Computing Science and Statistics

Volume 27

Statistics and Manufacturing with Subthemes in Environmental Statistics, Graphics and Imaging

Proceedings of the
27th Symposium on the Interface
Pittsburgh, PA, June 21-24, 1995

Editors

Michael M. Meyer

James L. Rosenberger

INTERFACE
FOUNDATION
OF NORTH AMERICA

EPA/600/A-96/103

Perspectives on the Analysis of Massive Data Sets

By

Daniel B. Carr
Center for Computational Statistics
George Mason University
Fairfax, VA 22030

Abstract: This talk presents perspectives on the analysis of massive data sets. These perspectives derive from past efforts to develop tools for the analysis of large data sets and speculations about the future. The perspectives touch on many topics: previous research, recurrent problems, standard tricks for addressing problems, new problems, and new opportunities. In looking forward, it seems that evolving notions about the man-machine interface, models, and model criticism will play an important role in massive data set analysis. The extent to which the statistical community responds this evolution will have far reaching consequence in terms of the influence of our profession with the scientific community.

1. Introduction

The analysis of massive data sets (AMDS) is an important topic. Massive data sets present new challenges and new opportunities for learning about the world around us. In fact massive data sets provide the opportunity to learn how the statistical community will respond to the challenges. The progression from large to massive data sets has occurred steadily over the last two decades. It might be thought that the statistics community has followed this progression and is well-prepared to address massive data sets. However, the analysis of large data sets has drawn little attention. Much of the computational statistics community has focused its attention on computationally intensive methods for small data sets. We can attack small problems with portable PC's and crunch medium problems with workstations. Addressing massive data sets requires more finesse, new computational environments and foreign (and sometimes simpler) ways of thinking.

In this paper I provide a sketch of the AMDS landscape. The approach uses a series of brief descriptions to provide a broad view of AMDS. A narrow view of AMDS might focus on a new algorithmic approach to complexity reduction or characterize the feasibility of applying a particular class of algorithms in terms of n, the number of cases, and p, the number of variables. A broader view also addresses complexity reduction issues for the analysis team that is pursuing answers to subject matter questions. As part of this broader view, I raise concerns about whether or not statisticians will be involved on such analysis teams. The sketch of the AMDS landscape provides a basis for developing an involvement strategy.

Behind my view of the research landscape lies experience from working on the Analysis of Large Data sets (ALDS) project (for some early papers see Carr 1979 and Nicholson et al. 1980). In this paper I relate perceived patterns to my historical experience. The perceived patterns are my interpretation and undoubtedly reflect my biases. Hopefully the patterns reflect the biases of other data analysts.

Why would anyone be interested in AMDS? After years of consideration I am not so sure that the exploratory analysis of large (massive) data sets is exactly the challenge that catches my fancy. To me, the computing revolution is not so much that data sets are massive as it is that more and more things are recorded electronically and can be treated as data. The word massive hides at least four types of opportunities:

> The word massive does not immediately convey the opportunity to integrate data from multiple sources nor the opportunity to pay attention to details.
>
> The word massive does not convey the notion that some types of data are new and deserving of new analysis methodology.
>
> The word massive does not convey the increased capability to transform information into forms more suitable for human consumption. Historically we have had to adapt just to be able to use computers.
>
> The word massive does not convey the opportunities provided by working in teams in the pursuit of scientific inquiry.

Massive data sets should be considered in the full context of the computing revolution. It is likely that few are interested in data sets just because they are massive.

2. How Large is Your Large?

The Analysis of Large Data Sets (ALDS) project was a DOE

project that flourished about fifteen years ago. The obvious motivation was the wide-spread existence of large data sets within the DOE community that were presumably worthy of analysis. Previous research such as the Bell Telephone Laboratory Analysis of proton data from the Telstar I Satellite (Gabbe, Wilk, and Brown 1967) and methodology developments such as projection pursuit (Tukey and Friedman 1974) provided inspiration. This led to a white paper entitled Exploratory Analysis of Large Data Sets (Nicholson and Hooper 1977) and subsequent DOE funding in 1978. By 1979 the ALDS project was under way with a Ramtek 9400 monitor (1280 x 1024 with bit-plane control) and a VAX to drive it.

Thinking about the analysis of large data sets was part of the spirit of the times. In 1977 the IMS devoted a special meeting to the topic. The following are some quotations from the 1979 DOE Statistics Symposium: Lou Gordon said, "No sufficient statistics, not reducible." Charlie Smith said "File update and analysis histories must be maintained on the computer," and "greater than 10^{12} bytes with more than half of that devoted to data base history." Leo Brieman noted, "Not large, but large and complex is the problem." Dick Beckman wondered, "Are large data sets more properly a question for computer science?" Many at the symposium asked ALDS team members what they meant by large.

The ALDS response for exploratory analysis consisted of two statements: (1) The quantity of data taxes our technical ability to organize, display and analyze. (2) Knowledge of the subject field is insufficient to define an analysis process that will summarize the information content of the data. This response got away from the game of "my data set is larger than your data set." However, a definition stated in relative terms proved problematic. When we learned to handle a data set with some ease, it was no longer large. Consequently we could never solve the problem of large.

Recently Huber (1994) provided a static definition in terms of bytes (tiny=10^2, small=10^4, medium=10^6, large=10^8 and huge=10^{10}). The labels seem reasonable by today's standards. If the labels don't change, perhaps we can master large data sets.

3. Early Issues

The ALDS project had a review panel to make sure that the project was aware of important issues and to steer us toward tasks we could address. Over the years panel members and contributing observers included Doug Bates, Peter Bloomfield, Mike Feldman, James Foley, Jerry Friedman, Jack Heller, Richard Quanrud, Roland Johnson, Benjamin Tepping, Marcello Pagano, Ari Shoshani, and David Wallace. John Tukey also participated in project review and Leo Brieman served as a consultant. Topics from the 1979 panel meeting included data quality problems, lack of homogeneity and data set dismemberment, sampling and cross-validation, integrating multiple related data bases and use of empirical Bayes methods, data base complexity and complex reduction methods, rapid data access and self-describing binary files, graphical displays and methods of exploiting the hardware (display list, pan and zoom and the video lookup table), extension of statistical packages such Minitab® and PSTAT to handle at least a million observations. We discussed issues like common tapes format just like researchers today discuss the need for common file formats. With the input of distinguished review panel members it was pretty hard not to be aware of the important issues of the time. (Hall, 1980 and 1983, provides panel review comments.)

Hardware and software environments have changed dramatically since the late 70's. Some of the ALDS discussions are no longer relevant. However, many of the issues remain the same. The analysis of large data sets brought out one of the most recurrent and frustrating of issues. Different disciplines have a hard time communicating. Still, a path to communication is clear: jointly funded research in AMDS.

4. Selected Research Areas and the ALDS Approach

Since the ALDS team did not have the database environment to address massive data sets, research turned to what is now called information visualization. In the early 1980's we worked with color anaglyph stereo, color encoding, rocking, 4-D rotation, and generalized glyph drawing. In the 1982 ASA graphics exposition, our detective work examples included anaglyph stereo and class-colored plots, from stem and leafs through generalized glyphs. Our large n adaption to the scatterplot was binned-data plots. Publications soon followed that used binned data in smoothes, scatterplot matrices, density different plots, animation, and in graphical subset selection and display. Figure 1 shows a scatterplot matrix of binned plots. (See also Carr 1991) Each little binned plot in the Figure 1 represents a roughly quarter of million points. This particular variation represents the log of counts in hexagon bins using the size of hexagon symbols. Carr et al. (1987) presented a color-based visualization approach that provides more detail for the counts. Note that the binning approach scales to much larger n and that the outliers, for example in band 1 versus band 2, are readily visible. Elementary sampling approaches may work for some massive data set problems but not for the task of learning about outliers. In later research we used binned-data in image processing algorithms (gray-level erosion and thinning), in multivariate versions of box plots and in cognostics algorithms.

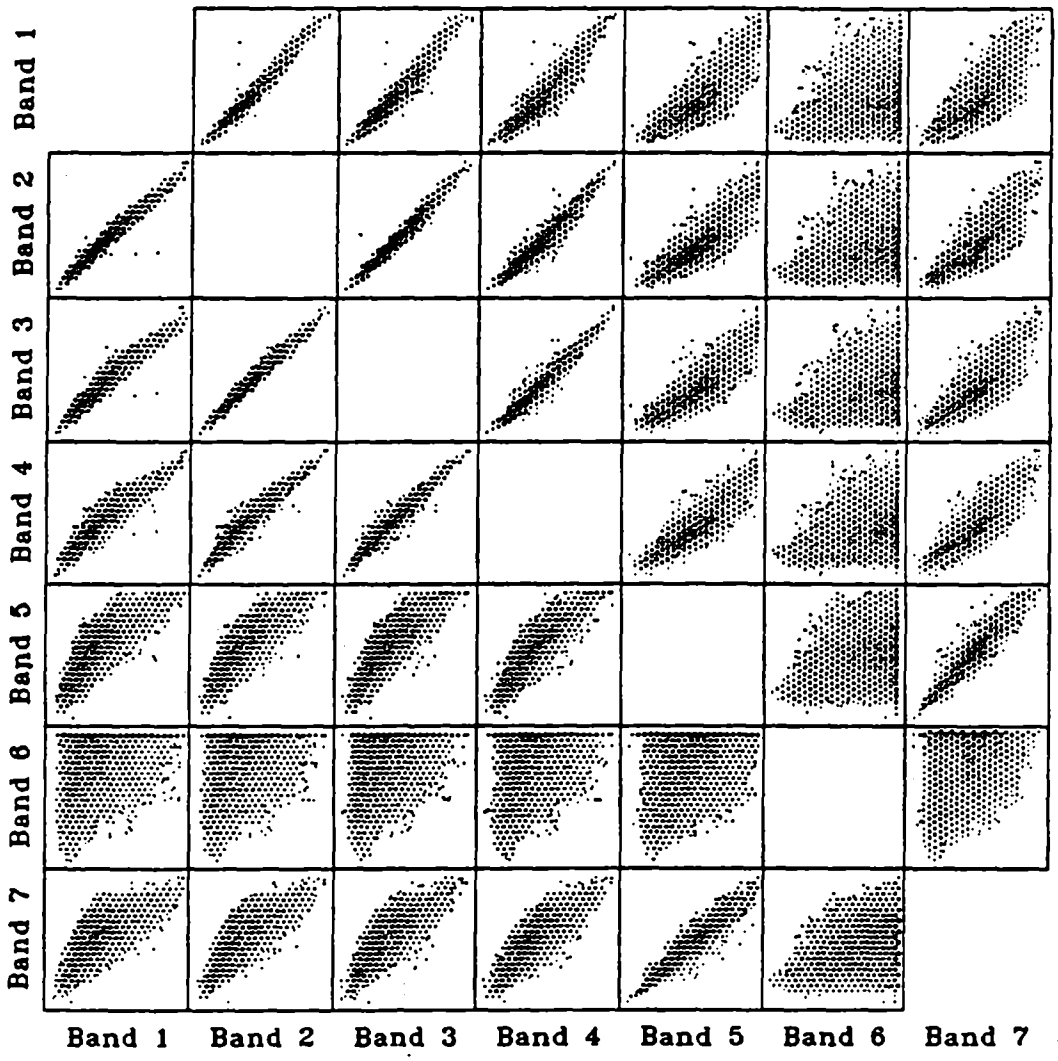

Figure 1. A Hexagon Binned Scatterplot Matrix. The data set is a 512 x 512 Landsat image. Each plot represents more than 250,000 pixels. The spectral bands correspond to different wavelengths of light. Band 1 covers .45 to .52 μm and Band 7 covers 2.08 to 2.35 μm. The bands are in wavelength order except Band 6 that covers 10.4 to 12.5 μm. Note the bivariate outliers that would not be found in typical univariate histograms.

The ALDS cognostics research warrants some description because it is so little known and so close to what is now called intelligent agents. John Tukey coined the term cognostics. One definition is computer-guiding diagnostics. (Ultimately the guidance is for data analysts.) Given an overwhelming number of plots to review, we wanted algorithms that would decide which virtual plots to store for later review.

Our application involved looking at computational fluid dynamics solutions sets as data. For example we would ignore the spatial index on vorticity vectors, bin the 3-D vectors into truncated octahedron cells and then use algorithms to decide if the density pattern was unusual enough to warrant attention. Features of interest included clusters, arm like appendages to the high density region and holes. Usual density patterns could be associated with low model resolution in turbulent areas or distinctive flow patterns (such as the jet stream). Fuzzy ellipsoids constituted a class of uninteresting views. The algorithms ranked 2-D and 3-D bin-based density summaries. The variable pairs and triples included intermediate quantities such as derivatives of velocity as well as paired scalar quantities such as temperature and pressure. Our implementation ran on a parallel processor one time step behind the CFD model.

Some of the plots we found were strange and fascinating. Figure 2, from Carr (1991), was an image flagged for its high degrees of "armness". The variables are partial derivatives of velocity. Presumably a higher resolution CFD grid would produce values that filled in the gaps. The solid black center region in the plot contains the high density part of the plot and 95 percent of the total counts. The gray level thinning algorithm operates on the remaining cells. The hexagons symbols that nearly touch indicate strings of two or longer that are left behind by thinning. Small filled hexagons are isolated cells. Small open hexagon are eroded cells, these may appear filled in this small plot but are identifiable because they are not isolated.) The measure of armness was the number of cells left behind relative to those subject to thinning and exceed 60 percent. We had looked at thousands of "statistical data" scatterplots and had never seen patterns like this.

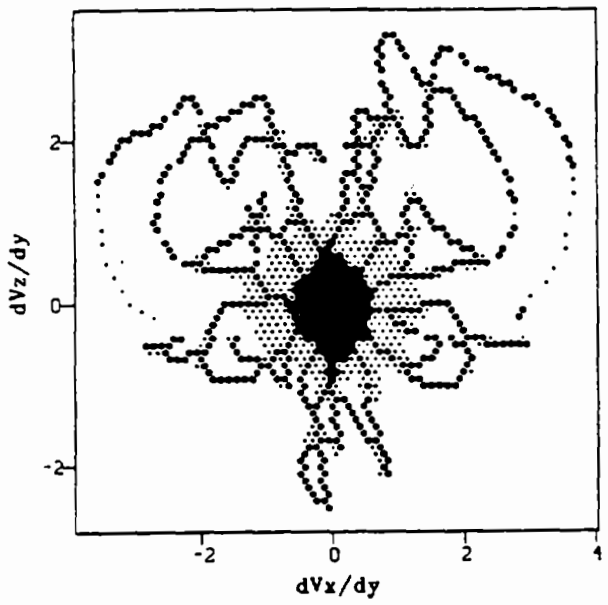

Figure 2. Gray-level Thinning of Hexagon Cells. The variables are partial derivatives of velocity components in a CFD solution set. The gaps between the high density region in the center and the loops suggest usual phenomena or lack of grid resolution in the model.

So far ALDS has had little impact on the statistics community. At first glance one might think that software vendors would snap up the public domain ALDS modifications of Minitab. (The user manual of Littlefield and Carr 1986 indicates the many graphics capabilities.) However, the community with color workstations was small back then. Even today much of the statistics community does not yet have 1024 x 1280 resolution monitors. Tools for really utilizing bit planes will not appear in force until typical computing systems have more than eight bit planes. Software vendors design for the mass market. The technology transfer rate depends on how quickly the hardware and software becomes affordable to the majority of the community. With affordable gigabyte drives, the statistics community may soon give binned scatterplots at try.

The ALDS approach was to analyze data sets. Since looking at data sets was our business, at least four analysis steps were obvious. 1) An early step toward understanding the data is to get the background, sometimes both literally and figuratively. Tasks include reading about the metadata, talking to the people who collected the data or designed the sensor systems, and gathering additional related data. 2) Much early work involves getting the data into the standard shape (typically n cases x p variables) to apply the standard algorithms. 3) Managing the analysis process is a major task, especially in a quality assurance environment. For example, the volume of derived data can easily exceed the original. This will be true of the Earth Observing Station (EOS) data even before the data is distributed around the world. 4) Content rich and scientifically important data sets are often of recurrent interest. However, human memory fades, and both hardware and software evolve. Historically the storage media has been subject to deterioration. Unfortunately analysis archiving (AA) is usually a last minute thought that occurs when project funds and energies are exhausted. Human nature does not change rapidly, and many facets of the analysis process have changed little over the two decades.

The ALDS project worked on managing the data analysis process. We had tree-based representation of the analysis process, mousable icons for "aha's" and even a computer-controlled cassette tape for voice annotation. We envisioned a day when the logs would be analyzed to gain a deeper understanding of the analysis process.

Our first papers were Carr et al. (1984) and Nicholson et al. (1984). Oldford and Peters (1985) and Becker and Chambers (1986) quickly noticed our beginning efforts and produced some really promising work. If the quality assurance movement for data analysis had caught on as we had expected, their work would have attracted much more attention. Now the field seems dormant. This field needs some serious attention. Computer programmers use software management tools for multiple person software engineering projects. AMDS teams need something similar.

Speed was a recurrent issue for ALDS. The notion of binning traded off resolution for speed. Reducing temporal resolution also increases speed. For example, weekly data can be aggregated to seasonal data. A 52 to 4 reduction is not a lot but factors of 10 here and there can help. Of course, different patterns appear at different scales of resolution. That is why scientists are always pressing for higher resolution. Something

is lost by summarization methods that reduce resolution. However, many data sets are never examined. The ALDS position was that it was better to reduce the resolution used in the analysis than to drop the whole data set.

AMDS teams will benefit from software and hardware engineers who can directly address speed problems. Software engineers typically design for generality. Once specific tasks are identified, significant speedups can be gained by removing layers of code. At a recent massive data workshop Albert Anderson of the Population Studies Center in Ann Arbor Michigan reported PC speed-ups on the order of 1000 for the simple task of computing a mean. One key was maximizing size and use of fast processor cache. This is an engineering task, not a statistical one.

AMDS will involve compromises. By agreement we avoid thinking about certain classes of finite precision limitations. For all the computed images, has anyone ever seen a correct view of the Mandelbrot set? Consider an example closer to home. Suppose a researcher wants to split a 200 case data set into two equal subsets of size, a training set and a validating set. Ideally the researcher would like the training set to be an equally likely selection from the approximately 2^{198} possible splits. Now the traditional accept/reject selection method will typically generate a different subset if the random number seed changes. On a 32 bit machine there are no more than 2^{32} such seeds. Clearly the majority of subsets are excluded from selection by the standard subset generating process. Perhaps the process is sufficiently random for our conceptual purposes. In not we are in deep trouble when in comes to selecting large equally-likely subsets from massive data sets. Perhaps we can simplify high-dimensional problems when variables seem sufficiently independent or sufficiently conditionally independent

5. Funding and Analysis Teams

DOE funding was known for its ebbs and flows. It became apparent that if ALDS was to survive the project needed to attach itself to a well-funded application area. We began applying our methodology to computational fluid dynamics (CFD) models. We soon discovered that we had little chance of being heard in the CFD community without a professional CFD modeler on our team. By the time we got a CFD modeler on our team it was too late to salvage the project. Nonetheless, during the last year of the project the modeler provided a wealth of experience and new perspectives. The modeler became our translator and we began getting opportunities to present our work in CFD forums.

For those making the transition from a pure statistics training into world of applications world, it is important to know how things work. The existing infrastructure often makes it difficult to engage in cross-disciplinary work. To publish in journals one needs to know the language, and often the key people to reference. Good ideas and solid research are not enough. Certain credentials can be crucial for funding. Common knowledge is that the letters "MD" are particularly important for NIH funding. There is little point in questioning this wisdom or lamenting that fact that major funding for data analysis research is provided without the requirement of having a professional statistician on the team. The point is that there are political as well as scientific advantages to working on an interdisciplinary team.

Little research occurs without funding. To be involve in AMDS requires serious computational resources and significant funding. Statisticians should be aware that our profession is a very small service profession, small in size and small in the amount of direct research funding. While the current interest in AMDS may bring some direct funding, most of the research opportunities will appears a spin-offs from well-funded applications area.

Funding for medically related research is significant. This provides numerous AMDS opportunity areas, from molecular modeling to the study of health care costs. Bill Eddie's session on functional magnetic resonance imaging (fMRI) session at this conference provided an outstanding example of statisticians collaborating on an interdisciplinary team in a medically related area.

The immediacy of profits suggests AMDS opportunities in the area of market analysis. The study of manufacturing provides the opportunity to increase reliability and profits. Profitable uses of administrative data will be identified. The major funding for and lack of statistician involvement with EOSDIS (Earth Observing Stations Data and Information System) suggests that it is yet another opportunity area. Reasons behind lack of statistician involvement so far includes competition and lack of background.

6. Competition and Credibility

Many disciplines are eager to provide analysis methodology. Many physicists and engineers do not perceive that they need help. If anyone needs help, the computer scientists will gladly provide it. They discovered DATA sometime during the last decade (anthropomorphized it on Star Trek) and now can mine knowledge from data at will. As a service discipline we have major competition, not only from computer science but from the directly funded disciplines themselves. The computing revolution means that scientists do things themselves. Software tools have replaced the graphic designers, secretaries, and statisticians.

How many statisticians have the training to make them credible candidates for an AMDS team? Years ago, ALDS project staff visited Karl-Heinz Winker at Los Alamos National Laboratory. He had a direct high-speed link to a CRAY, a large portion of a mobile home devoted to tapes, and special high-speed disks to drive his high resolution monitor. One of his papers suggested that if a visualization system could not display information at 3 gigabaud, the system was wasteful of his eye-brain resources. I heard that in one month during a particularly bad winter he managed to use two Cray-months of time by grabbing unused cycles. In contrast, our project had requested supercomputer time more on the order of one hour. Our visits to places such as NASA AMES, Lawrence Berkeley Laboratory, and NSA provided further cause for computational humility. Yes, we ALDS statisticians had developed nonlinear models to fit tens of thousands of observations at a crack, but was that large? Resources, such as tape robots, put our project resources (a VAX 11/780 and Ramtek display monitor) to shame. It would take many of our data sets to match the amount of data in one satellite image, and then there was NASA's black hole of tapes to consider. The situation today is a bit better than in the past. More statisticians are working in supercomputer environments, but the number is small.

7. Positioning and Education

Bill Eddy reports that fMRI project walked in the door as a question about t-tests. His opportunity to work on the project might be regarded as pure luck. However, Dr. Eddy was at Carnegie Mellon University where computational activities flourish and Dr. Eddy had done very interesting computational work (for example see Eddy and Schervish 1988). He was well-positioned to have the fMRI opportunity and to recognize it as something more than t-tests.

Writers like Hahn (1989) suggest changing the curriculum to prepare statisticians for the data-intensive environments in the work place. Clearly some facets of statistical education need to change if graduating students are to be well-positioned to work on AMDS teams. AMDS statisticians will need to have experience with supercomputers, parallel processing, database systems, data structures, algorithms and high end graphics software. Traditional statistics programs offer little in most of these areas.

When change occurs too slowly, alternatives emerge. Statistics students eager to respond to AMDS challenges may find their needs better served by emerging programs in computational science. Computational science programs provided experience in computationally advanced environments. Some programs even promote team research. This is different than short-term consulting on someone else's research project. The excitement of team research can be hard to forget. Contacts made in school lead to projects in later years.

As the teaching of statistics evolves, we should make efforts include data analysis methodology developed in other disciplines. By history we lay claim to data analysis. The Webster definition of statistics is: the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that by use of mathematical theories of probaliliy imposes order and regularity on aggreagates of more or less disparate elements. We should teach methods that work (despite their origins), debunk wild claims, and contribute our statistical wisdom to when others are working to develop new data-analytic methodology.

Professionals benefit from the education opportunities provided by professional communities. The longer scientists work at data analysis, they more they find in common with statisticians. The statistical community should work even harder at providing opportunities for cross-disciplinary sharing.

Some facets of tradition statistical training serve quite well. Each generation of scientists needs to be weaned away from the one at a time experimental approach and taught experimental design. Each generation needs to be taught that experimenter bias is not just a problem of other disciplines. Each generation needs to be taught to reexamine assumptions. The sophistication of people working with massive data sets usually guarantees thorough discipline-based grounding. However, single discipline grounding frequently leaves blind spots. Often as not, the most important contribution of the statistician is obvious from the statistician's perspective and scarcely worth a footnote in the research literature[1].

---

[1] Every statistician collects discovery stories. Ralph Kahn (JPL) and I co-chaired a session at the IGARRS '94 meeting. Afterwards we went to see the demonstrations at the IDL booth. At the booth David Stern selected a view of the Earth and showed us an animation of newly acquired sensor data. Ralph immediately understood and pointed out patterns in what we were seeing, such as swirling regions that built up behind continents and then broke off. In the session I had talked about the importance of showing differences (a statistical graphics thing) so I asked David to animate the difference between images. David quickly modified the demo program to show differences (impressive). After a brief period of watching differences, satellite trajectories appeared on the globe. Not all the data had been properly aligned. The staff responsible for the data just happened to be close at hand, so David and Ralph

Statisticians ask obnoxious questions like, "Can you verify that the sensor calibration is still applicable?"

Focusing attention on the currently perceived critical path is a common scientific trait. Scientists often presume to know where the problems are. This allows strategic allocation of resources to study the problem. The narrowing of attention can be cost effective but sometimes misses the mark. For a long time it was known that the atmospheric action happened at the equator. While polar phenomena may be localized, little details like the hole in the ozone layer are worth discovering. In general, narrow focus limits the ability to make discoveries and inferences about larger domains. For example, in U.S. environmental monitoring, a common strategy is to monitor at locations where pollution is known to be high. This complicates making discoveries about new regions of high pollution and complicates characterizing the state of the nation.

Sampling methodology for AMDS needs to be carefully considered. Environmental monitoring often uses stratified sampling methods. This provides some improvement of estimates when the strata are correct. Incorrect assumptions about strata can lead to poor estimates. For example, the Washington/Oregon salmon population has been over-estimated for years due to incorrect stratification assumptions. Now it appears that it will be too expensive to save the salmon. Narrow focus and incorrect weighting of information are more common in science than scientists like to admit. Sometimes the consequences are massive.

Skeptical statisticians can be very helpful when experts are ready to build their certain knowledge into the software that preprocesses massive data sets. Summarizing incoming information and passing on the summary is a common technique for controlling the amount of stored data. For example, calculations using the Doppler shift in laser beam reflections can produce estimates of atmospheric wind velocities at various altitudes. The Doppler shift information may exist only during the brief period when it is used to calculate wind velocities. Parameter calculations, such as wind velocity calculations, are not necessarily linear and may incorporate the results of calibration experiments. Statistical adjustments such as background and interference correction are often necessary. Statisticians trained in the traditional areas of calibration, background and interference correction will be able to help. Kahn et al. (1991) raise the issue of validating parameter estimates. It would seem that many statisticians have the tools to contribute, although the size and complexity of the data may take some getting used to.

Model criticism is an area of particular importance. The world of models continues to grow. A short list includes differential equations models, regression models, Monte Carlo models, genetic algorithms, neural nets, fuzzy set models, tree models and hierarchical Bayesian models. Statisticians can assess models for internal consistency, compare models against data (real or simulated), and compare models against other models. Statisticians may not make the best salespersons, but we can market our strengths once we have established computational credibility.

8. Types of Data and Challenge Areas

In the ALDS project, we thought about practical cardinality for data sets. Our first distinction was storable versus flow-through data. For some data sets there is no intention of ever storing all the data. Flow-through data requires the development of models that update as the data flows by. In terms of graphics an obvious problem is scaling. We want good resolution within each time window and comparable plots over time. Having a large buffer helps to achieve a compromise between the two conflicting goals. What can be done with large buffers? Our second distinction was repeatable versus unique data. Computational simulations may generate too much data to store in any detail, but are potentially repeatable. There can be two passes: one to determine global summarization scales and one do the summarization. Only a little research is being done in these area.

Statisticians will encounter increasingly complicated data objects. The notions dependent on Euclidean distance between objects may have to be forgone. For example, the application of molecular similarity analysis in drug discovery research involves binary vectors representing the presence or absence of up to 300 molecular fragments, labeled graphs representing the bonding structures of molecules and scalar fields in R^3 representing the electrostatic flow of molecules (Johnson 1989, Johnson and Maggiora, 1990). Such complex data types provide the opportunity to develop new approaches.

While remote sensing data is often nicely structured, many challenges remain. Consider a currently extreme case consisting of a 2^{14} by 2^{14} pixel image with intensities for 12^3 channels (bands). Our rule of thumb for multivariate analysis is that n should be greater the 3*p. Can we reliably perform

---

called them over to watch. I presume the staff fixed the problem at their first opportunity. It would have been fun to pretend that I, a mere statistician, had seen the problem in the original animation and had thought up the second animation to confirm my suspicions. Unfortunately the idea did not occur to me until later. Simple statistical graphics principles apply to some fairly large data sets.

change detection with two images? Atmospheric conditions and changes in sensor condition induce variability. Sensor position calculations and atmospheric corrections are less than perfect. With massive data sets even computers are not so reliable. I once had a client who went to considerable effort to develop confidence bounds for temperature estimates computed from pixel values in the thermal band. He based his bounds on data from one image and tedious-to-obtain "ground truth" values. The client was not happy with me when I said that he had only one case and that the bounds should incorporate image to image variability. He carefully explained to me that each image was expensive and that his image contained millions of observations. While I still maintain my position, his question made me wonder. How do we borrow strength from the internal structure of such large data sets?

Sensor data provides challenges beyond modeling its internal structure. Sensor data is useful not only in relationship to similar data sets, it is also useful for refining computational science models. For example, an ocean CFD solution set may include surface temperatures. Sophisticated regression models can use the mismatch between the modeled and observed surface temperatures to adjusted CFD model input values or to raise questions about the adequacy of the boundary conditions and grid resolution. While research refining ocean models may be well underway, the task of using sensor data to improve on computational science models provides many statistical opportunities.

I remember colleagues assessing a complicated Monte Carlo code. The clients needed simulation results for different sets of input values. For each set of input values they had to run the code. It turned out that a simple regression model characterized the relationship between inputs and outputs. Within the domain of the explored input parameter space there was no need to run the complicated Monte Carlo model. Trivial calculations using regression coefficients sufficed. The direct study of Monte Carlo simulation intermediate and final results can provide insights.

9. More Data, More Complexity and More Barriers

Key scientists from many disciplines responded to grand challenge queries concerning their information needs. Almost to the person, the answer was that crucial developments depended on having higher resolution, more detailed data. Science is competitive and scientists understand the advantages of being the first to have access to the "best" data. It is hard to imagine scientists saying that they had all the data they could handle for the next few years. The pressure for obtaining higher resolution data will continue.

The mass of available data continues to grow. For years sensors have produced time series with several millions of observations per second. Computer clock speed indexes the ever-increasing collection rate. Spatial resolution is heading toward overwhelming detail with higher spatial-resolution sensors always on the drawing boards. The resolution of measurements at space-time coordinates is increasing. It is scarsely a step from 7 spectral bands to 1728 energy channels. Soon we will have multiple versions of global AVHRR data sets, but we can accumulated such data sets faster by using more sensors. Atmospheric scientists will not be happy until there is complete global coverage at fixed time points. The mass of soon-to-be-available data defies imagination.

The complexity of data objects will increase. Consider the complete medical record for a human being, from the birth record through medical exams that include CAT scans and blood workups, to the death certificate. However does one compare groups of such data objects? How does one use the evolving data object for patient care? If we consider the problem in the abstract it becomes overwhelming. Being pragmatic allows progress to be made (see Powsner and Tufte 1994). The challenge is to refine the pragmatic assumptions to more closely reflect the relationships in the data and to possibilities provided by new computing tools and new understandings of human cognition.

Unfortunately the barriers to analysis are numerous. These include the interdisciplinary barriers and the lack of training described earlier. A big change since the ALDS project has been the emphasis on software and data as intellectual property. The ALDS project could get source code. Today it may be easier to link in special routines to commercial packages, but the source code is often unavailable. The source code is necessary for understanding exactly how things work, throwing out unneeded software layers, and optimizing.

Access to data and algorithms is becoming increasingly difficult. Companies treated data as private property. Researchers exercise their rights to new data until it is well worn. Algorithms such as marching cubes receive patents. Researchers wanting algorithms to remain in the public domain need to copyright their work to keep it from being privatized. Data and data processing methodology translate into money and power. Politicians have noticed. They influence what will and will not be collected. Some information will not be collected.

In the climate of secrecy and security it become difficult to develop an AMDS community. At a recent massive data set workshop, it was not excepted that NSA attendees would be among the featured speakers. A talk by a communication industry representative described the general nature of one massive data sets but said little about data specific and the

analysis that had been done. A representative from the computer chip manufacturing industry indicated that the data sets were massive and that their engineers have specific ways of using the data. Again data and analysis details were missing. One can envision an AMDS conference when everyone comes, not to share, but in the hope that someone else will reveal useful technology. This image may be an exaggeration, but it expresses a concern about the trend toward secrecy and security. The rate of progress in AMDS will be a function of the amount of shared learning. To promote scientific progress, agencies such as NSF will need to work to counter-balance the tendency of companies and universities to isolate AMDS researchers. Further, AMDS researchers need to quickly establish a forum AMDS papers. At least two statistics journals would welcome AMDS papers. Unfortunately the statistics journals reach only a small segment of the AMDS community. AMDS researchers have not yet coalesced into a community and need encouragement.

In many scientific applications, the lack of analysis funding is major barrier to AMDS research. It is particularly easy to be jealous of all the money spent for developing data collection systems and for gathering data. Sometimes the process of collecting data seems to have a life of its own. The purpose can be long forgotten. The data collection might be useful to someone sometime. Sometimes data collection serves as political proof that a problem is being studied. Studying the problem is a great delaying tactic. For example the U.S. still has a massive amount of nuclear material that still needs to be stored. If cursory analysis discovers that the data is lacking in resolution, there is justification for more data collection and the cycle is complete. The data collection industry often escapes notice. However, NASA's black hole of tapes has drawn attention. The EOSDIS is increasing the percentage of funds allocated for data analysis. It would seem that if analysis is really wanted, the percentage of funds for analysis would increase in other areas as well.

10. Data Analysis Environments and Graphics

The ultimate filter for massive data sets is the human mind. If the analyst is the weak link in the analysis process and it makes sense to devote resources toward improving analyst performance. For discussion purposes it might be helpful to imagine that a massive data set analyst was as important as a nuclear plant control room operator, an airplane pilot, or a commander in a command control center. Training programs would be considered extremely important. The environment would be structured toward making important decisions.

Comparing the imagined environment to typical data analysis environment raises many questions. Do our graphics workstations optimize human performance? Tufte (1990) calls attention to the low resolution of the CRT screen as compared to printed graphics. In recent times, readily available laser plotters have improved from 300 to 600 to 1800 line resolution but the workstation screen remains basically same. Further, I am not aware of any plans to extend the gamut of colors currently provided by RGB guns toward the perceptual limits of the human eye. While man-machine interface research is in progress for virtual reality environments, the financial driving force pushes the research toward entertainment applications. We know that virtual reality environments can provide loud sounds, bright flashes and disorienting time-lagged motion. The design of environments for depth of thought is not in the main stream. Current deficiencies in data analysis environments are literally glaring (see Tufte 1989).

Reexpression (transformation) of information is a key to human learning. As a simple psychological description, non-reflexive transformations of internal representations characterize learning. That is, learning has occurred when a new way of perceiving precludes going back to the old perception. Computers enhance our ability to make transformations. Computer transformations mediate our transformations between internal representations. Important computer transformations include mathematical symbols to visual representations and visual representations to other visual representations. Computers can help us take the steps from visual representations back to words, mathematical symbols and our other forms of internal coding. Human assessments and insights are the targets. Toward this end various sensory representations can be useful, including auditory, kinesthetic and olfactory senses. The whisper of a voice seemingly located near one's head can change awareness. We learn through our senses. We learn by doing. We will learn about massive data sets by reexpressing them.

11. Closing Remarks

Statisticians live on the edge of scientific research where purely deterministic models are inadequate and where stochastic approximations are needed to reduce complexity and speed processing. As the volume of scientific research increases, so should the corresponding statistical surface area. It is cause for concern that the profession is not growing more in the face of the current opportunity. I hope that the challenge provided by AMDS provides a rallying call to our profession.

Acknowledgments

Research related to this article by EPA under cooperative agreement No. CR8280820-01-0. The article has not been subjected to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred.

Minitab is a registered trademark of Minitab Inc.

References:

Becker, R. A and J. M. Chambers. 1986 "Auditing of Data Analyses." American Statistical Association 1986 Proceedings of the Statistical Computing Section, pp. 11-18.

Carr, D. B. 1980. "The Many Facets of Large." Proceedings of the 1979 DOE Statistical Symposium, pp. 201-204.

Carr, D. B. 1991. "Looking at Large Data Sets Using Binned Data Plots." Computing and Graphics in Statistics, eds. A. Buja and P. Tukey, pp. 7-39. Springer-Verlag, New York, New York.

Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. 1987. "Scatterplot Matrix Techniques For Large N." Journal of the American Statistical Association 82(398) pp. 424-436.

Carr, D. B., P. J. Cowley, M. A. Whiting, and W. L. Nicholson. 1984. "Organization Tools for Data Analysis Environments." American Statistical Association 1984 Proceedings of the Statistical Computing Section, pp. 214-218. American Statistical Association, Washington, DC.

Eddy, W. F. And M. J. Schervish. 1988. "Asynchronous Iteration." Computing Science and Statistics Proceedings of the 20th Symposium on the Interface. Interface Foundation of North America, Fairfax Station, VA. pp. 165-173.

Friedman, J. H., and J. W. Tukey. 1974. "A Projection Pursuit Algorithm for Exploratory Data Analysis." IEEE Trans. Computing, pp. 891-890.

Gabbe, J. D., M. B. Wilk, and W. L. Brown. 1967. "Statistical Analysis an Modeling of the High-Energy Proton Data from the Telestar I Satellite." Bell Systems Technical Journal, XLVI, 7, pp. 1303-1450.

Hahn, G. J. 1989. "Statistics-Aided Manufacturing: A Look Into the Future." The American Statistician, Vol 43, No. 2 pp. 74-79.

Hall, D. L. (Ed.) 1980. "ALDS 1979 Panel Review," PNL-SA-8781. Pacific Northwest Laboratory, Richland WA.

Hall, D. L. (Ed.) 1983. "ALDS 1982 Panel Review," PNL-SA-11178. Pacific Northwest Laboratory, Richland WA.

Huber, P. J. 1994. "Huge Data Sets." Compstat 1994: Proceedings, (R.Dutter and W. Grossmann, eds.), Physica Verlag, Heidelberg.

Johnson, M. A. 1989. "A review and examination of the mathematical spaces underlying molecular similarity analysis," Journal of Mathematical Chemistry, 3, pp. 117-145.

Johnson, M. A. and Maggiora, G. M. 1990. Concepts and Applications of Molecular Similarity Wiley Inter-Science, New York.

Kahn, R., R. D. Haskins, J. Knighton, A Pursch, and S. Granger-Gallegos. 1991. "Validating a Large Geophysical Data Set: Experiences with Satellite-Derived Cloud Parameters. Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface. Interface Foundation of North America, Fairfax Station, VA. pp 133-140.

Littlefield, J. S. and D. B. Carr. 1986. "MINIGRAPH, a Device Independent Interactive Graphics Package." PNL-SA-14366. Pacific Northwest Laboratory, Richland WA.

Powsner, S. M. and Tufte, E. R. "Graphical Summary of Patient Status." THE LANCET, Vol. 344 No. 8919 pp 386-389.

Nicholson, W. L., D. B. Carr, P. J. Cowley, and M. A. Whiting. 1984. "The Role of Environments in Managing Data Analysis." American Statistical Association 1984 Proceedings of the Statistical Computing Section, pp. 80-84. American Statistical Association, Washington, DC.

Nicholson, W. L., D. B. Carr and D. L. Hall. 1980. "The Analysis of Large Data Sets." American Statistical Association 1980 Proc. of Statistical Computing Section, pp. 59-65.

Nicholson, W. L. and R. L. Hooper, R. L. 1977. "A Research Program in Exploratory Analysis of Large Data Sets." (Copies available from D. Carr)

Tufte, E. R. 1989. Visual Design of the User Interface, IBM Corporation, Armonk, NY.

Tufte, E. R. 1990. Envisioning Information, Graphic Press, Cheshire, CN.

NHEERL-COR-2042A

TECHNICAL REPORT DATA
(Please read instructions on the reverse before c

1. REPORT NO. EPA/600/A-96/103	2.
4. TITLE AND SUBTITLE Perspectives on the Analysis of Massive Data Sets	5. REPORT DATE
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Carr, D.B.	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Computational Statistics George Mason University Fairfax, VA 22030	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS US EPA ENVIRONMENTAL RESEARCH LABORATORY 200 SW 35th Street Corvallis, OR 97333	13. TYPE OF REPORT AND PERIOD COVERED Symposium paper
	14. SPONSORING AGENCY CODE EPA/600/02

15. SUPPLEMENTARY NOTES
1995. Proceedings of the 27th Symposium on the Interface. Computer Science and Statistics, Statistics & Manufacturing with Subthemes in Environmental Statistics, Graphics, and Imaging, Pittsburgh, Pa, June 21-24, 1995.
24, 1995.

16. ABSTRACT
This talk presents perspectives on the analysis of massive data sets. These derive from past efforts to develop tools for analysis of large data sets and speculations about the future. The perspectives touch on many topics: previous research, recurrent problems, standard tricks for addressing problems, new problems, and new opportunities. In looking forward, it seems that evolving notions about man-machine interface, models, and model criticism will play an important role in massive data set analysis. The extent to which the statistical community responds this evolution will have far reaching consequences in terms of the influence of our profession with the scientific community.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Analysis of massive data sets, statistical graphics, statistical computation		

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (*This Report*)	21. NO. OF PAGES 10
	20. SECURITY CLASS (*This page*)	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE