects of interest in?

To that end let us build a model of science and physics: we identify objects, we fit them in a geometry, and then we develop estimating algorithms to estimate an unknown situation. We have then a set of objects, O, a geometry, G, and estimates E. Classically we treated O and G as given, and we put all our ingenuity in figuring out E so that we

achieve harmony between G-O-E. It would be interesting now to use an "equal opportunity" approach: to try and mold all the ingredients of the desired harmony for the purpose of achieving it.

We start by taking on geometry, G. Given a set of objects, O, given some estimating algorithms, E, what is the geometry that would achieve the desired G-O-E harmony?

To formalize: we define a set O comprised of o' objects $o_1, o_2,....o_{o'}$. Each object is associated with p' properties $p_1, p_2,....p_{p'}$.

We now analyze the concept of estimates. We focus on one property, p. We say the following: suppose that we know not the value of p for any of the o' objects in O. We do know though that p has a range r, namely p may assume the values: $v_1, v_2, ... v_r$. We shall then say that the probability for every object to have a p value v_i is 1/r for i=1,2,...r. Simply because we don't have any grounds to differentiate one value of p over any other.

Now let us examine the following case. For a given object in O the value of property p is known: v. The values of property p for all other objects are not known. We now ask ourselves what will be our best estimate for the other (o'-1) values.

The 'sterile' answer is that the information about the value of p for one object has no bearing on the estimate of the value of p for any other object, and so the most that can be said is that each object has a 1/r probability with respect to any of the r possible values of p. While this logic is reasonable, it is by no means the only way to think about this situation. We propose to introduce the causality assumption. Namely that there is a cause that effected the value of p for that object to be v. And that cause is our best estimate as to the value of p for the other objects. We know that the value v is feasible, and is desirable by that cause. We don't know that with respect to all the other r-1 possible values for p. This will lead one to conclude that the probability for the value of p for any

other object in O to be v is ever so slightly higher than 1/r. And if so, then given what we know, our best estimate for the value of p for every object in O should be v.

This "sameness assumption" is valid even if one does not know r, nor what are the other possible values for p.

We now describe a different case: Again, we have the object set O, only that this time there are two objects, calls them o_1 and o_2 for which we know the value of some property p: v_1 and v_2 respectively. If $v_1 = v_2$ then the above "sameness logic" applies, and all the objects in O will be estimates with the same value for p.

But what do we do for the case $v_1 \neq v_2$?

It is the answer to this question which is where geometry comes into play. Euclid based his geometry on points and lines. We may take another route: use the notion of 'distance' as the foundation of geometry.

2.0 The Notion of Distance

Conceptually distance implies a measure of being apart, of being different in some way, say location, or any other property. Given three objects o_1, o_2, o_3 for which the value of some property p are respectively v_1, v_2, v_3, if $v_1 = v_2 \neq v_3$ we shall intuitively consider the distance between o_1 and o_2 (d_{12}) to be smaller than the distance between o_1 and o_3 (d_{13}): $d_{12} << d_{13}$. In general we see similar objects as forming a cluster, and removed from fundamentally different objects. It stands to reason that the notion of distance is more fundamental, and earlier than the Euclidean notions of point, line, and planes.

If we measure the depth of a lake at one point in its middle, and read a measurement m. Now if we are asked to estimate the depth of the lake in any other spot in the middle

of the lake, we probably estimate m because we have no other anchor, and n reason to assume that the depth at the new location will be any higher or nay lower than m. But then when we are asked again: how confident are we about your estimate? Then our answer will depend on the distance between the new point and the m point. If the points are close to each other we will be more confident of the estimate.

Let the three objects above be points in the described lake, and let us be unaware of the depth of the lake at point 2 (o_2). If only knew the depth in point 1 we would have estimated $v_2=v_1$. If we knew only the depth at point 3, we would have estimated $v_2=v_3$. Now how do we make use of the double knowledge: knowing both v_1, and v_3? If points 1 and 2 are close to each other and far away from point 3 then we would like in some way to give v_1 a greater role, impact, in estimating v_2, relative to v_3.

In other words we propose to view distance as a measure of waning confidence. To say the value of A is x, therefore the value of B is also X (the sameness assumption) is of lower validity when the distance between A nd B is large.

We propose hence to answer the question of estimating values for some property p for objects in an object set, O, by the sameness principle. Estimated will be extended from the objects for which we have knowledge. Arising conflicts among the various estimates (by different sources -- different objects) will have to be sorted out in some mathematical way which accounts for the respective distances from the estimated object to the various estimating objects. And the engaged formula will have to express the principle of waning confidence over longer distances.

If the objects in O are already situated in a given space, in a geometry, and have mutual distances well defined, then this process is straight forward. That is the case with the example of the points in the lake.

When we do so, we may eventually compare our estimates to the estimated values, and come up with a *hit-ratio*, say *success ratio*. It may be that by and large we estimated well most of the properties we tried to estimate, or it may be the other way around.

3.0 Substance Geometry Harmony

We can also regard the existing geometry, G, which holds O (and determines the mutual distances between its elements), as one option out of many. We may search then for a better geometry -- with different distances. Perhaps distances that will lead to a higher estimating hit rate.

Namely, given a distance-based estimation algorithm E, then a set of objects, O, will dictate its 'most harmonious geometry' – the geometry with the highest hit-ratio.

We can then analyze the objects in O with respect to any number t of properties p_1, p_2,...p_t, and for each property build a geometry (a set of mutual distances) to make it harmonious with O. Mapping this to the common notion of multi dimensional space, we can say that every property of the objects in O can be fitted into a set of distances between the objects, such that the hit-ratio will be high. This will be viewed as a high harmony setup. Say then that every property of the O objects will match a dimension. If N properties of the objects are so analyzed then O will be fitted into a most harmonious N dimensional space.

4.0 Disharmony, and Problem Solving Strategy

We can recruit the hit-ratio for the estimates as the metric to indicate degree of harmony between the objects and their geometry. So a very poor hit-ratio will be

interpreted as a poor harmony between O and G, or regard G as a poor fit to carry O. Poor hit ratio indicates chaos, disorder, lack of pattern.

This notion suggests that the way to research any subject is to first express it as an object set O, and then to search for a most harmonious geometry G in which to fit it. One would expect for new insights to surface simply on account of a fitting geometry. Say, to solve a problem, first find its fitting geometry.

5.0 Point Less Geometry

Euclid's fundamental object is the "point" which is described by what it lacks (no length, no breadth, no height). While the physical notion of a point is very familiar and easy to relate to, the abstract notion of the same is an imagination stressor, surrounded by essential vagueness, which then projects to the derived notions of lines, planes, and three dimensional shapes. It is this cloud around the notion of a point that gives rise to irrational numbers, which in turn led to the paradox of having one infinite set being larger in size than another infinite set -- neither of these concepts has any life, or existence beyond the 'theology of math'. Say then that from a practical, from an engineering standpoint, the Euclidean point is problematic, and it may be of interest to by pass it, and define a geometry without it. The geometry described above does not make use of the definition of an Euclidean point, and is not bound to a drawing or a figure. But it can be super imposed on Euclidean or non Euclidean geometries at any desired degree of fitness.

The geometry described above does not make use of the definition of an Euclidean point, and is not bound to a drawing or a figure. But it can be super imposed on Euclidean or non Euclidean geometries at any desired degree of fitness.

The degree of abstraction used here raises the daring possibility that this approach to geometry may prove to be a productive research to by-pass the prevailing difficulties in sub-atomic quantum physics with respect to points with no size, locality, continuity, entanglement. It offers a generalization of Einstein general relativity in as much as it claims that the idea of interpreting space as curved to explain gravity is but a special case of the idea that space, geometry are shaped to simplify complexities and provoke insights.

Our senses can only deliver to us a curved space if it is limited to two dimensions. For higher dimensions, we rely on mathematical abstraction and logic. We don't really experience it. We are convinced of space curvature as a given because, as Einstein shows, it explains the mystery that baffled Newton: how does gravity work? Yet, a curved space is a special case of the harmony set described herein. The lesson from the theory of relativity is that geometry is not an empty firm container of things which is prescribed to us by the objective reality. Geometry is our tool, for us to shape.

6.0 A More Formal Presentation

We consider a set O comprised of o objects. Each object may be associated with t properties: p_1, p_2,....p_t. Let $N \in \mathbf{N}$. We define a set of No^2 variables to be called "distances" over O as follows: Let o_i, and o_j be arbitrary two objects in O. This pair will be associated with N types of distances: d^1_{ij} , d^2_{ij}.... d^N_{ij}.

For i=j we shall set $d^k_{ij}=0$ for every k=1,2,....N. The case where there is no direct distance of type k between objects i and j, we shall writeL $d^k_{ij}=\infty$

We agree to denote the set of No^2 distances as the "geometry" (G) of the set O. We also agree to regard all the distances of type k: d^k_{ij} as a distance of "dimension k". The

geometry G is defined via N dimensions. We also agree to call the combined O+G system as the Harmony Set: H=O+G.

Let "q" denote a set of dimensions. Let us define a "trip" (Tr^q_{ij}) as a path that leads from element i to element j, traversing through element k_1, k_2,.... k_t in between. "Traversal" is an effort exerted between two objects, and it is a ssociated with "cost" represented by the distance between the two objects. We shall define the "trip distance" between two elements i and j in $d\left(Tr^q_{ik_1k_2\ldots k_tj}\right)$as the sum distances traversed between i, to k_1, plus the distance traversed from k_1 to k_2, plus the distance from k_2 to k_3:... + the distance from k_t to j, where all the distances belong to dimensions listed in q.

$d\left(Tr^q_{ik_1k_2\ldots k_tj}\right) = d^q_{ik1} + d^q_{k1k2} + \ldots d^q_{ktkt} + d^q_{ktj}$

Let us define the smallest trip distance between i and j as δ^q_{ij}:

$\delta^q_{ij} = \text{Min}(d\left(Tr^q_{ik_1k_2\ldots k_tj}\right))$.

We shall continue our analysis over a simplified Harmony sets

6.1 A Simple Harmony Set

We consider now a harmony set (O+G) defined over a single dimension (N=1), and regarding a single property, p, defined over the o elements in O.

We introduce the **fundamental premise of the harmony set:** Given that the value of property p for some element i in O is v_i, then the best estimate for the value of property p for all other elements in O is $v_j = v_i$, for j=1,2,....o

We now distinguish between a ranked property p and an unranked property. First for unranked p:

We consider the case where for only $t < o$ elements in O, we know the value of property p. We regard these elements as the T estimation set. We wish to set forth our best estimate for an object i which is not included in the t known elements (not included in T). The fundamental premise of the harmony set as is does not help up because we may have $t > 1$ conflicting estimates. We now invoke **The Estimate sorting premise for unranked properties,** which says: the estimate of value of property p for some element o_i not included in T will be the value of property p for element $o_j \in T$, where δ_{ij} is the minimum, and if there is another element $o_k \in T$ such that $\delta_{ij} = \delta_{ik}$ then no estimate is being issued.

Given O+G and T, it is possible to apply the estimating principles set forth above, E, and issue estimates as to the values of property p over the (o-t) objects not included in T. Having done so, it is straight forward to evaluate a "success rate" or "hit-rate" (HR) defined as the ratio of correct estimates to the total estimates. Clearly $0 \leq HR \leq 1$. We write:

$$\boldsymbol{G+O+E+T \rightarrow HR}$$

There are C^t_o different combinations of t elements in O. For each of these combinations we can compute an HR as defined above, and then compute their average:

$$HRA = \frac{\sum_{i=1}^{i=C_o^t} HR_i}{C_o^t}$$

We can write: HRA =f(G,O,E,t). For any given O and t there exists one or more geometries G for which HRA is maximum. It is easy to see that for some (O,G,t) sets $HRA_{max} = 0$ and for some $HRA_{max}=1$. If each element in O has a different value for

property p then HRA_{max}=0, and if all the elements in O have the same value for p then HRA_{max}=1. Also if the range of values for p, r is larger then o (o < r) then at least two elements in O share a value for p and hence $HRA_{max} > 0$.

Practically, computing HRA per the above formula may be extremely tedious, and it can therefore be simulated via Monte Carlo technique.

A set {O,G,E,t} for which $HRA_{max} = 1$ will be defined as a perfect harmony set of degree t. And a set {O,G,t} for which HRA_{max}=0 will be defined as a perfect disharmony set for degree t. We then measure the harmony of a set of degree t by the value of its HRA_{max} value. The closer it is to one, the greater its harmony.

The question that rises now is: given an object set O, how to find the geometry G that would yield the maximum HRA possible under that O, at degree t.

It is believed that such a geometry is the most convenient geometry to handle O, and also the one that would yield more insight into O. If O represents any topic of research or investigation, or any organized data to infer upon, then it is likely that some, or even many of the properties of its elements will not be known, and the higher the HRA, the greater the chance that the unknown variables can be correctly estimated from the known variables.

For the case where the values of property t are rankable and hence it makes sense to average, then round up, a set of conflicting estimates, we may agree then on any averaging algorithm that would take into account the weight of each estimate according to the fundamental premise of harmonic geometry.

Illustration: Let's consider O to be comprised of six natural numbers:{33,5,16,9,12,22} Let us consider the property of MOD 3 for each number, so the values of this property for the six elements is Property MOD 3: {0,2,1,0,0,1}. Now let's fit these six numbers in a two dimensional geometry:

0	2	1
0	0	1

Let's agree that any two elements which share an edge will have a mutual distance of 1, all others will have no direct distance (distance of infinity). Let us now use the estimating procedure, E to estimate the designated property value of element 6 (rightmost on the bottom line). Accounting for the distance in an inverse, we write:

Element	Value	δ	weight
1	0	3	1/3
2	2	2	1/2
3	1	1	1
4	0	2	1/2
5	0	1	1

Which compute to an estimate of "0", which is incorrect. Considering another property: MOD 2, the values are {1,1,0,1,0,0}, using the same geometry the probability to estimate the parity of the 6^{th} element as odd is: 1/3+1/2+1/2= 1.33, and the probability for parity of even is: 1+1=2, so the estimate will correct: "0".

6.2 Mapping A Harmony Set into Euclidean Geometry

Considering a ranked property p with a range r equal to an arbitrary number of points on a Euclidean line segment, and with L as its lowest value, and H as its highest value. Given two objects one for which the value of this property is L, and the other for which that value is H, it is simple to construct an object set with o'=r elements, such that each will have a different value for p. We begin with the given two elements, and then establish a new element X such that its distance from its two constructing elements will

be the same: $d_{xl} = d_{xh}$. We shall repeat this process between X and L, and between X and H, and so on. The estimating rule will be that only the two closest elements to a given element participate in estimating the value of p. This will clearly apply to all the r elements in the constructed object set, O. Clearly O maps well onto a Euclidean straight line. However the harmony set may construct any number of parallel settings of elements. Much as element X was constructed and 'geometrized', so one can construct another element $X' \neq X$, where too: $d_{x'l} = d_{x'h}$. The distance between X and X' can be set at will, perhaps no direct distance: $d_{x,x'} = \infty$

A function y=f(x) is drawn as a curve on a 2D Euclidean plane, with x as its horizontal axis, and y as its vertical axis. One can set the x axis as indicator of distance (geometry), and the y axis as indicator of the value of some property 'f'. The harmony set will express this situation by setting the distances between three consecutive elements, (i-1),i,(i+1): $d_{i-1,i}$, $d_{i,i+1}$ to fit into the proportion:

$$\frac{d_{i-1,i}}{d_{i,i+1}} = \frac{f(x_{i+1}) - f(x_i)}{f(x_i) - f(x_{i-1})}$$

to insure that the harmony set estimate, including the two closest elements will be correct. The steeper f(x) becomes, the smaller become the distances between the elements. This can be extended to functions of more variables. Whatever is covered by analytic geometry is hence covered by a harmony set.

Clearly any Euclidean space, or a curved space may be mapped into a harmonized geometry. But it is not so in reverse. A Euclidean or curved "point" has definite distances to the other points in the space, which is what is needed for the harmony set with each point being an element of the set. Each such point may be associated with a scalar, a vector, or a tensor, defined as the properties of this point (element). And each such property is subject to the harmonized estimation procedure. Any harmony set which satisfies the metric conditions can be mapped into a Euclidean space, and any N-

dimensional harmony set where each property has distances which obeys the metric condition, may be mapped into an N-dimensional Euclidean space.

6.3 A Multi-Dimensional Harmony Set

We can repeat the base procedure over any number of properties, p_1, p_2, ...p_q by allocating a new dimension to each property. This will yield q HRA_{max} values: HRA_{1max}, HRA_{2max}, ... HRA_{qmax}. They can be averaged to HRA^q_{max} the value of which will determine the harmony of the set.

Another approach to multi-dimensional harmony may be a 'fusion' of the individual dimensions. For this approach one chooses N=1, namely one dimension for distances. Or say, every two elements, o_i and o_j in O will have only one distance d_{ij} between them. Now given a set Q of some q properties: p_1, p_2,....p_q, one would seek a geometry G that would maximize the hit-ratio which will be counted across all q properties. One would expect the maximal hit ratio for the fusion case to be much lower than the hit ratio for individual properties, so the resultant harmony will be less, but that may be compensated by allowing for hard-to-estimate properties to be a bit easier to estimate.

Multivariate analysis has become the foundation of BigData and AI. It is nominally handled via multi dimensional geometry where it faces a nagging challenge: how to proportion the various dimensions one to the other. Any arbitrary ratios will affect the inference, will redefine clustering etc. The 'fusion' option, by contrast, will operate with one shared set of distances for all variables, which should alleviate this contamination of the results.

Co-Factor Correlation is another widely used tool with established statistical constructs. It can be approached via the harmony set in two different ways: Let property p_1 be a well measured property, and let property p_2 be a poorly measured property of

interest. One wishes to establish the existence of correlation between these two properties, and then exploit it to deduce p_2 values from measured p_1 values. Correlation between these two properties will manifest itself by similarity of their optimized geometries. If $d^1_{ij} \sim d^2_{ij}$ where i,j =1,2,...o' then the geometry established for p_1 can be used to estimate missing p_2 values.

7.0 Extrapolation

By its essence the harmony estimates don't extrapolate. For unranked properties the estimates will be selected from the set of given values for each property, and for ranked properties the estimates will be limited by the range of the given values for each property.

This limitation can be overcome by recalling that any change of any variable (property, p) outside the domain of known values, can always be expressed by a derived property p* which expresses the constant rate of change for p (say from object to object). And if the rate of change is not constant then its rate of change is constant, and so on, at some point the deep derivative will be constant. And that derivative will be a proper property to analyze for the harmony set.

8.0 Applications

Living in the information age we are all bombarded by an enormous amount of information, which is only useful if it is organized so that it can have an impact on our decisions, conclusions, and actions. The idea of molding geometry to organize large bodies of data is promising on a very broad base. It may be applicable to the study of physics, and natural science in general. It can be used in intelligence analysis, economics, management, law enforcement -- avalanche of data is the style of modern activities. It is

hard to think of an area with no impact -- once (and if) these nascent ideas mature into useful tools.

Robotics, AI, and Big Data are intriguing areas of applications. To be useful these products must be able to absorb and digest large amounts of data at high speed. The general idea is to minimize the amount of arbitrary impositions on the inference process. A fixed geometry is certainly such an imposition. So to let the data flow in, analyze it in order to identify a most useful geometry to "hang it on" is an imposition-free alternative. Robots are not limited by the Euclidean geometry, we humans are locked into, by the Darwinian evolution we all share. So robots may view data and dimensionality with a fresh attitude, follow the data.

9.0 Conclusion

These few broad brush ideas are designed to provoke interest in the underlying notion of molding geometry to organize and put in useful ordre a given set of objects such that working with these objects will be more convenient, their study more insightful, and inference from their data more productive. This note essentially claims that we should view geometry as a moldable scaffolding, as a means to better express data to make it easy to handle, and make it more productive to infer upon.

Reference

Richard D. Hornung, Andrew M. Wissink, Scott R. Kohn "Managing complex data and geometry in parallel structured AMR applications" Engineering with ComputersDecember 2006, Volume 22, Issue 3–4, pp 181–195

Laurens van der Maaten, Geoffrey Hinton "Visualizing Data using t-SNE" JMLR 9(Nov):2579--2605, 2008

Hartshorne Robin, “Geometry: Euclid and Beyond” Springer, 2005

Mlodinow, Leonard “Euclid’s Wkindow” Simon & Schuster, 2001

Tufte, Edward R. The Visual Display of Quantitative Information, 1983, Cheshire, CT: Graphics Press.

David Fridovich-Keil, Erik Nelson, and Avideh Zakhor “AtomMap: A Probabilistic Amorphous 3D Map Representation for Robotics and Surface Reconstruction” University of Berkley http://www-video.eecs.berkeley.edu/papers/dfk/atom_map.pdf