Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls,M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

DATA FARMING: DISCOVERING SURPRISE

Gary E. Horne
Ted E. Meyer

The MITRE Corporation
2750 Killarney Drive, Suite 100
Woodbridge, VA 22192, U.S.A.

ABSTRACT

The development of models and analysis of modeling results usually requires that models be run many times. Very few modelers are satisfied with the computing resources available to do sensitivity studies, validation and verification, measurement of effectiveness analysis, and related necessary activities. Fortunately, *high performance computing*, in the form of distributed computing capabilities and commodity node systems, is becoming more pervasive and cost effective. In this paper the authors describe the concept and methods of ***Data Farming***, the study and development of methods, interfaces, and tools that make high performance computing readily available to modelers and allows analysts to explore the vast amount of data that results from exercising models.

1 BACKGROUND

In 1998, General Charles Krulak, then Commandant of the Marine Corps, recognized the inherent nonlinearity of war and that many of the existing combat models and simulations were archaic or inappropriate given the asymmetric nature of modern combat. For a few years leading up to his comments, the USMC led the development of ideas to capture answers not provided by traditional models. A fast running model called ISAAC was developed and the concept of ***Data Farming*** (Brandstein and Horne 1998) was invented. Congress expressed interest in combining the ideas above with high performance computing and other high tech capabilities.

Project Albert (named after Einstein in the same manner that the first model was named after Newton) has continued since then stressing the development of technology to capture and explore huge possible outcome spaces generated by Data Farming.

The capability development has emphasized and benefited from a wide range of interdisciplinary, joint, and coalition collaboration and the still developing experimental technologies have recently begun to be tested in various application areas in collaborative efforts.

Project Albert is *not* about running specific models for predicting a final "answer." Project Albert *is* about Data Farming any model to gain insight in potential outcomes and experimentation with emerging methods. Data Farming, by providing the ability to process large parameter spaces, makes possible the discovery of surprises (both positive and negative) and potential options.

Project Albert is addressing real questions. It has used Data Farming to seek insight into questions such as:

- When is decentralized (vs. centralized) command and control desired or preferred?
- What is the role of trust, or other so-called 'intangibles', on the battlefield?
- How can we best protect our homeland from a martyr-based offense?
- How can a bio-terrorist attack be mitigated in a free society?
- What system characteristics are important in military convoy protection systems?
- How can groups co-exists peacefully?

Of course, there are many other questions which are of interest, but these are but a few of the ones that teams have attempted to address using Data Farming. The Data Farming approach explores these types of questions from the perspective of the 'whole', vice from the perspective of the component parts. And finally, the desire in all of our efforts is to go well beyond point estimates, because the understanding we seek requires much more (Horne 2001).

Data Farming relies on a set of enabling technologies and processes that have been the focus of ongoing research and development efforts: distributed and high-performance computing; agent-based simulations and rapid model development; knowledge discovery methods; high-dimensional data visualization techniques; design-of-experiments methods, human-computer interfaces; team work and collaborative environments; and heuristic search techniques. Project Albert has been, and is continually, pursuing a program of developing, integrating, and applying this methodology and these technologies to problems in the military domain. Project Albert's mission is to: "Create the best Data Farming environment possible to collaboratively explore the vast space of possibilities inherent in the questions that our decision makers face in today's uncertain world."

2 WHAT IS DATA FARMING?

Data Farming can be thought of as nothing more than putting the advances mentioned earlier to work to engage the scientific method. Testing of models and a complete exploration of model output requires that a model of even a modest level of complexity be run a statistically significant number of times over a potentially large parameter space. Few models are examined as extensively as the modeler designers and the decision-makers who rely on them would wish. Even though high performance computing (HPC) is becoming cheaper and more available, the use of HPC currently requires specialized development and expertise. A lot of developmental resources are expended to build ad hoc model execution capabilities.

One objective of this research is to answer the question: "What interfaces, human and software, will allow modelers to easily submit and execute models, design experiments, collect results, explore the results, and support decision making?" Although much research has been done in HPC, most has been done in the areas of processing methodologies and improvements, not in human access and interfaces. As any computing capability becomes more of a commodity, however, ease of use and access takes on a premium and becomes the conduit for new opportunity. Typically, *Data Farming* is an iterative team process. Figure 1 presents the data farming process as a set of imbedded loops. The following steps are inherent in this figure and may be repeated until insight is gained.

- Question/topic research and definition
- Model development and gaming
- Parameter space exploration
- Data exploration and analysis

These steps normally require input and participation by subject matter experts, modelers, analysts, and decision-makers.

The "Scenario Creation" loop shown on the left side of the figure involves building a model that adequately represent the system that pertains to a question being asked by the decision-maker. The scenario creation loop involves both the creation of a model, but also the honing and refining of the question so that all participants understand the scope and intent and confirm that the designed model significantly addresses the question. The scenario or model is crafted so that the decision-maker knows that it is addressing his issue, the subject matter expert believes that the model adequately represents the real world processes at work, the analyst can acquire the data required to examine the outcome space, and the modeler has the resources to implement the model in software. The modeler iterates the implementation with all participants providing feedback into the development. In this phase the team must also define what measurements should be collected from the model in order to address the question being asked. The loops shown in Figure 1 all require the same participation and concurrence.

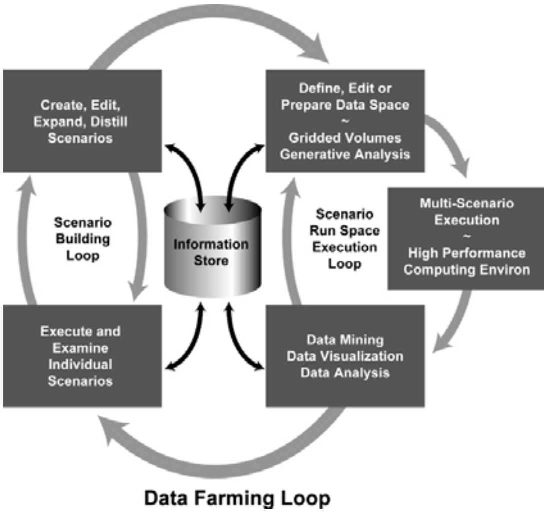

Create, Edit, Expand, Distill Scenarios
Define, Edit or Prepare Data Space ~ Gridded Volumes Generative Analysis
Scenario Building Loop
Information Store
Scenario Run Space Execution Loop
Multi-Scenario Execution ~ High Performance Computing Environ
Execute and Examine Individual Scenarios
Data Mining Data Visualization Data Analysis
Data Farming Loop

Figure 1: Data Farming Iterative Process

During the scenario creation loop the team may want to execute the model any number of times, examining the results to ensure the model is meeting their specific requirements. At this point, high performance computing may be used to "game" parameters. That is, to adjust them so that model appropriately emulates the system being examine.

Once the team agrees that the scenario is represented by a *basecase*, the "Scenario Run Space Execution" loop shown in Figure 1 is entered. In this loop the team determines which scenario parameters should be examined and what processes should be used to vary them. Here the team is exploring the possible variations (or *excursions*) in the initial conditions of the scenario. Specifically those parameters that address the question being posed are considered. The *basecase* provides a starting set of parameter values, but the team must decide what range and limits of variation are appropriate for the scenario. Additionally, the method for varying the parameters should be considered. "Gridded" Data Farming will result in a simple, but voluminous full factorial variation of all parameters by defining ranges and step sizes. Alternatively, evolutionary (or generative analysis) algorithms may be used to explore the parameter space using optimization processes (Lucas et al. 2002). Any number of other methods may be developed to optimally cover the potential parameter space to be examined.

If the model uses random seeds (e.g., to "randomize" initial conditions or in ongoing order or conflict resolution) then the model may be executed some number of times, each with a variation in the initial random seeds used. The number of *replicates* of each excursion must be defined. A study should be done over the parameter space to determine the sensitivity of the results to the random seeding. "Outliers" in the results may be dependent on this randomization and may provide insight into the potential volatility in real-world systems.

On defining the parameter space to be explored, the model is ported to a high performance computing environment and executed. At this point any number of analysis, data mining, and visualization methods may be applied to the output measurements. In Data Farming, there is not necessarily a predefined hypothesis that is being confirmed. The data is being "explored." Trends and relationships may become exposed, but outliers, cusps, and saddle points may also be found in the n-dimensional output space. The completion of this exploration step can result in various outcomes:

- the team may determine that additional areas or resolutions of parameter space should be explored (iteration of the "Scenario Run Space Execution" loop);
- the team may decide that the model needs to be adjusted (back to the "Scenario Creation" loop);
- the team may decide that sufficient insight has been gained or that circumstances require a completion of the effort.

The results of this process may be incorporated into other modeling and operation analysis activities. Insight may be used to adjust wargames, provide input to deterministic models and equations, or build higher verisimilitude simulations and models.

Data Farming provides a never-ending opportunity to explore our questions. The idea is to grow more data in the areas of interest. This growth within a particular definition of a particular distillation might be in the form of more runs or a different preparation of the sample space to include different parameters, finer gradations of parameter values, or greater ranges. After the execution of samples and analysis using data visualization and search methods, the data farmers are free to grow more data in interesting areas, integrate with information from other tools, prepare a different scenario using the same distillation, select another distillation, or any combination of these possibilities that might lead to progress on the question at hand or new questions that arise during this exploration process.

3 DISTILLING QUESTIONS

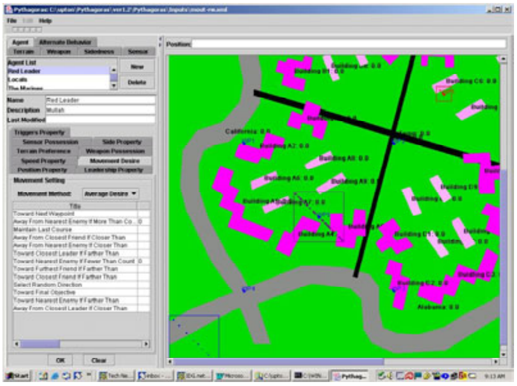
> *"Everything should be made as*
> *simple as possible, but not simpler."*
>
> *"Any intelligent fool can make things*
> *bigger, more complex… It takes a touch of genius*
> *and a lot of courage to move in the*
> *opposite direction."*
> ~ Albert Einstein

Models used within the paradigm of data farming are referred to as "distillations." It is recognized that all models are "distillations" or abstractions of the real world. It is only by judicious selection of specific aspects of a system that we can produce models that are helpful. The Einstein quotes above capture the intent of modeling within the realm of Data Farming. Distillations should be complex enough to address the question… and no more complex.

Ideally, distillations have the following characteristics:

- Intuitive–the team must be able to understand the parameters and rules that define the model and how they relate to the system being modeled;
- Transparent–the team must be able to understand how the behaviors that emerge in the model emerged from a set of parameters and rules; and
- Transportable–the model must be portable to a Data Farming environment.

Although any model could be data farmed, distillations are intended to be a bottom up reduction to the essence of a question. Typically, distillations are expected to be developed quickly–potentially in a matter of a few days, hours, or even minutes. Realistically, though, model development environments have not reached the ease of use required to produce models in minutes. As a result of these requirements, the current focus has been on the implementation of agent-based modeling environments such as Pythagoras, represented in Figure 2.



Name Red Leader
Description Mullah
Last Modified

Movement Setting
Movement Method: Average Desire

Title	
Toward Next Waypoint	
Away From Nearest Enemy If More Than Co...	0
Maintain Last Course	
Away From Closest Friend If Closer Than	
Away From Nearest Enemy If Closer Than	
Toward Closest Leader If Farther Than	
Toward Nearest Enemy If Fewer Than Count	0
Toward Furthest Friend If Farther Than	
Toward Closest Friend If Farther Than	
Select Random Direction	
Toward Final Objective	
Toward Nearest Enemy If Farther Than	
Away From Closest Leader If Closer Than	

Figure 2: Pythagoras Distillation Environment

Distillations use abstraction judiciously. "Weapons" can represent interchanges of various types such as food, resources, messages, and positive or negative messages or propaganda. Location or proximity in the model can be abstracted to represent relative aspects of other relational parameters. Modeled obstacles can represent walls, floors, borders, or sociological or psychological obstructions in non-geoterrain or combat interchanges. Another Einstein quote, "Imagination is more Important than Knowledge," is an important guideline in distillation development. Distillation modelers must often innovate and use imagination to define abstractions. For example, communication level has been used as a "proxy" for trust–do you use or ignore information provided?

4 WHY DATA FARMING?

The availability of Data Farming resources can be advantageous to any decision support process that is aided by modeling. Data Farming was developed to provide methods to address several phenomena that are not easily addressed using traditional methods of modeling:

- Non-linearity–including sensitive dependence on initial conditions and bifurcation events;
- Intangibles–"Fuzzy" parameters such as leadership, morale, and trust; and
- Adaptation–including opponent reaction and co-evolution.

Data Farming is a process that can address questions quickly. Distillation models are developed quickly and HPC allows results to be produced and collected quickly.

Data Farming allows the examination of whole landscapes of potential outcomes, not just a few cases. It provides the capability of executing enough experiments so that outliers might be captured and examined for insights. Data Farming in not intended to predict an outcome, it is used to aid intuition and to gain insight.

Recapping, Data Farming is a methodology and set of tools that provides modelers the ability to execute models or simulations hundreds of thousands or even millions of times. This capability can be used to support modeling and decision support in a number of ways:

- Sensitivity Studies – Models of any complexity are subject to chaotic or non-linear behaviors that may vary over the space of possible inputs to the model. Data Farming provides the ability to examine much larger and higher resolution areas of the parameter space to examine the statistical variability of the model.
- Validation and Verification – Data Farming allows modelers to fully test a model's reaction to various inputs over a broad space of possible and potentially unforeseen combinations of input parameters. Results may be examined to ascertain the correct execution of the model algorithms and to compare the results to the real world.
- Model Development and Gaming – "All models are simplifications of the real world." In order to hone the model and its parameters to better represent the real world, models are often repeatedly executed to "steer" parameters. Furthermore, the process of developing models requires innumerable executions of the model to aid in debugging and algorithm development. The ability to run models over a larger parameter space speeds the model development process. This process is greatly enhanced by Data Farming.
- Scenario Analysis – Once models are developed, they are executed. The results of the execution are studied to provide insight or to address real world questions. Data Farming allows the model to be executed over a much larger number of input parameters and a larger number of random variations, which can give decision-makers a more complete view of the possible outcomes and system dynamics.
- Trends and Outliers – Traditionally models are run a few times to do scenario analysis of a small window of the possible outcomes. A few summary statistics are generated to represent the results. If one examines a wider parameter space, however, trends and relationships between inputs and measurements of effectiveness can be studied. Of equal importance is the ability to identify which parameter combinations or random variations result in "outliers," special cases that may indicate model problems, or high risk or high opportunity domains of the parameter space.
- Heuristic Search and Discovery – Data Farming encompasses the ability to apply iterative methodologies for model analysis such as genetic algorithms and other sophisticated optimization and search methodologies.
- Generation of Massive Test Data Sets – Data Farming can be used in conjunction with models to generate massive data sets to test learning algorithms and other data mining tools. This is particularly valuable where actual data may not be available for security or privacy concerns.

5 A TOOLKIT FOR DATA FARMING

Since 1998, a suite of tools has been developed to implement Data Farming environments and distillation models that can be executed in these environments. In general these tools are expected to be openly available to the collaborative community that is involved in Data Farming development.

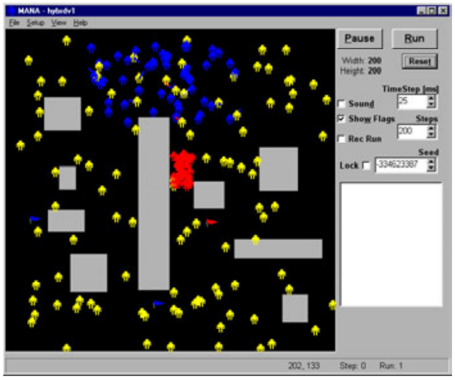
These tools fall into three categories: 1) implementations of Data Farming environments, 2) distillation modeling environments, and 3) data exploration tools.

Two distillation modeling environments have been developed: the Maui High Performance Computing Parallel Execution System (PES) and OldMcData.

The PES is accessed through a web based interface that allow the uploading of basecases to several supercomputing multi-node systems. The web interface also allows users to define the excursion space to be examined, the number of replicates, and the types of output to be produced and collected. The system includes software that distributes distillations to nodes, executes them, and collects output in a central repository. This system is currently being maintained and is undergoing a major developmental upgrade..

OldMcData (Upton 2004) is a smaller scale system that can be used to execute model excursions on a standalone computer or on a distributed set of nodes on a network. In combination with an application called Xstudy. It allows users to set up a Data Farming environment on any networked set of nodes.

Six (agent-based) distillation modeling environments have been integrated into OldMcData and the PES. These include ISAAC, Socrates, Pythagoras, Mana, PAX, and NetLogo. Of these, three are currently under development: Pythagoras by Northrup Grumman, Mana (displayed in Figure 3) by the New Zealand Defence Technology Agency, and the PAX Peace Support Operation Model by EADS, Germany.



Width: 200
Height: 200
TimeStep [ms] 25
☐ Sound
☑ Show Flags
Steps 200
☐ Rec Run
Seed
Lock ☐ -334623387
202, 133 Step: 0 Run: 1

Figure 3: A Sample Scenario in the Mana Model

Of the six integrated models three have source code available: ISAAC, Pythagoras, and Socrates. Each of these models have rich features sets.

Of the integrated models, netLogo is the most open ended, being a complete programming environment. NetLogo is an open source modeling system available online.

It should be noted, though, that *any* model that adheres to a fairly simple set of specifications can be integrated into the Data Farming environments. Models of various types could potentially be integrated: logistics, deterministic, network, game theory, predictive agent, and others.

To be integrated into a Data Farming environment models must minimally be able to define a basecase using an XML text file and adhere to a simple text-based delimited record/field output format.

Three visualization tools have been developed to be used in the Data Farming process. These tools include: the Playback Tool, the VizTool Landscape Plotter, and Avatar, all developed at MHPCC. The current visualization development is aimed at data exploration, not data presentation. The tools are aimed at fairly sophisticated analysts (Meyer and Johnson 2001). Development is intended to support the special needs of Data Farming community and the high dimensional data that is produced. The long term goal is to provide interfaces and tools to directly support decision makers.

The first visualization tool, the Playback tool, allows users to take time series output from a model and watch the model with VCR type stepping, rewind and fast forward controls. The utility of the software is currently limited to the ISAAC model, but it allows the playback of multiple excursions and replicates at the same time, giving users a powerful comparative view of multiple excursions or replicates at the same time.

The second visualization tool, the VizTool Landscape Plotter, is displayed in Figure 4. This is a powerful tool that allows users to extract 2D slices out of high dimension full factorial gridded Data Farming output to easily display the relationship of multiple parameters to the output measurements. Figure 4 depicts the maximum, mean, and minimum of replicates for a slice of data extracted from a 5 dimensional data set of 1000's of records.

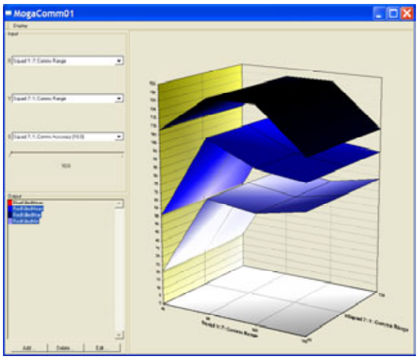


Figure 4: The MHPPC VizTool Landscape Plotter

Two visualization concepts in particular support the exploration of high dimension data: *focus* and *linking* (Buja et al. 1991). Focus refers to the ability to manipulate the view or perspective of a visualization interactively. Zooming, rotating, and subsetting/sampling to examine relevant data at varying resolution are examples of focus. Linking refers to being able to examine multiple perspectives/visualizations at the same time to discover relationships among parameters. Selecting or coloring data in one view results in linked selection or coloring in other views.

Avatar, the latest visualization tool, is designed to begin implementing focusing and linking for the Data Farming environment. Figure 5 shows two views of the same data in using the Avatar visualization modules. Avatar also provides for subsetting of data sets, can support non-gridded data, and is currently under ongoing development. It is intend to be integrated with model playback so that uers, when they have discovered interesting results can examine the behaviors that caused those results

Figure 5 represents two linked views of data: a 3D scatter plot and a parallel coordinate plot. Avatar allows the selection of any three parameters or output to be displayed in the 3D scatter. It provides for rotation, zooming, selection and subsetting as well.

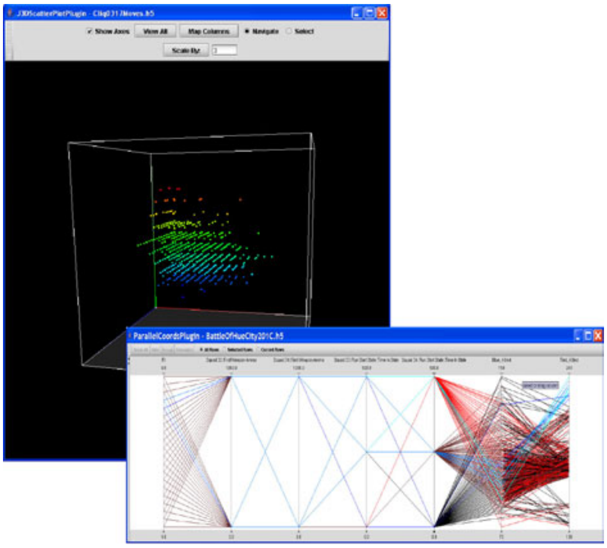


Figure 5: Avatar 3D Scatter and Parallel Coordinate Plot

The parallel coordinate plot is a powerful visualization that allows the examination of a large number of dimensions at one time. This plotter also provides for data subsetting, selection, and manipulation of scale and focus. Avatar has a plug-in architecture for adding visualization modules and can be easily extended. Modules for 2D scatter, 2D scatter with jitter, bubble plots, box plots, table view and statistical summary have also been developed.

6 OPEN SOURCE, OPEN INVITATION

This paper has two purposes. The first is to introduce the reader to the concept of Data Farming, its utility and value, and the toolkit that has been developed to support it. The second purpose of the paper is to provide an open invitation to collaborators. By November 2004 we expect that much of the software developed to support the Data Farming process will be hosted on SourceForge.net. Those familiar with Open Source projects will be aware that this means that we hope that other modelers can benefit from this software and that we can benefit by the expansion of our user and developer community. Please contact the authors with any questions regarding collaboration or use of these tools.

REFERENCES

Brandstein, A. and Horne, G. 1998. Data Farming: A Meta-Technique for Research in the 21st Century. *Maneuver Warfare Science 1998*, Marine Corps Combat Development Command Publication, Quantico, Virginia.

Buja, A., McDonald, J. A., Michalak, J., and Stuetzle, W. 1991. Interactive Data Visualization Using Focusing and Linking. *Proceedings of the 2nd conference on Visualization '91*, San Diego, California, SESSION: Multivariate visualization, Pages: 156 – 163. IEEE Computer Society Press. Los Alamitos, California.

Horne, G. 2001. Beyond Point Estimates: Operational Synthesis and Data Farming. *Maneuver Warfare Science 2001*. United States Marine Corps Project Albert. Quantico, Virginia.

Lucas, T., Sanchez, S., Brown, L., and Vinyard, W. 2002. Better Designs for High-Dimensional Explorations of Distillations. *Maneuver Warfare Science 2002*. United States Marine Corps Project Albert. Quantico, Virginia.

Meyer, T. and Johnson, S. 2001. Visualization for Data Farming: A Survey of Methods. *Maneuver Warfare Science 2001*. United States Marine Corps Project Albert. Quantico, Virginia.

Upton, Stephen. 2004. Users Guide: OldMcData, the Data Farmer, Version 1.0. United States Marine Corps Project Albert. Quantico, Virginia..

AUTHOR BIOGRAPHIES

GARY E. HORNE is the Executive Director of Project Albert. He currently works for the Marine Corps Warfighting Laboratory as an employee of the MITRE Corporation. His main research interest is the development and application of Data Farming which he invented to address shortcomings in our abilities to answer important questions. He served as a scientific analyst, project director, and field representative for the Center for Naval Analyses. Dr. Horne received his DSc in Operations Research from George Washington University. His email address is `<HorneGE.CTR@mcwl.quantico.usmc.mil>`.

TED MEYER is currently the Information Architect for Project Albert. He is a doctoral candidate in George Mason University's Computational Science and Informatics program. Previously, he was the CTO and product designer for Fortner Software LLC, where he managed the development of the company's visualization tools. From 1990 to

1996, Mr. Meyer worked as the Information Architect for NASA's Earth Science Data and Information System Project. Before NASA, he was a Geodesist and Physical Scientist at the Defense Mapping Agency. Mr. Meyer co-authored *Number by Colors: A Guide to Using Color to Understand Technical Data.* His email address is `<tedmeyer@mitre.org>`.