

# The Virginia Center of Excellence in Data Mining:

## A Proposal to establish a CIT Technology Innovation Center

**1. Introduction** Society is in the midst of a revolution as profound and far reaching in its consequences as the industrial revolution of the 19<sup>th</sup> century. The current revolution, an information revolution, has changed and is changing the nature of commerce and the way of life on a global scale. Manufacturing, although it will never disappear, is taking its place along side of farming, an industry once practiced by the majority of the population, but now practiced by a comparatively small percentage of the population. The information age is upon us and its consequences are profound. The creation and transmission of information and knowledge will increasingly be the occupation of the majority of workers, and will provide the largest number of high quality and high paying jobs. The regions of the country that seize the opportunities to act upon this fundamental societal change will be the regions that prosper.

It is the nature of the information revolution to level the playing field since knowledge workers require few natural resources and may in principle be located anywhere in the world. Presently Virginia has a natural advantage because of the proximity of the federal government and the proximity of the high tech industry and laboratories that support that government. Virginia has already taken the first steps to establish itself as a regional center for information technology. By establishing itself as an information technology intensive region early in the information revolution, Virginia will seize the opportunity to become a prosperous national center for the information age.

Computing technology has been accelerating steadily since the early 1980s when the first microprocessor-based personal computers and workstations were introduced to the scientists and the general public and became more than hobbyist/specialist tools. Network technology, likewise, has been particularly accelerating since 1993 when world wide web became widely available to scientists and the general public through web browsers. This convergence of computer resources connected via a global network has created an information tool of unprecedented power, a tool in its infancy. The global network is awash with data, uncoordinated, unexplored, but potentially containing information and knowledge of immense economic and technical significance. It is the role of data mining technologies arising from many discipline areas to convert that data into information and knowledge. It is our premise therefore that data mining technologies lie at the heart of the information revolution providing the mechanism for creating information and knowledge. These we contend are the commodities that increasingly are driving regional, national and global economies. Thus effective development of data mining strategies is central to the economic well being of the region.

**2. Data Mining** What precisely do we mean by data mining and what are the issues surrounding the effective development of data mining strategies? We distinguish four concepts: data, information, knowledge and wisdom. Data is raw, unprocessed information often in numerical form, but also possibly text, symbolic or images. It is usually untouched by the human mind or at least unprocessed by human thought. Information in contrast results from processed data in which structure and form are found. Processing can be visual or analytical and may be carried out with little or no human intervention. The structure may be found in terms of association, correlation, clustering, recognition of spatial and time patterns, exceptional or unusual instances, and the like. Knowledge results from synthesis of information into insight and into verifiable and extensible theory. Wisdom is the state of a human mind that is able to synthesize knowledge and experience into a cohesive and satisfying plan for living. The last is obviously not achieved by everyone and is certainly not the object of data mining. Data mining generically refers to the process of extracting information from data with the objective of creating knowledge.

For millennia, humans have been converting data into information. Geometry, statistics, astronomy, physics, demography are clearly disciplines that have had this as a fundamental goal. However, until recent times the scale of data has been comparatively small so that discovery processes within these disciplines have been of a scope that was comparatively easily accomplished with pencil, paper and comparatively simple computation engines. In most cases, experiments were designed in such a way that comparatively small, relevant data sets were obtained. However, electronic instrumentation and computer networks now offer the possibility of collecting large, opportunistic data sets that are information rich but which may not have been collected to answer the specific question at hand. Thus data mining is characterized generally by the exploration and exploitation of large collections of opportunistically collected data whose internal structure is unknown and unmodeled a priori. Data set size and complexity are usually key parameters in data mining. Data quality and algorithmic complexity are concomitants that impact upon the success of data mining efforts.

Descriptor	Data Set Size in Bytes	Storage Mode
Tiny	$10^2$	Piece of Paper
Small	$10^4$	A Few Pieces of Paper
Medium	$10^6$	A Floppy Disk
Large	$10^8$	Hard Disk
Huge	$10^{10}$	Multiple Hard Disks e.g. RAID Storage
Massive	$10^{12}$	Robotic Magnetic Tape Storage Silos

Table 1. Taxonomy of Data Set Sizes

Conventional methods are usually quite successful with tiny or small data sets. Analysis in these regimes usually focuses on statistical optimality, extracting maximal information from each observation. Medium to large data sets begin to severely tax traditional methodologies. Wegman (1995) shows that traditional visual exploration devices are stretched to their limit and responsiveness of present day computers is sufficiently slow that a high degree of human-computer interaction is not possible. Timely transfer of data over the network becomes impractical. Huge to massive data sets challenge all traditional techniques. Visual exploration is all but impossible. Complexity of some algorithms, for example traditional clustering algorithms, even with teraflop computers result in computations that could last for centuries. Transfer of data over the network is unfeasible so that any analysis must be done on the computer where the data is resident.

A natural question for the skeptic, then, do data sets on the larger end of the taxonomy really exist. The answer is an unequivocal yes. AT&T for example has a three terabyte ( $10^{12}$  bytes) database of data related to long distance phone calls made in the U.S.A. Sam's Warehouse, a discount retailer, regularly collects data on every retail financial transaction made daily, which results in a terabyte of data every several weeks. In Minamisaku, Japan, an array of eighty-four antennas monitors the surface of the sun for evidence of solar flares resulting in the need to process data at the rate of slightly over .5 teraflop (557 billion floating point operations per second). Data for the antenna array accumulates so quickly that it is impossible to store the data real time. NASA's EOSDIS (Earth Observing System Data Information System) project is expected to accumulate data at the rate of nearly .75 terabytes per day. Medical imagery from mammographies, MRIs and similar devices are often each a megabyte and accumulate at the rate of 100,000 per day. One year's worth easily becomes a terabyte database. High resolution military surveillance imagery also occupy 1 to ten megabytes each and are accumulated at the rate of 10,000 or more per day. Literally billions of bytes of data are passed over the world wide web network daily. Security considerations, both commercial and military, require that databases of network transactions be

explored for potential and actual security breaches. Thus it is clear that effective data mining methodologies are crucial for many sectors of society including commercial, financial, scientific, health-related, and military to name a few. It is also clear that present methodologies only begin to scratch the surface of needed developments. Within GMU, there are major resources, intellectual talents, and exciting developments within a variety of discipline areas focused on these data mining issues. Virginia is in an ideal position to exploit existing methods, to develop new methods and to become a national leader in data mining.

**4. Representative Task Areas** The foregoing discussion is intended to make the case that data mining is an arena with major economic and scientific impact in which the Commonwealth of Virginia is positioned to become a national and global leader. The potential impact is both short term, because the proposed center is in effect a metacenter bringing together groups already doing work in the area which are able to rapidly respond both individually and jointly to short-term needs, and also long term, because the research agenda is clearly driven by deep scientific issues which as they are addressed will position the economy of Virginia to become a national and global leader in the area. The following discussions are intended to indicate the scope of research undertaken by the proposed technology innovation center.

**Data Quality** Data quality issues are critical for data-intensive applications. In data mining and knowledge discovery, the quality of the underlying data sets impacts the reliability of the discovered conclusions. Most data mining techniques already quantify the significance (strength) of their conclusions. Such quantification must also reflect the a priori quality of the data sources used. In database applications, the quality of the answers issued by a database system is the primary concern of the database clients. In decision-support systems the quality of the supporting data is a major component in the decision-making process. Most database systems assume no responsibility for the quality of their contents, leaving the interpretation of their output to their clients.

In applications that integrate information from multiple, overlapping information sources, data quality estimates can help resolve cross-source inconsistencies. As a simple example, an integrator can resolve an inconsistency regarding the price of an item, on the basis of timestamps (an example of a quality parameter). More generally, information is a commodity, and the quality of an information product is the main parameter determining its utility and hence its price. For example, the value of a mailing list that targets potential customers of a specific item, must be proportional to its soundness (level of accuracy) and completeness (thoroughness of coverage).

Some of the research and development challenges in the area of data quality include: 1) establishment of a standard for the specification of data quality, so that information products can be rated with respect to their quality (not unlike hardware specification sheets). The information quality standard should rate not only the overall data quality, but also consider situations in which data quality is not homogeneous; i.e., some parts of the data set are better than others. 2) development of efficient and reliable methods for estimating data quality. This includes "laboratory" methods in quality estimation requires manual verification of samples from the data set, as well as "automatic" methods that can roughly assess the quality of data sets with only limited human guidance. For data sets that are dynamic, develop methods for efficient periodical reassessment of quality. 3) development of algorithms that estimate the quality of individual subsets of the data sets (e.g., answers to specific queries) from the overall data quality specifications. As mentioned above, the non-homogeneity of quality implies that the quality of answers varies from one query to another. 4) application of data quality information to the process of data cleansing; i.e., the removal of erroneous or noisy data, and the inference (imputation) of missing data. In particular, data quality information can be used to harmonize inconsistent data obtained from multiple, overlapping information sources.

**Pattern Recognition** No single model exists for all pattern recognition problems and no single technique is applicable for all problems. Rather what we have in pattern recognition is a bag of tools and a bag of problems. It is time to stop arguing over which type of pattern classification technique is best because that depends on our context and goal. Instead we should work at a higher level of organization and discover how to build managerial systems to exploit the different virtues and evade the different limitations of each of these ways of comparing things. The ability to predict and improve the performance on test data is crucial for data mining applications. As test data is not necessarily the same as training data, we have developed a Predictive Learning (PL) methodology using hybrid classifiers and active learning methods. Active learning, known also as corrective training, is related to recent training methods including bootstrap, bagging and arcing (Breiman, 1996). We have used hybrid classifiers and active learning techniques on surveillance and CBIR (contents based image recognition) {indexing and retrieval} tasks related to face and hand gesture recognition. We have successfully tested our methods on large image data bases (Takacs and Wechsler, 1998; Gutta and Wechsler, 1997).

*Data reduction and projection* The goal is to define a lower dimensional space with improved discrimination ability by combining PCA (principal component analysis) and LDA (Linear Discriminant Analysis). The method consists of two steps: first we project the data using PCA for dimensionality and data reduction, and second we use LDA to obtain a linear classifier. The basic concept here of combining PCA and LDA is to improve the generalization ability of LDA when the number of data is small. We have successfully combined PCA and LDA for face image data bases under our sponsored R&D FERET project.

*Predictive Learning* The goal is to address issues related to the bias-variance tradeoffs and improved performance on test data, lower the predictive risk error using (i) active learning, corrective training and resampling methods, and hybrid classifiers, i.e. mixture of experts and gating networks, (ii) complexity control via Support Vector Machines (SVM), and (iii) evolutionary computation as an optimization technique. Within the framework of Predictive Learning (PL), estimating a model from finite data requires then specification of three concepts:

PL(1) - a set of approximating functions (i.e., a class of models : DICTIONARY) - as an example we have used Radial Basis Functions (RBFs) and Decision Trees (DTs);

PL(2) - an inductive principle and an optimization (parameter estimation) procedure. The notion of inductive principle is fundamental to all learning methods. Essentially, an inductive principle provides a general prescription for what to do with the training data in order to obtain (learn) the model. There is just a handful of known inductive principles (Empirical Risk Minimization, Regularization, Structural Risk Minimization, Bayesian Inference, Minimum Description Length), but there are infinitely many learning methods based on these principles. We have used ERM, regularization and Bayesian Inference as inductive principles. We successfully used yet another inductive principle, based on high class separability and the concept of support vectors (SV), for discrimination of face poses as encountered during the task of face recognition.

PL(3) - a learning method as a constructive implementation of an inductive principle (i.e., an optimization or parameter estimation procedure) for a given set of approximating functions in which the model is sought (such as feed forward nets with sigmoid units, radial basis function networks etc.). In addition to standard optimization techniques we have used also stochastic optimization via Genetic Algorithms (GAs).

**Knowledge Discovery with Bayesian Networks** Bayesian Networks are graphical models that encode the joint and local distributions of a data set. The nodes of the network define the variables in the data set and

the directed-arcs represent the dependency among the variables. Each node has an associated conditional probability table (CPT). The structure of a Bayesian Network not only defines explicit dependencies among variables but also conditional independence. Unlike other models that can only classify or predict, a single Bayesian Network allows us to classify any variable based on observations (not necessarily all variables) or make prediction by querying variables.

The important implication for data mining is that there are algorithms that allow us to learn the Bayesian Network from a data set. From a single Bayesian Network, businesses or government agencies can visualize the relationships among the variables in their databases. They can ask what-if questions by setting variables in the network to certain values and receive a predicted outcome. They can make classifications based on observations. They can use the Bayesian network to fill in values of missing data. A Bayesian network associated with a database can provide highly sophisticated integrity checks. The Bayesian network can flag values for a variable that are unusual given the values of other variables for the record (i.e., a record with salary of \$150,000 for a custodian would be flagged, while the same salary for a physician would not cause concern).

The learning algorithm finds the relationships and builds the networks. The resulting network becomes a decision analysis tool. Even though algorithms for inducing Bayesian Networks exist there is plenty of room for improvement. Much research is still needed in search algorithms, scoring metrics, dealing with missing data, dealing with hidden variables, etc.. This research will focus on advancing the state of the art in Bayesian Network learning.

*Adaptive Feature Selection* In any data mining or pattern classification problem we are faced with the daunting task of deciding which features are important. There are several statistical techniques such as principal component analysis that aid in our decision. However, the number of features we consider in a very large database may be well over 100 or 200 variables. In addition, we may want to combine the features to get better predictive or classification results. We are now not only faced with the problem of finding "good" features, but are dealing with a combinatorial problem space. This research proposes to adaptively discover good features and combinations of features, to include discovering new operators for combining features. We will explore the use of different stochastic algorithms such as Evolutionary algorithms and Markov Chain Monte Carlo approaches to discovering "good" features and feature operators.

*Data Mining Small Datasets* Much research has gone into data mining applications where large amounts of data exist. With the combinatorial explosion of data storage, this is an important problem. Yet this problem seems to be aided somewhat by the law of large numbers. What happens when we are trying to discover knowledge from small datasets? This is not an uncommon problem for many enterprises. We may only have collected a few samples because that's all that's available. But yet we are forced to make a decision that could impact the entire organization. This research will investigate techniques for knowledge discovery from small samples and how "good" are the models we discover.

*Discovering Hidden Variables in Data* Databases contain only the data we collect and store in them. In many cases when trying to discover knowledge from data we are unaware of variables (not in the database) that clearly effect the values of observed variables (in the database). These variables, that are not observed or included in the database, are called hidden variables. It's important when trying to discover knowledge from a given database to find these hidden variables and understand their influence on other variables in the database. By discovering the hidden variables, we are better able to understand our model plus we may decide to explicitly represent these variables, if possible, in the database.

**Visual Designs For Human Comprehension of Multivariate Data** While the eye-brain system is one of

our most powerful tools of thought, studies of human perception and cognition and of human decision making under uncertainty are cause for humility. The task of designing graphics that communicate multivariate patterns to ourselves and others remains a challenge.

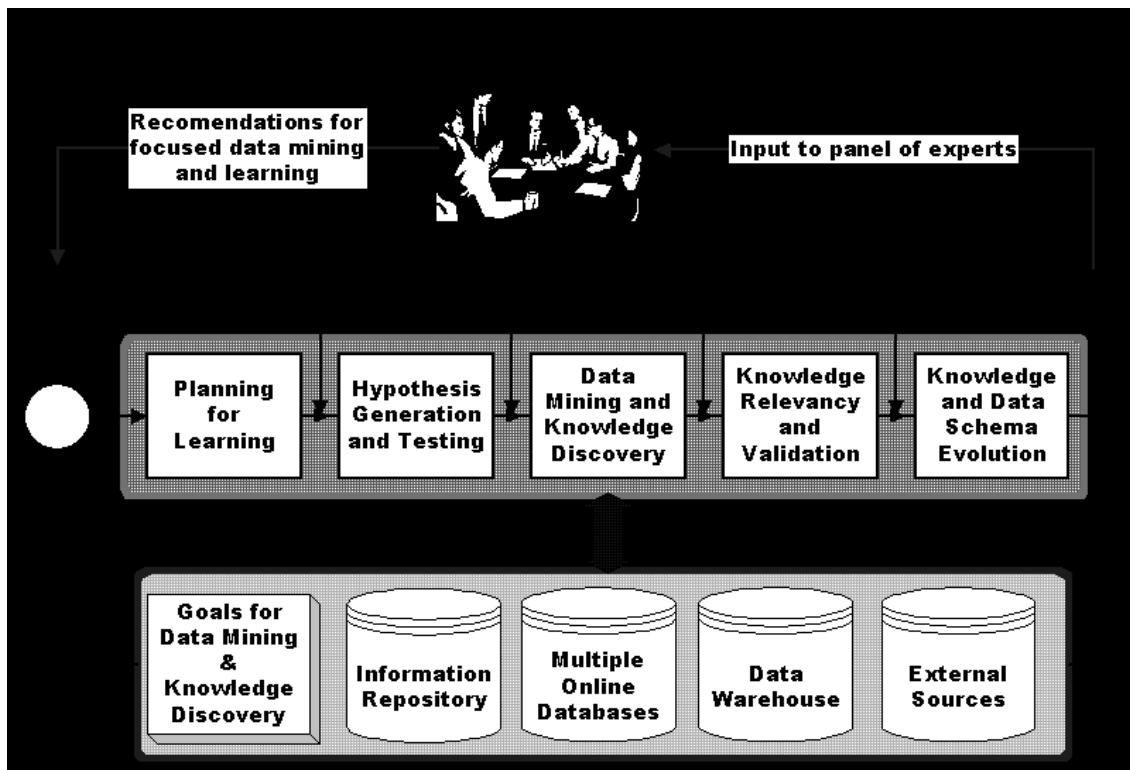
*Linked Micromap Plots for Spatial Data* Massive amounts of data (satellite imagery, environmental monitoring, human demographics, medical images, etc.) carry spatial indices. Typically statistical summaries reduce the amount of data. The challenge is to link quality graphics for statistical summaries to spatial coordinates. Since this work provides a serious alternative to choropleth maps, the potential market is huge. Applications include visualization of ecoregions for the continental U.S., digital elevation, and multiple AVHRR image-based assessment of vegetation class (8 million pixels, 159 vegetation classes).

*Visualization Methodology For Views in Dimensions 3-6* In 1995 we developed a multidimensional view using a scatterplot matrix with 2-D binned views for a 54 million pixel image. The plot represent over a half billion point pairs. With some density exploration tools (Carr et al., 1987), one can rapidly look for structure. Seeing higher dimensional structure via brushing is awkward with 54 million points per panel. Another type of work in the this general area relates to the study of gene expression data. The data for each gene is one or more short times series. We are gearing up for the jump from 100 genes to 50,000 genes. I have made progress in developing visualization methods to assess unsupervised clustering results. This includes published residual plots and my unpublished heuristic, close-neighbor-based, congestion-controlled multidimensional scaling.

*ISEA Global Grids* We have developed an equal area approach to global gridding used in the cartography community. The approach could potentially be adopted by NASA for level 3 satellite image products. The current NASA approach uses equal angle grids and has all the problems associated with the Mercator projection. Those working with polar models have developed discipline specific approaches. The equal area approach leads to hexagon cells. Hexagon cells have some distinct advantageous in terms of the Navier-Stokes equations and CFD modeling.

**Data Mining and Knowledge Discovery Life Cycle Model** Large-scale systems, such as the Human Genome Project, the Earth Observing System Data Information System (EOSDIS), and the Defense Information Systems Agency's efforts to integrate and evolve existing information systems, all require the integration and interchange of information from heterogeneous autonomous data systems. Moreover, there is a pressing need to train new information systems scientists and engineers to address multi- and interdisciplinary challenges in areas such as health care, advanced manufacturing technology, and the national information infrastructure, all of which require interdisciplinary approaches to the problem. This requires bringing together domain experts, engineers, and scientists to create new information system paradigms.

It is of utmost importance to convey to potential users of data mining tools and techniques that these instruments should be used within a framework and life-cycle model. It has been established that 80% of the cost associated with such projects involves the cleaning and quality assurance of data to be used for data mining. Moreover, one must plan for data mining. We propose to investigate the activities of the Data Mining and Knowledge Discovery Life Cycle Model, as depicted in Figure 1 below.



**Figure 1: Data Mining and Knowledge Discovery Life Cycle Model**

This process model would benefit a wide variety of companies in Virginia and could be used to implement a tool-suite to support one or more of the activities and data sources.

*Data Mining and Knowledge Discovery in Large Databases* We will investigate data mining tools and techniques within the context of very large databases. In particular we will extend and enhance the results reported in Yoon and Kerschberg (1998). We plan to develop a methodology and tools for **Query-Driven Data Mining** within the database management system. The approach consists of the user specifying an SQL query against a relational database. The query result and the relational tables accessed are used to create the set of positive examples, while the complement of that set is used to construct the set of negative examples. Association rules and mined and analyzed for their *relevance and interest* to the original query. The results are then presented to the user for feedback and tuning.

**An Automatic Improvement of the Representation Space for Data Mining** Current data mining methods assume that the attributes used in a database are relevant for the task of discovering important regularities and strategic patterns. This assumption may not hold in many applications, because data are often collected without knowing a priori what type of patterns or regularities may characterize the phenomena of interest. In such situations, unknown patterns may be complex and difficult to express directly in terms of the a priori known concepts and attributes that are already present in the data. When these attributes are insufficiently relevant to the given class of tasks, conventional methods, such as statistical multivariate analysis, classification tree builders, decision rule learners, and neural nets, do not perform well.

Research in the Machine Learning and Inference Laboratory aims at investigating several novel ideas for automatically improving the original representation space (i.e., space spanned over the given attributes), and for efficiently use them to implement systems for extracting patterns, exceptional events, and, generally, high level strategic regularities in multi-type data (symbolic and/or numeric). We plan for

develop flexible and portable tools applicable across platforms and scalable to very large databases and data warehouses. We will adapt advanced methods of machine learning for such problems, and integrate the methods with database and data warehouse technologies.

To this end, we will investigate a constructive induction approach, in which a data mining system performs iteratively two intertwined searches: one for the most adequate knowledge representation space (e.g., for the most relevant features), and the second for the "best" hypothesis in the found space. The first search will integrate several strategies, such as an automatic generation of new attributes, numeric or symbolic transformations of groups of attributes, conceptual clustering, attribute abstraction, as well as removal of irrelevant or weakly relevant attributes, as they may obscure the detection of strong patterns. The second search involves the application of an advanced inductive learning system, capable of generating high level data descriptions. While we will experimentally investigate the applicability of different learning systems to this problem, we will concentrate on the AQ-type progressive covering approach (aka "separate and conquer") due to its powerful knowledge representation and the ability to employ a wide range of inductive generalization operators. We have developed initial methods for data-driven, hypothesis-driven, and knowledge-driven constructive induction. These methods represent different strategies for transforming the original representation space into a new space that is tailored to the given data mining task.

The data-driven constructive induction method (DCI) proposes changes in the representation space based on the analysis of the data. The changes may involve attribute abstraction, attribute selection and attribute generation. The hypothesis-driven method (HCI) proposes changes in the representation space on the basis of analysis of the preliminary hypotheses generated in the original space. A knowledge-driven method (KCI) utilizes domain knowledge for improving the representation space.

The proposed research will investigate problems for integrating these methods for synergistically transforming the original representation space into a new space that is tailored to the given data mining task. The search for an improved space is guided by a pattern importance criterion (PIC) that initially represents a guess about what type of patterns or unusual events may be of interest to a data analyst (in the absence of such a guess, a predefined default PIC is used). The initial PIC is incrementally refined using a feedback from a search and show process. To test these ideas, we performed experiments on applying data-driven and hypothesis-driven attribute generation strategies to such problems as prediction of a country's energy consumption on the basis of past records, classification of texts in a large collection of documents, determination of demographic and economical patterns in the data in the World Factbook, and detection of patterns in a database of construction accidents. Initial results from these experiments were highly promising.

**Massive Data Set Methods** *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. Clustering is often accomplished by using distance measurements. As indicated above, for  $n$  observations this means  $n^2$  complexity.

We propose to examine density-based methods for clustering. In contrast with distance based methods,

most conventional nonparametric density estimators have a complexity of  $n$ . Thus if a suitable density based clustering algorithm can be developed, the computational complexity is likely to be reduced from  $n^2$  to  $n$ . A first approach to a density-based clustering algorithm might be to estimate the probability density declare a cluster to be any simply connected support region corresponding to bumps in the density, i.e. for which where is some predetermined threshold level. The idea here would be that if there are distinct clusters, the number of observations falling in the interstitial areas would be small and hence the corresponding probability density would have small magnitude in these regions.

*Visualization Complexity* The problem of visualizing large data sets is a vexing one. 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. We have advocated the use of immersive techniques (virtual reality) and three-dimensional techniques in the past. 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."

Visualization techniques probably cannot hope to examine truly massive data sets. However, progress can be made for larger data sets. One approach that makes sense is to examine replacing data plots with corresponding density plots, particularly in a immersive environment. However, these may not be fully satisfactory. First it is the case that the density plots are not particularly effective from a visualization point of view above three or possible four dimensions (using time, i.e. motion, as the surrogate for the fourth dimension). Dynamic adjustment of contour levels and nested contour levels help somewhat in visualizing the multidimensional shape of the data, however, as pointed out above even the 1% outliers of a huge data set are a large data set and hence even the outliers themselves become problematic to visualize. One mathematical approach we intend to explore is to investigate linked density plots over support subspace which are orthogonal.

**Naval Data Mining** The automatic identification of regions of interest in aerial imagery and video and the automatic detection of computer intruders are two recent research areas. We currently have implemented a system for the collection of network raw traffic. This data consists of a time stamp, source IP address, source port, destination IP, destination port, transfer protocol, and the number of bytes transferred per connection. This information is logged in a relational database and is subjected to operator analysis in a post mortem manner. We are currently logging several million records into the database daily and that number is expected to grow. In fact the number of records has more than doubled in a six month period.

A second application of interest focuses on the automatic segmentation of images into regions of homogeneous content, with an ultimate goal of classifying the different regions. Our research has focused on a wide variety of image modalities and desired classification systems. They include the detection of combat vehicles in natural terrain, the detection of man-made regions in aerial images, and the characterization of mammographic parenchymal patterns. Typical image sizes range from standard frame size of 640x480 up to sizes of 10K x 10K. This can lead to an enormity of data given that we may have tens to hundreds of computed features or spectral bands.

Our planned network security data mining efforts will focus on visualization of network activities and the identification of statistical measures that will aid us in intrusion detection. These measures include the detection of outliers or abnormal activity, the identification of spatially or temporally correlated attacks, and the detection of automated attacks based on time series analysis.

## **5. Benefit to Virginia Economy**

Data mining is often thought of as a business application and so-called data mining and data warehousing software and hardware is usually marketed as such by the larger computing companies. Most of this software and hardware arises out of commercial database vendors and contain fairly unsophisticated data exploration tools. The commercial market for the more sophisticated tools we anticipate developing in the Virginia Center of Excellence in Data Mining is enormous. Table 2 is an excerpt from a table supplied by the Virginia Employment Commission listing the 50 job sectors with the highest rate of growth in Virginia. We have highlighted in bold font those job categories which relate directly to data exploitation. Almost uniformly these are the highest quality jobs in the sense of being the highest paid jobs.

It should be noted that while many of these jobs do relate to the financial, management, and sales and marketing sectors of the Virginia economy and are thus the classical targets of the data mining technology, there is also a broad group of scientific and related technical professionals also included.

One source of massive databases traditionally exploited by current data mining technology arise from the accumulation of financial transaction point-of-sale data. Retail sales sectors such as discount stores, major retailers, airline and travel agencies, utility especially telecommunication companies, as well as credit card companies often have elaborate information networks and would like to exploit data mining technology to discover patterns and structure in their data. Representative of the commercial sector are companies like AMS and Synectics. We include expressions of support and partnering from these two companies indicating a strong desire to see a center such as the proposed Data Mining Center developed.

However, scientific, military and governmental sources of massive databases are emerging as likely targets for data mining technology. The proposed NASA EOSDIS system will generate nearly a terabyte of data each day. Even now, satellites generate sufficient data that much of the data is never seen by human eyes. The exploitation of data mining techniques of satellite remote sensing databases is likely to not only enhance the richness of the scientific process, but also provide new economic benefits. One could easily imagine discovery of commercial and residential development patterns, soil fertility patterns, pollution patterns based on satellite imagery which could have enormous economic impact. The proximity of Virginia to NASA Goddard, to NASA Langley and to the supporting sectors to the space agency clearly suggests a likelihood of positive economic benefit to the Virginia economy.

The Department of Defense collects an enormous amount of data from surveillance and intelligence sources. Again much of this data is not exploited because of its scale and also because of an inability to merge data from multiple sources. For example, the Navy's SOSUS array generates 500 megabytes of data every 10 minutes. A single high resolution image may be as much as 6000 by 6000 pixels or equivalently 100 megabytes of information. Problems associated with automatic target recognition (ATR) require location of objects in such an image which may only occupy 10 to 15 pixels. Military exploitation of data mining technology is again a major potential application. Virginia is host to a number of significant military operations including the U. S. Army Intelligence and Surveillance Command (INSCOM) housed at Ft. Belvoir, VA and the



VIRGINIA EMPLOYMENT COMMISSION

VIRGINIA EMPLOYMENT COMMISSION PROJECTIONS AND 1996 WAGE DATA

## FOR VIRGINIA OCCUPATIONS WITH THE GREATEST GROWTH

RANK	OES CODE	OCCUPATIONAL TITLE	EMPLOYMENT			WAGES	
			1994	2005	CHANGE	MEAN	
<b>1</b>	<b>25125</b>	<b>Systems Analysts/Computer Programm</b>	<b>41358</b>	<b>74435</b>	<b>33077</b>	<b>\$23.16</b>	<b>\$</b>
2	49023	Cashiers	90311	106868	16557	\$6.33	
3	49011	Salespersons, Retail	102959	118440	15481	\$7.94	
<b>4</b>	<b>19005</b>	<b>General Managers &amp; Top Execs</b>	<b>88292</b>	<b>103016</b>	<b>14724</b>	<b>\$25.33</b>	<b>\$</b>
5	67005	Janitors & Cleaners	54330	69009	14679	\$7.00	
<b>6</b>	<b>22127</b>	<b>Computer Engineers</b>	<b>9399</b>	<b>21696</b>	<b>12297</b>	<b>\$27.11</b>	<b>\$</b>
7	63047	Guards	21378	32208	10830	\$7.90	
8	65008	Waiters & Waitresses	38100	48241	10141	\$6.00	
9	32502	Registered Nurses	39162	49200	10038	\$18.00	\$
10	66008	Nursing Aides & Orderlies	23453	32918	9465	\$6.61	
11	31308	Teachers, Secondary School	30829	40020	9191	\$20.97	\$
12	55305	Receptionists & Information Clks	25767	34770	9003	\$8.36	
<b>13</b>	<b>39999</b>	<b>All Other Prof, Paraprof, Techns</b>	<b>18398</b>	<b>27223</b>	<b>8825</b>	<b>\$16.28</b>	<b>\$</b>
<b>14</b>	<b>22199</b>	<b>All Other Engineers</b>	<b>12165</b>	<b>20327</b>	<b>8162</b>	<b>\$25.20</b>	<b>\$</b>
15	66011	Home Health Aides	9331	17423	8092	\$8.29	
<b>16</b>	<b>41002</b>	<b>Marketing &amp; Sales, Supervisors</b>	<b>38488</b>	<b>45729</b>	<b>7241</b>	<b>\$14.94</b>	<b>\$</b>
17	51002	Clerical Supervisors	34088	41083	6995	\$14.72	\$
18	15026	Food Service & Lodging Mgrs	13710	20440	6730	\$12.71	\$
19	85132	Maintenance Repairers, Gen Util	32278	38617	6339	\$10.49	
20	55108	Secretaries, Ex Legal & Medical	71555	77691	6136	\$10.96	\$
21	32505	Licensed Practical Nurses	19295	25169	5874	\$11.64	\$
<b>22</b>	<b>13002</b>	<b>Financial Managers</b>	<b>19858</b>	<b>25212</b>	<b>5354</b>	<b>\$23.99</b>	<b>\$</b>
23	65038	Food Preparation Workers	28853	34197	5344	\$6.43	
24	31305	Teachers, Elementary	31031	36313	5282	\$19.78	\$
25	67002	Maids & Housekeeping Cleaners	23190	28344	5154	\$6.35	
26	31311	Teachers, Special Education	9390	14485	5095	\$21.35	\$
<b>27</b>	<b>13017</b>	<b>Engineering, Math, Nat Sci Mgrs</b>	<b>16518</b>	<b>21482</b>	<b>4964</b>	<b>\$29.75</b>	<b>\$</b>
28	63017	Correction Officers	9564	14499	4935	\$11.12	\$
29	98999	All Other Helpers, Laborers	43333	48124	4791	\$8.48	
30	53123	Adjustment Clerks	9136	13511	4375	\$11.55	
31	31302	Teachers, Preschool & Kindergtrtn	6123	10251	4128	N/A	
32	31521	Teacher Aides, Paraprofessional	11317	15416	4099	\$6.20	
<b>33</b>	<b>25199</b>	<b>All Other Computer Scientists</b>	<b>2940</b>	<b>6881</b>	<b>3941</b>	<b>\$21.72</b>	<b>\$</b>
34	32102	Physicians	10791	14723	3932	N/A	
<b>35</b>	<b>21114</b>	<b>Accountants &amp; Auditors</b>	<b>21340</b>	<b>25193</b>	<b>3853</b>	<b>\$18.15</b>	<b>\$</b>
36	97105	Truck Drivers, Light	21314	24901	3587	\$9.52	
<b>37</b>	<b>21905</b>	<b>Management Analysts</b>	<b>10499</b>	<b>14009</b>	<b>3510</b>	<b>\$20.73</b>	<b>\$</b>
38	53905	Teacher Aides & Education Assts	5643	9090	3447	N/A	
39	19999	All Other Managers & Adminstors	21586	25028	3442	\$23.95	\$
40	98902	Hand Packers & Packagers	20832	24246	3414	\$6.68	
<b>41</b>	<b>13011</b>	<b>Marketing Adver Public Rel Mgrs</b>	<b>10557</b>	<b>13971</b>	<b>3414</b>	<b>\$23.91</b>	<b>\$</b>
42	65041	Combined Food Prep & Serv Wrkrs	23271	26626	3355	\$5.92	
43	55347	General Office Clerks	74891	78162	3271	\$9.42	
<b>44</b>	<b>25302</b>	<b>Operations Research Analysts</b>	<b>4776</b>	<b>7960</b>	<b>3184</b>	<b>\$21.41</b>	<b>\$</b>
45	85302	Automotive Mechanics	16708	19853	3145	\$13.05	\$
46	65026	Cooks, Restaurant	16265	19401	3136	\$8.07	
47	61099	All Other Service Supervisors	17155	20244	3089	N/A	
<b>48</b>	<b>22126</b>	<b>Electrical &amp; Electronics Engrs</b>	<b>11132</b>	<b>13994</b>	<b>2862</b>	<b>\$27.58</b>	<b>\$</b>
49	31321	Instructors & Coaches, Sports	8059	10830	2771	\$8.62	
50	49017	Counter & Rental Clerks	8082	10807	2725	\$6.39	

Table 2. The 50 highest growth rate jobs in Virginia. Jobs potentially exploiting data and data mining technology are highlighted in a bold font.

Naval Surface Warfare Center located in Dahlgren, VA. We include letters of support from these two agencies as well as a Cooperative Research and Development Agreement (CRADA) in the final stages of

development with the Army’s White Sands Missile Range. Proximity to other major DoD operations such as the National Security Agency and the Navy Research Laboratory, operations that are likely to exploit data mining technology suggests additional strong economic benefit to Virginia.

Finally, we note that the federal and state governments are also generators of massive amounts of data. Agencies such as the U. S. Census Bureau, the Bureau of Labor Statistics, the Internal Revenue Service, the Federal Aviation Administration, the Food and Drug Administration, and the U. S. Department of Agriculture all in close proximity to Virginia all deal with massive data sets. It is likely that successful development of advanced data mining technologies will bring additional research and commercial revenues to the Commonwealth. MITRE is representative of Virginia based companies supporting with general government data interests. Similarly the World Bank is a quasi-governmental agency with strong data mining interests. We include letters from each of these two organizations.

## 6. Structure and Administration of the Center

The proposed Virginia Center of Excellence in Data Mining (VCEDM) is extraordinarily gifted with talented and enthusiastic participants. Short biosketches and lists of publications are included in Section 7 of this proposal. The collection of individuals represented there include virtually all of the senior figures on the George Mason campus involved with data exploitation and information technology. The biosketches and lists of publications speak for themselves.

We have suggested earlier that the proposed center is in some sense a metacenter. By this was meant that the individuals involved are key players or indeed the lead figures in other research units on campus. Thus they bring more than themselves to the table in the development of the Data Mining Center; they bring the resources of other units to the table which are already quite considerable. Table 3 summarizes the other units involved in the Virginia Center of Excellence in Data Mining by means of participation of those units leadership.

Because so many of the key participants of the proposed Virginia Center of Excellence in Data Mining are senior investigators in their own right, the proposed administration structure will be a director and an Administrative Council. The Director will be Professor Wegman and the Administrative Council will be the senior members listed in Table 3 above. In addition the Dean of the School of Information Technology and Engineering, Lloyd Griffiths, and the Director of the Institute for Computational Sciences and Informatics, Murray Black will be ex-officio members of the Administrative Council.

Campus Unit	VCEDM Participant	Role
Center for Computational Statistics	Edward J. Wegman	
Department of Applied and Engineering Statistics	Edward J. Wegman	
Center for Distributed and Parallel Computation	Harry Wechsler	
Center for Secure Information Systems	Sushil Jajodia	
Department of Systems Engineering	Kathryn Laskey	
Center for Information Systems Integration and Evolution	Larry Kerschberg	
Machine Intelligence Laboratory	Ryszard Michalski	
Center for Earth Observing and Space Research	Menas Kafatos	
Institute for Computational Science and Informatics	Menas Kafatos	
Center for Computational Statistics	Daniel B. Carr	

Table 3. Key Participants and their GMU Campus Units

In addition, because we wish to have strong corporate/government interactions, there will be an Extramural Advisory Panel. The Extramural Advisory Panel will be composed of senior management from

industry and government at the VP level or higher or federal senior executive service members. Initially, the panel will be drawn from those organizations supporting this proposal. However, as the Center develops, we anticipate expanding and restructuring Advisory Panel as appropriate.

## 7. Short Biosketches and Publications

**Edward J. Wegman** Professor Edward J. Wegman received his B.S. in mathematics degree from St. Louis University in 1965. He received the M.S. and Ph.D. degrees in mathematical statistics from the University of Iowa, the latter degree in 1968. Subsequently, he spent 10 years on the faculty of the world-class Department of Statistics at the University of North Carolina. Dr. Wegman's early career focused on the development of aspects of the theory of mathematical statistics. In 1978, Professor Wegman went to the Office of Naval Research (ONR) where he was the Head of the Mathematical Sciences Division. In this role, he had responsibility Navy-wide for basic research programs in applied mathematics, statistics and probability, systems theory, operations research, discrete mathematics, communication theory, and numerical analysis and computational architectures. In addition, he was responsible for a variety of cross-disciplinary areas including such projects as mathematical models of biological intelligence, mathematical methods for remote sensing, and topological methods in chemistry. As part of his duties at the Office of Naval Research, coined the phrase, computational statistics, and developed a high profile research area around this concept. The idea was to focus on techniques and methodologies which could not be achieved without the capabilities of modern computing resources. This program led to a revolution in contemporary statistical graphics. Dr. Wegman was the original program director of the basic research program in Ultra High Speed Computing at the Strategic Defense Initiative's Innovative Science and Technology Office (Star Wars Program). As the SDI program officer, Dr. Wegman was responsible for programs in software development tools, highly parallel architectures and optical computing.

Dr. Wegman came to George Mason University with an extensive background in both theoretical statistics and computing technology, with an extensive knowledge of the considerable data analytic problems associated with large scale scientific and technical databases and with a strong motivation to develop the computational and methodological

tools to address these problems. In 1986, he launched the Center for Computational Statistics and developed the M.S. in Statistical Science degree program. More recently he has been involved with the development of the Institute for Computational Science and Informatics and the new Ph.D. program in Computational Sciences and Informatics at George Mason University.

He has been consultant to a variety of governmental and private sector organizations including the states of North Carolina and Ohio, the U.S. Navy and the Executive Office of Management and Budget. He has organized some fifteen major workshops and conferences, including the 1988 Symposium on the Interface of Computing Science and Statistics. He has served as associate editor of *the Journal of the American Statistical Association*, *Statistics and Probability Letters* and *Communications in Statistics*. He presently serves on the editorial boards of *the Journal of Statistical Planning and Inference*, *Naval Research Logistics*, *the Journal of Nonparametric Statistics* and *Computational Statistics and Data Analysis*. Dr. Wegman completed a four-year term as the Theory and Methods editor of the prestigious *Journal of the American Statistical Association*. He is the founder of the Interface Foundation of North America, Inc. which is the host organization for the Symposia on the Interface of Computing Science and Statistics. The Interface Foundation in conjunction with the American Statistical Association and the Institute for Mathematical Statistics has also launched the interdisciplinary *Journal of Computational and Graphical Statistics*. Dr. Wegman served in national office in the Institute of Mathematical Statistics, the American Statistical Association and the American Association for the Advancement of Science. He has published some 110 papers and five books. His professional stature has been recognized by his election as Fellow of

the American Statistical Association, the American Association for the Advancement of Science, the Washington Academy of Science and the Institute of Mathematical Statistics. In addition he was elected as a Senior Member of IEEE. Dr. Wegman has been elected to membership in the International Statistical Institute. Dr. Wegman has also received numerous Navy awards including the Navy's Meritorious Civilian Service Medal. He is listed in Who's Who in America, Who's Who in the South and Southeast, Who's Who in American Education, Who's Who among Entrepreneurs, Who's Who in Leading American Executives, Who's Who in Frontier Science and Technology, Who's Who in Science and Engineering, and American Men and Women of Science. Dr. Wegman came to George Mason University in 1986 and is the Bernard J. Dunn Professor of Information Technology and Applied Statistics, the Chair of the Department of Applied and Engineering Statistics and the Director of the Center for Computational Statistics. A full resume is available at

<http://www.galaxy.gmu.edu/stats/faculty/wegman.resume2.html>.

### **Papers published by E. Wegman over last five years.**

"A spectral representation for the class band-limited functions," with H. T. Le, *Signal Processing*, 33(1), 35-44, 1993

"Visualizing multivariate data," with D. B. Carr and Q. Luo, in *Multivariate Analysis: Future Directions*, (Rao, C. R., ed.), Amsterdam: North Holland, 423-466, 1993

"Statistical graphics and visualization," with D. B. Carr, in *Handbook of Statistics 9: Computational Statistics*, (Rao, C. R., ed.), Amsterdam: North Holland, 857-958, 1993

"Three-dimensional Andrews plots and the grand tour," with J. Shen, *Computing Science and Statistics*, 25, 284-288, 1993

"History of the Interface since 1987: The corporate era," *Computing Science and Statistics*, 25, 27-32, 1993

"Simulating a multi-target acoustic array on the Intel Paragon," with C. A. Jones, *Proceedings of the Intel Supercomputer Users Group Conference*, 108-117, 1994

"Parallel simulation of an active vision model," with B. Takacs and H. Wechsler, *Proceedings of the Intel Supercomputer Users Group Conference*, 202-207, 1994

"Fast multidimensional density estimation based on random-width bins," with L. B. Hearne, *Computing Science and Statistics*, 26, 150-155, 1994

"Correlation estimators based on simple nonlinear transformations," with M. Sullivan, *IEEE Transactions on Signal Processing*, 43(6), 1438-1444, 1995

"Estimating spectral correlations with simple nonlinear transformations," with M. Sullivan, *IEEE Transactions on Signal Processing*, 43(6), 1525-1526, 1995

"A new visualization technique to study the time evolution of finite and adaptive mixture estimators," with J. L. Solka and W. L. Poston, *Journal of Computational and Graphical Statistics*, 4(3), 180-198, 1995

"Huge data sets and the frontiers of computational feasibility," *Journal of Computational and Graphical*

*Statistics*, 4(4), 281-295, 1995

"Immersive methods for mine warfare," with W. L. Poston and J. L. Solka, *MASEVR '95: Proceedings of the Second International Conference on the Military Applications of Synthetic Environments and Virtual Reality*, 203-218, 1996

"Wavelets and nonparametric function estimation," with W. L. Poston and J. L. Solka, in *Research Developments in Statistics and Probability*. (Brunner, E. and Denker, M., eds.), Utrecht: VSP, 257-274, 1996.

"Massive data sets in Navy problems," with J. L. Solka, W. L. Poston, D. J. Marchette, in *Massive Data Sets: Proceedings of a Workshop*, Washington, D. C.: National Academy Press, 157-167, 1996.

"Moments and wavelets in signal estimation," with H. T. Le, Wendy L. Poston and Jeffrey L. Solka, in *Statistics of Quality*, (Ghosh, S., Schucany, W. R. and Smith, W. B., eds.) New York: Marcel-Dekker, 253-274, 1997. Also a short unrefereed version in *Moments and Signal Processing*, (Purdue, P. and Solomon, H., eds.), Monterey, CA: Naval Postgraduate School, 270-294, 1992

"A normalized correlation estimator for complex data based on a quadruplex transformation," with Mark C. Sullivan, *IEEE Signal Processing Letters*, 4(1), 26-28, 1997.

"Geometric modeling methods for facial landmark detection and recognition," with Barnabas Takacs and Harry Wechsler, *Computing Science and Statistics*, 28, 278-286, 1997.

"Geometric modeling of vehicle paths and confidence regions," with Celesta G. Ball, *Computing Science and Statistics*, 28, 287-292, 1997

"High dimensional clustering using parallel coordinates and the grand tour," with Qiang Luo, *Computing Science and Statistics*, 28, 352-360, 1997

"A mixed measure formulation of the EM algorithm for huge data set applications, with George Rogers and Bradley C. Wallet, *Computing Science and Statistics*, 28, 492-497, 1997

"A genetic algorithm for best subset selection in linear regression," with Bradley C. Wallet, David Marchette, Jeffrey Solka, *Computing Science and Statistics*, 28, 545-550, 1997

"A new iterative adaptive mixtures type estimator," with Jeffrey Solka, Wendy Poston and Bradley Wallet, *Computing Science and Statistics*, 28, 573-578, 1997

"Nonparametric density estimation using wavelets and the method of sieves," with S.-C. Li, *Computing Science and Statistics*, 28, 641-646, 1997

"The filtered mode tree," with David Marchette, *Journal of Computational and Graphical Statistics*, 6(2), 143-159, 1997

"Statistical software, siftware and astronomy (with discussion)," with Daniel B. Carr, R. Duane King, John J. Miller, Wendy L. Poston, Jeffrey L. Solka, and John Wallin, in *Statistical Challenges in Modern Astronomy II* (Babu, G. J. and Feigelson, E. D., eds.) New York: Springer-Verlag, 185-206, 1997

"Parallel coordinate plot analysis of polarimetric NASA/JPL AIRSAR imagery," with J. L. Solka, G. W.

Rogers, W. L. Poston, *Automatic Target Recognition VII - Proceedings of SPIE*, 3069, 175-184, 1997

"A deterministic method for robust estimation of multivariate location and shape," with W. L. Poston, C. E. Priebe, and J. L. Solka, *Journal of Computational and Graphical Statistics*, 6(3), 300-313, 1997

"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, 1998

**Ami Motro** Professor Amihai Motro received the B.Sc. in mathematics from Tel Aviv University in 1972, the M.Sc. degree in computer science from the Hebrew University of Jerusalem in 1976, and the Ph.D. degree in computer and information science from the University of Pennsylvania in 1981. From 1981 to 1990 he was on the faculty of the Computer Science Department at the University of Southern California. Since 1990 he has been on the faculty of the Information and Software Systems Engineering Department in George Mason University.

Dr. Motro's main areas of research are *information systems, database management, and information retrieval*. Within these areas he has specialized in subjects such as

- intelligent integration of information from multiple sources,
- data mining and information quality,
- imprecision and uncertainty in information systems, and
- intelligent and cooperative database systems.

Some recent accomplishments in these areas are listed below.

In the area of intelligent integration of information, Dr. Motro has designed and led the implementation of an innovative system called Multiplex, that can quickly and flexibly integrate multiple, heterogeneous, and inconsistent information sources. The most recent version is available at <http://www.isse.gmu.edu/~philipp/multiplex>. He was a recipient (as co-PI) of two grants from DARPA: "Information Integration and Interchange: A Federated Systems Approach", and "Knowledge Rovers: A Family of Intelligent Software Agents for Logistics for the Warrior."

In the area of data quality, he has supervised a recent Ph.D. Dissertation entitled "Data Quality and Its Use for Reconciling Inconsistencies in Multidatabase Environments", and is an invited speaker in a forthcoming workshop in Denmark, where he will talk about "Estimating the Quality of Databases and the Answers they Issue."

In the area of uncertainty in databases he has organized two NSF-ESPRIT workshops and is the co-editor of a recent book entitled "Uncertainty Management in Information Systems: From Needs to Solutions". He is the Program Committee Vice-chair of the 1998 14th International Conference on Data Engineering (Management of Uncertainty and Information Quality).

In the area of intelligent and cooperative database systems he has designed and implemented several novel database interfaces, was the invited speaker in an international conference, and wrote several survey articles. In this area he was awarded research grants from the National Science Foundation for "Unified Paradigm for Informal Retrieval of Data and Knowledge" and for "New Applications of Database Knowledge."

Dr. Motro has published over 50 research papers in leading journals and in international conferences. He was the editor and co-editor of several books, conference proceedings, and special issues of journals. He has been involved in about two dozen international conferences and workshops as organizer, program committee chair and vice-chair, program committee member, panelist, and invited speaker. He has received several research grants from agencies such as the National Science Foundation, DARPA, Amoco, and AT&T. Dr. Motro is on the editorial boards of the Journal of Intelligent Information Systems and the Journal of Knowledge Discovery and Data Mining. He is a member of the Association for Computing Machinery and the IEEE Computer Society, and has been serving as ACM Lecturer since 1991.

### **Selected publications of A. Motro over the last five years**

A. Motro. Responding with Knowledge. *International Journal of Expert Systems (Special Issue: Artificial Intelligence and Databases)*, Vol. 6, No. 1, 1993, pp. 121-138.

A. Motro. Intentional Answers to Database Queries. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 6, No. 3, June 1994, pp. 444-454.

A. Motro. Management of Uncertainty in Database Systems. In *Modern Database Systems: the Object Model, Interoperability and Beyond*, Addison-Wesley/ACM Press, 1994, pp. 457-476.

A. Motro and M. Tennenholtz, Editors. *Proceedings of the NGITS-95, Second International Workshop on Next Generation Information Technologies and Systems*, Naharia, Israel, June 1995, 218 pages.

A. Motro. Imprecision and Uncertainty in Database Systems. In *Fuzziness in Database Management Systems*, Physica-Verlag, Heidelberg, 1995, pp. 3-22.

A. Motro. Panorama: A Database System that Annotates its Answers to Queries with their Properties. *Journal of Intelligent Information Systems*, Vol. 7, No. 1, September 1996, pp. 51-73.

A. Motro. Cooperative Database Systems. *International Journal of Intelligent Systems*, Vol. 11, No. 10, October 1996, pp. 717-732.

A. Motro and I. Rakov. Estimating the Quality of Data in Relational Databases. In *Proceedings of the 1996 Conference on Information Quality*, Cambridge, Massachusetts, October 1996, pp. 94-106.

A. Motro and P. Smets, Editors. *Uncertainty Management in Information Systems: from Needs to Solutions*, Kluwer Academic Publishers, 1996, 480 pages.

A. Motro and I. Rakov. Not All Answers are Equally Good: Estimating the Quality of Database Answers. In *Flexible Query Answering Systems*, Kluwer Academic Publishers, 1997, pp. 1-21.

**Harry Wechsler** Harry Wechsler received the Ph.D. in Computer Science from the University of California, Irvine, in 1975, and he is presently Professor of Computer Science at George Mason University and Director of the Center for Parallel and Distributed Computation. His research, in the field of intelligent systems, has been in the areas of perception: Computer Vision (CV), Automatic Target Recognition (ATR), Signal and Image Processing (SIP), machine intelligence: Pattern Recognition (PR), Neural Networks (NN), Machine Learning (ML), Information Retrieval, Data Mining and Knowledge Discovery, evolutionary computation: Genetic Algorithms (GAs) and Animats, multimedia and video processing: Large Image Data Bases, Document Processing, and human-computer intelligent interaction

(HCII) : Face and Hand Gesture Recognition, Biometrics and Forensics. He was Director for the NATO Advanced Study Institutes (ASI) on "Active Perception and Robot Vision" (Maratea, Italy, 1989), "From Statistics to Neural Networks" (Les Arcs, France, 1993) and "Face Recognition: From Theory to Applications" (Stirling, UK, 1997), and he served as co-chair for the International Conference on Pattern Recognition held in Vienna, Austria, in 1996. He authored over 150 scientific papers, his book *Computational Vision* was published by Academic Press in 1990, and he is the editor of the book, *Neural Networks for Perception* (Vol 1 & 2), published by Academic Press in 1991. He was elected as an IEEE Fellow in 1992. He holds grants from the Army Research Laboratory (ARL) and Defense Advanced Research Project Agency (DARPA) (DAAL01-97-K-0118) for research on "Face Recognition II." He is a principal in the Center for Parallel and Distributed Computation and is involved with the development of expertise in both theoretical, developmental, and applied areas of the field of parallel and distributed computation. Specific objectives of the center are: 1) to perform fundamental research on algorithms, architectures, and system software for parallel and distributed computation. 2) to develop and test unique parallel architectures and special purpose applications of parallel and distributed computation. 3) to advance computational intelligence studies, including brain and cognitive modeling using the connectionist and distributed approaches. 4) to serve as a knowledge base in the area of parallel and distributed computation.

### **Selected Publications of H. Wechsler in the last five years**

J. Bala and Harry Wechsler (1993), Shape Analysis Using Genetic Algorithms, *Pattern Recognition Letters*, 14(12), 965-973.

H. Wechsler (1993), A Perspective on Evolution and the Lamarckian Hypothesis Using Genetic Algorithms, *Revue Internationale de Systemique*, 7(5), 573-592.

D. X. Le, G. R. Thoma, and H. Wechsler (1994), Automated Page Orientation and Skew Angle Detection for Binary Document Images, *Pattern Recognition*, 27(10), 1325-1344.

J. Bala, K. DeJong, J. Huang, H. Vafaie, and H. Wechsler (1995), Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification, *14<sup>th</sup> Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Montreal, Canada.

B. Takacs and H. Wechsler (1995), Face Location Using A Dynamic Model of Retinal Feature Extraction, *1<sup>st</sup> Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland.

S. Gutta, J. Huang, I. Shah, D. Singh, B. Takacs, and H. Wechsler (1995), Benchmark Studies on Face Recognition, *Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland.

D. X. Le, G. R. Thoma, and H. Wechsler (1995), Classification of Binary Document Images as Textual and Non-Textual Data Using Neural Networks, *Machine Vision*, 8, 289-304.

D. Le, G. R. Thoma, and H. Wechsler (1996), Automated Borders Detection and Adaptive Segmentation for Binary Document Images, *13<sup>th</sup> Int. Conf. on Pattern Recognition (ICPR)*, Vienna, Austria.

S. Gutta, J. Huang, B. Takacs, and H. Wechsler (1996), Face Recognition Using Ensembles of Networks, *13<sup>th</sup> Int. Conf. on Pattern Recognition (ICPR)*, Vienna, Austria.

J. Bala, K. DeJong, J. Huang, H. Vafaie, and H. Wechsler (1996), Visual Routine for Eye Detection Using

Hybrid Genetic Architecture, *13<sup>th</sup> Int. Conf. on Pattern Recognition (ICPR)*, Vienna, Austria.

J. Bala and H. Wechsler (1996), Shape Analysis Using Multistrategy Learning, *Pattern Recognition*, 29(8), 1323-1333.

V. Concepcion and H. Wechsler (1996), Detection and Localization of Objects in Time-Varying Imagery Using Attention, Representation, and Memory Pyramids, *Pattern Recognition*, 29(9), 1543-1557.

J. Huang, S. Gutta, and H. Wechsler (1996), Detection of Human Faces Using Decision Trees, *2<sup>nd</sup> Int. Workshop on Automatic Face and Gesture Recognition*, Killington, VT.

S. Gutta, J. Huang, H. Wechsler, and V. Chen (1997), Automatic Video-Based Person Authentication Using the RBF Network, *1<sup>st</sup> Int. Conf. on Audio - and Video - Based Biometric Person Authentication*, Crans-Montana, Switzerland.

J. Bala, K. DeJong, J. Huang, H. Vafaie, and H. Wechsler (1996), Using Learning to Facilitate the Evolution of Features for Recognizing Visual Concepts, *Evolutionary Computation*, 4(3), 297-312.

S. Gutta and H. Wechsler (1997), Face Recognition Using Hybrid Classifiers, *Pattern Recognition*, 30(4), 539-553.

S. Gutta, I. Imam, and H. Wechsler (1997), Hand Gesture Recognition Using Ensembles of Radial Basis Functions (ERBFs) and Decision Trees (DTs), *Int. Journal of Pattern Recognition and Artificial Intelligence*, 11(6), 845-872.

**Sushil Jajodia** Sushil Jajodia is Professor of Information and Software Systems Engineering and Director of Center for Secure Information Systems at the George Mason University, Fairfax, Virginia. He joined GMU after serving as the director of the Database and Expert Systems Program within the Division of Information, Robotics, and Intelligent Systems at the National Science Foundation. Before that he was the head the Database and Distributed Systems Section in the Computer Science and Systems Branch at the Naval Research Laboratory, Washington and Associate Professor of Computer Science and Director of Graduate Studies at the University of Missouri, Columbia. He has also been a visiting professor at the University of Milan, Italy and at the Isaac Newton Institute for Mathematical Sciences, Cambridge University, England.

Dr. Jajodia received his Ph.D. from the University of Oregon, Eugene. His research interests include information security, temporal databases, and replicated databases. He has published more than 150 technical papers in the refereed journals and conference proceedings and has edited or coedited ten books, including *Advanced Transaction Models and Architectures*, Kluwer (1997), *Multimedia Database Systems: Issues and Research Directions*, Springer-Verlag Artificial Intelligence Series (1996), *Information Security: An Integrated Collection of Essays*, IEEE Computer Society Press (1995), and *Temporal Databases: Theory, Design, and Implementation*, Benjamin/Cummings (1993). He received the 1996 Kristian Beckman award from IFIP TC 11 for his contributions to the discipline of Information Security.

Dr. Jajodia has served in different capacities for various journals and conferences. He is the founding co-editor-in-chief of the *Journal of Computer Security*. He is on the editorial boards of *IEEE Concurrency and International Journal of Cooperative Information Systems* and a contributing editor of the *Computer & Communication Security Reviews*. He serves on numerous conference program committees including 1997 IEEE Symposium on Security and Privacy, 1997 Computer Security

Foundations Workshop, and 1997 Very Large Data Base Conference. He has been named a Golden Core member for his service to the IEEE Computer Society. He is a past chairman of the IEEE Computer Society Technical Committee on Data Engineering and the Magazine Advisory Committee. He is a senior member of the IEEE and a member of IEEE Computer Society and Association for Computing Machinery.

### **Selected Publications of S. Jajodia over the last five years**

X. S. Wang, C. Bettini, A. Brodsky, and S. Jajodia, "Logical design for temporal databases with multiple granularities," *ACM Trans. on Database Systems*, Vol. 22, No. 2, June 1997, pages 115-170.

C. Bettini, X. S. Wang, and S. Jajodia, "Testing complex temporal relationships involving multiple granularities and its application to data mining," *Proc. 15th ACM PODS Symp.*, Montreal, Canada, June 1996, pages 68-78.

C. Bettini, X. S. Wang, and S. Jajodia, "A general framework and reasoning models for time granularity," *Proc. 3rd Int'l. Workshop on Temporal Representation and Reasoning*, Key West, FL, May 1996.

X. S. Wang, S. Jajodia, V. S. Subrahmanian, "Temporal Modules: An Approach Toward Federated Temporal Databases," *Information Sciences*, Vol. 82, 1995, pages 103-128.

G. Wiederhold, S. Jajodia, and W. Litwin, "Integrating temporal data in a heterogeneous environment," in *Temporal Databases*, A. Tansel et al., eds., Benjamin/Cummings (1993), pages 563-579.

P. Ammann, S. Jajodia, and I. Ray, "Applying formal methods to semantic-based decomposition of transactions," *ACM Trans. on Database Systems*, Vol. 22, No. 2, June 1997, pages 215-254.

P. Ammann, S. Jajodia, I. Ray, "Using formal methods to reason about semantics-based decompositions of transactions," *Proc. VLDB Conf.*, Zurich, Switzerland, September 1995, pages 218-227.

P. Ammann, S. Jajodia, and I. Ray, "Ensuring atomicity of multilevel transactions," *Proc. IEEE Symp. on Research in Security and Privacy*, Oakland, Calif., May 1996, pages 74-84.

O. Wolfson, S. Jajodia, and Y. Huang, "An adaptive data replication algorithm," *ACM Trans. on Database Systems*, Vol. 22, No. 2, June 1997, pages 215-314.

O. Wolfson, S. Jajodia, "An algorithm for dynamic data allocation in distributed systems," *Information Processing Letters*, Vol. 53, No. 2, 1995, pages 113-119.

P. Ammann, S. Jajodia, C. D. McCollum, and B. T. Blaustein, "Surviving information warfare attacks on databases," *Proc. IEEE Symp. on Research in Security and Privacy*, Oakland, Calif., May 1997, pages 31-42.

Neil F. Johnson and Sushil Jajodia, "Exploring Steganograph: Seeing the unseen," *IEEE Computer*, Vol. 31, No. 2, February 1998, pages 26-34.

P. Ammann, S. Jajodia, and P. Frankl, "Globally consistent event ordering in one-directional distributed environments," *IEEE Trans. on Parallel and Distributed Systems*, Vol. 7, No. 6, June 1996, pages 665-670.

P. Ammann, S. Jajodia, and P. Mavuluri, "On-the-fly reading of entire databases," *IEEE*

*Trans. on Knowledge and Data Engineering*, Vol. 7, No. 5, 1995, pages 834-838.

P. Ammann, V. Atluri, and S. Jajodia, "The partitioned synchronization rule for planer partial orders," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 7, No. 5, 1995, pages 797-809.

P. Ammann, S. Jajodia, "Distributed timestamp generation in planar lattice networks," *ACM Trans. on Computer Systems*, Vol. 11, No. 3, August 1993, pages 205-225.

S. Jajodia, P. Samarati, V. S. Subrahmanian, and E. Bertino, "A Unified Framework for Enforcing Multiple Access Control Policies," *Proc. ACM SIGMOD Int'l. Conf. on Management of Data*, May 1997, pages 474-485.

S. Jajodia, P. Samarati, V. S. Subrahmanian, "A logical language for expressing authorizations," *Proc. IEEE Symp. on Security and Privacy*, Oakland, Calif., May 1997, pages 31-42.

E. Bertino, S. Jajodia, and P. Samarati, "Supporting multiple access control policies in database systems," *Proc. IEEE Symp. on Research in Security and Privacy*, Oakland, Calif., May 1996, pages 94-107.

K. S. Candan, Sushil Jajodia and V.S. Subrahmanian, "Secure mediated databases," *Proc. 12<sup>th</sup> Int'l. Conf. on Data Engineering*, 1996, pages 28-37.

P. Samarati, E. Bertino, and S. Jajodia, "An authorization model for a distributed hypertext system," *IEEE Trans. on Knowledge and Data Engineering*, August 1996, Vol. 8, No. 4, August 1996, pages 555-562.

E. Bertino, P. Samarati, and S. Jajodia, "An Extended Authorization Model for Relational Databases", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 9, No. 1, 1997, pages 85-101.

P. Samarati, P. Ammann, and S. Jajodia, "Maintaining replicated authorizations in distributed database systems," *Data & Knowledge Engineering*, Vol. 18, No. 1, February 1996, pages 55-84.

E. Bertino, S. Jajodia, and P. Samarati, "A non-timestamped authorization model for relational databases," *Proc. 3rd ACM Conf. on Computer and Communications Security* New Delhi, India, March 1996, pages 169-178.

P. Samarati, P. Ammann, S. Jajodia, "Propagation of authorizations in distributed database systems," *Proc. 2nd ACM Conf. on Computer & Communications Security*, November 1994, pages 136-147.

**Kathryn Blackmond Laskey** Kathryn Laskey earned her Ph.D., Statistics and Public Affairs, Carnegie Mellon University, her M.S., Mathematics, University of Michigan, and her B.S., Mathematics, University of Pittsburgh. She is presently an Associate Professor of Systems Engineering, George Mason University. In 1990-1991 she held the position of Research Associate Professor of C3I, George Mason University where she performed research in probabilistic inference, automated construction of Bayesian inference networks, and information fusion for situation assessment. Prior to that from 1985-1991, she was a Principal Scientist, Decision Science Consortium, Inc. where she managed projects to develop decision and inference support systems, developed methodologies for decision and inference support systems, provided consultation on statistical methods and performed statistical analyses for a national survey of pesticide contamination of well water.

Her research interests include knowledge representations and inference strategies for automated reasoning under uncertainty, decision and inference support systems, Bayesian probabilistic inference and its relationship to other theories of inference, automated construction and revision of probability and decision models, fusion of sensor information with higher level knowledge, and intelligent tutoring. She has served as a member of the Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics, National Academy of Sciences, May 1994-present, an instructor, Summer Institute on Probability in Artificial Intelligence, Oregon State University, July, 1994, as a discussant, Workshop on Statistical Issues in Defense Analysis and Testing, Committee on National Statistics, National Academy of Sciences, September 1992, as a member of Panel to Evaluate Studies in Bilingual Education, Committee on National Statistics, National Academy of Sciences, 1991.

### **Selected Publications of K. Laskey**

Laskey, K. B. and P. E. Lehner. "Meta Reasoning and the Problem of Small Worlds," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 24, No. 11 (1994), 1643-1652.

Laskey, K. B. "Bounded Rationality and Search over Small World Models," *International Journal of Approximate Reasoning*, Vol. 11, No. 4 (1994), 361-384.

Adelman, L. A., M. S. Cohen, T. A. Bresnick, J. O. Chinnis, and K. B. Laskey. "Real-Time Expert System Interfaces, Cognitive Processes, and Task Performance: An Empirical Assessment," *Human Factors*, forthcoming.

Laskey, K. B. "The Bounded Bayesian," *Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference*, Morgan Kaufman, 1992.

Laskey, K. B. "Bayesian Meta-Reasoning: Determining Model Adequacy From Within a Small World," *Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference*, Morgan Kaufman, 1992.

Laskey, K. B. "Conflict and Surprise: Heuristics for Model Revision," *Uncertainty in Artificial Intelligence: Proceedings of the Seventh Conference*, Morgan Kauffman, 1991.

Laskey, K. B. "A Probabilistic Reasoning Environment," *Uncertainty in Artificial Intelligence 6*, North Holland Press, 1991.

Laskey, K. B. and V. Cambell. "Evaluation of an Intermediate Level Decision Analysis Course." In Baron, J. and R.V. Brown (eds.). *Teaching Decision Making to Adolescents*. Lawrence Erlbaum and Associates, 1991.

Laskey, K. B. and V. Cambell. "Implementation Strategy for a Course in Decision Making." In Baron, J. and R.V. Brown (eds.). *Teaching Decision Making to Adolescents*. Lawrence Erlbaum and Associates, 1991.

Laskey, K. B. "Adapting Connectionist Learning to Bayes Networks," *International Journal of Approximate Reasoning* (May 1990).

Laskey, K. B., M. Cohen, and A. Martin. "Representing and Eliciting Knowledge About Uncertain Evidence and its Implications," *IEEE Transactions on Systems, Man and Cybernetics* (1989).

Laskey, K. B. and P. Black, "Hierarchical Evidence and Belief Functions," *Uncertainty in Artificial*

*Intelligence 4*, North Holland Press, 1989.

**Larry Kerschberg** Dr. Larry Kerschberg is Professor and Chairman of the Department of Information and Software Systems Engineering in the School of Information Technology and Engineering at George Mason University. He is also Director of the Center for Information Systems Integration and Evolution. Dr. Kerschberg holds a Ph.D. in Engineering from Case Western Reserve University, an M.Sc. in Electrical Engineering from the University of Wisconsin-Madison, and a B.Sc. degree in Engineering Science from Case Institute of Technology.

Dr. Kerschberg's research is in the areas of data and knowledge models, database design, active data dictionaries, distributed query processing, object-oriented systems, software architecture, knowledge discovery in databases and expert database systems. His current research projects include "Information Integration and Interchange: A Federated Systems Approach" and "Knowledge Discovery in Databases," both sponsored by ARPA, and "Process Model Generation and Tailoring in the Evolutionary Spiral Process Model," sponsored by the Software Productivity Consortium, the Virginia Center of Excellence in Software Reuse and Technology Transfer (VCOE), and the Virginia Center for Innovative Technology.

Dr. Kerschberg is also President of KRM, Inc., a small-business enterprise dedicated to knowledge resource management for the design and construction of intelligent information systems. Dr. Kerschberg has consulted for the Banco Real of Sao Paulo, Brazil where he developed the Functional Data Model, IBM do Brasil, Blue Cross and Blue Shield of South Carolina, Computer Sciences Corporation, and Dove Electronics. In his work with BC/BS he served as Project Director for the MEDCLAIM system, an operational expert system for medical claims review. Past research projects include Active Data/Knowledge Dictionaries sponsored by Rome Laboratory, and Knowledge Based Approaches to Telecommunication Network Management sponsored by MCI, Inc.

Dr. Kerschberg serves as Coordinating Editor-in-Chief of the *International Journal of Intelligent Information Systems*. He served as General Chair of the 1993 ACM SIGMOD Conference held in Washington, DC, in May 1993. Dr. Kerschberg organized and has served as Program Chairman of both the First and Second International Conferences on Expert Database Systems. These conferences helped to provide impetus for the integration of AI and Database technologies. He is past Chairman of the IEEE Computer Society's Technical Committee on Data Engineering.

### **Selected Reference of L. Kerschberg**

C. Bosch, H. Gooma, and L. Kerschberg, "Design and Construction of a Software Engineering Environment: Experiences with Eiffel," in *IEEE Readings in Object-Oriented Systems and Applications*, D. Rine, Ed. Piscataway, NJ: IEEE Computer Society Press, 1995.

A. Brodsky, L. Kerschberg, and S. Varas, "On Optimal Constraint Decomposition in Distributed Databases," Center for Information Systems Integration and Evolution, George Mason University, Fairfax, VA, Technical Report October 1997.

H. Gooma and L. Kerschberg, "Domain Modeling for Software Reuse and Evolution," presented at Proc. Computer Assisted Software Engineering Workshop (CASE 95), Toronto, CA, 1995.

H. Gooma, L. Kerschberg, and D. Menascé, "A Software Architecture Design Method of Large-Scale Distributed Data Intensive Information Systems," George Mason University, Fairfax, CISIE Technical Report 1, April 1995.

- H. Gomaa, L. Kerschberg, V. Sugumaran, C. Bosch, and I. Tavakoli, "A Prototype Domain Modeling Environment for Reusable Software Architectures," presented at International Conference on Software Reuse, Rio de Janeiro, Brazil, 1994.
- H. Gomaa, L. Kerschberg, V. Sugumaran, I. Tavakoli, and L. O'Hara, "A Knowledge-Based Software Environment for Reusable Software Requirements and Architectures," *Journal of Automated Software Engineering*, vol. 3, 1996.
- H. Gomaa, D. Menascé, and L. Kerschberg, "A Software Architectural Design Method for Large-Scale Distributed Information Systems," *Journal of Distributed Systems Engineering*, 1996.
- S. Jajodia and L. Kerschberg, "Advanced Transaction Models and Architectures," . Norwall, MA: Kluwer Academic Publishers, 1997.
- L. Kerschberg, "Knowledge Rovers: Cooperative Intelligent Agent Support for Enterprise Information Architectures," in *Cooperative Information Agents*, vol. 1202, *Lecture Notes in Artificial Intelligence*, P. Kandzia and M. Klusch, Eds. Berlin: Springer-Verlag, 1997, pp. 79-100.
- L. Kerschberg, "The Role of Intelligent Agents in Advanced Information Systems," in *Advanced in Databases*, vol. 1271, *Lecture Notes in Computer Science*, C. Small, P. Douglas, R. Johnson, P. King, and N. Martin, Eds. London: Springer-Verlag, 1997, pp. 1-22.
- L. Kerschberg, H. Gomaa, D. Menascé, and J. P. Yoon, "Data and Information Architectures for Large-Scale Distributed Data-Intensive Information Systems," George Mason University, Fairfax, CISIE Technical Report CISIE Technical Report, April, 1995.
- L. Kerschberg, H. Gomaa, D. A. Menascé, and J. P. Yoon, "Data and Information Architectures for Large-Scale Distributed Data Intensive Information Systems," presented at Proc. of the Eighth IEEE International Conference on Scientific and Statistical Database Management, Stockholm, Sweden, 1996.
- L. Kerschberg, S. W. Lee, and L. Tischer, "A Methodology and Life Cycle Model for Data Mining and Knowledge Discovery for Precision Agriculture," Center for Information Systems Integration and Evolution, George Mason University, Fairfax, Final Report 1997.
- K. Massey, L. Kerschberg, and G. Michaels, "VANILLA: A Dynamic Data Model for a Generic Scientific Database," presented at International Conference on Statistical and Scientific Database Management, SSDBM, Olympia, WA, 1997.
- D. A. Menascé, H. Gomaa, and L. Kerschberg, "A Performance-Oriented Design Methodology for Large-Scale Distributed Data Intensive Information Systems," presented at First IEEE International Conference on Engineering of Complex Computer Systems, Florida, 1995.
- R. S. Michalski, L. Kerschberg, K. Kaufman, and J. Ribeiro, "Mining for Knowledge in Databases: The INLEN Architecture, Initial Implementation and First Results," *Journal of Intelligent Information Systems*, vol. 1, pp. 85-113, 1992.
- L. J. Milask, T. Guynup, C. Hammel, L. Kerschberg, and G. Michaels, "An Integrated Scientific Database System and Value-Added Support Center: Application to Ecological Research of the Forest Canopy and Biosphere Interface," presented at International Conference on Statistical and Scientific Database Management, SSDBM, Olympia, WA, 1997.

J. Ribeiro, K. Kaufman, and L. Kerschberg, "Knowledge Discovery in Multiple Databases," presented at ISMM International Conference on Intelligent Information Management Systems, Washington D.C., 1995.

J. Ribeiro, K. Kaufman, and L. Kerschberg, "Knowledge Discovery in Multiple Databases," presented at First International Conference on Knowledge Discovery and Data Mining, Montreal, CA, 1995.

L. Seligman and L. Kerschberg, "An Active Database Approach to Consistency Management in Heterogeneous Data- and Knowledge-based Systems," *International Journal of Cooperative and Intelligent Systems*, vol. 2, 1993.

L. Seligman and L. Kerschberg, "Knowledge-base/Database Consistency in a Federated Multidatabase Environment," presented at IEEE RIDE — Interoperability in Multidatabase Systems, Vienna, Austria, 1993.

L. Seligman and L. Kerschberg, "Federated Knowledge and Database Systems: A New Architecture for Integrating of AI and Database Systems," in *Advances in Databases and Artificial Intelligence, Vol. 1: The Landscape of Intelligence in Database and Information Systems*, vol. 1, L. Delcambre and F. Petry, Eds.: JAI Press, 1995.

L. Seligman and L. Kerschberg, "A Mediator for Approximate Consistency: Supporting "Good Enough" Materialized Views," *Journal of Intelligent Information Systems*, vol. 8, pp. 203 - 225, 1997.

J. P. Yoon and L. Kerschberg, "A Framework for Knowledge Discovery and Evolution in Databases," *IEEE Transactions on Knowledge and Data Engineering*, 1993.

J. P. Yoon and L. Kerschberg, "Semantic Query Reformulation in Object-Oriented Systems," in *International Conference on Deductive and Object-Oriented Databases*, vol. 760. Phoenix, AZ: Springer-Verlag, 1993.

J. P. Yoon and L. Kerschberg, "Semantic Update Optimization in Active Databases," presented at Proceedings IFIP WG2.6 Working Conference on Database Semantics (DS-6), Atlanta, 1995.

J. P. Yoon and L. Kerschberg, "Query-Driven Data Mining and Knowledge Discovery in Very Large Relational Databases," Center for Information Systems Integration and Evolution, George Mason University, Fairfax, Technical Report 1998.

**Ryszard S. Michalski** Ryszard S. Michalski received the B.S. degree from the Warsaw University of Technology, the M.S. degree from St. Petersburg Polytechnical University, and the Ph.D. degree from the University of Silesia in Poland, the latter in 1969. Before emigrating to the United States in 1970, he was a research scientist at the Polish Academy of Sciences. From 1970 to 1987, Dr. Michalski was at the University of Illinois at Urbana-Champaign where he became a Professor of Computer Science and more recently the Director of Artificial Intelligence Laboratory. In 1988 he joined George

Mason University. He is currently a Planning Research Corporation Professor of Computer Science and Systems Engineering, and Director of the Center for Machine Learning and Inference. Dr. Michalski is co-founder of the field of machine learning and is world-renowned for his contributions to this field. He has introduced a number of novel

ideas to artificial intelligence, pattern recognition, cognitive science and multiple-valued logic, and originated or co-originated several research subareas, such as symbolic methods of inductive inference, conceptual clustering, expert systems with learning capabilities, constructive induction, conceptual clustering, and variable-precision logic. Dr. Michalski has authored or co-authored over 250 publications in his areas of interest, and co-edited several books in machine learning. He is a co-founder of the Journal of Machine Learning and a member of the editorial board of several journals. He is a Fellow of the American Association for Artificial Intelligence.

***Menas Kafatos*** Dr. Menas Kafatos is a University Professor affiliated with the Institute for Computational Science and Informatics and is Director of the Center for Earth Observing and Space Research. He received his 1967 B.A. in Physics from Cornell University, and in 1972, he earned his Ph.D. in Physics from M.I.T. His thesis advisor was Professor Philip Morrison. His research fields are theoretical and computational astrophysics, observational astronomy, space sciences, computational fluid dynamics (CFD), foundations of quantum theory, informatics, Earth Observing System Data Information System architecture, visualization of Earth and space science data, interdisciplinary studies in Earth science. He was also the principal investigator of the EOSDIS Independent Architecture Study in 1994. His topics of research include black holes, active galaxies and quasars, accretion hydrodynamics in curved spacetime, gamma-rays from active galaxies, ultraviolet astronomy, symbiotic stars, cosmological observations and their limitations, atomic physics and atomic calculations, time-dependent cooling in the interstellar medium, evolution of supernova remnants inside superbubbles, stellar winds and superbubbles, universal diagrams, mass loss from Mira variables, visualization of structure of the universe, fractal distribution of galaxy clusters, foundations of quantum theory, Virtual Domain Application Data Centers, EOSDIS architecture, visualization of Earth and space science data, interdisciplinary studies in Earth science.

Dr. Kafatos was principal investigator or co-principal investigator on more than 50 observing programs and multiple grants in astronomy and Earth science. He was a principal in establishing the Institute for Computational Sciences and Informatics and served as its founding director. He helped establish a new doctoral program in Computational Sciences with emphasis in space sciences, Earth science and global changes, computational chemistry, computational physics, bioinformatics, computational mathematics and computational statistics, and he secured external funding for the operation of the research program of the Institute from a variety of sources. He is author of more than 125 publications and author or editor of 9 books.

### **Selected Publications of M. Kafatos**

M. Kafatos, "High-Energy Emission in Accretion" Flows in AGN", in *Testing the AGN Paradigm*, ed. by S. Holt, S.G. Neff, and C.M. Urry, *AIP Conference Proc.* 254, 333 (1992).

A.G. Michalistsianos, M. Kafatos and S.R. Meier, "Fe II Fluorescence and Anomalous C IV Doublet Intensities in Symbiotic Novae", *Ap. J.*, 389, 649 (1992).

S.R. Meier, and M. Kafatos, "Correlated UV Line Fluxes of Two Symbiotic Stars", in *Nonisotropic and Variable Outflows from Stars*, ed. by L. Drissen, C. Leitherer, and A. Nota, *ASP Conference Series*, 22, 311 (1992).

S.N. Shore, A.G. Michalitsianos, and M. Kafatos, "Long-Slit UV Spectroscopy of the Circumstellar Environment of the Symbiotic Star R Aqr", *ibid.* 308 (1992).

M. Kafatos, S. Meier and I. Martin, "Extended Variability of the Symbiotic Star AG Dra", *Ap. J. Suppl.*, 84, 201 (1993).

- M. Kafatos, and P. Becker, "Gamma-Rays from Hot Accretion Disks in AGN", *Proceedings of CGRO Symposium*, St. Louis (1993).
- M. Kafatos, and R. Yang, "Transonic Thin Disk Flows in the Schwarzschild Metric", *M.N.R.A.S.*, 268, 925 (1994).
- A.G., Michalitsianos, M. Perez, and M. Kafatos, "Evidence Signaling the Start of Enhanced Counterjet Flow in the Symbiotic System R Aqr", *Ap. J.*, 423, 441, (1994).
- P. Becker, M. Kafatos and M. Maisack, "Relativistic Particle Transport in Hot Accretion Disks", *Ap. J. Suppl.*, 90, 949 (1994).
- M. Kafatos, Lead PI, "The GMU ECS Federated Client-Server Architecture", Report to Hughes Applied Information Systems (August 1994).
- P. Becker, and M. Kafatos, "Implications of Gamma-Ray Transparency Constraints in Blazars", *Ap. J.*, 453, 83 (1995).
- D.J. Macomb et al., "Multiwavelength Observations of Markarian 421 During a TeV/X-Ray Flare", *Ap. J. Lett.*, 449, L99 (1995).
- R. Yang, and M. Kafatos, "Shock Study in Fully Relativistic Isothermal Flows. II", *Astron. Astrophys.*, 295, 238 (1995).
- S.R. Meier, and M. Kafatos, "UV Temporal Variability of the Peculiar Star R Aqr", *Ap.J.*, 451, 359 (1995).
- M. Kafatos, and H. Wolf, "The EOSDIS Core System Architecture: Earth Science Challenges to Information Technology Implementation", white paper (1995).
- M. Kafatos, "Knowledge Limits in Cosmology", in *Examining the Big Bang and Diffuse Background Radiations*, edit. M. Kafatos and Y. Kondo, *IAU Symposium* 168, 431, Kluwer Academic Press (1996).
- M. Kafatos, "Jet and MHD Flows Associated with Symbiotic Stars", in *MHD Flows in Astrophysical Plasmas*, edit. K. Tsinganos (1996).
- P. Subramanian, P. Becker, and M. Kafatos, "Ion Viscosity Mediated by Tangled Magnetic Fields: An Application to Black Hole Accretion Disks", *Ap. J.*, in press.
- M. Kafatos, E. Ramos, P. Becker, P. Subramanian, and R. Yang, "Unified Models of AGN Accretion Disks and Blazars", workshop on Blazar Continuum Variability, in press.
- E. Ramos, and M. Kafatos, "EUVE Observations of 3C273", workshop on Blazar Continuum Variability, in press.
- M. Kafatos, and R. Nadeau, *The Conscious Universe: Part and Whole in Modern Physics Theory*, Springer-Verlag (1990).

**Jim X. Chen** Dr. Jim X. Chen is an Assistant Professor of Computer Science in the Department of Computer Science at George Mason University. He joined the Department as an Assistant Professor in

Fall 1995. Previously, he was a Research Associate (Visual Systems Scientist) at the Institute for Simulation & Training, University of Central Florida and worked on graphical modeling and distributed interactive simulation for 6 years before he came to GMU. He was the director of the Artificial Intelligence Lab in the Department of Computer Science, Southwest Jiaotong University in Sichuan, China during 1987 and 1989.

Chen has over 15 years of experience in teaching, research, and software/hardware design and analysis. Over the years he has worked on robotics, artificial intelligence, computer education methodology, physical modeling, real-time animation, and networked virtual environments. His current research interests are in physically-based modeling, real-time simulation, distributed interactive simulation, information visualization, and virtual reality. Chen served as a guest editor for IEEE Computational Science and Engineering. He was the originator of the GMU Upsilon Pi Epsilon Chapter and is currently serving as a faculty advisor. He is a member of ACM and IEEE Society.

Dr. Chen earned his Ph.D. in computer science from the University of Central Florida with the dissertation *Physically-based Modeling and Real-time Simulation of Fluids*. He earned the M.S. in computer science from Southwest Jiaotong University in Sichuan, China and a B.S. in computing science from the same institution.

### **Selected Publications of J. Chen**

J. Wang, J. X. Chen, and E. J. Wegman, "Physical Modeling of Dust Behaviors," *International Conference on Scientific Computing and Mathematical Modeling, IMACS '98*, Alicante, Spain, June, 1998.

Y. Zhu and J. X. Chen, "Establishing a 3D Human Gait Model," *The Sixth International Conference in Central Europe on Computer Graphics and Visualization*, University of West Bohemia, Czech Republic, Feb, 1998.

J. X. Chen, E. J. Wegman, and J. Wang, "Animation of Dust Behaviors in a Networked Virtual Environment," *The Sixth International Conference in Central Europe on Computer Graphics and Visualization*, University of West Bohemia, Czech Republic, Feb, 1998.

E. J. Wegman, Q. Luo, and J. X. Chen, "Immersive Methods for Exploratory Analysis," to appear *Computing Science and Statistics*, 1997.

M. C. Salzman, C. J. Dede, R. B. Loftin, and J. X. Chen, "Understanding How Immersive VR Learning Environments Support Learning through Modeling," submitted to *PRESENCE: The Journal of Teleoperators and Virtual Environments*.

J. X. Chen, "Multiple Segment Line Scan-Conversion," *Computer Graphics Forum*, Vol. 17, No. 5, 1997, pp. 257-268.

J. X. Chen, "An Improvement on Line Scan-Conversion," *SIGGRAPH'97 Visual Proceedings*, 1997.

J. X. Chen, N. V. Lobo, C. E. Hughes and J. M. Moshell, "Real-time Fluid Simulation in a Networked Virtual Environment," *IEEE Computer Graphics and Applications*, May, 1997.

J. X. Chen, D. Rine, and H. D. Simon, "Advancing Interactive Visualization and Computational Steering," *IEEE Computational Science and Engineering*, Vol. 3, No. 4, 1996.

J. X. Chen, "Physically-based Modeling in Information Categorization," *Siggraph'96 Technical Sketches*, New Orleans, Louisiana, August 1996.

Lobo., N. V. and J. X. Chen, "Real-Time Simulation of Fluids," US Patent number 5537641, July 16, 1996.

J. X. Chen and O. Frieder, "Information Retrieval Using Hyper-image Visualization," the *ACM Conference on Information and Knowledge Management (CIKM'95)*, Baltimore, Maryland, Nov., 1995.

Chen, J. X., "Fluid Simulation and Synchronization of Fluids," *Southeastern Simulation Conference*, University of Central Florida, October, 1995, pp. 318-324.

Chen, J. X., N. V. Lobo, C. E. Hughes, and J. M. Moshell, "Simulation and Synchronization of Fluids in a DIS." *First Workshop on Simulation and Interaction in Virtual Environments (SIVE)*, University of Iowa, Iowa City, Iowa, July 13-15, 1995, pp. 159-167.

Chen, J. X. and N. V. Lobo, "Toward Interactive-rate Simulation of Fluids with Moving Obstacles by Navier-Stokes Equations." *CVGIP: Graphical Models and Image Processing*, March, 1995, pp. 107-116.

Chen, J. X. and M. Sartor, "An Approach to Implementing Fluids on Dynamic Terrain." *12th DIS Workshop*, Orlando, Florida, March 1995, pp. 309-318.

Chen, J. X., J. M. Moshell, et al., "Distributed Virtual Environment Real-Time Simulation Network." *Advances in Modeling and Analysis B*, AMSE periodicals, 31:1, 1994, p.1-7.

**Jeffrey L. Solka** Jeffrey L. Solka was born in Harrisonburg, Virginia, on January 31, 1955. He earned the B.S. degree in Mathematics and Chemistry from James Madison University in 1978, the M.S. in Mathematics from James Madison University in 1981, the M.S. in Physics from Virginia Polytechnic Institute and State University in 1989 and his Ph.D. in Computational Sciences and Informatics (Computational Statistics) at George Mason University, working under the direction of Prof. Edward J. Wegman, in May of 1995. Since 1984, Dr. Solka has been working in nonparametric estimation and statistical pattern recognition for the Naval Surface Warfare Center, Dahlgren, VA. He has published over 100 journal, conference, and technical papers, has won numerous awards, and holds 4 patents. His research interests include statistical visualization and mixture based probability density estimation. He has developed Navy systems for such diverse areas as image and video exploitation, acoustic signal processing, and chemical agent detection. Since 1995 Dr. Solka has served in the role of Adjunct Professor of Computational Sciences and Informatics at George Mason University. In Fall of 1997 he was appointed as a permanent part-time faculty member of the department.

### **Selected Publications of J. Solka**

J. L. Solka, G. W. Rogers, and W. L. Poston, "Application of Statistical Visualization Techniques to Image Classification", presented at and appearing in the *Proceedings of Interface 97* (1997).

J. L. Solka and D. J. Marchette, "A New Data Driven Mixture Estimator for Spatially Dependent Observations," presented at and appearing in the *Proceedings of Interface 97* (1997).

J. L. Solka, G. W. Rogers, W. L. Poston, and E. J. Wegman, "Parallel Coordinate Plot Analysis of Polarimetric NASA/JPL AIRSAR Imagery," *Automatic Target Recognition VII - Proceedings of SPIE*, Vol. 3069, pp. 175-184 (1997).

D. J. Marchette, J. L. Solka, "A mixture based MAP estimator for image segmentation,"

to be presented at and appear in the *Proceedings of AeroSense 97* (1997).

D. J. Marchette, J. L. Solka, R. Guidry, J. Green, "The Advanced Distributed Region of Interest Tool," submitted to *Pattern Recognition* (1997).

J. L. Solka, D. J. Marchette, B. C. Wallet, V. L. Irwin, G. W. Rogers, "Identification of Man-made Regions in Unmanned Aerial Vehicle Imagery and Videos," under review *PAMI* (1997)

J. L. Solka, W. L. Poston, D. J. Marchette, and E. J. Wegman, "Massive Data Sets in Navy Problems," *Massive Data Sets: Proceedings of a Workshop*, National Academy Press, Washington, D. C., pp 157-167 (1996).

J. L. Solka, D. J. Marchette, G. W. Rogers, E. C. Durling, J. E. Green, and D. Talsma, "Region of Interest Identification in Unmanned Aerial Vehicle Imagery," Presented at and Appearing in the *Proceedings of Applied Imagery and Pattern Recognition 96: Emerging Applications of Computer Vision*, (1996)

W. L. Poston, E. J. Wegman, J. L. Solka, "A New Finite Mixtures Parameter Initialization Procedure," Presented at and Appearing in the *Proceedings of JSM '96* (1996)

B. C. Wallet, D. J. Marchette, J. L. Solka, "A Matrix Representation for Genetic Algorithms," To appear in the *Proceedings of the 28th Symposium on the Interface* (1996).

B. C. Wallet, D. J. Marchette, J. L. Solka, E. J. Wegman, "A Genetic Algorithm for Best Subset Selection in Linear Regression," To appear in the *Proceedings of the 28th Symposium on the Interface* (1996).

D. J. Marchette, C. E. Priebe, G. W. Rogers, and J. L. Solka, "Filtered Kernel Density Estimation," *Comp. Stat.* vol 2, 95-112 (1996).

Wegman, Edward J., Daniel B. Carr, R. Duane King, John J. Miller, Wendy L. Poston, Jeffrey L. Solka, and John Wallin, "Statistical software, software and astronomy (with discussion)," in *Statistical Challenges in Modern Astronomy II* (Babu, G. J. and Feigelson, E. D., eds.) New York: Springer-Verlag, 185-206 (1997).

J. L. Solka, E. J. Wegman, W. L. Poston, "A New Order Independent Adaptive Mixtures Type Estimator," Presented at and Appearing in the *Proceedings of JSM96*

J. L. Solka, R. A. Lorey, D. J. Marchette, G. W. Rogers, V. L. Irwin and B. C. Wallet, "Evaluation of Multiple Feature Sets for the Identification of Man-made Regions in Unmanned Aerial Vehicle Imagery" , Presented at and appearing in the *Proceedings of the 5th Automatic Target Recognition Systems and Technology Symposium* (1996).

J. L. Solka, W. L. Poston, E. J. Wegman, and B. C. Wallet, "A New Iterative Adaptive Mixtures Type Estimator," *Computing Science and Statistics*, 28, 573-578 (1997).

R. A. Lorey, J. L. Solka, G. W. Rogers, D. J. Marchette, and C. E. Priebe, "Promising Gains Made in Mammography Thanks to Dual-Use of Military Technology in Statistical Analysis and Image Processing," *NSWCDD Technical Digest*, pp. 152-165 (1995).

- G. W. Rogers, D. J. Marchette, and J. L. Solka, "Wavelet Based Segmentation and Resistive Grid Averaging for Local Feature Extraction," *CRC Handbook on Industrial Electronics*, to appear. V75.
- G. W. Rogers, C. E. Priebe, And J. L. Solka, "A PDP Approach To Localized Fractal Dimension Computation With Segmentation Boundaries," *CRC Handbook on Industrial Electronics*, to appear.
- E. J. Wegman, J. L. Solka, and W. L. Poston, "Immersive Methods for Mine Warfare," *Proceedings of the Military Applications on Synthetic Environments in Virtual Reality '95* (1995)
- Wegman, E. J., Poston, W. L. and Solka, J. L. "Wavelets and nonparametric function estimation," in *Research Developments in Statistics and Probability*, (Brunner, E. and Denker, eds.), Utrecht: VSP, 257-274 (1996).
- E. J. Wegman, H. T. Le, W. L. Poston, and J. L. Solka, "Moments and Wavelets in Signal Estimation," in *Statistics of Quality* (Ghosh, S., Schucany, W., and Smith, W., eds.), Marcel Dekker, 253-274 (1997).
- W. L. Poston, E. J. Wegman, C. E. Priebe and J. L. Solka, "A Recursive Deterministic Method for Robust Estimation of Multivariate Location and Shape," accepted pending *Journal of Computational and Graphical Statistics* (1995).
- J. L. Solka, J. C. Perry, B. R. Poellinger, and G. W. Rogers, "Fast Computation of Optimal Paths Using a Parallel Dijkstra Algorithm With Embedded Constraints," *Neurocomputing*, Vol. 8, pp. 195-212 (1995)
- G. W. Rogers, C. E. Priebe, and J. L. Solka, "A PDP Approach to Localized Fractal Dimension Computation with Segmentation Boundaries," *Simulation* Vol. 65, No. 1, pp. 26-36 (1995).
- O. T. Holland, W. L. Poston and J. L. Solka, "Using Fractal Geometry to Determine the Roughness of Cast Ductile Iron," presented at Interface '95, June (1995).
- J. L. Solka, W. L. Poston, E. J. Wegman, and D. J. Marchette, "Visualization of Adaptive Mixtures Estimates of DNA Flow Cytometry," invited paper Interface '95, June (1995).
- J. L. Solka, W. L. Poston, and E. J. Wegman, "A Visualization Technique for Studying the Iterative Estimation of Mixture Densities," *Journal of Computational and Graphical Statistics*, 4(3), pp.180-197,(1995).
- J. L. Solka, E. J. Wegman, C. E. Priebe, W. L. Poston, and G. W. Rogers, "A Method to Determine the Structure of an Unknown Mixture Using the Akaike Information Criterion and the Bootstrap," *Statistics and Computing*, accepted pending revision (1995).
- R. L. Lorey, J. L. Solka, G. W. Rogers, D. J. Marchette, and C. E. Priebe, "Mammographic Computer Assisted Diagnosis Using Computational Statistics Pattern Recognition," *Real-Time Imaging*, Vol. 1, pp. 94-104 (1995).
- C. E. Priebe, H. I. Hayes, E. G. Julin, G. W. Rogers, D. J. Marchette, and J. L. Solka, "Improved texture classification and image segmentation with boundary incorporation,"
- W. L. Poston and J. L. Solka "Choosing data sets that optimize the determinant of the Fisher information matrix," *Proceedings of the 1994 IEEE-IMS Workshop on Information Theory and Statistics*, Alexandria,

VA, October 27-29, p. 73 (1994).

J. L. Solka, C. E. Priebe, G.W. Rogers, W.L. Poston and D.M. Marchette, "The Application of Akaike Information Criterion Based Pruning to Nonparametric Density Estimates," *Proceedings of the 1994 IEEE-IMS Workshop on Information Theory and Statistics*, Alexandria, VA, October 27-29 (1994).

K. S. Woods, J. L. Solka, C. E. Priebe, W. Philip Kegelmeyer, Jr., C. C. Doss, and K. W. Bowyer, "Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography," *State of the Art in Digital Mammographic Image Analysis*, K.W. Bowyer, and S. Astley, Eds., pp. 213-231 (1994).

C.E. Priebe, E.G. Julin, G.W. Rogers, D.M. Healy, J.L. Solka, and D.J. Marchette, "Incorporating Segmentation Boundaries into the Calculation of Fractal Dimension Features," *Proceedings of the 26th Symposium on the Interface: Computing Science and Statistics*, Interface'94, Research Triangle Park, NC, June 15-18, (1994).

W. L. Poston, G. W. Rogers, C. E. Priebe, J. L. Solka, "A Qualitative Analysis of the Resistive Grid Kernel Estimator," *Pattern Recognition Letters*, 15, pp. 219-225 (1994).

C. E. Priebe, J. L. Solka, R. A. Lorey, G. W. Rogers, W. L. Poston, M. Kallergi, W. Qian, P. Clarke, and R. A. Clark, "The Application of Fractal Analysis to Mammographic Tissue Classification," *Cancer Letters*, 77, pp. 183-189 (1994).

E. G. Julin, G. W. Rogers, C. E. Priebe, and J. L. Solka, "Calculation of Power Law Features in the Presence of Segmentation Utilizing a Dijkstra Potential Based Algorithm," *Proc. High Performance Computing; 1994 Simulation Multiconference*, April 10-15, San Diego, Calif. pp. 357-362 (1994).

W. L. Poston, J. L. Solka, "A Parallel Method to Maximize the Fisher Information Matrix," presented at and appearing in *Proceedings of Intel Supercomputer User's Group*, San Diego, June (1994).

J. L. Solka, C. E. Priebe, G. W. Rogers, W. L. Poston, R. A. Lorey, "Maximum Likelihood Density Estimation With Term Creation and Annihilation," presented at and appearing in the *Proceedings of the 26th Symposium on the Interface '94* (1994).

J. L. Solka, W. L. Poston, C. E. Priebe, G. W. Rogers, R. A. Lorey, D. J. Marchette, K. Woods, K. Bowyer, "The Detection of Micro-Calcifications in Mammographic Images Using High Dimensional Features," presented at and appearing in *Proceedings of the IEEE Seventh Symposiums on Computer-Based Medical Systems* June 10-12, Winston-Salem, North Carolina, pp. 139-145 (1994).

C. E. Priebe, G. W. Rogers, D. J. Marchette, and J. L. Solka, "Change Point Analysis With Adaptive Mixtures Models," presented at and appearing in the *Proceedings of Int'l Geoscience and Remote Sensing Symposium (IGARSS'94)*, Pasadena, CA 8-12 Aug. (1994).

C.E. Priebe, R.A. Lorey, D.J. Marchette, J.L. Solka, and G.W. Rogers, "Nonparametric Spatio-Temporal Change Point Analysis for Early Detection in Mammography," *Proceedings of the 1994 Second Int'l Workshop on Digital Mammography (SIWDM)*, York, UK, July 10-12, pp. 111-120 (1994)

G. W. Rogers, J. L. Solka, C. E. Priebe, and D. S. Malyevac, "Self-Organizing Network for Computing A Posteriori Conditional Class Probabilities," *IEEE Systems, Man, and Cybernetics*, Vol. 23, No. 6, Nov/Dec, pp. 1672-1682 (1993).

J. L. Solka, C. E. Priebe, and G. W. Rogers, "A Probabilistic Approach to Fractal Based Texture Discrimination," *Adaptive and Learning Systems II*, F. A. Sadjadi, Ed., Proc. SPIE 1962, pp. 209-218 (1993).

**Daniel B. Carr** Dr. Daniel Carr is Professor of Statistics in the Department of Applied and Engineering Statistics, at George Mason University and serves as Associate Director of the Center for Computational Statistics and as a member of the Institute for Computational Science and Informatics. He receives his Ph.D. in Statistics from the University of Wisconsin, Madison in 1976, the M.S. Statistics from Oregon State University in 1972, the M.Ed. in Counseling from Idaho State University also in 1972 and the B. A. Mathematics and Psychology from Whitman College in 1968. Dr. Carr is a Fellow of the American Statistical Association, a Fellow of the Washington Academy of Sciences and received the Washington Academy of Sciences 1992 Award for Outstanding Achievement in Mathematics and Computer Science. He holds or has held grants from the U. S. Environmental Protection Agency, the BLS/NSF/ASA Fellowship Program, the National Agriculture Statistical Service. His research interests include statistical graphics and knowledge visualization, exploratory analysis of large and massive data sets, computational environments for data analysis, and reasoning systems. He is a member of the American Statistical Association, Washington Academy of Science, American Association for Artificial Intelligence, IEEE Computer Society, Biometric Society. He has served as Chair, on the Nominating Committee, on the Committee on Fellows, as Newsletter Editor, and as a Newsletter Column Writer for the ASA Statistical Graphics section. He has served as associate editor for *Computational Statistics*, *Journal of the American Statistical Association*, and the *Journal of Computational and Graphical Statistics*.

### **Selected Publications of D. Carr**

Carr, D. B. 1998. "Multivariate Graphics," in *Encyclopedia of Biostatistics*, Peter Armitage of Oxford, and Theodore Colton Boston University Eds. (in press - 23 pages).

George S. Michaels, Daniel B. Carr, Manor Askenazi, Stefanie Fuhrman, Xiling Wen, Roland Somogyi. 1997. "Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data", *Proceedings of the Pacific Symposium on Biocomputing*, edited by Russ Altman, World Scientific Publishing Co, River Edge, NJ, USA (in press)

Carr, D. B., R. Kahn, K. Sahr, and A. R. Olsen. 1997. "ISEA Discrete Global Grids," *Statistical Computing & Graphics Newsletter*, Vol. 8 No. 2 in press.

Wen, X., S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. 1998. "Large-scale Temporal Gene Expression Mapping of Central Nervous System Development," *Proceedings National Academy of Science*. Vol. 95, pp. 334-339.

Carr, D. B., R Somogyi and G. Michaels. 1977. "Templates for Looking at Gene Expression Clustering," *Statistical Computing & Graphics Newsletter*, Vol. 8 No. 1 pp. 20-29.

Wegman, E. J., D. B. Carr, R. D. King, J. J. Miller, W. L. Poston, J. L. Solka, and J. Wallin. 1997. "Statistical Software, Siftware and Astronomy," *Statistical Challenges in Modern Astronomy II*, Eds. G. J. Babu and E. D. Feigelson, Springer-Verlag, New York, pp. 185-206.

Carr, D. B. and S. Pierson. 1996. Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps. *Statistical Computing & Graphics Newsletter*, Vol. 7 No. 3 pp. 16-23.

Carr, D. B., R. Valliant, and D. Rope. 1996. "Plot Interpretation and Information Webs: A Time-Series Example From the Bureau of Labor Statistics" *Statistical Computing & Graphics Newsletter*, Vol. 7 No. 2 pp. 19-26.

Carr, D. B. and A. R. Olsen 1996. "Simplifying Visual Appearance By Sorting: An Example Using 159 AVHRR Classes," *Statistical Computing & Graphics Newsletter*, Vol. 7 No. 1 pp. 10-16.

Carr, D. B. 1995. "Perspective on the Analysis of Massive Data Sets". *Computing Science and Statistics, Proceeding of the 27th Symposium on the Interface*. Vol. 27. p410-419.

Kwang-Su Yang, D. B. Carr and R. J. O'Connor. 1995. "Smoothing of Breeding Bird Survey Data To Produce National Biodiversity Estimates". *Computing Science and Statistics, Proceeding of the 27th Symposium on the Interface*. Vol. 27, pp. 405-409.

Carr, D. B. and S.A. Nusser 1995. "Converting Table to Plots, A Challenge from Iowa State," *Statistical Computing & Graphics Newsletter*, Vol. 6 No. 3 pp. 11-18.

Carr, D. B. and A. R. Olsen. 1995. "Parallel Coordinate Plots For Representing Distribution Summaries in Map Legends." *Proceedings 1 of the 17<sup>th</sup> International Cartography Association Conference 10<sup>th</sup> General Assembly of the ICA*. pp. 733-742.

Carr, D. B. 1995. "Scanning a 4-D Domain for Local Minima: A Protein Folding Application," *Statistical Computing & Graphics Newsletter*, Vol. 6 No. 2 pp. 8-12.

Carr, D. B. and A. R. Olsen. 1995 "Parallel Coordinate Variants of CDF and Quantile Plots," *Statistical Computing & Graphics Newsletter*, Vol. 6 No. 1 pp. 13-18.

Carr, D. B. 1994. Comments on "Prosection Views: Dimensional Inference though Sections and Projections", *Journal of Computational and Graphical Statistics*, pp 369-376.

Carr, D. B. and K. Yang. 1994. "Variations on Row-Labeled Pot for Reexpressing Tabular Summaries." *Computing Science and Statistics, Proceedings of the 26th Symposium on the Interface*. pp. 436-440.

Carr, D. B. 1994. "A Colorful Variation on Boxplots" *Statistical Computing & Graphics Newsletter*, Vol. 5, No. 3, pp. 19-23.

Carr, D. B. 1994. "Using Gray in Plots." *Statistical Computing & Graphics Newsletter*, Vol. 5, No. 2, pp. 11-14.

Carr, D. B. 1994. "Color Perception, The Importance of Gray and Residuals on a Choropleth Map." *Statistical Computing & Graphics Newsletter*, Vol. 5, No. 1, pp. 16-20.

Wegman, E. J. and D. B. Carr. 1993. "Statistical Graphics and Visualization." in *Handbook of Statistics, Computational Statistics*, Vol. 9. ed. C.R. Rao, North Holland, New York.

pp. 857-958.

Wegman, E. J., D. B. Carr and Qiang Luo. 1993. "Visualizing Multivariate Data,"

*Multivariate Analysis: Future Directions*, ed. C.R. Rao, North Holland, New York.

pp. 423-466.

Carr, D. B. and L. W. Pickle. 1993. "Plot Production Issues and Details: Smoothed Cancer Rates and Hexagon Mosaic Maps." *Statistical Computing & Graphics Newsletter*, Vol 4, No. 2, pp. 16-20.

Carr, D. B. 1993. "Production of Stereoscopic Displays for Data Analysis." *Statistical Computing & Graphics Newsletter*, Vol.4 No. 1, pp. 2-7.

The proposal copyright (c) 1998, Center for Computational Statistics, George Mason University. All rights reserved.