REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE dd-mm-yyyy	3. REPORT TYPE AND DATES COVERED
	08-05-2002	Final 01-05-1998 to 31-10-2001

4. TITLE AND SUBTITLE
Final Report on Massive Data Sets: Visualization and Analysis

5. FUNDING NUMBERS
DAAG55-98-1-0404

6. AUTHOR(S)
Edward J. Wegman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
George Mason University, MS 4A7
4400 University Drive
Fairfax, VA 22030-4444

8. PERFORMING ORGANIZATION REPORT NUMBER Final 2002-01

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
098056

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This project argued that the downsizing of U.S. Army implies a profound change in the way the Army carries out its war-fighting mission. There will be an increased reliance on technology. Because these are technology based warfare areas, there will be massive data sets generated electronically as part of the normal operations. The ability to analyze such data sets is crucial to the prosecution of a military engagement in 2000 and beyond since it is in the structure of these data sets that crucial information about the timing, size and nature of enemy attacks is likely to be found. We have developed several methods for carrying out the analysis of massive data sets, in particular we have carried out research to reduce the computational complexity of clustering algorithms, to reduce the complexity of visualization methods and to increase the ability of visualization methods to handle massive data sets, and research on quantization methods for massive data sets.

14. SUBJECT TERMS
visual data mining
quantization
clustering
nonparametric density estimation

15. NUMBER OF PAGES
12

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	**UNCLASSIFIED**	**UNCLASSIFIED**	**UL**

Final Report on Massive Data Sets: Visualization and Analysis

by

Edward J. Wegman

George Mason University, MS 4A7
Center for Computational Statistics
157 Science-Technology Building #2
Fairfax, VA 22030

Table of Contents

Massive Data Sets: Visualization and Analysis - Statement of Problem Studied

1. Introduction

The U.S. Army faces radical changes in its operations over the next decade. The continuing downsizing and the changing nature of ground warfare threats imply an increasing reliance on technology. Two principal examples of this are the emergence of an enlarged engagement theatre and the consequent emphasis on theatre level anti-air and antimissile defense (THAAD) and the increasing emphasis on information warfare. Both of these technology enhancements imply the need to deal with the rapid analyses of massive data sets and subsequent decision making based on these analyses.

The THAAD theatre-level defense system is based on the premise that enemy forces will have increased capabilities to wage air and missile warfare against ground forces as was evidenced by the capabilities of the Iraqis with their SCUD missiles in the Gulf War. Consequently, there is a need to provide theatre-level defense. The THAAD system is a radar-based, two-tier anti-air, antimissile defense system. THAAD itself provides a high-level, upper-tier defense which is intended to eliminate most of the incoming threats. Patriot is intended as a lower-level, lower-tier supplementary defense to eliminate remaining threats. Because of the high information content supplied by the detection and interception electronics, rapid processing is required. The incoming missile systems are likely to have penetration aids (PENAIDS) supplying false targets to the radars and other target detection systems such as IR. Incoming planes are likely to have their own electronic countermeasures in the form of standoff jammers. They are also likely to have anti-radiation missiles (ARM) which would ride the radar beams down to the friendly forces radar antennas. Incoming tactical ballistic missiles are likely to have missile-borne jammers (TBMBJ). In addition, there are electromagnetic environmental effects (E) such as lightning, EMP, and co-site interference. There is the threat of standoff tactical nuclear strikes which while not targeted in the theatre, could have significant EMP and initial nuclear radiation effects on both the defense system and soldier survivability. All of these factors imply an extremely intensive classification and discrimination load arising from sensor collected data.

While THAAD is oriented to a comparatively limited theatre of engagement, information warfare is a global, increasingly important threat. The modern Army runs on information as much as it runs on fuel, weapons, and soldiers. Attacks against information systems and C systems can destroy the Army's ability to project force and wage war as effectively as any weapons system. This is particularly the case since much of the Army's information technology relies on COTS hardware and software systems, perhaps somewhat modified for increased physical survivability. Moreover the majority of the Army's communications system traffic is carried over the commercial network infrastructure. These imply that Army information systems and computers are subject to the same viruses, worms, trojan horses, trap doors and other hacker threats that commercial machines are subject to. In a GAO report dated 8 May 1996, the results of a DISA vulnerability assessment are published. (See also “Information Systems Threat Assessment (U)”, DIA PC-1750-4-93, dtd November 93.) DISA carried out 16,840 attacks on DoD systems. Of that total, 14,819 attacks were successful, only 593 attacks were detected, and only 30 of those were reported. This is .178% reported attack rate. DISA estimates there are some 250,000 attacks annually on DoD computers. Again, it is clear that network traffic flow represents an enormous information database and that statistical clustering and discrimination techniques are key elements to detecting and tracking information and C attacks.

Motivated by these two examples, I am proposing to conduct research on certain aspects of the analysis of massive data sets, particularly focusing on certain algorithmic and visualization issues. The proposal is organized as follows. Section 2 contains a discussion of the computational and visualization impact of massive data sets Section 3 contains several closely related proposed topics for research: a) clustering complexity, b) visualization complexity, and 3) effects of quantization of data in massive data sets Section 4 details the results of previous support including publications and interactions with Army personnel and commands.

2. Impact of Massive Data Sets

Wegman (1995) has outlined a taxonomy of data set sizes and discussed the implication of large data sets in terms of computational complexity and visualization limits. He points out that complexity in both computation and visualization increase not linearly in the sample size, but more as proportional to the order of magnitude of the data set size.

Traditional statistical methods usually focus on data set sizes in the range characterized as tiny to small (up to 10,000 observations). Indeed, even modern exploratory data analysis techniques rarely consider data set sizes larger than those characterized as medium (1,000,000 observations). And yet, as we have seen in the case of THAAD and information warfare scenarios, it is likely that huge or massive data sets (10^{12} observations) would be involved. Indeed, much of the data would require real-time processing to be effective in these military scenarios. Wegman (1995) analyzes the complexity of a number of algorithms and concludes that algorithms of or even may not be computational feasible. Consider for example a n^2 complexity algorithm and a huge data set. The most ambitious supercomputer goals announced are the teraflop computers, which would take 3.17 years to compute this case. Clustering algorithms (cf. Everitt, 1993) are usually distance-based, hence for clustering points, they require distance computations. Thus it is computationally infeasible to use conventional clustering algorithms for huge data sets even with teraflop computers

From the visualization point of view, it is clear that in dealing with large to massive data sets, conventional methods become problematic. Consider for example a data set of size 10^{10} . The 1% outliers themselves amount to 10^8 observations. Suppose that we could invent an extremely efficient encoding of the data, say one pixel per data item. This is sometimes called a scatter plot. The question is with the resolution of the normal human eye, how many pixels could we see. Wegman (1995) suggests that even under the most wildly optimistic scenario we are unlikely to be able to visualize more than 10^7 observations.

From the discussion of Section 1, it is clear that the Army faces significant operational issues that depend on the ability to analyze and visualize massive data sets. From the discussion just given in Section 2, it is clear that real-time analysis and visualization of massive data sets is a nontrivial problem.

3. Research Tasks and Results

Clustering Complexity - Clustering is probably the single most important problem in discovering structure in data, i.e. in contemporary language, the most important data mining issue. Cluster analysis, while comparatively hard to define, refers to a process of dividing a data set into relatively homogeneous subsets where a priori the number and nature of the subsets is unknown. Classification, ordinarily viewed as an easier problem, refers to the association of data points with predefined groups or subsets of data. Often in clustering, there is a training data set in which the clusters or subsets are known by some external criterion. An adaptive procedure is

developed based on the training set which classifies new data into the clusters discovered in the training data. This is sometimes called supervised learning. Unsupervised learning is accomplished when there is no training data.

I proposed to examine density-based methods for clustering. In contrast with distance based methods, most conventional nonparametric density estimators have a complexity of O(n) although for multivariate data, the multiplier subsumed in may be large. Thus if a suitable density based clustering algorithm can be developed, the computational complexity is likely to be reduced from O(n^2) to O(n). A number of successful attacks on this problem has been made. Items 2, 3, 14, 17, 19, 23 and 53 in the publication list below refer to nonparametric density estimation and issues of computational complexity. In addition, a Ph.D. dissertation has been completed by my student, Amrut Champaneri entitled: *Multivariate Probability Density Estimation: Some Statistical Properties* and by my student, Sung Ahn entitled: *A Maximum Likelihood Method for Density Estimation*. In the former dissertation, Dr. Champaneri created a computational algorithm for tessellating multidimensional space with complexity O($n \log n$). This tessellation not only leads to a histogram like maximum likelihood estimator, but also leads to a natural clustering algorithm of complexity O($n \log n$). Dr. Champenari also showed consistency and asymptotic distribution results based on an empirical determination of the creation rate of tiles as a function of dimensions and the number of tessellating points. Dr. Ahn's work focused adaptive normal mixture models for nonparametric density estimation and created a Bayesian penalty function for creation of too many mixture terms. The adaptive normal mixture methodology identifies clusters by identifying mixture terms. The adaptive mixture methodology is a recursive technique that requires one pass through the data.

Visualization Complexity - The problem of visualizing large data sets is a vexing one. As indicated above, the standard high-resolution screen has about 10^6 pixels, so that at best we could hope to represent 10^6 observations. Even if more pixels were available, the ability of the eye to distinguish pixels is limited by the distance between foveal cones within the eye. Alternative strategies have to be discovered. I have advocated the use of immersive techniques (virtual reality) and three-dimensional techniques in the past. Much of our previous Army sponsored research has focused on these techniques. The reason for using these techniques is that the third dimension moves from a pixel to a voxel setting which potentially moves us from 10^6 pixels to 10^9 voxels. This gives us three orders of magnitude extra "screen real estate."

Our visualization work has focused on methods for expanding the scope of data that may be visualized. In part this is closely aligned with density methods discussed in the previous section because densities are a representation of data when the overplotting is too severe. Items 2, , 7, 11, 12, 13, 14, 15, 16, 18, 19, 21, 24, 25, 26, 27, 28, 34, 35, 39, 40, 41, 42, 44, 45, 46, 47, 50, 51, 52, 54, 55, 66, 69, 70, and 71 are all items related to data visualization and computer graphics. Perhaps the most important results here fall into two categories: 1) visual data mining and 2) low cost immersive environments. In the former category we have combined four basic techniques: a) parallel coordinates, b) grand tour, c) saturation brushing and d) stereoscopic displays to provide integrated techniques for large scale data analysis. Visual data mining has been demonstrated on several data sets of quite large magnitude e.g. 130,000 items in 8 dimensions and 58,000 items in 14 dimensions. Strategies we have devised include what we have called a BRUSH-TOUR strategy for high dimensional clustering and a TOUR-PRUNE strategy for constructing tree-based decision rules. In the second item, low cost immersive environments, we have constructed what we have called the MiniCAVE, essentially a PC-based voice-controlled immersive environment for approximate $20,000. We have recently been notified that a US Patent is being issued on this system. Also of interest, a Ph.D. dissertation has been completed by my student,

Rida E. A. Moustafa entitled: *Fast Conceptual Clustering Algorithm for Data Mining and Visualization.*

Quantization - Quantization, roundoff, and binning are aspects of a similar process arising in different disciplines. Quantization is the language normally used electrical engineering/signal processing referring to the discretification of signals. This has been a very successful strategy for dealing with the digitization of signals, for example with digital audio. Roundoff is of concern in the numerical analysis community and refers to the truncation of real numbers for purposes of numerical computing in a fixed word length computer. Roundoff analysis again is a success story in the numerical analysis community. Binning is used sometimes in the statistics community and refers to grouping data in representative groups. Often coupled with comparatively elementary notions like histograms, binning, in contrast with the other related concepts, does not seem to enjoy a great reputation within the statistics literature. One perhaps significant difference between binning and grouping in the statistics community and quantization and roundoff in the other communities is that conventionally binning and grouping have been thought to be comparatively coarse approximations whereas quantization and roundoff ideas are connected fine approximations. For example, quantization of audio is usually done at 16 bit level while quantization in images is usually done at the eight to twenty-four bit level, i.e. 266 to 64 million colors. In contrast, in binning for histograms we often think in terms of 10 to 20 class intervals. I would conjecture few statisticians have ever thought of constructing a histogram with 64 million bins. Yet this would not be entirely unreasonable in dealing with a terabyte of data.

There are distinct theoretical and computational advantages to binning. Binning or quantization seems particularly appropriate to setting in which there are huge to massive data sets because with a data set of this size, the bins can be sufficiently small that they are smaller than the limits of perception. Moreover, once quantized the storage requirements for the data are likely to be considerably smaller since the only information required is the bin and the count of items in that bin. My Ph.D. student, Martin Khumbah completed his dissertation entitled: *Mathematical Quantization for Massive Datasets* on this topic. Among the interesting things we have shown is that geometric quantization can be accomplished effectively up to about 5 or 6 dimensions and in this range there is almost no theoretical loss associated with quantization. If the representors of the quantized data are chosen appropriately, the quantized data is self-consistent and bias remains unchanged while variance is reduced. We developed results on computational complexity - O(n), storage complexity - 3k, where k is the number of quantized regions, and strategies for minimizing distortion. Items 49, 56, 62 and 63 refer to this research. Item 62 presented at the Interface meeting was selected for the session “Best of the Army Research Office.”

Summary of the Most Important Results

Density Estimation and Clustering

- Developed a computational algorithm for tessellating multidimensional space with complexity O($n \log n$).
- This tessellation not only leads to a histogram like maximum likelihood estimator, but also leads to a natural clustering algorithm of complexity O($n \log n$).
- Showed consistency and asymptotic distribution results based on an empirical determination of the creation rate of tiles as a function of dimensions and the number of tessellating points.
- Constructed an algorithm for adaptive normal mixture models for nonparametric density estimation and created a Bayesian penalty function for creation of too many mixture terms.

- Identified clusters by identifying mixture terms.
- The adaptive mixture methodology is a recursive technique that requires one pass through the data.

Visualization

- Created visual data mining strategies combining four basic techniques: a) parallel coordinates, b) grand tour, c) saturation brushing and d) stereoscopic displays to provide integrated techniques for large scale data analysis.
- Visual data mining has been demonstrated on several data sets of quite large magnitude e.g. 130,000 items in 8 dimensions and 58,000 items in 14 dimensions.
- Devised strategies we have called a BRUSH-TOUR strategy for high dimensional clustering and a TOUR-PRUNE strategy for constructing tree-based decision rules.
- Created a low cost immersive environment that we have called the MiniCAVE, essentially a PC-based voice-controlled immersive environment for approximate $20,000.
- MiniCAVE is voice controlled and feature stereoscopic capability.
- We have recently been notified that a US Patent is being issued on our MiniCAVE system.

Quantization

- Demonstrated a fast algorithm for quantization of massive datasets.
- Demonstrated that geometric quantization can be accomplished effectively up to about 5 or 6 dimensions and in this range there is almost no theoretical loss associated with quantization.
- Showed that if the representors of the quantized data are chosen appropriately, the quantized data is self-consistent and bias remains unchanged while variance is reduced.
- Developed results on computational complexity - O(n), storage complexity - 3k, where k is the number of quantized regions, and strategies for minimizing distortion.

Listing of All Publications and Technical Reports Supported Under this Contract

(a) Papers Published in Peer Review Journals

1. Wendy L. Poston, Edward J. Wegman, and Jeffrey L. Solka (1998) "A parallel algorithm for subset selection," *Journal of Statistical Computation and Simulation*, 60, 1-17.

2. Michael Minnotte, David Marchette, and Edward J. Wegman (1998) "The bumpy road to the mode forest," *Journal of Computational and Graphical Statistics*, 7(2), 239-251.

3. J. L. Solka, E. J. Wegman, C. E. Priebe, W. L. Poston and G. W. Rogers (1998) "A method to determine the structure of an unknown mixture using the Akaike information criterion and the bootstrap," *Statistics and Computing*, 8, 177-188.

4. Wendy L. Poston, Edward J. Wegman, and Jeffrey L. Solka (1998) "D-optimal design methods for robust estimation of multivariate location and scatter," *Journal of Statistical Planning and Inference*, 73, 205-214.

5. Wilhelm, A. F. X., Wegman, E. J., and Symanzik, J. (1999) "Visual clustering and classification: The Oronsay particle size data set revisited," *Computational Statistics*, 14(1), 109-146.

6. Wegman, E. J. (1999), "Visions: The evolution of statistics," *Research in Official Statistics*, 2(1), 7-19.

7. Chen, J. X., Fu, X, and Wegman, E. J., (1999), "Real-time simulation of dust behaviors generated by a fast traveling vehicle," *ACM Transactions on Modeling and Computer Simulation*, 9(2), 81-104.

8. Wegman, E. J. (2000) "On the eve of the 21st century: Statistical science at a crossroads," *Computational Statistics and Data Analysis*, 32, 239-243.

9. Wegman, E. J. (2000) "Visions: New techniques and technologies in statistics," *Computational Statistics*, 15, 133-144.

10. Martinez, W. and Wegman, E. (2000) "An alternative criterion useful for finding E-optimal designs," *Statistics and Probability Letters*, 47, 325-328.

11. Wegman, E. J. (2000) Book Review of *The Grammar of Graphics* by Leland Wilkinson, *Journal of the American Statistical Association*, 95(451), 1009-1010.

12. Wegman, E. J. (2000) "Affordable environments for 3D collaborative data visualization," *Computation in Science and Engineering*, 2(6), 68-72, 74.

13. Chen, Jim X., Wang, J. and Wegman, E. J. (2000) "Physical model of dust behaviors behind a moving object," *International Journal of Applied Science and Computations*, 7(2), 1-12.

14. Wegman, E. J. and Luo, Q. (2002) "On methods of computer graphics for visualizing densities," *Journal of Computational and Graphical Statistics*, 11(1), 137-162

15. Wegman, E. J. and Symanzik, J. (2002) "Immersive projection technology for visual data mining," *Journal of Computational and Graphical Statistics*, 11(1), 163- 188

(b) Papers Published in Non-Peer-Reviewed Journals or Conference Proceedings:

16. Edward J. Wegman, Qiang Luo, and Jim X. Chen (1998) "Immersive methods for exploratory analysis," *Computing Science and Statistics*, 29(1), 206-214.

17. David J. Marchette and Edward J. Wegman (1998) "Finding modes with the filtered kernel," *Computing Science and Statistics*, 29(1), 498-507.

18. Wendy L. Poston, Edward J. Wegman and O. Thomas Holland (1998) "Ultrasonic imaging of cast ductile iron projectiles," *Computing Science and Statistics*, 29(1), 292-298.

19. Michael C. Minnotte, David J. Marchette and Edward J. Wegman (1998) "New terrain in the mode forest," *Computing Science and Statistics*, 29(1), 473-477.

20. Bradley C. Wallet, Edward J. Wegman and David J. Marchette (1998) "Evolutionary subspace pursuit," Computing Science and Statistics, 29(1), 402-406.

21. Edward J. Wegman, Wendy L. Poston and Jeffrey L. Solka (1998) "Image grand tour," *Automatic Target Recognition VIII - Proceedings of SPIE*, 3371, 286-294.

22. Edward J. Wegman (1998) "Visions: New techniques and technologies in statistics (keynote talk)," *Proceedings NTTS '98: International Seminar on New Techniques and Technologies for Statistics*, 1, 23-34.

23. Sung Ahn and Edward J. Wegman (1998) "A penalty function method for simplifying adaptive mixtures density estimates," *Computing Science and Statistics*, 30, 134-143.

24. J. Solka, E. Wegman, L. Reid and W. L. Poston (1998)"Explorations of the space of orthogonal transformations from to using space-filling curves," *Computing Science and Statistics*, 30, 494-498.

25. J. X. Chen, J. Wang and E. Wegman (1998) "Animation of dust behaviors in a networked virtual environment," *Proceedings of the Sixth International Conference in Central Europe on Computer Graphics and Visualization*, University of West Bohemia, Plzen, Czech Republic, pp. 487-494.

26. J. Wang, J. X. Chen and E. Wegman (1998) "Physical model of dust behaviors behind a moving vehicle," *Proceedings of the International Conference on Scientific Computing and Mathematical Modeling, IMACS'98*, Alicante, Spain, June, 1998.

27. Wegman, E. J., Symanzik, J., Vandersluis, J. P., Luo, Q., Camelli, F., Dzubay, A., Fu, X., Khumbah, N.-A., Moustafa, R., Wall, R. and Zhu, Y. (1999), "The MiniCAVE - A Voice controlled IPT environment," *Proceedings of the Third International Immersive Projection Technology Workshop*, (H.-J. Bullinger and O. Reidel, eds.), Springer-Verlag: Berlin, 179-190.

28. Wegman, E. J. (1999), "Data mining and visualization: Some strategies," *Bulletin of the International Statistical Institute*, Tome LVIII, Book 3, 223-226.

29. Moustafa, R. E. A, and Wegman, E. J. (1999), "Using genetic algorithms (GAs) for the gene mapping problem," *Computing Science and Statistics*, 31, 487-492.

30. Wegman, E. J. and Solka, J. S. (1999), "Implications of distance learning methodologies for statistical education," *ASA Proceedings of the Sections on Statistical Education, Teaching Statistics in the Health Sciences, and Statistical Consulting*, 13-16.

31. Moustafa, R. E. A., Wegman, E. J., and DeJong, K. (1999), "Adaptive numerical approximation based on genetic algorithms," *Proceedings of the 1999 Genetic and Evolutionary Computing Conference (G ECCO).*

32. Moustafa, R. and Wegman, E. (2000) "Mining evolutionary models to multidimensional scaling of gene measurements," *Computing Science and Statistics*, 32, /HTMLProceedings/ RMoustafa/moustafa.pdf (CD-based publication).

33. Moustafa, R. and Wegman, E. (2000) "A GA-based method for function approximation using adaptive interpolation," *Proceedings of the 2000 Genetic and Evolutionary Computing Conference (GECCO).*

34. Wegman, E. (2000) "Authenticating vulnerability measurements," *Computing Science and Statistics*, 32, 284-293.

35. Wegman, E. J. and Symanzik, J. (2001) "Data visualization and exploration via virtual reality: An overview," *Bulletin of the International Statistical Institute*, LIX(2), 76-79.

36. Dorfman, A. H., Lent, J., Leaver, S. G. and Wegman, E. J. (2001) "On sample survey designs for consumer price indexes," *Bulletin of the International Statistical Institute*, LIX(2), 421-424.

(c) Papers presented at Meetings but not published in Conference Proceedings

37. Wegman, E. J. (1998) "Data mining: How hard is it to find the gems," IMS Meeting, Pittsburgh, PA, April, 1998

38. Aghevli, B., Wegman, E. J. (1998) "Virtual manipulatives on the internet: Mathematics education for all using dynamic visualization," National Council of Teachers of Mathematics, Washington, DC, April, 1998

39. Wegman, E. J. (1998) "Data graphics for tactical decision making," Army After Next Tactical Decision Aids Technology Symposium, El Paso, TX, April, 1998

40. Wegman, E. J. (1998) "Statistical graphics in C3: Tactical decision aids," ONR Workshop on Statistics and Probability in C3, Arlington, VA, June, 1998

41. Wegman, E. J. and Symanzik, J. (1998) "MiniCAVE," AT&T Workshop on Data Visualization in Statistics, Morris, NJ, July, 1998

42. Wegman, E. J. (1998) "Mining the sands of time," Joint Statistical Meetings/ASA Annual Conference, Dallas, TX, August, 1998

43. Wegman, E. J. (1998) "Geometric Methods in Statistics," Five lecture series at the University of Naples, Naples, Italy, November, 1998

44. Wegman, E. J. (1999), "Roundtable - New Graphics Environments," presented at the Joint Statistical Meetings, Baltimore, MD. August, 1999.

45. Wegman, E. J. (1999), "Visual data mining," keynote talk presented at the Army Conference on Applied Statistics, West Point, NY, October, 1999.

46. Wegman, E. J. (1999), "Assessing vulnerability measurements," presented at the INFORMS meeting, Philadelphia, PA, November, 1999.

47. Wegman, E. (2000) "Visual data mining," Conference in Honor of the 80th Birthday of Professor C. R. Rao, Austin, TX, March, 2000.

48. Wegman, E. (2000) "Statistical data mining," Two-day Short Course Organized by the Washington Statistical Society, Washington, DC, April, 2000.

49. Wegman, E. (2000) "Data reduction by quantization," 5th World Congress of the Bernoulli Society and the Institute of Mathematical Statistics, Guanajuato, Mexico, May, 2000.

50. Wegman, E. and Luo, Q. (2000) "The MiniCAVE and Crystal Vision DataMining Software," Invited Technical Demonstration, Joint Statistical Meetings, Indianapolis, IN, August, 2000.

51. Wegman, E. (2000) "CrystalVision: A new visual data mining software," Joint Statistical Meetings, Indianapolis, IN, August, 2000.

52. Luo, Q. and Wegman, E. (2000) "Visual data mining," Joint Statistical Meetings, Indianapolis, IN, August, 2000.

53. Wegman, E. (2000) "Multivariate density estimation: adaptive mixtures" and "Multivariate density estimation: geometric approaches," Workshop on Nonparametric Model Building, Splines and other Smoothing Techniques, State College, PA, October, 2000.

54. Wegman, E. (2000) "Visual data mining," Graduiertenkolleg "Angewandte Statistik," Herbstkolloquium, Dortmund, Germany, November, 2000.

55. Wegman, E. (2000) "Crystal Vision: A new visual data mining software," Conference on Data Mining and Statistics, Augsburg, Germany, November, 2000.

56. Wegman, E. (2000) "Data Reduction by Quantization," Nonparametrics in Large, Multidimensional Data Mining Conference, Dallas, TX, January, 2001

57. Wegman, E. (2001) "Visual Data Mining," 8th Biennial CDC/ATSDR Statistics Symposium, Atlanta, GA, January, 2001

58. Wegman, E. (2001) Short Course on Statistical Data Mining, ENAR Meeting, Charlotte, NC, March, 2001

59. Wegman, E. (2001) Five Lectures on Geometry, Visualization and Data Mining, University of Aalborg, Denmark, May, 2001

60. Wegman, E. J. (2001) "Visual Data Mining," Keynote Talk, Danish Society of Theoretical Statistics, Aalborg, Denmark, May, 2001

61. Wegman, E. J. (2001) "Visualizing Cereal World," DataViz II Workshop, Fairfax, VA, May, 2001

62. Wegman, E. J. (2001) Short Course on Statistical Data Mining, Interface '01, Orange County, CA, June 2001

62. Wegman, E. J. (2001) "Data Reduction by Quantization," Interface '01, Orange County, CA, June, 2001

63. Wegman, E. J. (2001) "Data Reduction by Quantization," Joint Statistical Meetings, Atlanta, GA, August, 2001

64. Wegman, E. J. (2001) "Pixel Tours," IMA Workshop on Geophysics and Statistics, Minneapolis, MN, November, 2001

65. Wegman, E. J. (2001) "Pixel Tours," American Geophysical Union Meeting, San Francisco, CA, December, 2001

(d) Special Issues or Books

66. Edward J. Wegman (1998) "Parallel coordinate and parallel coordinate density plots," *Encyclopedia of Statistical Sciences*, Update Volume 2, (Kotz, S., Read, C. B., and Banks, D. L., eds.), 518-525+color plates.

67. Wegman, E. (ed.) (2000) *On the Eve of the 21st Century*, Special Issue of *Computational Statistics and Data Analysis*, guest edited, Vol. 32, Nos. 3 and 4.

68. Wegman, E. and Martinez, Y. (2000) *Computing Science and Statistics: Proceedings of the 32nd Symposium on the Interface* (issued as a CD), Fairfax Station, VA: Interface Foundation of North America, Inc.

(e) Manuscripts Accepted but not yet Published

69. Wegman, E. and Solka, J. (2002) "On the mathematics of visualizing high dimensional data," to appear *Sankhya*

70. Wegman, E. J. (2002) "Visual data mining," to appear *Statistics in Medicine*

71. Wegman, E. J. and Dorfman, A. H. (2002) "Visualizing cereal world," to appear *Computational Statistics and Data Analysis*

List of all participating scientific personnel

1. Edward J. Wegman
2. Martin Khumbah, earned Ph.D.
3. Rida Moustafa, earned Ph.D.
4. Xin Wang, earned M.S. in Statistical Science

Report on Inventions

1. A Voice-Controlled Immersive Virtual Reality System, Provisional Patent Disclosure #37067, Patent award notification April, 2002.