Data Farming and the Exploration of Inter-agency, Inter-disciplinary, and International “What If?” Questions

Steve Anderson  
Naval Surface Warfare Center  
steven.e.anderson@navy.mil

Gary Horne  
Naval Postgraduate School  
gehome@nps.edu

Ted Meyer  
Naval Postgraduate School  
temeyer@nps.edu

Larry Triola  
Naval Surface Warfare Center  
larryтриola@navy.mil

Data farming uses simulation modeling, high performance computing, and analysis to examine questions of interest with large possibility spaces. This methodology allows for the examination of whole landscapes of potential outcomes and provides the capability of executing enough experiments so that outliers might be captured and examined for insights. This capability may be quite informative when used to examine the plethora of “What If?” questions that result when examining potential scenarios that our forces may face in the uncertain world of the future. Many of these scenarios most certainly will be challenging and solutions may depend on interagency and international collaboration as well as the need for inter-disciplinary scientific inquiry preceding these events. In this paper we describe data farming and illustrate it in the context of application to questions inherent in military decision-making as we consider alternate future scenarios.

1.0 INTRODUCTION

What if energy sources became more scarce and suitable replacements are not cost effective? (Call this possibility: Out of Gas.)

What if renewable energy such as solar or wind became reasonably practical and widespread? (Good and Green.)

What if nuclear fusion became available to supply energy efficiently, safely, and at a fraction of the cost of current sources? (Safe and Cheap.)

Certainly these three possibilities are not the only energy futures that our world faces. And each of the three in and of itself poses challenges and opportunities for our military forces. But then also consider another very large global question we name Coastal Flooding: What if climate change factors develop and result in an increase in the number and severity of calamities such as floods? Or what if it didn’t? (No Floods).

Combining these two sets of what-if questions results in 3 times 2 = 6 possibilities already. One of which would be Out of Gas and Coastal Flooding, certainly a stark challenge. But even Safe and Cheap and No Floods might result in other challenges such as instability in former oil producing regions.

And with these six possibilities we have only begun to scratch the surface of the plethora of what-if questions that result when examining potential scenarios that our forces may face in the uncertain world of the future. Many of these scenarios most certainly will be challenging and solutions may depend on interagency and international collaboration as well as the
need for inter-disciplinary scientific inquiry preceding these events.

In this paper we describe a way forward to examine future possibilities using the methods of data farming. Data farming uses simulation modeling, high performance computing, and analysis to examine questions of interest with large possibility spaces. This methodology allows for the examination of whole landscapes of potential outcomes and provides the capability of executing enough experiments so that outliers might be captured and examined for insights.

1.1 What If?
The title of figure 1 provides the overarching philosophy of the What If? Network we are developing. All disciplines have strengths, but for the large challenges and broad scope of the questions we would like to grapple with, we need a truly multi-disciplinary approach.

A Multi-disciplinary/multi-agency/multi-national What If? Network of Questions

Finally, the challenges we face are global in nature and our collaborators in places such as Sweden and Finland have shown great interest in the kinds of challenges that we as a world community share as well as great acumen in modeling, simulation, and data farming.

In figure 1 we have outlined three sets of what-ifs? The first block represents the big issues and our introduction gave you a brief look into the kinds of questions we have been establishing as a starting point. And we will present our outline of a number of them in section 3.

The second block represents the detailed questions that will allow us to integrate possibilities on a level where insights and solutions might become more clear. For example, consider a more specific and technology driven question: What if a new military craft was developed that would be a significant technological advancement over previous ship-to-shore transport capabilities? That development may have impact in the Coastal Flooding what-if mentioned earlier. And what if this craft had a power supply stemming from the Safe and Cheap what-if?

Finally, the third block represents the data farming of the questions and we will give a very brief overview of data farming in the next section, although a deeper discussion can be found in our MODSIM World 2010 paper “Data Farming and Defense Applications.”

1.2 Data Farming Overview
Data farming is a collaborative and iterative process that requires input and participation by inter-disciplinary teams to be most effective. It allows for an examination of a more complete landscape of outputs rather than one particular answer. Data farming also allows for the discovery of outliers that may be even more instructive than any general patterns that are discovered.

Data Farming has been described as including six domains. The six domains

And although we (the authors) work within the United States Department of Defense, these large challenges require a broadening of possible solution spaces that comes with looking across agencies. In our initial efforts we have focused on Department of Energy agencies. But certainly our network must grow beyond these two cylinders of excellence and indeed find ways of connecting the ideas found across agencies.
make up the focus of the six sub-groups in a NATO Modeling and Simulation Group effort (NMSG-088) called “Data Farming Support to NATO” that is now in its second year of existence. MSG-088 members are performing two case studies, one in the area of humanitarian assistance / disaster relief and the other in the area of force protection. The MSG-088 subgroups are in the process of defining the six domains in detail as they apply to data farming, and the domains were described in our previous MODSIM Paper that we mentioned, but they are listed here for reference.

- Model Development
- High Performance Computing
- Rapid Prototyping of Scenarios
- Analysis and Visualization of simulation output
- Design of Experiments
- Collaborative processes

We believe all of the domains of data farming listed above will be important in the examination of the what-if questions we are considering. However, the domain of analysis and visualization promises to be a key domain and in the next section we explain why and give some detailed information regarding our proposed approaches in this domain.

2.0 WHAT IF? ANALYSIS AND VISUALIZATION

In general, analysis, and visualization in particular, have two broad purposes: 1) answering questions and 2) determining what questions to ask. Answering specific questions is accomplished by a traditional and ever-growing suite of statistical and graphical techniques, both well-known (e.g., averages, line plots) and more exotic (e.g., Yuen’s modified t-tests, trilinear graphs).

Often times, though, an analyst may be involved in a what-if process that is open-ended... the data being examined is complex, voluminous and not well understood. Before well-defined questions can be asked, “exploratory” analysis of the data may be undertaken to establish an understanding of the data “landscape.”

The class of analyses that are to be undertaken by the proposed What-If? Network requires that collaborative multi-disciplinary teams undertake exploratory-type analytic processes. These multidisciplinary teams of modelers, decision-makers, analysts, subject-matter experts are what will form the What-If? Network.

These teams are not looking to simply provide basic summary statistics and trends from models. The intent is for these teams: to exercise a variety of scenario development models (e.g., subject-matter judgment, agent-based simulation, scenario network mapping, role playing, etc.); to data farm these resultant models; to integrate the models’ results into a coherent landscape of potential outcomes; and explore the landscape to gain an understanding of alternative futures. Individual modeling processes, undertaken by teams with limited disciplinary and collaborative breadth, are unlikely to provide results with verisimilitude that covers the breadth of real-world potentialities.

Real-world scenarios encompass outliers, second and third order effects and unintended consequences, runaway trends from feedback loops and dampening effects, emergent behaviors, social and human response, and nonlinearities and chaos. To generate potential real-world alternative futures multiple modeling techniques are required to move past the limitations of single classes of models.

Data farming a single simple model often results in analytic challenges in examining potential voluminous results. Exploratory visualization techniques offer capabilities in exploring landscapes of potential results. Integrating and exploring the results of multiple models becomes a greater
Perhaps just as important as exploring the results of modeling efforts is the exploration of HOW and WHY the results are attained. Some modeling methods allow analysts to examine the state of the system over time, the evolution of the system, the interaction and behaviors of its components, and the changes to it inherent networks. This need to examine the “how” and “why” significantly increases the data to be examined beyond the data volume impact of data farming.

2.1 Visualization Methods and Examples

Visualization can play two roles in accessing this potential mountain of data that faces what-if analysts: aiding in the discovery of insights in the integrated whole of what-if scenario studies and in the compelling presentation of these results to policy makers. Our current focus is on exploration of the alternative future space, as finding the important insights is a more difficult problem.

Three exploration tools are powerful techniques to optimize the examination of high-dimension, high-volume data sets:

1) Linked Displays

Typically information graphics provide a view of low dimensional data in easily understood representation. For example, histograms present univariate data, simple scatter plots represent bivariate data, and color and size can be added to a scatter plot to provide a third and fourth dimension. Linked displays tie two or more representations of the same data across multiple graphic displays. Each display can add to the overall dimensionality being presented at a single time. In the case of simulation, a playback of data can be tied to a representation of network state or other performance metrics.

2) Variable Focus

A data display can be adjusted in various ways to change the perspective being used to represent the data. In the simplest form the position and scale can be adjusted to zoom into detailed features or to zoom out to an overview. More complex forms of focus can include three dimensional rotation, geographic projection, and axis and parameter selection.

3) Interactivity

User interactivity can be used to adjust linking and focus in order to explore the relationships of data, select which portions of the data to examine, or determine how some parameters impact metrics.

These three techniques in combination can be integrated into the what-if process to interactively iterate model results and model development to hone insight into potential future outcomes.

As an illustrative example, a changing “focus” can drill down into more detail and a better understanding of a scenario. In this example a time-stepped simulation of a combat scenario is data farmed over 50 replicates.

1) Single Numeric Statistical Summary Value in Text Form: A model scenario results in 45 Blue casualties, on average, per run.

2) Numeric Statistical Summary Values in Text Form: the model indicates 45 Blue casualties, on average, per run; 65 Maximum Blue Casualties over all runs; 11 Minimum Blue Casualties over all runs.

3) Time Series of Numeric Statistical Summary Value in Line Plot Form: Figure 2a shows the average number of casualties each hour in 50 executions of the scenario.
4) **Time Series of Numeric Statistical Summary Values in Line Plot Form:** Figure 2b shows both the average as well as the maximum and minimum number of casualties for each hour in the 50 replicates.

5) **Jitter Plot of ALL Values Providing Distribution Details:** Figure 2c shows the average, minimum and maximum, but it also shows a full distribution of all casualty values for each hour for all 50 replicates.

Figures 2a to c demonstrate how more detail can reveal important information. A close examination of the distribution of casualties in Figure 2c reveals a bimodal distribution that occurs at about 7.5 hours into the scenario. An examination of this time frame shows a bifurcation event in the scenario that results in a split in the results.

Often, analysis techniques will *reduce* the amount of available information to the analyst or decision maker. Visualization provides methods for increasing the available information.

Another example of using visualization techniques to increase the display of information has been experimented with for sometime. Density Playback is a technique for examining multiple replicates or scenarios in an overlaid fashion to highlight similarities or differences in the model.

In density playback, data of interest is plotted in whatever state space is desired using scaled transparency. The amount of transparency is dependent on the number of model executions. An obvious example of this technique is represented in the Death Star™ scenario represented in Figure 3a and b where the spatial position of agents are represented using density playback.

In this scenario a central target is extremely well guarded by a ring of well-armed blue agents in the top right of the figures. Blue does not move and is well positioned. Fifty unarmed Red agents only need to penetrate the ring to win the scenario. 26000 runs of this scenario resulted in less than 100 Red wins.
Figure 3a is a random selection of 50 executions of the scenario from the 26000 executed displayed at timestep 350. Transparency is used to display the 50 red agents' trails. Dark trails represent locations where many agents traveled over multiple executions. Note that no agents penetrate the ring.

Figure 3b represents 50 executions selected from the small set of Red wins. Note that only two very specific paths led to Red winning. This example shows that changing focus to filter only on data associated with an outlier result results in a highlighted display of the path Red needs to take to victory.

The potential scenarios that are to be addressed by the What-If? Network are far more complex than the simplified examples shown. Additionally, the “terrain” used by these potential scenarios are likely to be abstractions of a political/social/economic landscape rather than the simple spatial examples shown. The techniques shown can be applied to the more complex cases and can be further explored using linked displays, interfaces that provide interactivity, and interfaces that link results to specific scenarios and replicates.

The techniques to be used by the What-If? Network will range from simple descriptive statistics to sophisticated machine learning algorithms. All analytic techniques have strengths and weaknesses, as do all modeling and techniques. And, of course, the combination of methods to best understand the potential outcomes will depend on the questions at hand. Thus, at this point we will now turn to describing some of the question areas to be considered by our growing What-If? Network.

3.0 THE QUEST FOR WIN-WIN SOLUTIONS

“The trouble with military force structure is that it typically outlives the geopolitical context that called it into existence.”

Whether the U.S. government faces future fiscal constraints, or makes changes to its overseas commitments, the U.S. Department of Defense may choose to reevaluate its force structure. Deer hunters can attest that the most effective hunters will aim at where the deer will be, rather than where the deer was. Leading the target, is as important to DoD force planning as it is to hunting deer. Getting future geopolitical contexts right, more or less, is an important part of any future debate.

The study of future geopolitical contexts is an obscure and delicate form of defense analysis that requires evaluating interdisciplinary trends and data, tracking science and technology investment areas, monitoring acquisition programs, conceptualizing current and emerging operational needs and missions, and translating all of this into robust recommendations for future concepts of operation, development portfolios and force structure.

Whereas most military planning scenarios are focused upon political-military actions and reactions of nation-states, the second half of the Post Cold War Era is replete with evidence and examples of trends, actions, and events that have shaped the international stage, and yet are neither

---

1 This observation is attributed to retired defense analyst, Mr. James S. O’Brasky.
initiated by, nor responded to by nation-states. A set of highly informative future contexts arise from examining certain global mega trends and wild cards such as climate change, wild card disasters, wild card revolutionary technology developments, global resource limitations and other trends. National security relevant illumination comes by relating the geopolitics and U.S. mechanisms of planning to address such events.

As civilization enters the second decade of the 21st century it is confronted with an abundance of diverse pressures the sum of which seems to outweigh each as individual problems. The United States and the U.S. Department of Defense must develop methodologies to enhance the positive synergies and mitigate the negative synergies of this collision of crises. Terrorism, tribalism, fossil fuel shortages, water resource shortages, challenges in health care and education, new technologies both helping and threatening, economic challenges, globalism and the flat-world syndrome, disruption of old alliances and the formation of new ones, climate change, mass migrations, resource limitations on numerous fronts and failures of governance, policies and leadership and many other issues wash wave upon wave in cascading calamities of white water of our changing times. Halal and Marien have called this situation the “Global MegaCrisis”.

While the authors this paper work for the Department of the Navy within the U.S. Department of Defense, the issues we uncover consistently require coherent and integrated whole-of-government response plans, many of which are time-critical in nature.

Looking at each issue individually and trying to develop individual solutions almost ensures that some solutions will interfere with others or that solutions will simply be too many and too costly to implement. What is here proposed is to construct a framework for the various crises such that they may be addressed in a coordinated, cross-domain, cross-disciplinary approach which leverages the emerging capabilities of data farming.

3.1 The World MegaCrisis Framework

The World MegaCrisis Framework (see the appendix) contains a set of ten categories of challenges with a list of more detailed topics and sub-topics listed under each.

1. Global Physical and Biological Dynamics such as climate change, weather and geological disasters, biodiversity loss and biological change in habitat

2. The limitations of natural and perhaps man-made resources and the limitations on or help to civilization which emerge from such resources and their limitations

3. Changing of demographics and populations such as the radical drop in Russian population and native European population and the rise of population in Europe from the Islamic countries stretching from Tunisia to Indonesia

4. How societies within nations and regions view themselves and choose or not to act as coherent groups to include the range of effects from socialized democracy to genocide

5. Economics and manufacturing issues include currency, national and international debt, trade imbalances, ability or lack thereof to manufacture goods

6. Knowledge includes educating the young to function in society but also includes the context of information or misinformation which drives public opinion and decisions

7. Infrastructure covers the range from electrical power grids, to water and sewage, to health care facility, to
8. Transportation is also suggested as its own topic to cover methods and the societal, logistical, and transnational implications.

9. Technological breakthroughs can destroy cities or provide them endless electricity and much more depending on what they are and how they are used and many potential world-changing wild cards exist in this category.

10. Leadership and governance at the human level and the institutional and legal levels can either be facilitation to or inhibitors of solving crises.

This is not an exhaustive set, yet all of which have significant world wide impacts. However, note that some challenges are supersets of others and that challenges listed under one heading often offer the potential to exacerbate or mitigate challenges listed under another heading. These interdependencies are a critical feature of this framework. As an example thread, consider that Climate Change shown under the heading of Global-Physical/Biological Dynamics leads to such things as Water Cycle Changes and Biodiversity Loss and impacts the availability of Resources such as Arable Farm Land. Fossil Fuel use shown under Resources Limitations is seen by many including the International Panel on Climate Change to directly impact Climate Change and provides a feedback loop for Resource loss. Under Technological and Scientific Breakthroughs alternative energy options and new material options create Resources and perhaps present new Resource limitation challenges.

The topics listed in the MegaCrisis Framework, each have self-organized communities of interest. Community cohesion can vary from topic to topic. In some instances, a given community may have significant shared outlooks and beliefs, while other communities are highly polarized on central points. To remain objective, data farming must focus on “possibilities” rather than probabilities and predictions.

Many threads can be drawn from populating a data base rich in diverse contexts such as suggested by the MegaCrisis Framework. These can be perhaps too complex and interwoven for simple brute force human interpretation. Thus, automated methods to mine this data and to develop additional information from the cross-connecting of the threads are essential. Looking through only one lens will not provide realistic nor optimum approaches for addressing the many challenges. Connecting the dots may in fact provide a clear synergistic set of solutions which may be balanced and implementable. It is hoped that resources will be made available to farm this data and connect the dots as well as to identify and flesh out other contexts which can help governments to successfully confront this tumult of challenges, and thrive.

Single-issue agendas are often fraught with unintended consequences. The so-called “Biofuel Controversy” is an example of ethanol producers who seek to increase the availability of ethanol to consumers at a reasonable cost. While this is an admirable single-issue agenda, there are scores of documented examples around the world where arable land that is used to generate crops for biofuels is not available to produce food – creating a situation where biofuel production causes famine. Biofuel production that leads to the starvation of impoverished peoples is a “win-lose” situation, at best.

If you are a senior decision maker, it almost does not matter where you sit, in any government in the world. You will be faced with the growing demand to make highly-informed cross-disciplinary cross-agency decisions that invariably reach outside of any given historic stovepipe. Ideologically,
the goal is to make coherent and integrated decisions that foster far reaching benefits, and minimize adverse consequences. The quest is to achieve “win-win” situations wherever possible. To do this will require new patterns of thinking, new questions, new methodologies and techniques, more data, and new and better tools.

4.0 SUMMARY
Pragmatically, how do we achieve these ends? There are some mandates, all of which can benefit from data farming practices, techniques and approaches.

- You must get to the point where you are asking the right questions.
- The path to the right questions will be iterative.
- We must resist the temptation to delete, discount, or discard “outliers” data and trends.
- Audacity, integrity, humility.
- While top-down approaches would be very useful, you will start with the data and tools you already have in hand. As top-down, middle-out, bottom-up strategies interact and normalize, you will need new data and new tools to effectively answer emergent questions.
- There is a sense of urgency. Thoroughly exhaustive approaches may not be practical or ultimately useful to solve near-term challenges.

But, we are still just beginning to ask the detailed what-if questions in the second block of figure 1 and only setting ourselves up for the modeling, simulation, and data farming efforts at this point. And that is why we are coming to MODSIM World now with this work: because inter-agency, inter-disciplinary, and international expertise will be needed as we move to the third block in figure 1. Thus we would like to invite you, members of the modeling & simulation community, to join our What-If? Network and contribute to the quest for win-win solutions.

5.0 ACKNOWLEDGMENT
The authors would like to thank the members of the growing inter-agency, inter-disciplinary, and international network of people who contributed their ideas in the quest for discovering win-win solutions through data farming.

6.0 REFERENCES
Appendix

<table>
<thead>
<tr>
<th>The World Mega Crisis—Collision of Crises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global-Physical/ Biological Dynamics</td>
</tr>
<tr>
<td>Climate Change</td>
</tr>
<tr>
<td>Solar Activity/Electric Grid Failure</td>
</tr>
<tr>
<td>Volcanic Eruptions</td>
</tr>
<tr>
<td>Tsunamis</td>
</tr>
<tr>
<td>Volcanic Weather</td>
</tr>
<tr>
<td>Biodiversity Loss</td>
</tr>
<tr>
<td>Disease and Pandemics</td>
</tr>
<tr>
<td>Crop Yields</td>
</tr>
<tr>
<td>Crop Yields</td>
</tr>
</tbody>
</table>


The Institute for Food and Development Policy, 398 - 60th Street, Oakland, CA 94618. (www.foodfirst.org)