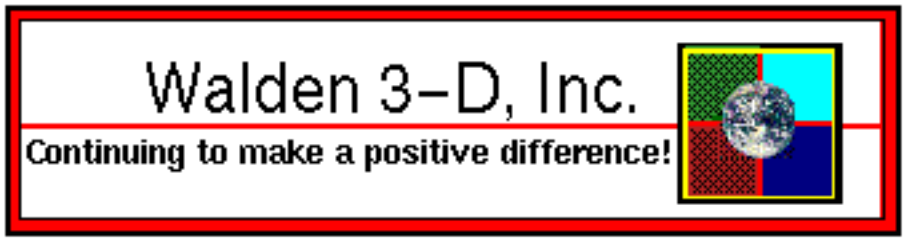

Welcome to Edition 13 of the Walden 3-D Journal.

Periodic publication of the W3D Journal will focus on science and technology as:

1. **comments on relevant news items;**
2. **observations on scientific and technology developments;**
3. **in-depth description of the pros and cons of relevant new products; or**
4. **a summary of a technical advance.**

Once every 12 editions a Virtual Seminar[SM], our name for a customized real-time web-enabled presentation, will be developed and presented to subscribers. The subject of the Virtual Seminar[SM] will be selected by subscribers from material in this set of Walden 3-D Journal Editions. Edited by H. Roice Nelson, Jr., some editions will be invited articles. Distribution is only to those who sign the Subscription Agreement (http://www.walden3d.com/journal/subscription.html).

As described in the Subscription Agreement, the Information in the W3D Journal is Confidential for 24 months from publication, is not to be used, licensed, published, or otherwise disclosed to any person not an employee, unless:

- **A. it can be shown to have been already known to the Receiving Party;**
- **B. it is already in the public domain;**
- **C. it is required to be disclosed under an applicable law;**
- **D. it is acquired independently from a third party who has the right to disseminate it; or**
- **E. it can be shown to have been independently developed by the Receiving Party.**

After 24 months, W3D Journal editions will be made available at walden3d.com.

The Information in the W3D Journal may be distributed to:

- **A. employees, officers, and directors of the Receiving Party;**
- **B. employees , officers, and directors of any Affiliated Companies; and**
- **C. any professional consulting or agent retained by the Receiving Party to establish technical feasibility or development of ideas**

as long as these persons or entities are bound under a confidentiality agreement.

The Receiving Party is free to use Information in the W3D Journal for Internal Projects, as long as the Information is treated the same as if it were Internal Proprietary Information.

W3D Journal Edition 13: Advanced Pattern Finding

Executive Summary

Walden 3-D's work in advanced pattern finding has been conceptual to data. This White Paper and the work being initiated by Dynamic Oil & Gas Corporation are the first efforts to apply these concepts. Bill Bavinger, who prior to his untimely death in a car accident in Oklahoma was a researcher at Rice University, developed the framework within which Walden 3-D and Dynamic plan to apply advanced pattern finding techniques. A key concept of this framework is that data is deterministic, in that it does speak, and it is a precursor to decision making (Bavinger, 14 December 1989).

A few years ago Walden brought together Bill Bavinger and Dr. Robert Ehrlich for an evening. Dr. Ehrlich has been applying advanced mathematical concepts to industrial problems, and specifically to geotechnical problems, for over three decades. His techniques were recognized as being much more sophisticated that the cluster analysis approaches Bavinger had used to develop his concepts. Dr. Ehrlich has retired from his research chair and with two colleagues has formed Residuum Energy, Inc. based out of Salt Lake City, Utah. Walden and Dynamic have reciprocal support agreements with Residuum, such that two or all three of the companies will come together as a virtual team to solve customer problems.

This White Paper is divided into five major sections, and a series of sub sections:

1. Data Collection
   - A. Collect Spatial Data
     - i. Collect Text
     - ii. Collect Points, Lines, Areas, and Volumes
   - B. Collect Temporal Data
     - i. Collect Temporal Text
     - ii. Collect Animations
     - iii. Collect Vectors, Ribbons, Flows, and Time-Lapse Data
     - iv. Collect Velocity and Acceleration Data
     - v. Collect Pressure and Temperature Data
2. The Bavinger Model for Pattern Finding
   - A. Sort Data
   - B. Region Growing
   - C. Cluster Analysis
   - D. Factor Analysis
   - E. Automated Self-Classification
3. Information Models
4. Five Somewhat Wild Scenarios
   - A. Text Pattern Finding Scenario
   - B. Numerical Pattern Finding Scenario
   - C. Spatial Pattern Finding Scenario
   - D. N-Dimensional Pattern Finding Scenario
   - E. Internet Classification Scenario
5. Acknowledgements and Next Steps

Introduction

Pattern finding is a key component of the Walden 3-D Design Process (see Edition 01). All the research in pattern finding was done by Bill Bavinger, who died in a tragic car accident in January of 1998. Walden has notes from discussions about pattern finding techniques and applications with Bill between 1989 and 1997. Walden has not yet applied these ideas. We expect there to be some successes and some failures.

The first significant application of using state-of-the-art pattern finding and data mining technologies will be through the Walden 3-D incubated company Dynamic Oil & Gas Corporation (see http://www.walden3d.com/dynamic). Dynamic is positioned to and intends to lead the next significant cycle of increasing hydrocarbon production by using advanced pattern finding technologies, From patterns found in raw data Dynamic expect to find new exploration concepts, identify leads (where and how to look for hydrocarbons), and define prospects (places to drill).

One of Bill Bavinger's concepts was that if we do not understand the process of our business, there is no core to the business (20 November 1989). Conversations along these lines led to the development of the Knowledge Backbone℠ (see W3D Edition 09). In one of the early Walden 3-D planning meetings, the complexity of advanced pattern finding was felt to be a barrier to general application. One of the participants stated "It doesn't matter how wonderful new technologies are, if they do not fit in the new world. How do you "dumb it down to bubba?" (Bill Atkin, 12 December 1989)

This Edition of the Walden 3-D Journal takes the ideas of Bill Bavinger and Dr. Robert Ehrlich of Residuum Energy, Inc. on advanced pattern finding and describes their probable value to business. Note that Dr. Ehrlich, has been successfully applying advanced pattern finding techniques in the geological sciences for over three decades. This White Paper also presents 5 scenarios, some of which are a bit futuristic, describing how advanced pattern finding technologies are most likely going to make a significant difference to society over the coming decades.

Data Collection

The future of things we find and build is based on information processes. Dynamic is focused on using information technologies to maximize the probability of finding new hydrocarbon reserves (see Edition 05: Dynamically Replenishing Hydrocarbon Reserves). Advanced Structures Incorporated (another Walden 3-D incubated company, see http://www.asidesign.com) is focused on using information technologies to design and build new space frame, tent, glass, and other advanced types of structures. Although these processes are completely independent at this stage, Walden anticipates there will prove to be considerable overlap, and hopefully a common spatial language derived over the coming months and years. Particularly as more data is collected and more patterns are found in the divergent areas of hydrocarbon exploration and production and the design of responsive urban and rural systems.

The data used in exploration and in building is spatial and comes in the same basic forms: text, points, lines, areas, and volumes. Basic units of temporal data are also common, namely animations, vectors, ribbons, flows, time-lapse volumes, velocity volumes, and acceleration volumes. In order to do advanced pattern finding, data defining these basic units of space and time are stored in relational database tables (see W3D Edition on Object Oriented Scene Graphs). Data, which is defined as instances of specific meanings occurring in the real world (see Figure 1), are the variables at each cell of a relational data base table. The cells form the patterns (Bavinger, 26 July 1991). Data does not have to be totally accurate to be useful. However, databases need concurrency, or in other words to have internal agreement. They also need updateability, and integrity. Data management can not be political. Yet it is important to realize enterprise planning provides a way to reconcile multiple vies of an object. One of Bavinger's goals for advanced pattern recognition is that user interface will be derived from the data. This means the widget is the knob, and vice versa (Bavinger, 14 December 1989).

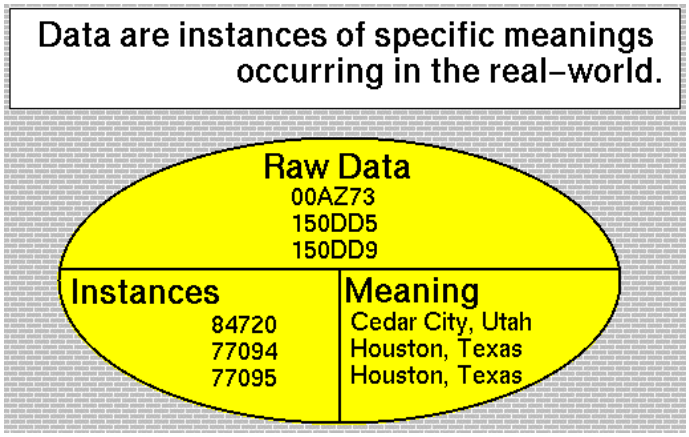

Figure 1. Definition of Data

There is a big difference between information and an image. Images are words, numbers, and algorithms. Adding process to images, defines the relationships between images. This combination of image and process forms a central design data type, or a meaning or a set of meanings, which are given to a set of computer variables (Bavinger, 14 December 1989).

Data types are the internal computer representations of entities humans interact with. For instance once the data types for ASCII characters were developed, they were mapped against characters and keyboards became data entry ports. The Infinite Grid℠ defines a one-word data type for spatial location (see W3D Edition 07). The Timedex℠ defines a one-word data type for temporal location (see W3D Edition 08). The Infinite Grid℠ defines a one-word data type for process (see W3D Edition 09).

The purpose of data collection is derivation of information, knowledge, and wisdom, each of which enables better decision making (Blaine Taylor, 1994). The elementary unit

of information is the difference (Bavinger, 15 March 1997). Information is defined as data in context, related to a specific purpose. Knowledge is the progressive gathering of bits of experience, along with the links which associate these disparate parts into a unified whole. Information is cumulative data, while knowledge is cumulative experience. The capture and dissemination of data maximizes information. The capture and dissemination of experience maximizes knowledge. Data requires a data model for optimal storage from which data is retrievable. Knowledge is immediately useful. Wisdom is knowledge of what is true or right coupled with good judgement, and it is embodied in those who remember the recipe and can tell the stories. Decisions are a judgement resolving a course of action. Better decisions optimize this course of action across time, in space, and in regards to related activities.

Collect Spatial Data

In Walden's primary domains of geotechnical consulting and designing responsive environments most of the data to be collected and worked with is spatial. The issue is that spatial coordinates are often not recorded with the data. Ownership changes, cities annex their neighbors, and boundaries are often in flux. A key contribution from Walden is the methodology for easy formalization of indexing spatial data types (see W3D Edition 07). These spatial index data types can easily be added as various data types are entered into a database.

Collect Text

With the explosion of the Internet there are unlimited amounts of text available to be placed in data bases being built to be processed by advanced pattern finding algorithms. The key usability issue is the spatial indexing of this data. By building a cross-reference between spatial keywords and coordinates in the Infinite Grid℠ (spatial data types), this type of spatial indexing process can become automated. For instance, Iron County, Utah covers E26.56.25, E26.57.11,12,21,22,31,32 (see Figures 2A-F). It is easy to foresee a process which searches text files about Utah, and every time Iron County is mentioned places an XML-tag, which enables easy browser search and retrieval, and automatic downloading of text passages into a database. The spatial location of text can be at a point, along a line, or enclosed by an area.

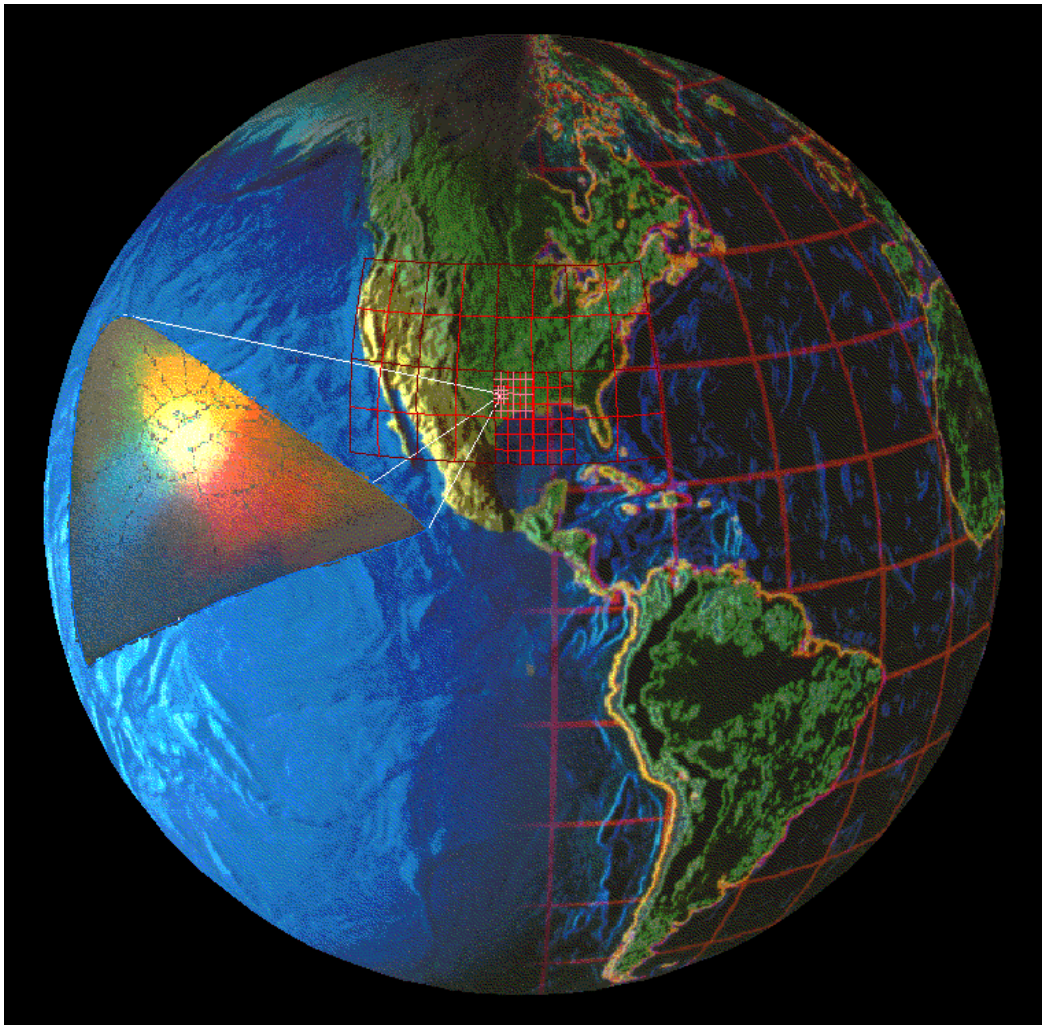

Figure 2A. Bavinger's stylized introduction to the Infinite Grid℠.

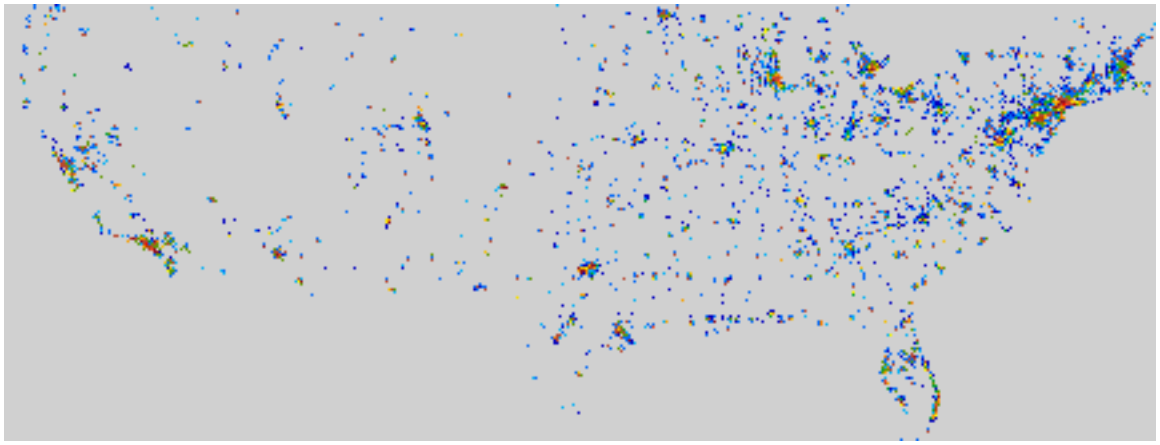

Figure 2B. Cumulative occurrence of SIC Code 15 extracted from the Select Phone CD and plotted using the Infinite Grid℠ spacing of 7.5 minutes Latitude and Longitude.

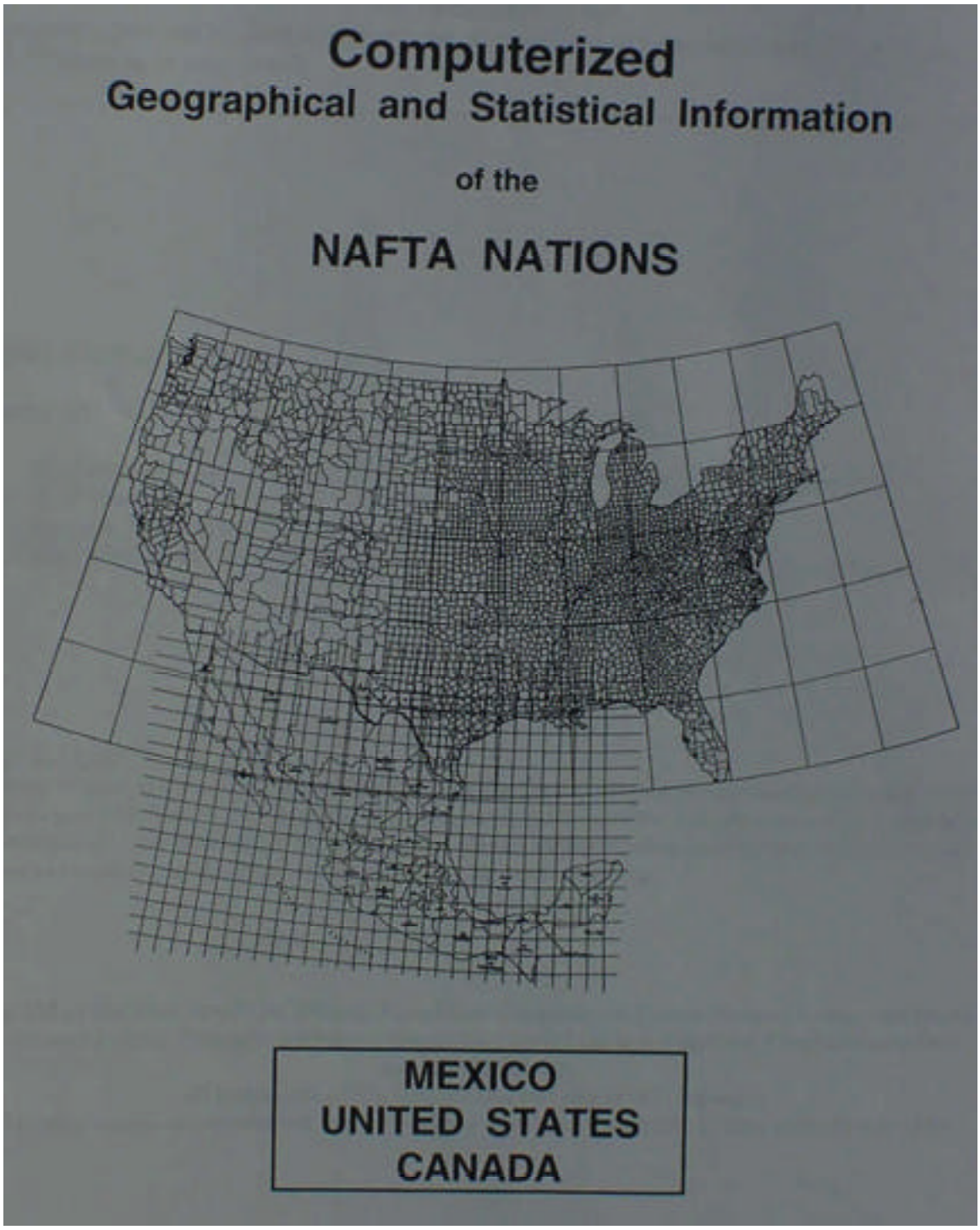

Figure 2C. Proposed Application of the Infinite Grid℠ for statistical studies for NAFTA.

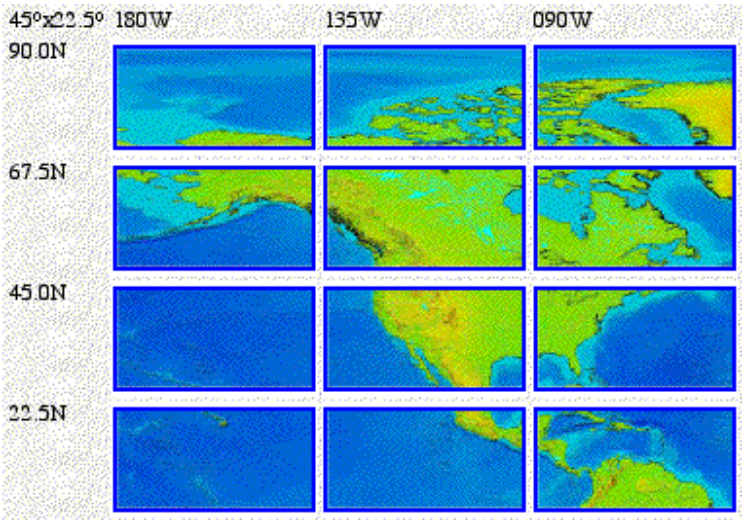

Figure 2D. 22.5°X45° Infinite Grid℠ E15 to E38 (see http://www.walden3d.com/E/E.html.

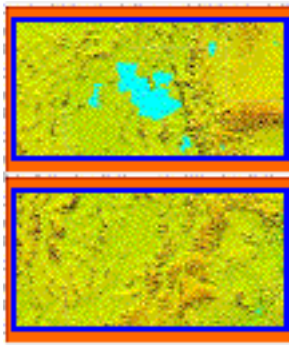

Figure 2E. 5°X2.5° Infinite Grid℠ E26.56-57.

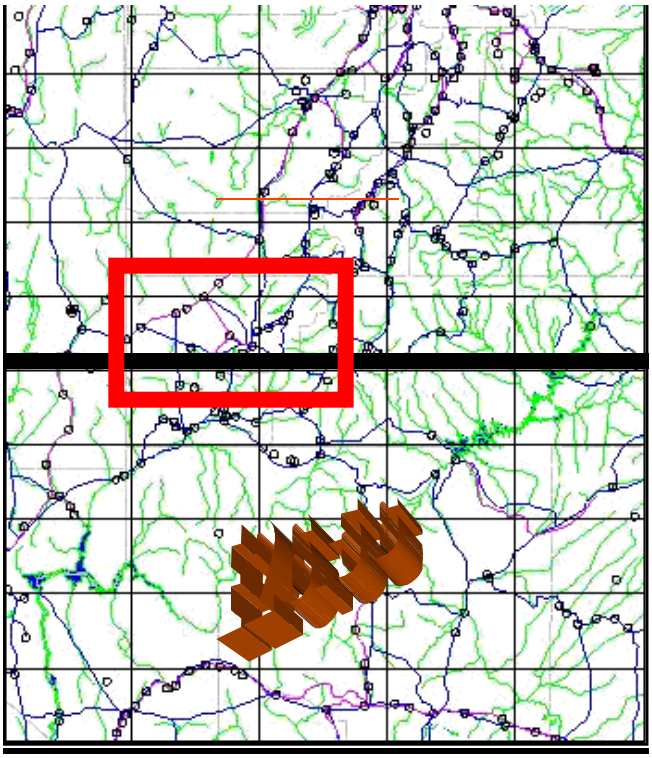

Figure 2F. 1°X0.5° Grids E26.56.25 and E26.57.11-2,21-22,31-32.

Collect Points

In both the geosciences and design there are numerous data types which can be consider points in space. Consider, for instance, in the subsurface 2-D seismic event terminations, well derived formation tops, perforation locations, bottom hole locations, etc. Consider on the surface rock sample locations, well locations, section corners, survey points, weather station locations, stream flow measurement locations, etc. Automation of spatial indexing of these points will develop as the value of advanced pattern finding techniques are proven.

Collect Lines

For every line located in space there are two or more points. In oil & gas exploration, these lines include a termination noodle, a well path, a well log, a check-shot, a seismic trace, location of a 2-D seismic line surface location, etc. These lines can be spatially indexed by locating the line end-points, or by identifying the range of Infinite Grid℠ cells encompassing the line. In design, a line forms a boundary between two entities: i.e. a road and a pasture, a river and a field, counties, incorporated areas, leased acreage, etc.

Collect Areas

G. Spencer-Brown, in his book "Laws of Form" taught:

> "A universe comes into being when a space is severed or taken apart. The skin of a living organism cuts off an outside from an inside. So does the circumference of a circle in a plane. By tracing the way we represent such a severance, we can begin to reconstruct, with an accuracy and coverage that appear almost uncanny, the basic forms underlying linguistic, mathematical, physical, and biological science, and can begin to see how the familiar laws of our own experience follow inexorably from the original act of severance. The act is itself already remembered, even if unconsciously, as our first attempt to distinguish different things in a world where, in the first place, the boundaries can be drawn anywhere we please."

Boundaries are critical spatial information in both the geosciences and design. For each spatial boundary there are three or more points and the or more lines defining that boundary. Therefore collection and indexing of boundaries and the areas they enclose is simply an extension of the spatial collection of points and lines. Each individual point or line on the boundary can be indexed, or the area can be indexed by identifying the range of Infinite Grid℠ cells encompassing the area.

In both geosciences and design significant data are captured as photographs, images, or some form of image from above. The boundaries of "maps" are often not as important as the relationships of data within the boundaries. Therefore collecting areas not only involves indexing area boundaries, it also involves indexing points and lines of interest within the area. This process can be tedious, and yet much of it can be automated. In

order to do advanced pattern finding it is key to have collected and organized the data so pattern finding algorithms can be used.

Image processing algorithms for semi-automated interpretation of satellite imagery have been around for decades. GRASS is one of the more comprehensive public domain image processing toolkits. These tools are able to identify linears, which can be related to subsurface expressions of faulting, as well as circulars, which can be related to the surface expression of salt domes or volcanic intrusions, etc. Reduction of images to points and lines of interests illustrates how there can be spatial data collection and indexing associated with the process of collecting areas. This sub indexing is directly tied to the processes of advanced pattern finding. Infinite Grid℠ cell size is dependent on the size of the objectives. SPOT and other satellite data, as well as high altitude photos, are useful for analysis. The data is deterministic, in that it does speak, and it is a precursor to decision making (Bavinger, 14 December 1989).

Maps are only important as distinctions and as stated earlier differences are all that matter in an information management system (Bavinger, 12 October 1989). For example, in designing a new wing for Herman Hospital, advanced pattern finding techniques were used to automatically identify the two rooms needing lead walls out of 27,000 relationships. Identifying these two rooms allowed them to be placed on either side of the same wall. Architects can not hold this much information in their heads. Blueprints, the designers traditional map, are obsolete (Bavinger, 20 November 1989).

Collect Volumes

Expanding on comments above, collection and spatial indexing of volumes is simply an extension of points, lines, and areas. For each volume there are four or more points and four or more lines defining the volume boundaries. Therefore collecting volumes of data is simply an extension of collecting points, lines, and areas. Indexing the location of a volume is simply the indexing of the area covered by the volume, and having separate same cell sized overlays defining the vertical axes as a function of elevation, depth, depth intervals like isochrons or isoliths or isopachs, and defining attributes at various depths or in various intervals. These attributes are interpretations of lithology, fluid content, geologic age, formation, etc. For some of the advanced pattern finding techniques a regular lattice is better than an arbitrary or even a finite element-type data organization.

Collect Temporal Data

Collect Temporal Text

Just as there are unlimited amounts of text with spatial implications available today, there are also unlimited amounts of text with temporal context. By building a cross-reference between keywords and Timedex℠ data types, the temporal indexing process can also become automated. For instance, the Eugene Island 330 Field was discovered in 1972, and had a normal decline in production until 1985 when there was a recharging event. The Timedex℠ equivalent for discovery is 2.7.0, and for the recharging is 1.0.0. As

mentioned above, data types are one word representations, which the computer can quickly understand, and which will be translated into a more standard notations for human use. Just as it is easy to foresee processes that place an XML tag for a spatial index, the Timedex℠ allows an equivalent methodology for temporal indexing (see W3d Edition 08).

Collect Animations

Animation sequences cover some period of time. They can also be spatially referenced, and thus tied to an area or a volume. Collection of Areas and Volumes has already been described, and so this section focuses on temporal indexing of animation.

We live in a world of animation, highlighted by the movies and television, cartoons and video sequences, as well as simulations and immersive environments (see W3D Edition 04). These animations have two components: (1) when they were created; and (2) what they represent. Either, or both components can be tagged with a Timedex℠ data type. In order to be able to use temporal or time data in advanced pattern finding it is necessary to capture the time index, as well as relevant characteristics of the animation in a relational database field. Once this is completed, the process of finding patterns is exactly the same as working with spatial point data.

Collect Vectors

To the advanced pattern finding algorithms, a vector, which is defined a s a point with a direction and magnitude, is the same as a line. There are at least two points, a starting point and an ending point or magnitude. Magnitude defines how far to go in a specified direction. These points can be spatial, temporal, or a combination of both. Examples include plate movement vectors, structural deformation vectors, stratigraphic depositional vectors, fluid flow vectors, fluid migration pathways, etc.

Collect Ribbons

Ribbons are equivalent to a line vector. These line vectors have magnitude and direction. There are at least three points, and the data structure can be considered to be equivalent to an area. There are one or more starting points and one or more ending points, or magnitudes. If there is one starting point and one ending point, as well as only one point along the entire vector string, then the ribbon reduces to a vector. Like areas ribbons can have points and lines, which represent sources and sinks. These sources and sinks should be indexed in order to be recognized by the advanced pattern finding tools. Examples of ribbons include plate tectonic reconstruction, chronotectonic cross-sectional reconstructions, chronostratigraphic cross-sectional reconstructions, stratigraphic depositional cross-sections, fluid flow cross-sections along or across flow directions, etc.

Collect Flows

Flows are equivalent to an area vector. These areas have magnitude and direction, as typically defined by three or more starting points and three or more ending points or magnitudes. Since there are four or more points and four or more vectors defining the boundaries of any flow, say once source and three ending points or magnitudes, a flow is structurally equivalent to a volume in data storage. The vectors defining a flow can also be spatial, temporal, or both. Examples of flow include 3-D palinspastic reconstruction, sediment flow simulations, 3-D geochronostratigraphic reconstructions, reservoir fluid flow, etc.

Collect Time-Lapse Data

Time Lapse data most commonly refers to two or more volumes of data collected over the same area. In geoscience, the most common time-lapse data is known as 4-D seismic data. This data can be created by reconciling the location and processing characteristics of different 3-D surveys shot at different times, and then differencing them in order to evaluate how the acoustic impedance has changed over time. This science is based on the fact that as hydrocarbon fluids and gases are produced, it changes the seismic energy reflected from various producing horizons, and these differences can be monitored by collecting multiple seismic surveys over the same area and differencing them. For areas with a high signal-to-noise ratio, these different surveys can be shot with completely different data collection parameters and still result in useful data. For areas with a low signal-to-noise ratio, it is best to plant the geophones, or otherwise minimize the variables changing as different 3-D seismic surveys are collected.

Back in 1994, Saudi Aramco was considering a plan to bury geophones across key fields, build cement platforms so the vibrators always were at the same place, and then to recollect 3-D seismic surveys every six months in order to continuously monitor the reservoir, and to insure just-in-time delivery of fluids to transportation ships. Since the plan was not implemented, it must not have been economically or technically feasible.

The medical community is doing a lot of work with time-lapse MRI (Magnetic Resonance Imaging) and PET Scans (Positron Emission Tomography). This data is used to identify the spatial location of brain activity, This activity is mentioned in this report because Walden is preparing a proposal to do a research study, involving several hundred geoscientists, seeing if this technology can be used to quantitatively measure spatial intelligence. Spatial intelligence is the ability to conceptually understand complex data relationships in space. For instance, geoscientists regularly integrate dozens of data types in their mind in order to predict and explain subsurface geology. Designers of cars, equipment, cities, and urban areas have this same capability. The idea is that if it is possible to accurately predict a geoscientists ability to think and work in 3-D and 4-D space, it is possible to predict who are the better oil finders. The data collected for this study will be used for calibrating and testing advanced pattern finding technologies.

Collect Velocity Data

Geophysical velocity data comes as vectors (check-shots or even sonic logs), areas (cross-sectional definition of constant velocity intervals), or volumes (RMS, stacking, migration, DMO, or other related seismic processing velocity cubes). Although velocity data is typically organized like a line (vector), area (section), or volume, it is included in the section on collecting temporal data because these data allow translation between space and time. In the world of 3-D seismic volumes, these data allow seismic travel-time volumes to be converted to depth volumes and vice-versa.

In the world of fluid flow, velocity volumes define fluid and migration pathways. One of the best examples of this work in geosciences, is the work of Deitrich Welte and Bjorn Wygrula of IES GhmH. This software allows modeling of full 3D hydrocarbon generation, migration, and flowpaths (See Figure 3).

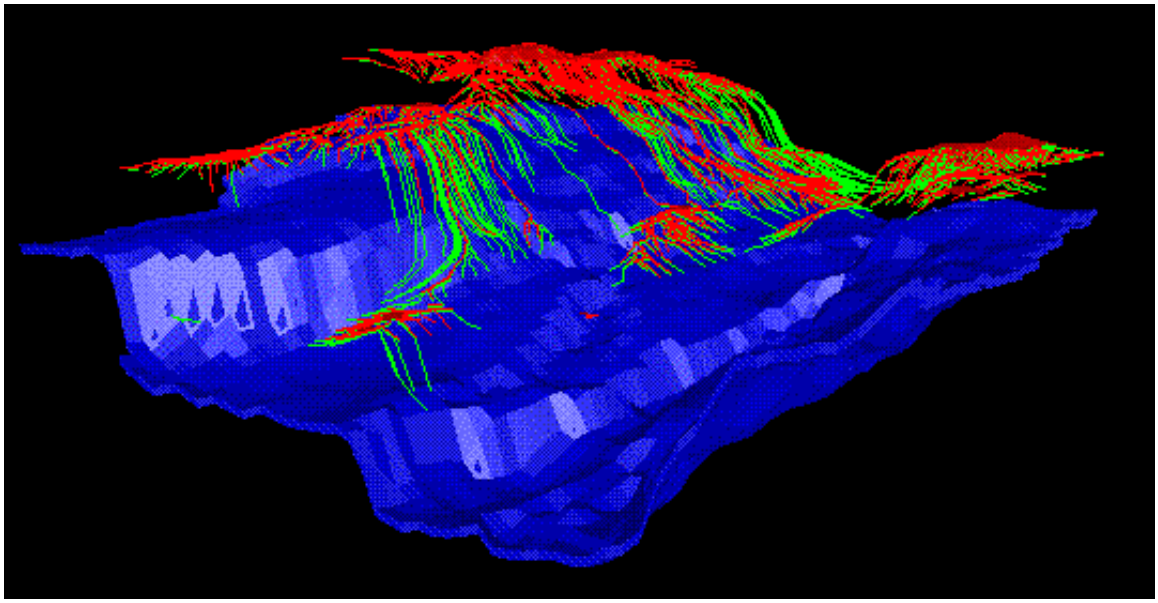

Figure 3. Modeling of primary and secondary hydrocarbon migration (oil - green; gas - red)

Collect Acceleration Data

Acceleration data can be thought of as the first derivative of individual velocity data cells per unit time. Velocity data cells can be related to either seismic travel-time or to fluid flow velocities. These data cubes are directly related to collection of temporal data, in that acceleration data describes how fast something is changing. For example, where and when migration pathways open up, allowing fluid pulses, which are the basis of the dynamic replenishment of hydrocarbon reserves (see W3D Edition 05). In terms of production of replenishing reserves, those which are being recharged by deeper undrillable (with current technologies) sources or pools, an acceleration data volume provides a new way to look at production. The whole production process is turned upside down. The faster the hydrocarbons are produced from the "river," the faster the river flows. Instead of coning and watering out, faster production simply produces more hydrocarbons. Acceleration seismic volumes across subsalt plays establish salt sediment interfaces, improving target definitions.

Collect Pressure Data

Pressure data also comes from several different sources. Engineers measure pressure in well bores. These pressure vectors tend to decrease in amplitude over time. Drillers are very cognizant of geopressure, as it can cause significant drilling problems. Roger N. Anderson at Lamont-Doherty Earth Observatory of Columbia University in New York has demonstrated (and patented) how the Cepstrum seismic attribute (reflection strength of a reflection strength volume) creates a seismic facies (texture or appearance) which correlates nicely with geopressure sediments.

Geopressure correlates with a large percentage of identified hydrocarbon reserves worldwide. Recent studies have shown an electrical characteristic of some shales results in packing of water molecules so close together they can have a density three times the normal density of water. This creates extra pressure, appears to be associated with hydrocarbons coming out of solution, and might be related to large hydrocarbon accumulations at hydrocarbon boundaries. Advanced pattern finding techniques will help to unravel these types of processes, if there is sufficient pressure data, which can be correlated to lithology, production, etc.

Applying advanced pattern finding on pressure data sets will result in identification of connected fault blocks, connected stratigraphic intervals, as well as highlighting migration pathways, and bipassed. pays.

Collect Temperature Data

Temperature data has value relative to understanding the timing and kerogen cracking to form liquid hydrocarbons. In the subsurface, temperature is a low frequency change. Subsurface temperatures appear to change in human time-frames if, for example, hot fluids from depth have migrated into a reservoir. Walden anticipates correlation between thermal anomalies and hydrocarbon accumulations.

Roger N. Anderson once compared the relationship between temperature and pressure, large faults and active fluid flow, to taking a can of carbonated soda out of a refrigerator. When the can is opened the pressure instantly equalizes with the surrounding atmosphere. However, the can is still cold to touch, and it takes much longer for the temperatures to equilibrate. Where there are significant vertical offsets of both isotemperatures and isopressures, say where a major fault exists, it is a prime location to look for hydrocarbon recharging. This is important, in that by owning a very small acreage can still result in tremendous amounts of production. Walden anticipates these examples will fall out of applying advanced pattern finding techniques to the appropriate data volumes; i.e. temperature data volumes.

The Bavinger Model for Pattern Finding

The advanced pattern finding techniques described in this White Paper are presented under an umbrella concept named the Bill Bavinger Model for Pattern Finding. Because Walden does not yet have hands-on experience at applying these techniques, we are quick to acknowledge there could be significant holes in this umbrella. However, if these holes are in the umbrella, it still does not invalidate the individual advanced pattern finding techniques described below, since they have been used by scientists for decades with tremendous success.

Bavinger taught patterns are the DNA of information (14 December 1989). Therefore it is important to organize and optimize data according to patterns. Patterns can be simple clusters of existence or absence (see Figure 4A). They can also be compound interrelationships or derived from regressions (see Figure 4B). Or patterns can be complex (see Figure 4C), derived from clustering, factoring, or ordination (Bavinger, 14 December 1989).

Using these techniques for pattern finding, there are three basic steps in the Bavinger Model for Pattern Finding (as reconstructed from Walden notes taken 20 November 1989 and 26 July 1991). The process starts with raw data in the form of interlocking matrices or data base tables. As an example, Figures 5A-C summarize how Bavinger used interlocking matrices to relate disparate data for land use planning.

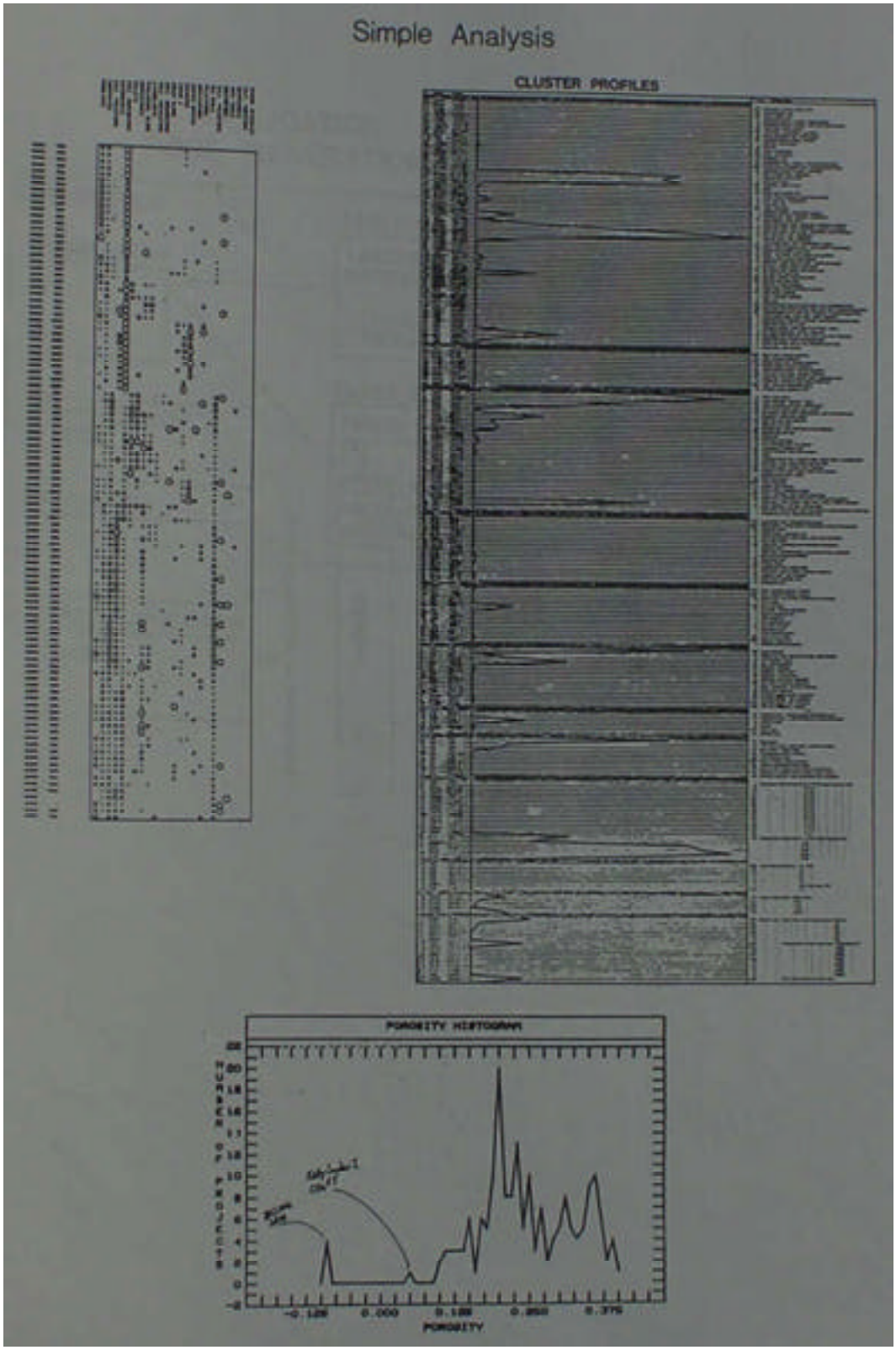

Figure 4A. Bavinger's simple analysis is based on putting the data in interlocking matrices, and then clustering the number of occurrences to find patterns.

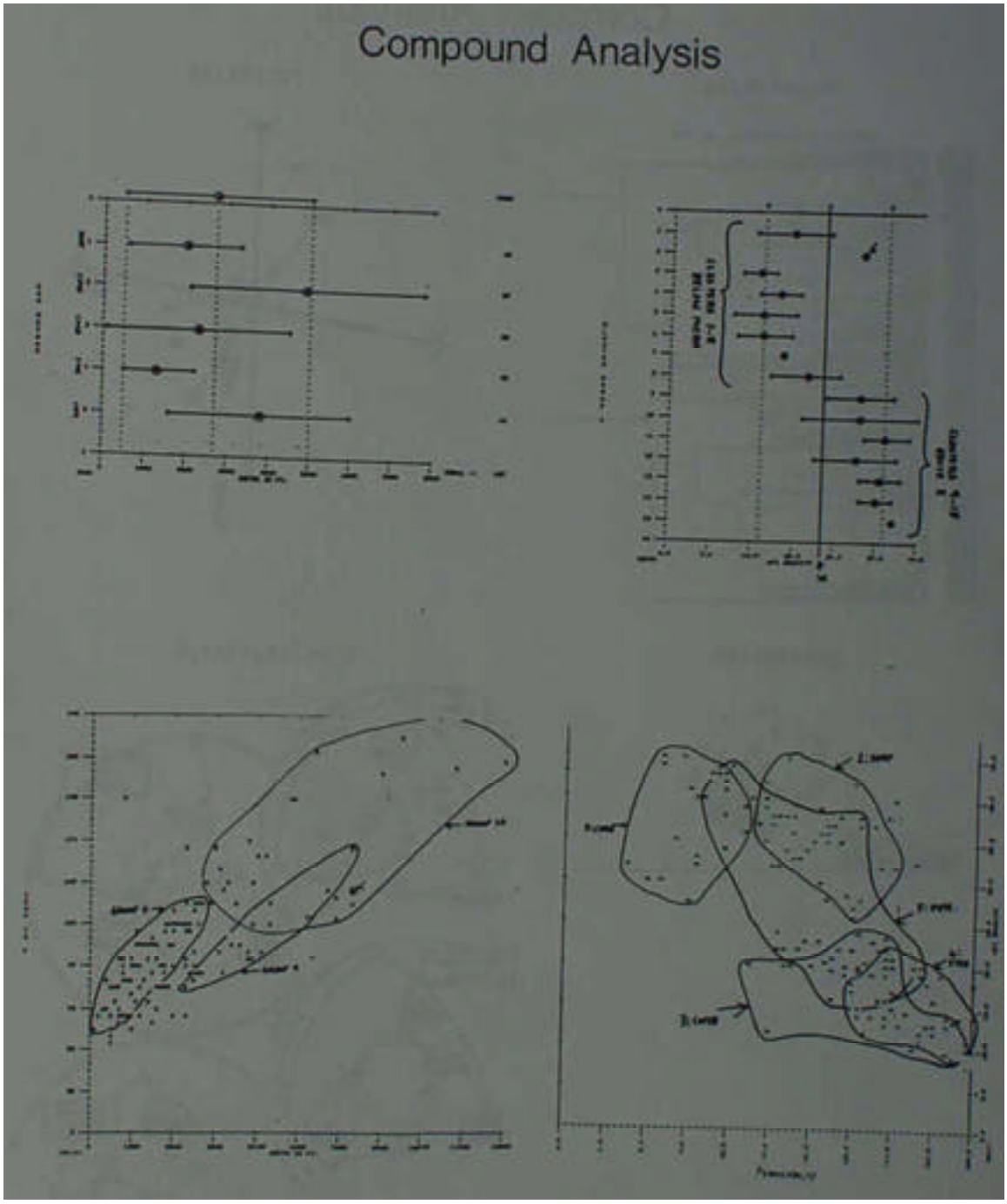

Figure 4B. Bavinger's compound analysis is based on regression studies of the data.

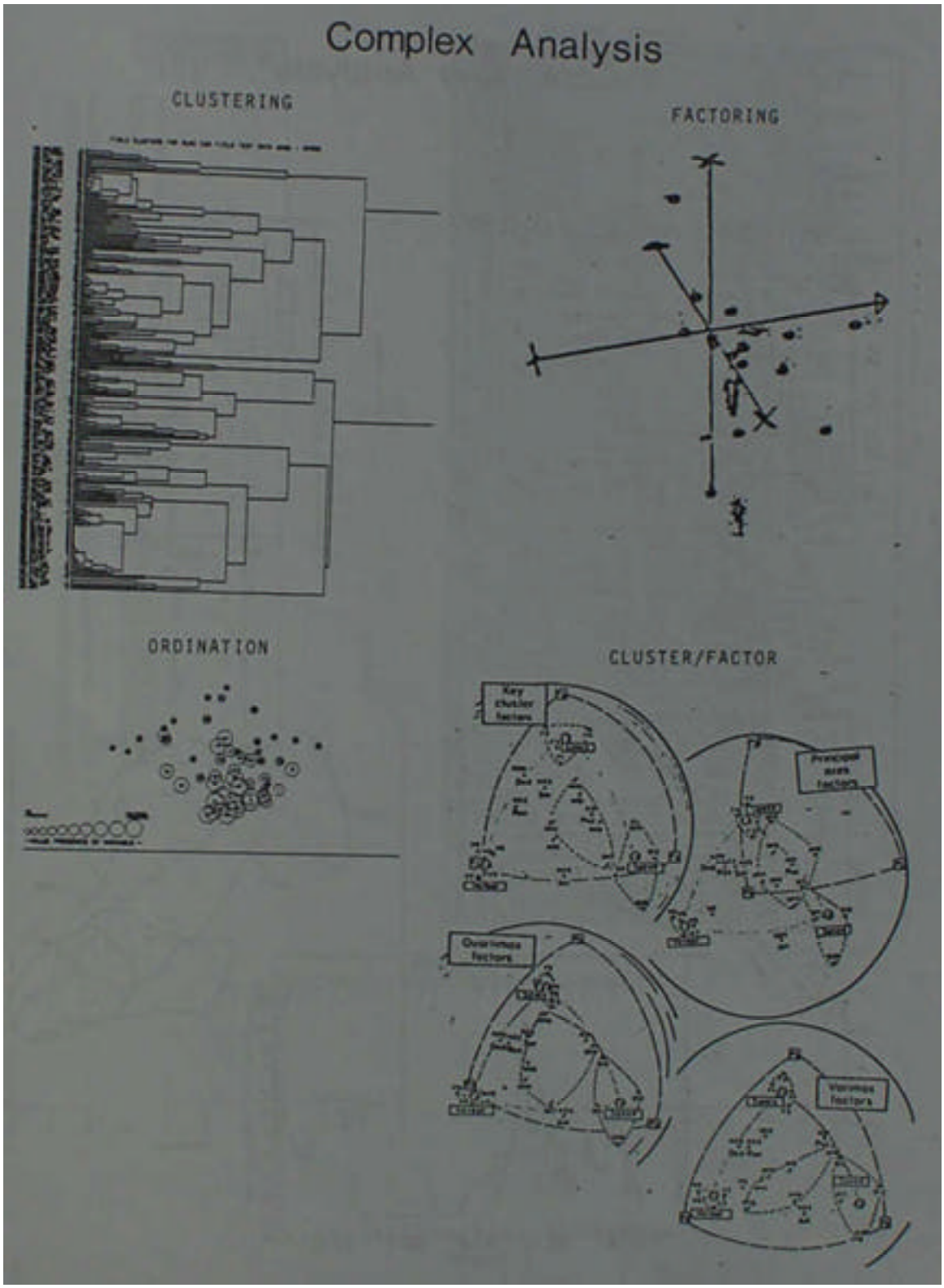

Figure 4C. Bavinger's complex analysis uses clustering, factoring, ordination, and manifold mapping from a hypersphere.

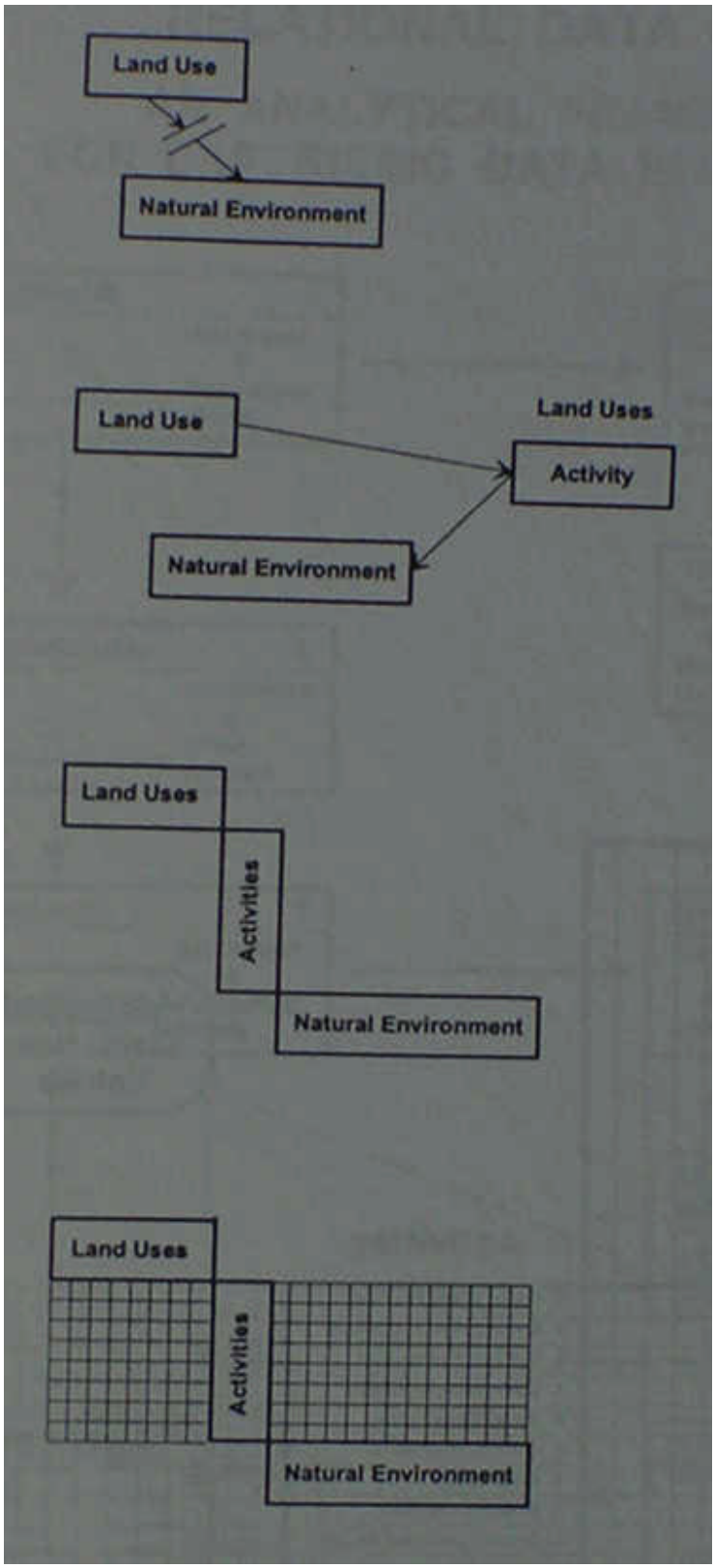

Figure 5A. Activities define the relationship between Land Uses and Natural Environment.

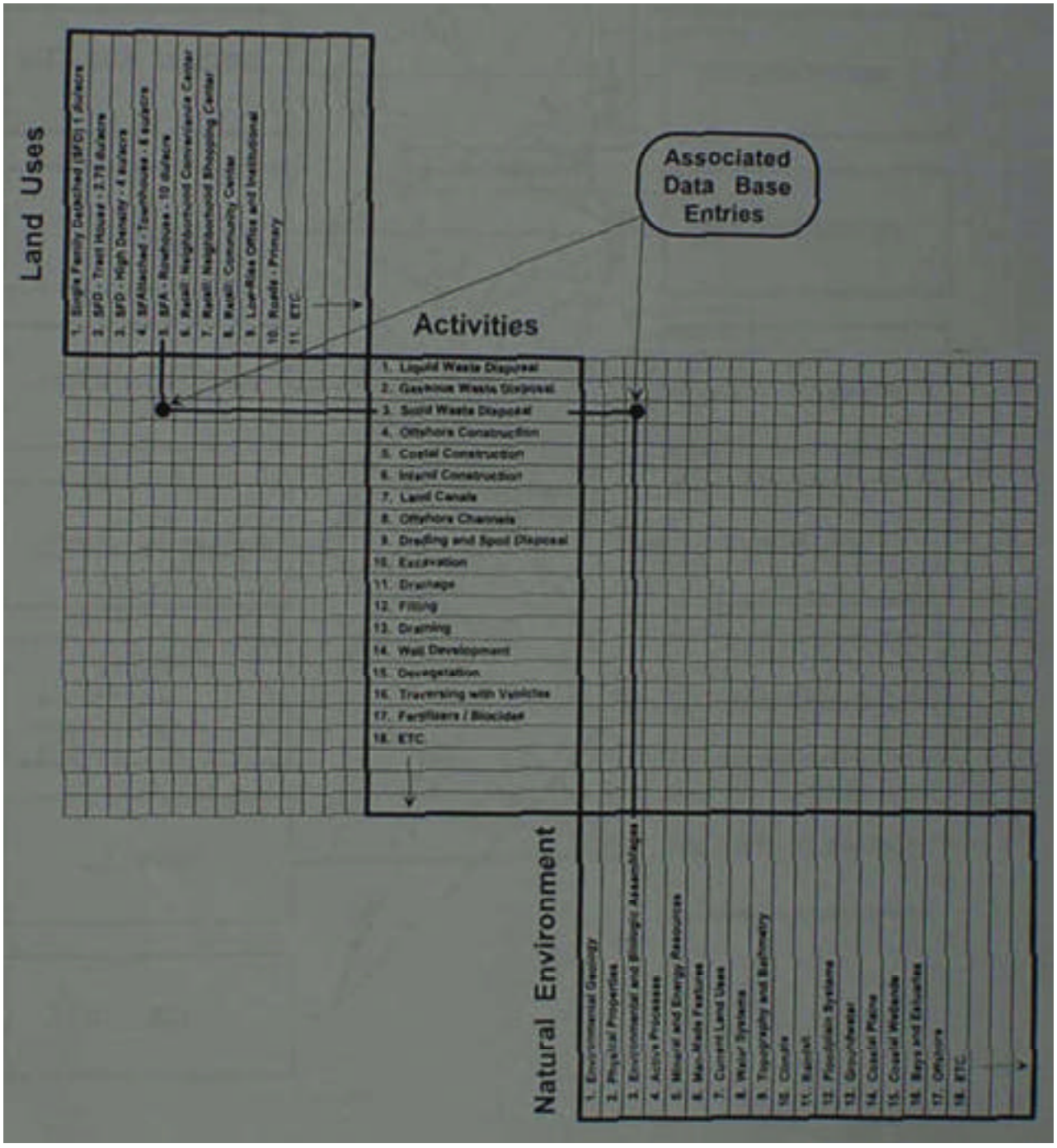

Figure 5B. Detail database entries, relating Land Uses to the Natural Environment by means of interlocking matrices.

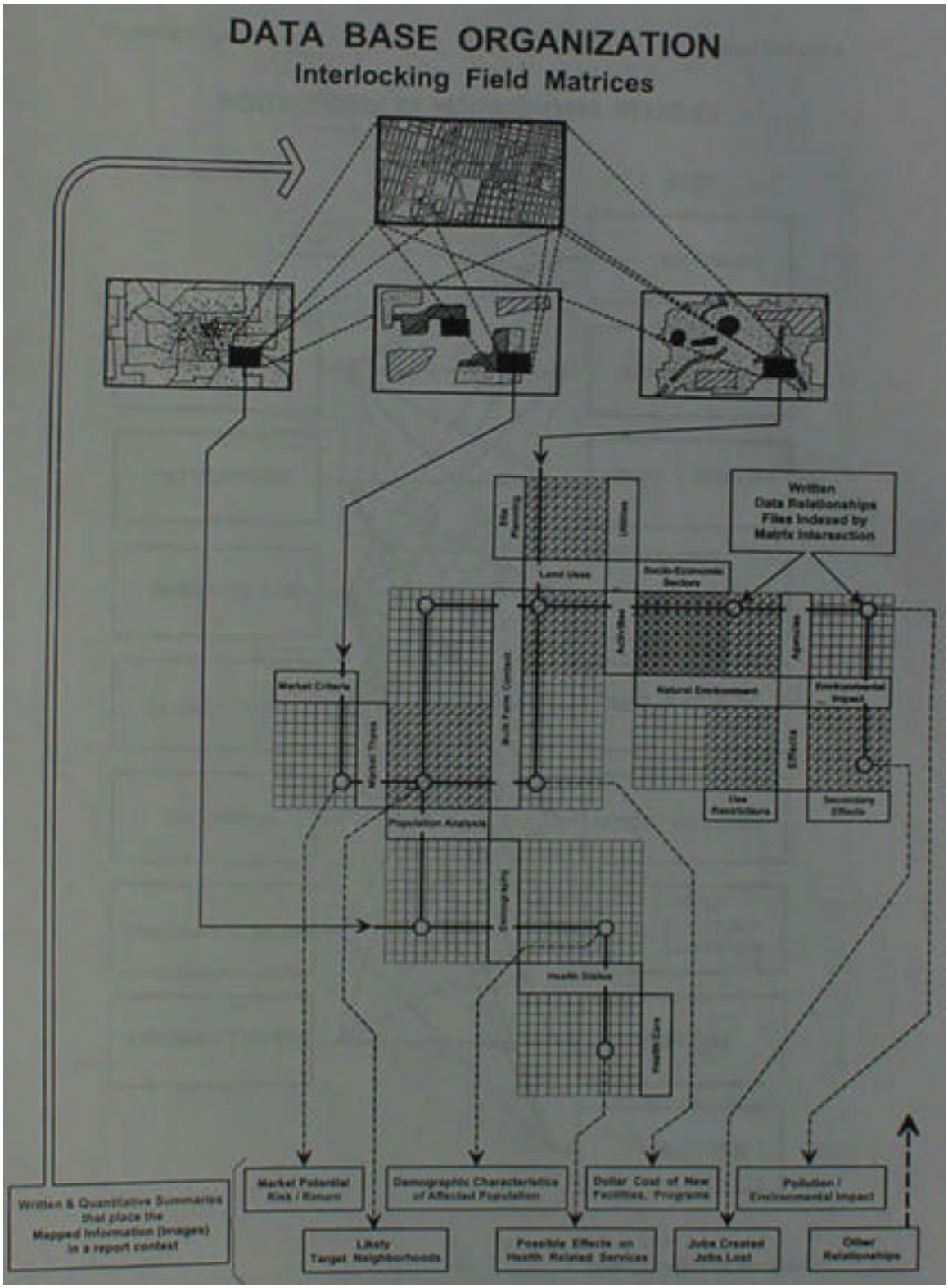

Figure 5C. Database organization via interlocking field matrices.

The three steps in the Bavinger Model for Pattern Finding are:

Step 1: Derive information or 1st order patterns. These patterns are data dependent.
A. Find Patterns of Raw Data, called Clusters (see Figures 6A-6C)
B. Make Factors of Raw Data, by identifying three principal axes and putting the data into a manifold. A manifold is a 2-D map of a hypersphere (see Figure 6D), or an eighth of a sphere, where each of the three principal component axes of the have been normalized. Factors are variables on a linear line forming the center of a hypersphere. A manifold is a combination of any three factors, and the projection to the center of the sphere.

Step 2: Derive knowledge or 2nd order patterns. These are stable, almost independent derived polynomials.
A. Make Clusters of Factors using advanced pattern finding techniques.
B. Make Factors of Clusters of Factors by identifying three principal axes of clusters and projecting the clusters into a manifold using ordination (see Figure 6E).

Step 3: Derive invariant polynomials or 3rd order patterns of polynomials (see Figure 6F).
A. Do advanced pattern analysis of 2nd order patterns by clustering the polynomials
B. Factor the clusters of polynomials.

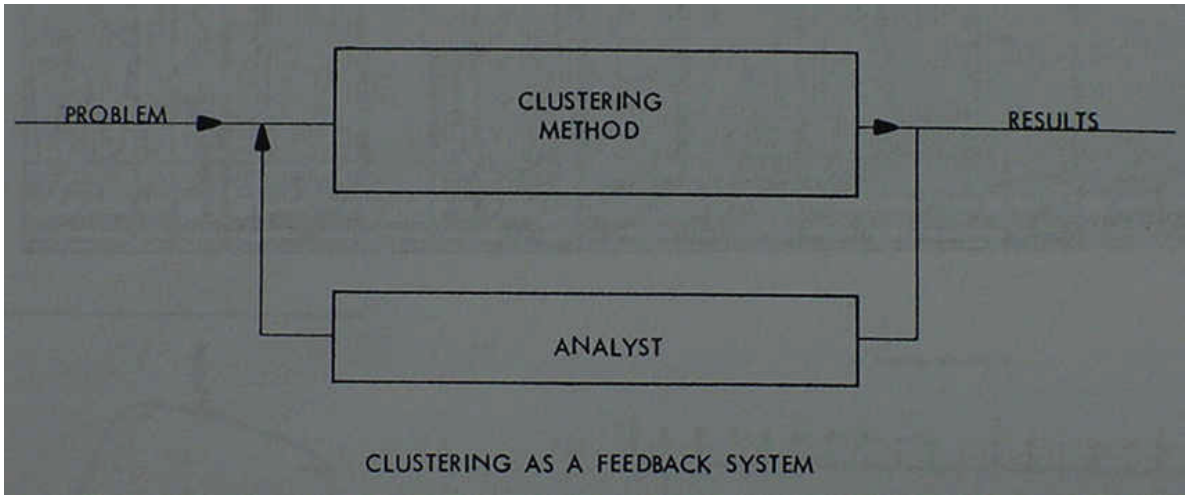

Figure 6A. Clustering as a feedback system.

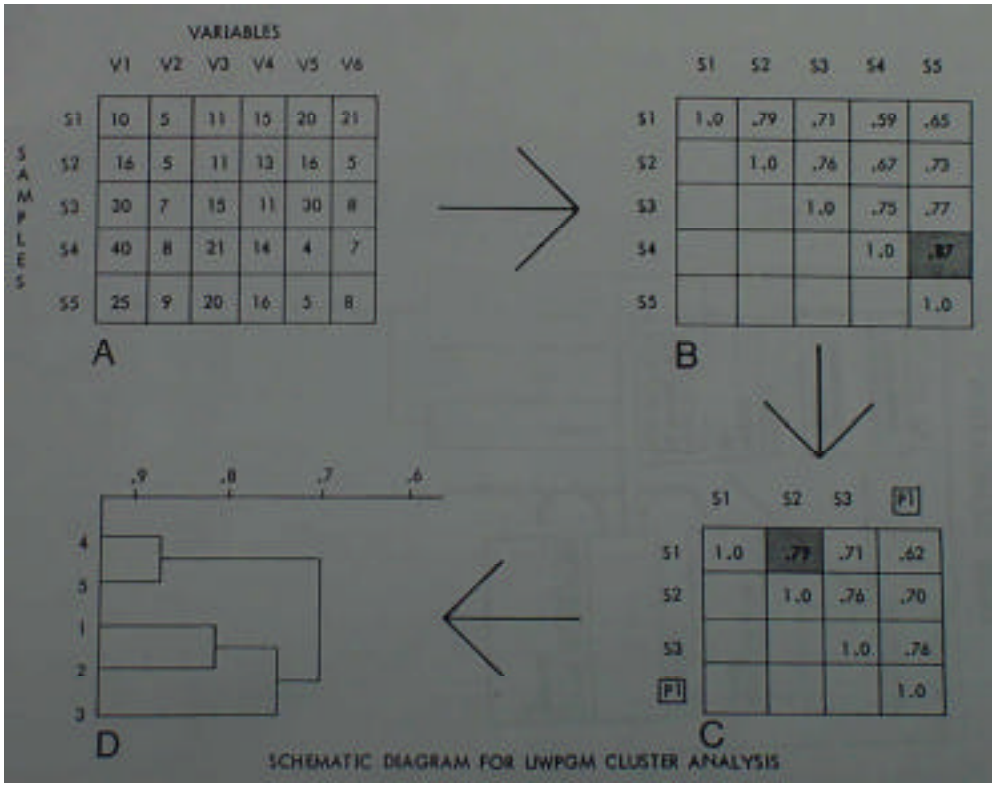

Figure 6B. Schematic diagram for UWPGM Cluster Analysis

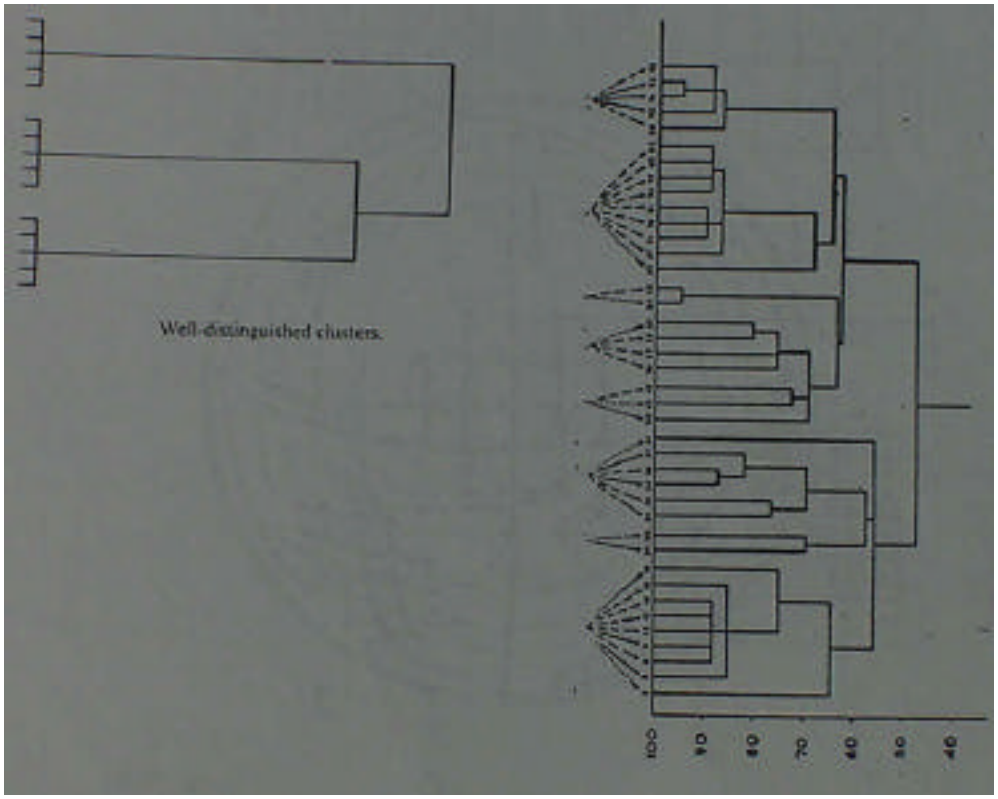

Figure 6C. Well-distinguished Clusters.

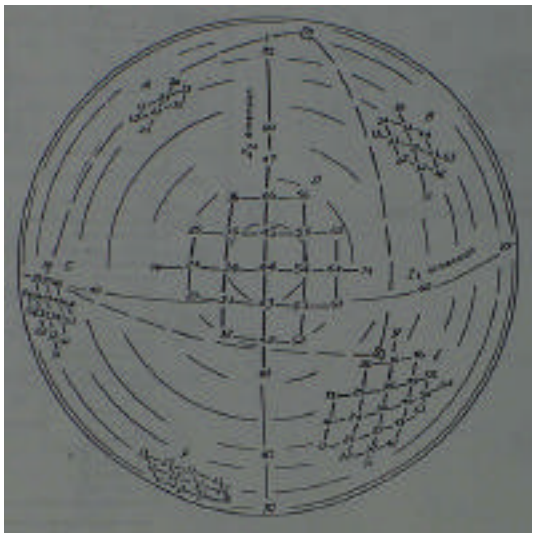

Figure 6D. Data plotted in a hypersphere.

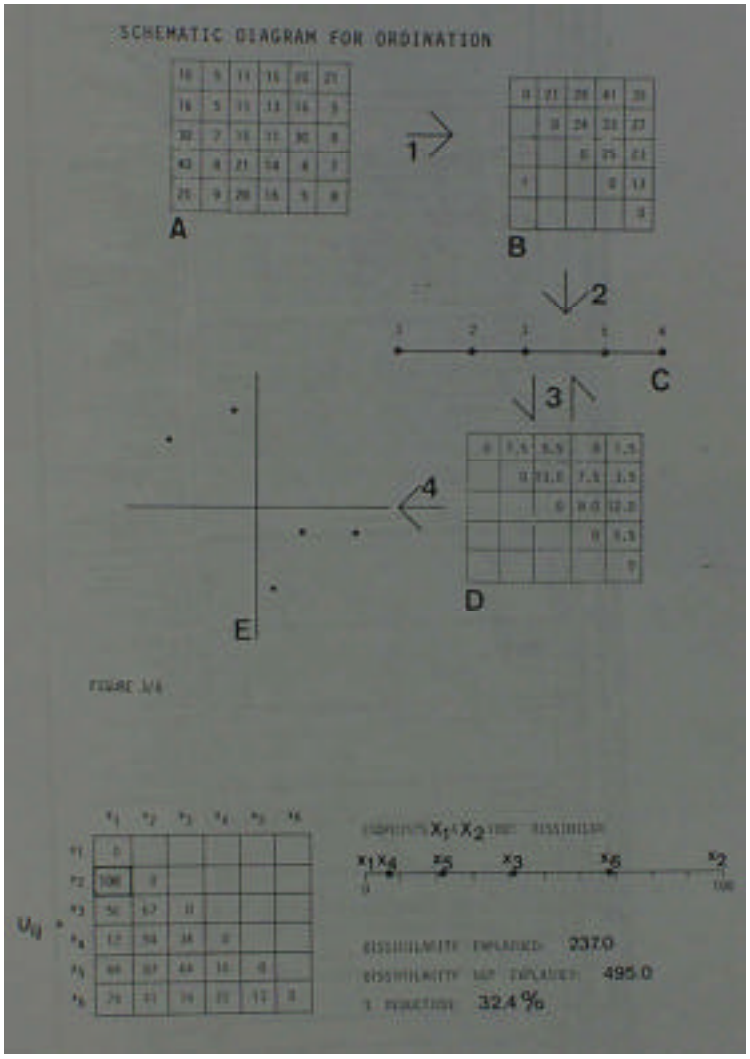

Figure 6E. Schematic diagram for ordination.

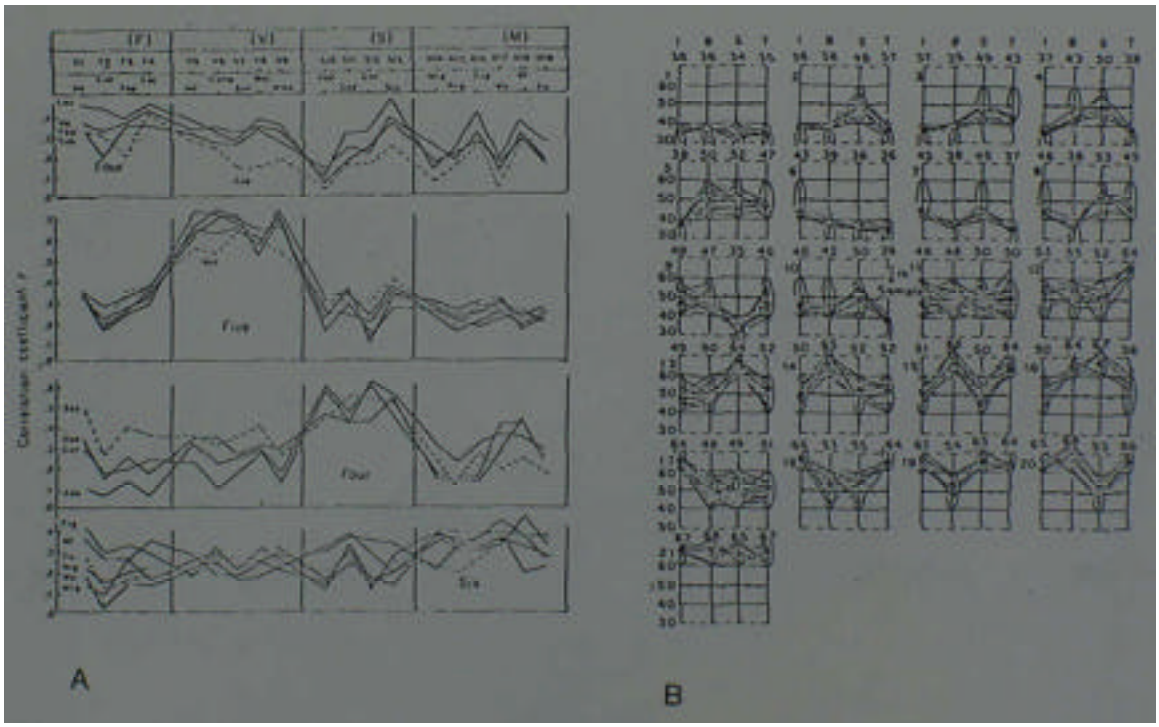

Figure 6F. Polynomials and pattern finding derived Invariant Polynomials.

Sort Data

Bavinger quoted Alan Kay as stating the "3-D spread-sheet should be the universal language for problem definition" (Scientific American, September 1984). Once the data is collected and entered into appropriate databases, the first step in advanced pattern finding is to sort the data. Text files are sorted by the number of occurrences of words and phrases, as well as spatially, temporally, and by activity. Spatial data is formally organized against the Infinite Grid℠ data types. Temporal data is formally organized against the Timedex℠ data types. Activities or processes are formally organized against the Knowledge Backbone℠ data types. Patterns begin to emerge simply by reviewing data stored in one of these "3-D spread-sheets."

Region Growing

Automatic horizon generation from 3-D seismic surveys has proven to be a significant enhancement to traditional seismic interpretation. Landmark Graphics Corporation's ZAP (Zoned Auto Picker) has led in this revolution. One of the specific enhancements it provided, when compared to picking line by line and cross-line by cross-line, is retention of a hierarchy of picks. Given a seed point, ZAP looks at surrounding samples in the 3-D lattice, and selects those samples which within user supplied parameters match this seed point. Retention of these parent-child relationships proved to be a significant advantage for seismic interpreters. This allows mispicks to be identified and deleted simply by identifying the location where the mispick occurred (at a fault or stratigraphic pinchout), selecting the mispick and thus all of the children of the mispick, and selecting delete. It

is important to note that although these picks occur in 3-D space, ZAP is set up so the picks are single-valued in the vertical axis. This “encourages” the picking algorithm to follow horizontal stratigraphic layers. The data between boundaries is more important than the data within the boundaries (Bavinger, 30 November 1990). It is important to stress ZAP and related algorithms are a useful form of advanced pattern finding.

Most region growing algorithms are similar to ZAP. Specifically because they need a seed sample to start and then need something to use as a benchmark when looking for similar waveform characteristics on adjacent samples. Adjacency is what defines region growing, as compared to visualization of all voxels (volume pixel elements) with similar characteristics within a volume of data. A major difference between most region growing algorithms and ZAP is that the generalized region growing algorithms are not single valued in any axes. This allows these algorithms to determine the surface of 3-D geologic bodies like salt intrusions, turbidite channels, fans, deltas, etc. However, these algorithms do not, nor will they in the foreseeable future, know geology. Therefore they are only useful when used by an experienced geoscientist.

Cluster Analysis

A cluster of data is a non-random segregation of products into separate groups. Although there is no way to determine the correct (or even optimal) number of clusters, Dynamic anticipates identification of first order trends using cluster analysis. One must specify the number of clusters as an input parameter in order to run the program. Alternatively, the number of specified clusters can be progressively increased through several iterations. Each iteration of the analysis will deliver a "solution"; however, there are no strong criteria to determine which is the correct solution. The major problem in cluster analysis is "cluster validity." Another problem is that the results are commonly "assessed by visual inspection of a graph" (a dendrogram) which, by virtue of its two dimensional nature, cannot accurately portray the relationships between clusters. We anticipate our access to N-D immersive environments will minimize this issue. A final problem is that the output is produced in a form difficult to understand or use by decision-makers. Although Walden does see an opportunity for visualization to assist in communication of results. today.

Factor Analysis

Although Principal Component Analysis has similar problems to cluster analysis, except for a change in jargon, Dynamic will also use these technologies when appropriate. Mathematical purists can substitute "cluster validity" with "determination of significant eigenvectors" and "assessed by visual inspection of a graph" with "problems tied to projection" in the above discussion. Problems are further exacerbated due to the inability to describe and defend the concept that an eigenvector or a "factor" in any fundamental context of business decisions. However, Dynamic intends to use correlate any “factor” derived from mathematical or concept space into holistic synergistic data models in order to identify trends and to be better able to rank new exploration Concepts, Leads (places to look and ways to look), and prospects (CLPs).

Automated Self-Classification

Self-training classification can be accomplished in at least three very different ways: cluster analysis, principal component / factor analysis, and Polytopic Vector Analysis (PVA). The first two procedures have been used and refined for decades. As described above both have inherent problems that limit generalized application.

Residuum's primary data mining tool is based on the concept of self-training classification of data. Ten years in development and testing, this new, more powerful, "data mining" software and procedures to produce maps and cross-sections using data sets that are cleared of the artifacts and inconsistencies commonly present in large data bases, removing a major obstacle to construction of databases containing millions of data points.

Proprietary "data cleaning" technologies insure reliable data is used in the analysis and when necessary can provide a "best estimate" correction based on analysis of the entire data matrix. Utilizing these tools Residuum Energy, Inc. has completed individual studies each involving 20,000+ wells and 40+ formation tops The software has the capacity to handle much larger data sets.

Data mining, pattern recognition, and self-training classifiers produce superior understanding of the data, types of activities, and customer trends. Dynamic sees a unique opportunity to use these data mining technologies in locating subtle traps in semi-mature hydrocarbon basins. This is accomplished by exploiting the giant databases that now exist for such basins. Well-based information, stratigraphic, engineering and production data, can be combined with other sources of commonly available information, such as gravity, aeromagnetic, and seismic, to yield a database with rich potential for understanding future hydrocarbon development.

Introduction to Automated Self-Classification

Competitiveness in the modern business environment is increasingly driven by information technology that permits strongly based timely decisions. The formation revolution was made possible first by the spreadsheet concept and then by the readily availabile large capacity data storage, fast data retrieval and transmission, and good database design. The first stage in information technology centered on rapid assessment of large data volumes following spreadsheet and relational database paradigms. The second stage in the information revolution concerns development of digital "explorers" whose function is to:

1. Detect complex patterns and relationships among the data and
2. Report these findings in a manner useful to decision-makers.

This section discusses a new sort of data explorer program called a "Self-Training Classifier". Self-training classifiers are designed to determine complex relationships in

large databases without being driven by pre-existing hypotheses. This seems at first glance to represent a leap backwards in that it is not driven by pre-existing theory concerning the underlying root causes of business dynamics. However, macro and micro economic theory is neither very robust nor complete enough to answer the day-to-day needs of corporate practice.

Dynamic has constructed a numerical procedure tailored to both the complexity of data and the objectives of the data analysis. Tailored to overcome the obstacles encountered by previous attempts to extract information from data complexes it is designed to explore an n-dimensional data space and return with significant insights easily understood by decision- makers.

Delineation of Data Structure

Corporate data consists of a complex of inter-correlated data. The simplest representation of this is displayed in a simple spreadsheet consisting of rows and columns. Commonly rows represent a physical or economic entity and columns define a set of attributes of that entity. For instance, rows can represent products and column represents various attributes of each product. An attribute (such as weight) commonly varies from product to product, as does another attribute (say cost). In addition the cost of a product may in some way depend on its weight. The pattern of non-random relationships present in a spreadsheet among attributes and products is defined as the data structure. Exploratory data analysis is the procedure designed to ferret out these relationships and report them in a manner appropriate for decision-making. Each row in a spreadsheet represents an entity (a product, a person, a company) and the columns represent values of a set of variables associated with that entity. Graphically, an entity represents a point defined by its location with respect a set of axes. The value of an entity can be displayed by its location on an axis (labeled for instance "cost"), each axis representing a variable. If only cost and weight are the only variables, a point on graph paper can represent a product. However, the number of variables (columns in the spreadsheet) is commonly far more than two. If we have "n" variables (columns) then a point defined by its position measured against "n" axes can represent the product.

Self-Training Classification

We cannot visualize the location of a points graphed into a space containing ten axes that are all mutually perpendicular for this would require an ability to "see" in more than three-dimensions. The relationships between samples in this n-dimensional space carry large amounts of information. Dynamic's proprietary technology is designed to allow us to explore this hyperspace by proxy and then report to us in a framework intelligible to us mere mortals and represents the most highly evolved procedure as an n-dimensional self-training classifier. Self-training implies that the data structure itself defines the output rather than an a priori assessment of the important underlying factors. Thus, patterns in n-space determined by relative positions of product locations define the associations between products and variables. The Dynamic approach differs in many ways from other self-training classification procedures. A major difference is that the analysis does not

equate the degree of practical importance of a class with the fraction of the variability of the total data set that it accounts for. We recognize, for example, that a few parts per million of dioxin on an industrial site impacts a corporation far more than several tens of percent of iron. In this respect, it differs from procedures such as principal component analysis or factor analysis, which were described above.

General Characteristics of Archetypes and Hybrids

The basic idea behind this technology is that entities such as products or people can be analyzed with respect to archetypes or "end-members". An end member is defined in the same terms as a product, that is, as a set of values of the same variables as used in the original spreadsheet. An end member therefore may be represented by a real product or may be defined by a set of variables that can potentially be a product. An archetype is defined as "a model of all things of the same type". Commonly real entities may approach an archetypal state but seldom attain it. Also many entities may be hybrid, i.e., a mixture of archetypes. For example, there may be three sorts of customer, each sort represented by an archetype defined by their buying habits; and all customers may be defined by a characteristic set of proportions that represent the "mix" of archetypes in that individual.

Polygons, Polyhedrons... Polytopes-the basis for classification

A polytope is a polygon of any dimensionality. A two-dimensional polytope is a polygon; a three-dimensional polytope is a polyhedron, and so on. We are interested in a certain sort of polytope generalized into any number of dimensions. In two dimensions it is a triangle, in three dimensions it is represented by a tetrahedron (a four sided pyramid), and so on into higher dimensions. The number of corners (vertices) of each polytope is one higher than the number of dimensions in which it resides. Thus a triangle (a two dimensional polytope) has three vertices. In the context of the following discussion, each vertex can represent an archetype or kind of entity and any point within or on the Polytopes represents classification of an entity in terms of relative proportions of an archetype. Obviously, an infinite number of Polytopes can enclose a cloud of data. The challenge we faced was to derive a special or unique Polytopes, sensitive to the data structure and carrying the most easily interpretable information.

The developers worked, on a variety of "bottom line" problems such as environmental fingerprinting, litigation support, petroleum exploration and production strategy, medical image analysis, and mineral exploration. Success or failure in any of these fields is predicated on sound data analysis coupled with an inherently effective means to transmit the results to decision-makers. This technology is designed to require few assumptions about data structure. Therefore it is not necessary to assume the existence of, normal frequency distributions, of the linkage of the magnitude of variance to the degree of importance, or that the data is clustered.

Capacity

The maximum number of archetypes is defined by the nature of the data matrix. The maximum potential number of archetypes derived must be equal to or less than the number of columns (cases) or rows (attributes) whichever is the lesser in the data matrix. Practice has shown that at least 15 to 10 cases should be included. The maximum number of cases that can be analyzed in a single analysis is virtually unlimited.

Information Models

Pattern finding and classification are of no use unless the results can be used by someone. This section describes the key views of an information model, and how these types of models expand on traditional information tools like a map, a blueprint, a balance sheet, or an income statement. The Infinite Grid℠, the Timedex℠, the Knowledge Backbone℠, and immersive reality scenes are each different views of the information model, focusing on space, time, activity, and visualization respectively. None of these information handling tools are part of the genetic makeup of humans. We use information to make decisions, and there is only so much information we can hold in our heads at one time. However, primitive societies had existed for millennia without the benefit of a map, a blueprint, a balance sheet, or an income statement. Their continued sustainability is directly tied to their stewardship over the resources they have, and their recognition of sufficiency. With the complexity of modern society, and specifically with the complexity of the type of problems Walden undertakes in geotechnical consulting and designing responsive environments, it is important to have tools which enable gigantic amounts of data to be captured, sorted, classified, and presented in an understandable way.

This section describes information models professionals can use and interact with, in much the same way as we interact with other people, in order to make better decisions. For example, Boeing had 220 departments and 4,000 computers. By putting an information modeler in each department they cut the departments to 140, and made similar cuts in computers, programs, and other entities. Still everyone gets their own view of the model. And these models are then transformed to built form in the real world. (Bavinger, 14 December 1989).

Walden refers to the type of pattern derived information modeling Boeing implemented as Pantopia. “Topia” means space, like in topology, and “Pan” also has a spatial connotation, in that it means “all or everywhere.” Put together into the word Pantopia, it is like “Pan” has a temporal connotation, so “Pantopia” means all places, at all times, in all dimensions. Walden believes there is a morality in this model driven data approach, because by bringing all of the data to bare on a problem, users are naturally minimizing risk. Minimizing risk, by the objectivity of bringing everything to bare on the project being studied, is the pragmatic version of the philosopher’s utopia.

Pantopia models integrate spatial, temporal, and activity indices (see W3D Edition 10). Initial analysis of these information models is performed by creating trails through interlocking data bases. Once the trails are identified, the next data set to be analyzed uses the same trails. A foreground binary view of the data base, allows quick reviews of data pulled along with various queries. The notion of smart data is that it is data that performs better as the information model is used to simulate real-world processes. (Bavinger, 14 December 1989). Pantopia models are also ideally reviewed in an immersive environment (see W3D Edition 04).

Pantopia models are multivariate, multicomponent, pattern finding derived sets of samples and variables. These information, mathematical, time, and space models are simple to understand once the space is visualized. Once we understand the models, new vistas open for us. For instance, we are limited in natural resources only because we have not found new ways to explore for them (Bavinger, 25 July 1996).

Bavinger taught (26 July 1996) that explicit systems require links to be made and implicit systems provide a context. The Infinite Grid℠, the Timedex℠, the Knowledge Backbone℠, and scenes are each tied to context. The map, the timeline, the activity model, and the database are all a matrix. These matrix displays allows automatic identification of sources and sink within the information model. Static layers are explicit. The flow between layers is implicit. Displaying field data this way is the difference between a conceptual model and mother nature. It is an interesting result that by reducing the error functions we find more information in patterns of what isn't there than what is there. In other words money is in the gaps.

Optimization	= sources	= growth
Distribution	= sinks	= decay
Production	= sweetspots, or information that doesn't fit the trend, which is always in the white spots. (Bavinger, 26 July 1996).	

Five Somewhat Wild Scenarios

It seems the most successful people wrap their problem solving in a story. Since this Bavinger comment fits the definition of wisdom presented above, the last section of this White Paper is a series of five "pro forma' scenarios of anticipated application of advanced pattern finding techniques. A follow-up on this White Paper will be to use spatial econometric modeling to measure sustainability, based on decisions derived from using advanced pattern finding techniques.

1. Text Pattern Finding Scenario

Digital books, which Walden calls ELDOs (ELectronic DOcuments) are becoming available. It doesn't take much imagination to see the day when PC's will have a series of filters which start out by automatically indexing and counting each word and phrase in a document. Here a phrase is defined as "N" words which appear within "M" words of

each other, where “N” and “M” are user defined variables. The next filter, the words would be matched against a dictionary to identify verbs (actions) and nouns (things). The next filter would match the words against a thesaurus, identifying synonyms, removing articles (like a, an, the, etc.), and then reindexing and recounting the words and phrases in the document based on this new mapping. Applying clustering or the automated self-classifier to these data will create a new type of index for the book. This index will have the statistical basis of an author’s word print, and will allow classification of the book based on the words the author(s) actually used.

Applying this process to all of the books in a library, whether personal or public, and then clustering or classifying these clusters or classifications, will create an extensible scientific alternative to the Dewey Decimal System used for current library storage and retrieval. The key is that the storage and retrieval would be directly related to context. In fact, it is easy to imagine an emitter attached to a book spine in a traditional hardcopy library which, when asked if the book contains data relative to a specific, place, time, or activity (as defined by Infinite Grid℠, Timedex℠, and Knowledge Backbone℠ data types), responds by lighting up. The ramifications of this concept are significant.

2. Numerical Pattern Finding Scenario

The example picked for this section is more a case history than a future scenario. However, implications in regards to applying this same technology to other areas, say the Stock Market, have significant implications.

Example from Baseball of Self-Training Classification

An Example from Baseball Analysis of the hitting statistics of Baseball players is a useful way to demonstrate the strengths of Residuum Energy, Inc. technology. In addition to ready availability of statistical data, there is a consensus regarding types of players and the value of each type. Therefore this context serves a means to understand and verify the significance of our results. The data consists of the detailed hitting statistics of a sampling of baseball players including members of the Hall of Fame, current players who bear promise of being elected to the Hall of Fame, as well as a sample of players new to the major leagues. For many of the more famous players, their entire career, year by year, has been entered. The data for each player's year includes six measures of batting prowess. A batting average includes two parts. If a player bats, say, .300, he also has an "out" average of .700. The data includes all the components of the batting average, the components of the batting average arising from singles, doubles, triples, and home runs plus two components of the "out average', strike outs and non-strikeouts. This yields six variables that can represent the hitting of a player in each season. These six variables do not exhaust those available in baseball in that similar numbers are available in terms of fielding performance, walks, etc. However we will use the smaller set of variables because the analysis started to evolve for us from a demonstration to an obsession. We show that all of the outfielders can be classified in terms of four "archetypes". Of course no single player may be a pure type but commonly is a hybrid between two or more types.

Table I Archetypal Players*

	Player 1	Player 2	Player 3	Player 4
Singles	26.2	0	0	18.9
Doubles	5.2	0	0	7.1
Triples	0	0	16.2	0
Home Runs	0	6.8	0	9.6
Strikeouts	0	93.2	83.8	0
Other Outs	68.6	0	0	64.4

* Note: Values in percentage. The sum of the first four rows in any column = batting average.

The object of the analysis is not to necessarily provide new insights into baseball but to illustrate the validity of Energy, Inc. analysis capabilities within a well-known framework. We argue that any business-related data will similarly be easily interpretable with the customary framework shared by business executives. Table 1 lists the batting attributes of four archetypal players. Baseball addicts can easily describe each archetype.

In the paragraphs below we discuss the characteristics of each archetypal player and compare their resemblance to actual players. Before this, however, we point out that many archetypes share common characteristics but to different degrees. Both Players I and 4 do not strike out but have dissimilar batting averages (.314 and .356 respectively). Player 1 is a singles hitter and Player 4 hits for extra bases. Both Player 1 and 4 hit about the same number of doubles. Players 2 and 3 tend to strike out as opposed to flying out. Both have miniscule averages (0.007 and 0.016 respectively) but player 2 tends to hit home runs and Player 3 hits triples. Player 2 is the only archetype that hits triples, which is a result of superior base running speed.

Player 1

Player 1 (column 1, table 1) has a batting average of .314 (the sum of rows 1-4 in column 1, table 1). The batting average is composed of singles (.262) and doubles (.052). Player 1 flies out but never strikes out. The real player resembling Player 1 most closely is Pete Rose, especially in the latter part of his career. In fact, Rose actually is located at a vertex of the classification tetrahedron in 1983 (a value of 1.00 Player 1) and so defines Player 1.

Other players with high degrees of Player 1 include John Cangelosi, Tony Gwynn and Lou Brock. Cangelosi has been consistent in this regard for his entire career examined in this analysis (I986-1998) having a greater than 0.8 resemblance, whereas both Gwynn and Brock display some variability over time. Tony Gwynn has strongly resembled Player 1 in some years (e.g. 1982-1986, 1990-1993, 1995-1995) where his score on Player 1 is about 0.8. In the intervening years his score declines coincident with a significant rise in his resemblance to Player 4 (Hall of Fame Superstar, as discussed below). Lou Brock, in contrast, has only an intermediate degree of resemblance to Player 1 in his early years (.5-.7) along with a relatively strong resemblance to Player 2 (.1 -.15,

strikes out a lot). In his later years his resemblance to Player 1 soars (.8-.9) as his resemblance to Player 2 decreases. Players with very low resemblance to Player 1 (>O.I) are generally superstars at the height of their careers: Greg Vaughn (0.08, 1996 (before and after his trade from Milwaukee to San Diego), Sammy Sosa (0.08,1996 and 0.00,1998) and Willy Mays (0.00 1995, 0.08 1962, 0.06 1964 and 0.02 in 1965).

Player 2

Player 2 represents an archetype that is impossible to exist in pure form in the Major Leagues. Player 2 hits exclusively home runs but has an abysmal batting average (0.007) and strikes out rather than flies out. Players who have a relatively high resemblance to Player 2 include Willy Mays, Reggie Jackson and, in some years, Sammy Sosa. Thus all tend to strike out if they don't hit the home run. They differ significantly from Joe DiMaggio and Ted Williams who have very low resemblance to Player 2.

Player 3

Player 3 in his pure form has never existed in the Major Leagues. He is purely a triples hitter with a modest batting average of 0.162. Generally, hitting a triple requires superior base running speed and so achieving a triple is more than a function of just power hitting. The actual player most resembling Player 3 is Willy Mays in 1958 (0.21) and it is not a coincidence that Mays is second in triples in the major league record book (behind Mike Tieman) and led the league, in stolen bases during four seasons. The high value for Mays for Player 3 as well as a high resemblance to Player 4 (home run superstar) indicates the uniqueness of Will Mays. Other players with similar values of Player 3 include Joe DiMaggio and Lou Brock.

Player 4

Player 4 is the archetype of the baseball superstar. He has a batting average of 0.359 and hits more home runs than any other archetype. He however never hits triples indicating a lack of speed on the bases. Only when combined with significant values of Player 3 does he become the epitome of a hitter. Of all hitters, Ralph Kiner most resembles Player 4. In 1949 he essentially was Player 4 with a value of 0.93. If we combine his scores that year with Player 3, Kiner achieves a score of 1.00, which by these standards makes him the greatest hitter in the set of players analyzed. Bill James in The Historical Baseball Almanac echoes this evaluation. Other players who strongly resemble Player 4 include: Greg Vaughn (I998), Joe DiMaggio (entire career), Willy Mays (1954-1965), Ted Williams and Mel Ott (1929).

Most of the Major League Players can be considered to be mixtures of archetypes 1,2, and 4. The player who represents the most uniform blend of these three archetypes is Reggie Jackson, "Mr. October', one of the most maddeningly inconsistent stars in baseball. Jay Buhner is a similar mixture but unfortunately has been unable to display the late season spud of Jackson.

Discussion

From this demonstration we can see that successful players are not alike with respect to a single characteristic but that they all contain high amounts of various combinations of three of the four archetypes. Some players are binary mixtures (Sosa), while others may contain significant amounts of Player 1 (the journeyman) but like DiMaggio, achieve maximum values in other categories (Player 2, the home run king). Comparison between players can turn up surprises. For instance, in terms of hitting, both Joe DiMaggio and Ralph Kiner are similar but the values suggest that Kiner was the better hitter. Kiner, albeit in the Hall of Fame, has garnered little recognition because, we imagine, he played for Pittsburgh and did not have the luxury of having played for the storied Yankees. Any two players can be compared on the basis of the relative amounts of only four archetypes. Commonly, the number of archetypes does not increase as rapidly as the number of variables. That is, if we further refine the at-bat record by including bunts, sacrifice flies, hitting into double plays, advancing the runner, etc., the number of archetypes will change little, if at all. In many cases, a characteristic that is "minor' to a statistician can be of major importance in the "real" world whether it be baseball or commerce. For instance, the maximum value of Player 3 (triples) is 0.13 for Willy Mays (I952) and Mays consistently scored above 0.05 in this category. The reason for the low variance is that triples are a rare event in any baseball game and so do not account for much of the batting average. In statistical jargon, this characteristic absorbs little variance and so is likely to be ignored in variance-driven decision making. Yet this characteristic alone, serves to differentiate Mays from all of the other players in the analysis. The logic behind this demonstration can be easily generalized into a number of business related situations. A balance sheet subdividing revenues and expenditures by category by month or year is an exact analogy to the baseball example. The spending decisions of customers are another. Comparing a set of companies within an industry is yet another. A business can be classified in terms of archetypal units of income and expenditure and the "mix" of archetypes can be monitored continuously. Similarly, customers can be classed in terms of archetypal membership (derived from the data itself.) and this, in turn, can be used to predict their future behavior. The gist of this report is that this technology represents a new way to evaluate data complexes and to report the data in a fashion that is simple and understandable to decision-makers

3. Spatial Pattern Finding Scenario

As in the previous example, this scenario is more a historical than a future scenario. It is based around using a commercial CD database of phone numbers, which costs less than $300. This set of CDs has every published residential and business phone number in the US on it. In addition, the CDs have the latitude and longitude (+/- 5 meters) of each phone jack, which between portable phones and 25+ foot cords exceeds the accuracy of the actual location. In addition, the CDs have four levels of the 6-diget SIC Code (United States Standard Industrial Code used for taxing and Yellow Pages organization). These CD's have been updated quarterly for several years. Bavinger and his team took these data and worked with one SIC code of interest to the design process: SIC 15 for Building Construction – General Contractors and Operative Builders (see Figure 7A-B).

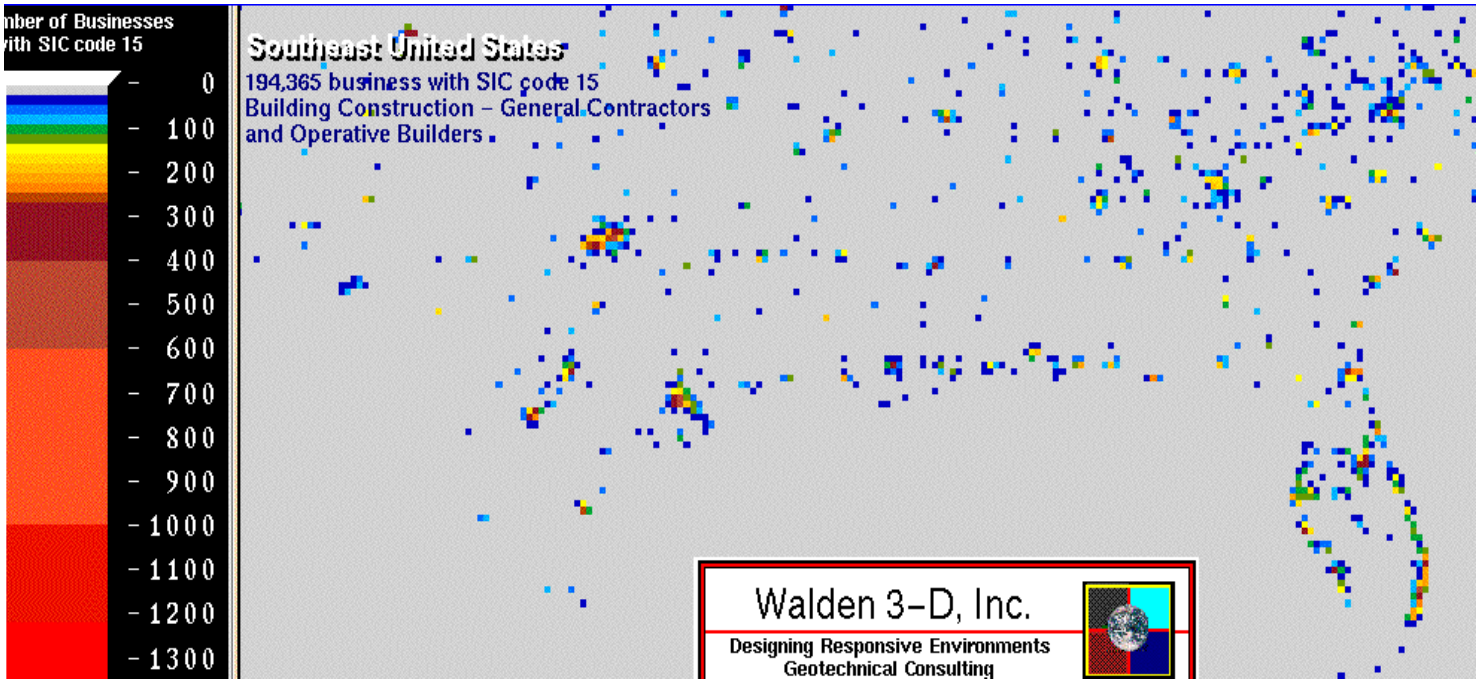

Figure 7A. Display of SIC Code 15 in a 7.5 minute (~7.5 kilometer) Infinite Grid℠ for SE US.

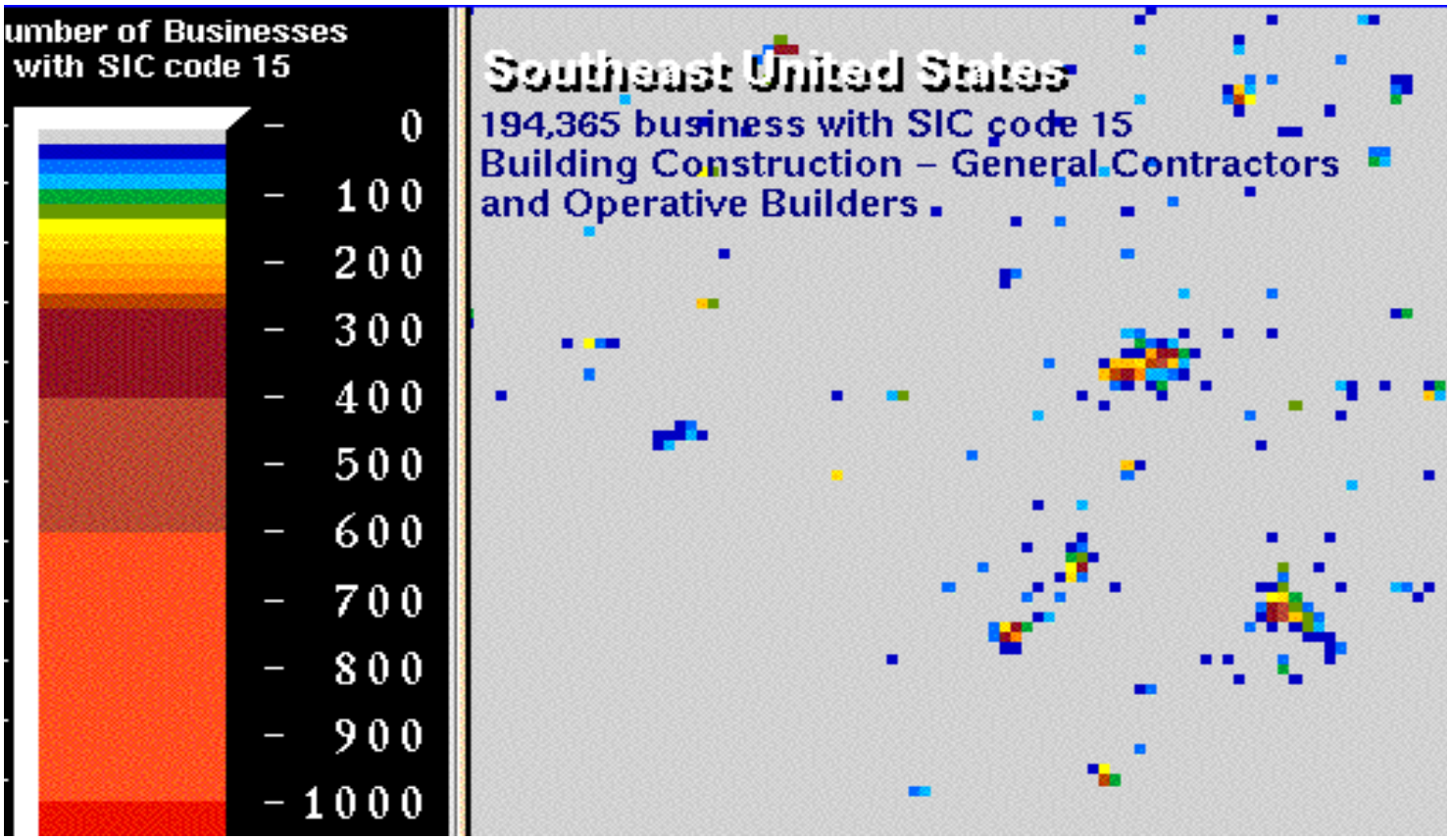

Figure 7B. Close-up on the Texas portion of Figure 7A (note how San Antonio, Austin, Dallas-Ft. Worth, Corpus Christi, and Houston-Galveston show up from just raw data)

The only advanced pattern finding techniques applied to this data was to sort and accumulate occurrences spatially into 7.5 degree cells (~7.5 km cells) and then to display the results as Infinite Grid℠ cells. As seen in Figures 7 and 7A the results are as easy to understand as a photograph. The bottom line is if someone wants to open a plywood store, or something else catering to the Building Construction industry, locate it within one of the redish blocks in order to be close to customers.

Dynamic Oil & Gas will be using this same approach to manage data from 3,000+ wells in the shallow Federal Lease Blocks Offshore Eastern Louisiana. The Infinite Grid℠ organization of the data is a key data organization aspect of how advanced pattern finding techniques will be used to maximize and optimize the number of new exploration Concepts, Leads (where to look and how to look), and drillable Prospects (CLPs) to be found as the main part of this US$2 million study.

4. N-Dimensional Pattern Finding Scenario

North American Industry Classification System (NAICS) has 1800 lines of industry classifications. Of these, 102 are related to the petroleum industry. Imagine the same approach that was taken in scenario 3 above is applied for each of the 102 petroleum industry related SIC/NAICS Codes. Then imagine using an enterprise model, like the Walden 3-D Knowledge Backbone℠, to show the movement of raw data from satellite images to gravity data to magnetic data to geologic field studies to seismic data to wells to water wells for injection to production to transportation to refining to end-use. Given that the density of cells in the Infinite Grid℠ for each activity will change, by animating through these "maps" in Knowledge Backbone℠ space, the spatial movement from idea to raw material to processed material to end use will be obvious. Because the Infinite Grid℠ is a matrix, the differences between these grids define a matrix mathematical model which can be used to predict flow across the grid. This flow might be supply and demand, or it might be shortages and excess inventory.

Bavinger described a generalized implementation of this process into two phases (15 March 1997):

Phase 1: Sort combinations and dynamics of businesses. Derive empirical relationship between different SIC codes to allows study of the dynamics between cities. Attempt to see material flow through business systems and to individuals. Then relate material flow to economics and pollution. The goal is to obtain a balance between natural and human systems on a performance basis. The emphasis is to model materials through the systems in order to get to a comprehensive understanding of material flows as the currency of exchange between humans and natural systems. And then to tie this to economics and pollution. This allows solving complex societal problems from a fresh viewpoint.

Phase 2: Residential integration of demographic and other census data in order to: dynamically understand supply and demand, monitor epidemics, predict and control crime, reduce gerrymandering to semi-natural boundaries, etc.

5. Internet Classification Scenario

Walden anticipates development of several products to classify web sites, once the following scenario is recognized by a few more people. There are currently a variety of search engines which allow retrieval of data from a web site by <metatags> or <keywords>. However, there is currently no tools available for automatic classification of web sites and of the web. This seems strange, especially since science is based on the concept of classification.

Walden anticipates clustering, factor analysis, and automated self-classification will create a scientific basis for web searching. In other words, when an oil man or woman searches for "stripper wells in Harris County, Texas" the results will be relevant to his or her professional work.

Imagine a tool which goes to a web site and catalogs all of the words at that site. There are numerous tools doing this today, including Alta Vista, Go To, Google, Infoseek, Lycous, and Microsoft Network. Now imagine the tool does advanced classification of these words using clustering, factor analysis, or automated self-classification. Specifically imagine words are indexed against the Infinite Grid℠ data types, the Timedex℠ data types, or the Knowledge Backbone℠ data types. This provides a map of the site as a function of locations referenced, times referred to, as well as activities discussed. Next imagine the tool does an automatic classification of all of the data tied to the site. This would involve text pattern finding per page, per site, and per region. It would also involve mapping links between pages at a site and in a region. Automatic classification would involve clustering or factoring or automatically self-classifying the text and the links in order to derive information models describing the site. The results of this work would be quite a revelation for most site webmasters and their managers.

For instance, Walden recently manually used a few of the procedures described in this White Paper to classify the Rice University web site. There is no way the mission of Rice, which is to be one of America's best campus-based teaching and research universities, could be automatically derived from the Rice web pages.

Among other things, Walden anticipates classification of site web pages will help site administrators recognize what the actual vision, mission, strategy, and tactics are. Once site web pages are classified, an automated classification of sites within different spatial, temporal, and activity boundaries will allow an understanding of regions. Correlating sociological classifications against the natural environment will help decision makers understand where people and the natural environments are walking on each other. The potential impact from a best practice, documentation, and implementation standpoint (see W3D Edition on Best Practice Documentation), as well as from the point of view of continuous improvement, is truly staggering.

In summary, advanced pattern finding techniques are important and proper implementation of them in web site classification will make a significant difference.

Acknowledgements and Next Steps

Large sections of this White Paper are quotes from notes taken when talking to Bill Bavinger. Most of these meetings occurred when he was at Rice University. Bill was an angry, hard man. He was also a genius, and he is missed. The clustering, factoring, automated self-classification, and baseball example sections were written by Dr. Bob Ehrlich. Although many of the ideas discussed in this White Paper are conceptual, there has been enough prototype development to be comfortable the basic concepts are correct.

The next step is to support Dynamic Oil & Gas and Residuum Energy, Inc. as they test these concepts in the oil and gas industry. Dynamic prepared a prospectus in October 2000 to raise $2 million to be used to apply these advanced pattern finding techniques in Offshore Eastern Louisiana. Residuum independently is working with several oil companies to raise $6 million to apply the automated self-classification system to cratonic basins in order to find hydrocarbon traps tied to far-field vertical stress faults. In both cases, it is anticipated these advanced pattern finding technologies will provide a significant competitive edge. As funds are generated through the application of these ideas, each point discussed will be tested, evaluated, documented, and publicized, as appropriate. Walden 3-D, Inc. intends to focus on application of these ideas in the spatial world of geotechnical consulting and designing responsive environments.

PostScript

Thank you for reading the W3D Journal Edition 13.

The W3D Journal is priced at US$3,000. per 12 Editions per subscribing companies, payable in advance. Each subscriber has a contact person who is responsible for distributing passwords within their entity and for answering questions about access the W3D Journal. There is a 20% discount for companies who purchase 72 Editions up front; i.e. instead of US$15,000., the price is US$12,000. Payment can be made by credit card on-line at http://www.hgol.net, and the Subscription Agreement is available at http://www.walden3d.com/journal/subscription.html .

The editor has struggled with pricing, especially since everything is free on the web. However, the cost is less than many companies will spend on legal fees evaluating the innocuous Subscription Agreement, which is necessary to share the documentation of Intellectual Property. Furthermore, within the framework of the editors values of sufficiency, sustainability, and stewardship (see: http://www.walden3d.com/values), excess beyond what is sufficient for a modest western lifestyle (if 10 kids can be called modest) will be fed back into the system to enable sustainability as fulfillment of the self-imposed stewardship to continue to make a positive difference. “Fed back into the system” means funding research projects, paying experts to document their work in science and technology, doing pilot projects, cutting subscription prices, and in all cases making the results available through the Walden 3-D Journal.

Current W3D Journal Editions:

D. http://www.walden3d.com/journal/j004/01.html · W3D Design Process ·
E. http://www.walden3d.com/journal/j004/02.html · Maps ·
F. http://www.walden3d.com/journal/j004/13.html · Advanced Pattern Finding Techniques ·

Upcoming Editions in the W3D Journal:

G. 03 · Models ·
H. 04 · Immersive Reality ·
I. 05 · Dynamically Replenishing Hydrocarbon Reserves ·
J. 06 · Knowledge Compilers ·
K. 07 · The Infinite Grid℠ as a Spatial Index ·
L. 08 · The Timedex℠ as a Temporal Index ·
M. 09 · The Knowledge Backbone℠ as an Activity Index ·
N. 10 · Integrating Spatial, Temporal, and Activity Indices ·
O. 11 · Data Mining ·
P. 12 · Using the W3D Design Process to Create a Process for Hydrocarbon Prospect Lead Generation ·
· 2001 · The Bill Bavinger Urban Machine ·
· 2001 · Best Practice Documentation ·
· 2001 · Continuous Improvement ·
· 2001 · IDEF Technologies ·
· 2001 · Object Oriented Scene Graphs ·

Send questions, comments, ideas for future topics, and other suggestions to:

H. Roice Nelson, Jr.
W3D Journal Editor
Walden 3-D, Inc.
P. O. Box 382
Barker, TX 77413-0382

Call: 281.579.0172; facsimile: 281.579.2141; or e-mail: rnelson@walden3d.com.