Database-to-Brain Bandwidth

John G. Boland, president
VisiBit Corporation
One Parker Square Suite 408
2525 Kell Boulevard
Wichita Falls, Texas 76308
USA

Phone: 940.322.9922
Fax: 940.723.1478
Email: john.g.boland@visibit.com
www.visibit.com

ORIGINAL PUBLICATION

This paper was originally presented at the USP 2001 (University Synergy Program) in Lubbock, Texas on 26 March 2002 and was published in the proceedings of that conference.

Keywords:
Productivity Improvement, Profitability Improvement, Data Management, Data Visualization, Data Analysis, Execution Process, Process Optimization, Software

ABSTRACT

In the broadest interpretation of "process" and "automation", any business activity is comprised of processes which can be monitored and, if appropriate, automated to improve overall profitability.

Representing processes with data and increasing process-to-database bandwidth, the rate of gathering and storing data, have been major computing initiatives for a quarter of a century. The volume of data collected now hinders understanding of intra-process and inter-process relationships and, hence, timely intervention.

Database-to-brain bandwidth, the rate of presenting and comprehending execution process data, now limits the overall process-to-brain bandwidth. Operators, supervisors, and managers struggle to correctly identify trends, problems, and discontinuities in extremely large quantities of execution process data, before taking corrective action.

Real processes are comprised of many variables with non-linear sensitivities and asymmetrical statistical distributions. These, and appropriate mathematical and graphical software applications to deal with them, are beyond the comprehension of those who are not mathematically proficient.

Representing data as abstract images and automating the image generating process relieves these constraints and vastly increases database-to-brain bandwidth.

INTRODUCTION

In the broadest interpretation of "process" and "automation", any business activity is comprised of processes which can be monitored and, if appropriate, automated to improve overall profitability. Automation has brought about the obvious manufacturing control systems, HMI (human-machine interfaces), and many other computer applications ranging from simple accounting packages to sales support, MRP, and MES.

Theoretically, good operations and management practices, relying on these software tools, make experiencing a given problem or sequence of events increasingly unlikely with time. However, training and simulation grow progressively more challenging as processes are optimized. The opportunity to improve shrinks. The risk of damaging the process increases. Problem solving becomes a search for events not seen before, a distinctly human, data-intensive task that is difficult to define and teach.

Representing processes with data and increasing process-to-database bandwidth, the rate of gathering and storing data, have been major computing initiatives for a quarter of a century. The volume of data collected now hinders understanding of intra-process and inter-process relationships and, hence, timely intervention.

Dr. Shoshana Zuboff, a Harvard social scientist, studied the early computer invasion of a wide variety of clerical and industrial workplaces. In the dissimilar environments, similar struggles arose over the choices between automating, which dehumanized the workplace, and informating, which gave workers more and increasingly accurate information with which to make decisions (1). The automation physically distanced operators, supervisors, and managers from the execution processes which they tended and managed. As hard as it is to imagine today, the early HMI clumsily replaced the sensuously rich seeing, hearing, touching workplace experience with cold, numerical, textual information and suffered, deservingly, from credibility problems. The familiar GUI (graphical user interface) metaphors and almost photo-realistic renderings have now addressed those early credibility problems so well that belief in "what appears on the screen" can supercede reality and intuition. Unfortunately, this realism limits the computer display screen space available for data presentation and constrains problem solving by guiding users with only information that the original designer feels is critical. Realistic representation of the process actually becomes a barrier to problem solving.

Database-to-brain bandwidth, the rate of presenting and comprehending execution process data, now limits the overall process-to-brain bandwidth. Operators, supervisors, and managers struggle to correctly identify trends, problems, and discontinuities in extremely large quantities of execution process data, before taking corrective action.

Petroleum production and glass melting are two representative, though very different, processes where automated control systems have both improved workers' ability to adjust their processes and simultaneously overwhelmed them with gigabytes of process data (2).

A simple HMI screen in Figure 1 represents petroleum process equipment with symbols and the process flow with lines and directional arrows. Text areas and bar charts display status and five pressures, a flow rate, three valve percent openings, and six tank levels. These few points and the general arrangement diagram consume the entire screen space.

What are the interactions between these variables? Are operating conditions nominal at this moment, beginning to drift away from nominal, or returning to nominal after an upset? Are these conditions unusual? What will the conditions be in the next few moments? Lastly and most importantly, what action, if any, should be taken?

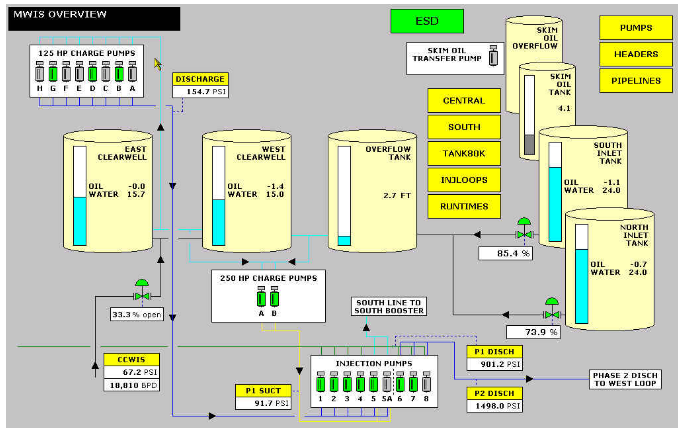

Figure 1 - Simple HMI Display Screen

Methodologies such as statistical process control help answer these questions, but compete with routine manufacturing tasks for time. While much of the workforce was not mathematically proficient in secondary school, even those skills have atrophied.

DESCRIPTIVE STATISTICS IN THE WORKPLACE

It is easy to implicitly assume a symmetrical Gaussian "bell curve" distribution of datapoint values (the numerical value of a variable at a particular datetime) about a nominal, or preferred, value with common statistical software. Figure 2 shows what appears to be such a distribution of hourly datapoint values for air-to-gas ratio, an important environmental and manufacturing cost parameter in glass melting. It is a typical bivariant variable, where the largest number of datapoints occurs at the nominal value, with deviations below to the minimum and above to the maximum. Careful study, however reveals that the distribution above nominal falls off much more rapidly than the distribution below. That is, there is a larger "tail" on the lower side of the curve. There is also a secondary distribution of datapoints with values around 12.6. These two subtle details are significant. The distribution is usually skewed below nominal because low air-to-gas ratios are better for the process. The secondary distribution of datapoints around 12.6 is due to an accidental setpoint shift.

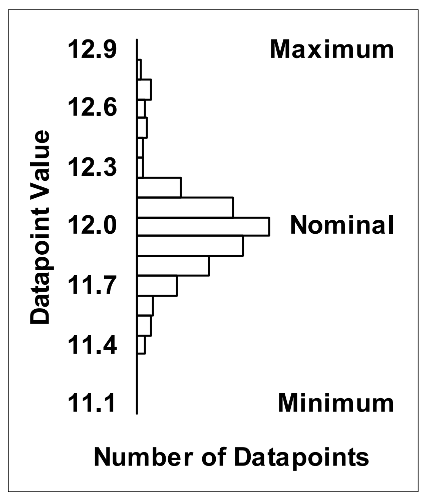

Figure 2 - Distribution of Hourly Datapoint Values for Air to Gas Ratio

PROBLEMS WITH ASYMMETRICAL DATA DISTRIBUTIONS

Beyond the subtle problems in the apparently normal data of Figure 2, process sensitivity to deviations above nominal is generally not the same as sensitivity to deviations below. This is often due to different constraints, for example, burst pressure above versus lower yield or throughput below. The situation becomes disproportionately worse as the process deviates from nominal.

A univariant variable, where either the maximum or minimum is preferred, is an extreme example of this asymmetry. Mathematically Poissonian or error function, its distribution is problematic for most statistical software and leads statisticians to false conclusions when they, often unknowingly with their software, apply the simplifying assumption that the distribution is Gaussian.

Figures 3 and 4 are examples of asymmetrical distributions of common univariant data. Efficiency is an important manufacturing cost parameter, where the largest number of datapoints occurs at the maximum value, with deviation only below nominal, as shown in Figure 3. Defects per hour is an important quality parameter that is sensitive to any increase. The largest number of datapoints occurs near the minimum value, with deviations predominantly above nominal, as shown in Figure 4.

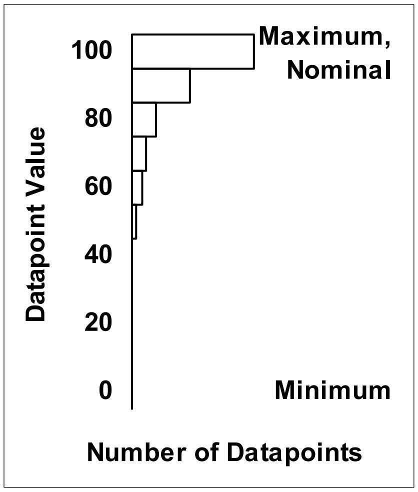

Figure 3 - Distribution of Hourly Datapoint Values for Percent Efficiency

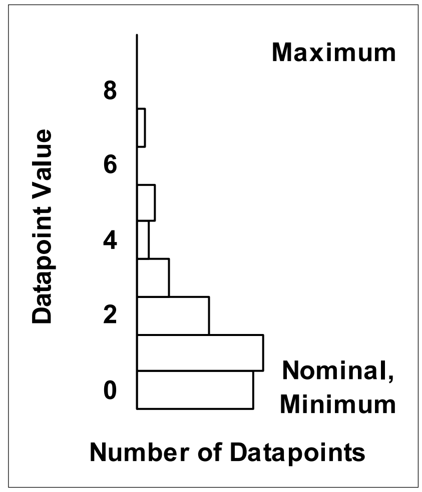

Figure 4 - Distribution of Hourly Datapoint Values for Defects per Hour

In business transactions, error rate and rework increase rapidly as the number of items handled per hour increases above nominal. This introduces costs that must be balanced with increased labor cost per item, from operator inactivity, below the nominal workload. The three variables correspond exactly to Figures 2 through 4. Figure 2 is the number of items handled per hour, Figure 3 is the percentage of time that the clerk is busy, and Figure 4 is the number of errors committed per hour. The three variables interact. As a clerk handles an increasing number of items per hour, the clerk is busy an increasing percentage of the time. This reduces the labor and capital costs per item, but the error rate increases simultaneously, requiring rework and with attendant, strongly increasing additional costs.

Unfortunately, real processes are comprised of many variables with non-linear sensitivities and asymmetrical statistical distributions. These, and appropriate mathematical and graphical software applications to deal with them, are beyond the comprehension of those who are not mathematically proficient.

Dr. Edward Tufte, a Yale professor, defined cognitive art and envisioning information as the intersection of image, word, number, and art (3). His studies further explain why representing data as abstract images removes the above constraints and vastly increases database-to-brain bandwidth.

THE PRINCIPLE OF ABSTRACTION

Abstraction consists of obscuring the technical and mathematical details surrounding each datapoint and arraying screen elements representing many thousands of datapoints so that patterns in the data are obvious. The line chart of Figure 5, for example, shows twelve datapoints for each of eighteen variables, a total of only 216 datapoints. Neither considerable study of the chart nor close examination of the raw data in the table of Figure 6 yields much understanding of the data.

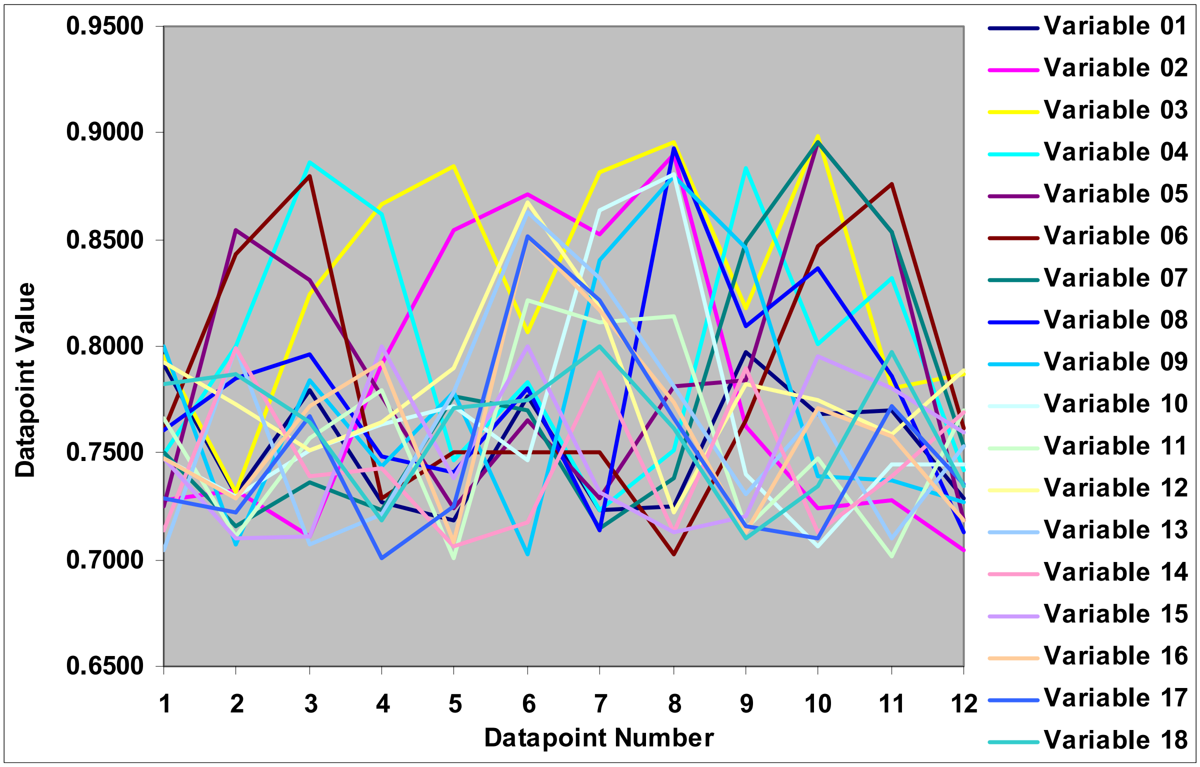

Figure 5 - Line Chart Representing Variables 01 Through 18

Datapoint	1	2	3	4	5	6	7	8	9	10	11	12
Variable 01	0.7915	0.7304	0.7795	0.7271	0.7188	0.7767	0.7233	0.7254	0.7969	0.7677	0.7701	0.7283
Variable 02	0.7278	0.7325	0.7109	0.7915	0.8545	0.8710	0.8521	0.8896	0.7625	0.7244	0.7278	0.7044
Variable 03	0.7955	0.7311	0.8248	0.8665	0.8844	0.8062	0.8818	0.8958	0.8176	0.8986	0.7801	0.7870
Variable 04	0.7495	0.8000	0.8859	0.8619	0.7463	0.7828	0.7234	0.7514	0.8833	0.8014	0.8317	0.7421
Variable 05	0.7252	0.8542	0.8312	0.7766	0.7238	0.7655	0.7290	0.7811	0.7845	0.8957	0.8534	0.7180
Variable 06	0.7605	0.8428	0.8799	0.7291	0.7499	0.7505	0.7502	0.7025	0.7653	0.8465	0.8761	0.7615
Variable 07	0.7505	0.7159	0.7361	0.7233	0.7767	0.7701	0.7151	0.7377	0.8489	0.8956	0.8538	0.7533
Variable 08	0.7605	0.7851	0.7964	0.7488	0.7409	0.7804	0.7141	0.8924	0.8094	0.8363	0.7855	0.7129
Variable 09	0.8000	0.7076	0.7845	0.7437	0.7781	0.7025	0.8403	0.8794	0.8456	0.7387	0.7371	0.7272
Variable 10	0.7476	0.7301	0.7516	0.7637	0.7721	0.7468	0.8636	0.8806	0.7404	0.7065	0.7443	0.7443
Variable 11	0.7667	0.7112	0.7573	0.7790	0.7008	0.8217	0.8117	0.8144	0.7149	0.7479	0.7011	0.7698
Variable 12	0.7926	0.7730	0.7512	0.7640	0.7899	0.8674	0.8184	0.7220	0.7824	0.7748	0.7589	0.7890
Variable 13	0.7047	0.7988	0.7070	0.7213	0.7782	0.8624	0.8315	0.7810	0.7308	0.7689	0.7098	0.7531
Variable 14	0.7134	0.7988	0.7395	0.7426	0.7061	0.7172	0.7878	0.7124	0.7898	0.7132	0.7394	0.7691
Variable 15	0.7472	0.7103	0.7106	0.7997	0.7382	0.8000	0.7319	0.7126	0.7202	0.7951	0.7800	0.7595
Variable 16	0.7473	0.7290	0.7721	0.7929	0.7084	0.8512	0.8171	0.7737	0.7131	0.7707	0.7576	0.7187
Variable 17	0.7290	0.7219	0.7670	0.7010	0.7246	0.8517	0.8214	0.7663	0.7157	0.7100	0.7721	0.7353
Variable 18	0.7823	0.7866	0.7645	0.7187	0.7707	0.7748	0.7997	0.7615	0.7101	0.7342	0.7975	0.7340

Figure 6 - Datapoint Values for Variables 01 Through 18

	Average	StdDev
Variable 01	0.7530	0.0298
Variable 02	0.7791	0.0693
Variable 03	0.8308	0.0540
Variable 04	0.7966	0.0575
Variable 05	0.7865	0.0597
Variable 06	0.7846	0.0599
Variable 07	0.7731	0.0602
Variable 08	0.7802	0.0511
Variable 09	0.7737	0.0577
Variable 10	0.7660	0.0523
Variable 11	0.7580	0.0441
Variable 12	0.7820	0.0362
Variable 13	0.7623	0.0511
Variable 14	0.7441	0.0340
Variable 15	0.7504	0.0355
Variable 16	0.7627	0.0434
Variable 17	0.7513	0.0464
Variable 18	0.7612	0.0302

Figure 7 - Statistics of Datapoint Values for Variables 01 Through 18

Given time, one might hope to find something with the simple statistics, average and standard deviation, as shown in Figure 7. Again, as with the line chart and the raw data, no pattern is evident, even with considerable time expended.

Datapoint	1	2	3
Variable 01	0.7915	0.7304	0.7795
Variable 02	0.7278	0.7325	0.7109
Variable 03	0.7955	0.7311	0.8248
Variable 04	0.7495	0.8000	0.8859
Variable 05	0.7252	0.8542	0.8312
Variable 06	0.7605	0.8428	0.8799

Figure 8 - White Screen Elements Represent Datapoints With Values Greater than 0.8000

Figure 8 demonstrates the principle of abstraction with an excerpt of the Figure 6 data. White screen elements represent datapoints with values greater than 0.8000 and black screen elements represents all other datapoints. The underlying numbers and technical details are unnecessary, since the brightness of the screen elements now signifies the value.

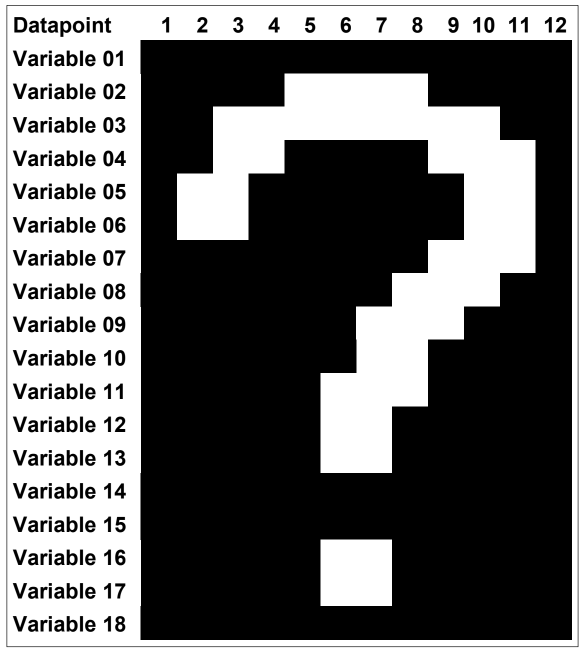

Figure 9 - Abstract Image Representing Variables 01 Through 18

Applying the principle to the entire 216 datapoints produces the abstract image of Figure 9, in which a pattern is clear, without any mathematical analysis. The individual screen elements are considerably smaller than in the other figures, and can be made even smaller, while still conveying the information. Hence, an image can represent many datapoints in a given screen space, meeting the challenges of Dr. Zuboff and Dr. Tufte.

While this is an example of applying a single grading criterion (go / no go, or pass / fail), most variables require a representation that indicates degree of deviation and whether the deviation is above or below nominal.

BRIGHT IS BAD

A line chart and a graphic strip in Figure 10 represent the same 480 datapoints of air-to-gas ratio as the histogram of Figure 2. Older data is at the graphic strip's "left" and newer is at its "right". Darker places in the strip represent data with values close to nominal, or desired. Brighter places are farther from nominal. Blues indicate deviation below nominal and reds are above nominal.

The graphic strip is shown much wider than a strip is in an actual abstract image, for clarity. This particular representation is scaled symmetrically, so screen elements representing datapoint values with deviations above nominal are equally as bright as those elements representing equal deviations below nominal. Choosing a maximum value closer to the nominal, for example, would represent datapoints with values above nominal with much brighter screen elements than those datapoints with similar deviations below nominal.

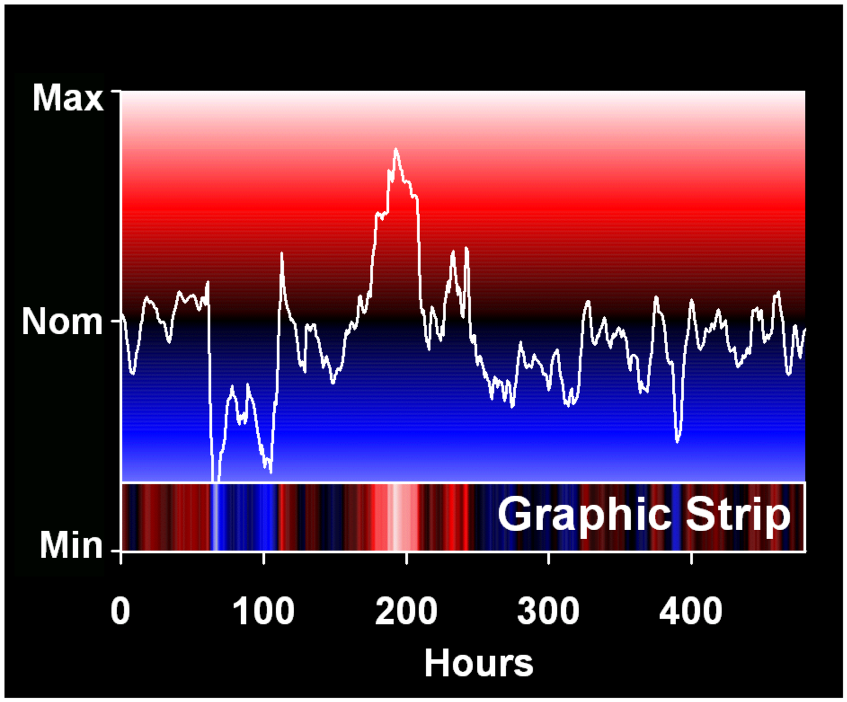

Figure 10 - Line Chart and Graphic Strip for Air to Gas Ratio

In cases where deviations above nominal are desirable, the screen elements representing those datapoints should be dark. An example is the percent efficiency data of Figure 3. Conversely, deviations below nominal are desirable with the defects per hour data of Figure 4 and screen elements representing datapoints with values below nominal should be dark.

An abstract image consists of up to 50 of such narrow horizontal graphic strips, one for each variable, bivariant and univariant data alike, adjacent to each other. Vertical patterns in an image are events that affect several variables at the same time. Horizontal streaks are individual variables.

An operator viewing an abstract image instantly recognizes the screen element brightness and color patterns without understanding the scaling and statistical issues for individual variables and without understanding how the variables are physically related.

DATA DENSITY

A single abstract image, representing forty pages of percent efficiency data, illustrates the data density that can be achieved. The image has a consistently blue hue, because all data is of the same univariant type and all deviations are below nominal. The general bivariant case, as discussed previously, will have red and blue patterns. Operators, supervisors, and managers can rapidly and accurately identify opportunities for improvement and ignore insignificant data in the abstract image.

The top caption of Figure 11 shows that the image is the "Wichita Falls" facility, "Line One" area, "Primary Operation" process, and "Efficiency" dataset. It displays 20 days of hourly data. Altogether, 24,000 datapoints make up this single image, 480 datapoints for each of 50 variables. Pattern recognition is simple. Analysis is immediate. Darker areas represent better performance. Brighter areas show poor performance, especially the bright central blotch.

Identifying variables and datetimes for an ROI (region of interest) must be instantaneous. This is the essential information to involve existing control or maintenance systems or to obtain help from technical departments.

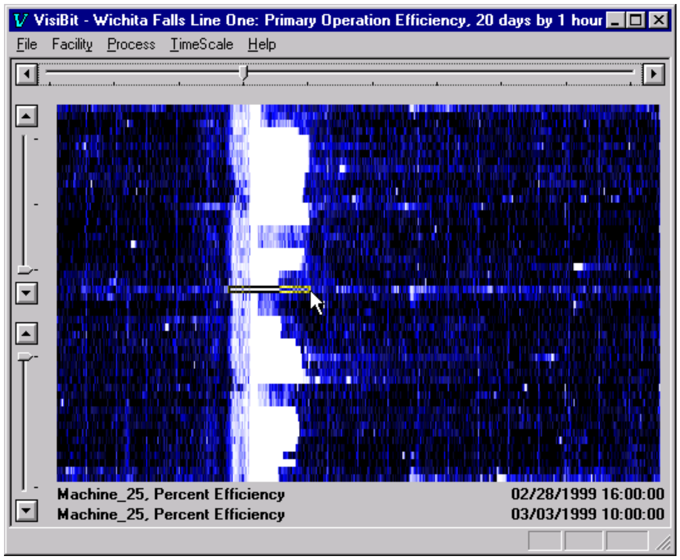

Figure 11 - 24,000 Datapoints Showing Globally Disastrous Percent Efficiency from 28 February Through 3 March and Machine_25 Bad Throughout

An ROI with poor performance is selected with the mouse pointer in Figure 11 and the variable is shown to be Machine_25, Percent Efficiency.

While Machine_25 is consistently worse than the other machines, there is a globally disastrous loss in efficiency from 16:00:00 28 February 1999 through 10:00:00 03 March 1999.

Faint vertical lines are evenly spaced across the image. These mathematical quirks occur at exactly midnight and are due to a miscalculation in the data acquisition routine. They are not noticeable in the raw data.

CONTINUOUS PROCESSING

Database-to-brain bandwidth also depends on having the abstract images ready for immediate delivery to the operator. Interactive database queries and generating graphs, charts, and reports are a common data analysis practice. While this, alone, restricts the number of potential users to those who can perform the technical tasks, the interactivity makes operators wait while databases cough up data and software processes it. Some delays are surprising. For example, while data insertion rates are emphasized, database query responses are not always as good as might be expected.

Lastly, then, increasing the database-to-brain bandwidth requires that the data processing be automated to continuously generate and archive small image files for rapid presentation across a network. Automation further improves profitability, as the database server experiences a steady load, not peaks from operator queries, and the image server and view clients can be minimal computers.

An image generator automatically and continually queries ODBC databases and converts the data to images, "rolls up" short timescale images into long timescale summary images, and stores the images on an image server for rapid retrieval by view clients. Figure 12 shows the general arrangement.

Query volumes are relatively low, since they are only for new data in the time interval between image updates, typically between two minutes and one hour.

View clients "live update" as new data is available. An image displaying current data "shifts left" as new data appears in the most recent screen elements at the right of the image.

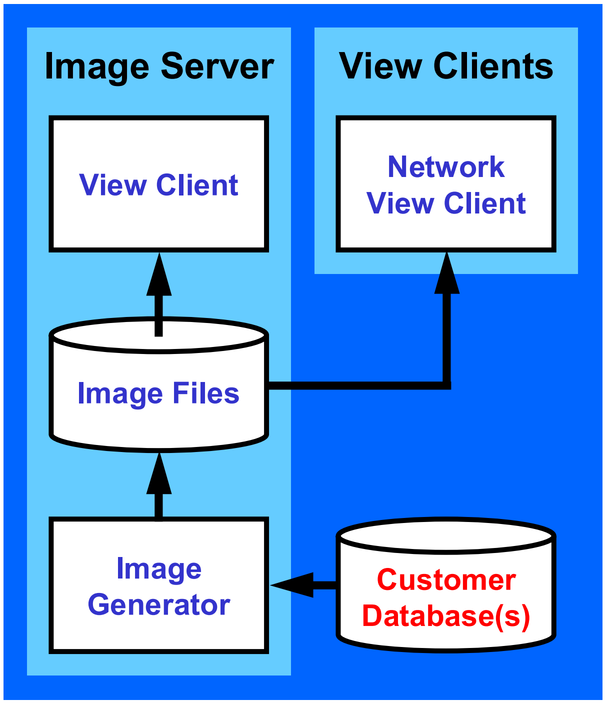

Figure 12 - General Arrangement of Image Server and View Clients

CONCLUSION

The past quarter century of computing has brought the two changes depicted in Figure 13. The data error rate has fallen dramatically, with the effect of increasing the credibility of machine-generated data and displacing human interactions with processes. Secondly, the data collection rate has risen dramatically, compounding the problems that workers face with absorbing and using the data to understand their increasingly distant processes.

Database-to-brain bandwidth, the rate of presenting and comprehending execution process data, now limits the overall process-to-brain bandwidth. Operators, supervisors, and managers struggle to correctly identify trends, problems, and discontinuities in extremely large quantities of execution process data, before taking corrective action.

Representing data as abstract images and automating the image generating process relieves these constraints and vastly increases database-to-brain bandwidth.

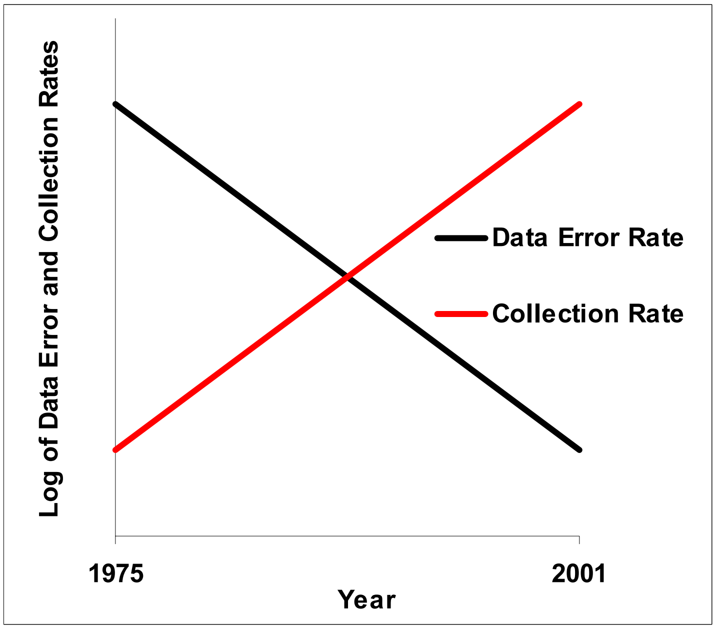

Figure 13 - Falling Data Error Rate and Rising Data Collection Rate

ACKNOWLEDGEMENTS

Figure 1 used by permission of Chevron USA Production Company. Figures 2 through 13 used by permission of VisiBit Corporation.

REFERENCES

1. Zuboff, Shoshana, In the Age of the Smart Machine: The Future of Work and Power, Basic Books, Inc., New York, New York, 1988.

2. Boland, John, Garcia, Ray, and Johnston, Jeff, “Abstract Presentation of Process Data”, ISA2001 Technical Conference, Houston, Texas, September, 2001.

3. Tufte, Edward, Envisioning Information, seventh edition, Graphics Press, Cheshire, Connecticut, August, 1999.