Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

DATA FARMING: BETTER DATA, NOT JUST BIG DATA

Susan M. Sanchez

Naval Postgraduate School
Operations Research Department
1411 Cunningham Rd.
Monterey, CA 93943-5219, USA

ABSTRACT

Data mining tools have been around for several decades, but the term "big data" has only recently captured widespread attention. Numerous success stories have been promulgated as organizations have sifted through massive volumes of data to find interesting patterns that are, in turn, transformed into actionable information. Yet a key drawback to the big data paradigm is that it relies on observational data, limiting the types of insights that can be gained. The simulation world is different. A "data farming" metaphor captures the notion of purposeful data generation from simulation models. Large-scale experiments let us grow the simulation output efficiently and effectively. We can use modern statistical and visual analytic methods to explore massive input spaces, uncover interesting features of complex simulation response surfaces, and explicitly identify cause-and-effect relationships. With this new mindset, we can achieve tremendous leaps in the breadth, depth, and timeliness of the insights yielded by simulation.

1 INTRODUCTION

What can be done with big data? Most people immediately think of data mining, a term that is ubiquitous in the literature (see, e.g., James et al. 2013 or Witten et al. 2017 for overviews). The concept of data farming is less well known, but has been used in the defense community over the past two decades (Brandstein and Horne 1998). The building blocks of data farming are a collaborative approach to rapid scenario prototyping, modeling platform development, design of experiments, high performance computing, and the analysis and visualization of the output – all with the intent of providing decision makers with timely insights about their problems of interest (NATO 2014). Rather than beginning with technical details, we find it useful to compare and contrast data mining and data farming with the following metaphors.

- *Data mining:* Miners seek valuable nuggets of ore buried in the earth, but have no control over what is out there or how hard it is to extract the nuggets from their surroundings. As they take samples from the earth, they gather more information about the underlying geology. Similarly, data miners seek to uncover valuable nuggets of information buried within massive amounts of data. They may look at "all" the "big data" available at a particular point in time, but they typically have no control over what data are out there, nor how hard it will be to separate the useful information from the rest of the data. Data mining techniques use statistical and graphical measures to try to identify interesting correlations or clusters in the data set.
- *Data farming:* Farmers cultivate the land to maximize their yield. They manipulate the environment to their advantage, by using irrigation, pest control, crop rotation, fertilizer, and more. Small-scale designed experiments can help them to determine whether these treatments are effective. Similarly, data farmers manipulate simulation models to their advantage – but using large-scale designed experimentation. This allows them to learn more about the simulation model's behavior in a structured way. In this fashion, they "grow" data from their models, but in a manner that facilitates

identifying the useful information. For large-scale simulation experiments, this often results in data sets that, while big, are far smaller than what would be needed to gain insights if the results were observational (i.e., obtained using *ad hoc* or randomly generated combinations of factor settings). The data sets are also better, in the sense that they let us identify root cause-and-effect relationships between the simulation model input factors and the simulation output.

In this paper, we posit that data farming provides a natural "better big data" mindset that can help to improve simulation studies. At the same time, it can make it easier for simulation professionals to stake out the area of inferential big data, and articulate simulation's benefits to new communities. On one hand, those who have conducted simulation experiments in the past may feel there is nothing really new in many of the topics in this paper. However, similar skepticism arose about the terms "data mining" and "big data." Some argue that we have always had big data, whenever we had more available than we could analyze in a timely fashion with existing tools. From that perspective, big data can be viewed as any data set that pushes against the limits of currently available technology. Running even a single multiple regression in the 1940's was very difficult, since matrix multiplication and inversion for even a small problem (say, a handful of independent variables and a few hundred observations) is a non-trivial task when it must be done by hand. As recently as 2014, many were arguing that big data was simply hype (Horgan 2014; Dodson 2014). Today, big data is largely taken for granted, as many sectors of the economy (marketing, social media, medicine, and more) are no longer debating *whether* to invest in big data and data analytics, but either made or are preparing for this leap. Our view that just as the "big" in big data changed both the perception and practice of data analysis, so the "big" in simulation experiment data has the potential to change the perception and practice of simulation.

In Section 2, we provide brief characterizations of "bigness" and relate them to data mining and data farming. In Section 3 we revisit the concepts of causation and correlation to clarify how a data farming approach – that is, an approach that explicitly allows for investigating and identifying cause-and-effect relationships within the context of the simulation model – offers benefits to decision makers beyond those that could be obtained without such a structure. In Section 4 we provide guidance about setting up a data farming environment, along with some graphs and brief examples to spark your interest in taking a "think big" approach to your next simulation study. We close with some thoughts about the future..

2 CHARACTERIZING BIGNESS

In this section, we discuss what characterizes bigness, whether we gather or grow our data. We assert that just as a different mindset can be beneficial for dealing with big (vs. small) data in observational settings, so a different mindset can be beneficial for generating and dealing with big data in simulation settings.

2.1 The 3 (or more) V's of Big Data

Big data is most often characterized by the 3 V's – volume, velocity, and variety – mentioned in the title of the article by Laney (2001). All three V's have increased at an astonishing rate during the last decade. Volume refers to the amount of data, with much being generated, captured from, or stored on the internet. Velocity refers to the speed at which data arrives. Variety refers to the nature of the data itself – it need not be highly structured data in tables or databases, but can involve unstructured data in many formats.

For a brief but informative history of big data up through 2013, we refer the reader to Press (2013). Numerous reports from 2013 to 2018 (see, e.g., SINTEF 2013; Marr 2018) all claim that the volume increases by 2.5 quintillion bytes per day, and that 90% of the world's data has been generated in the past two years. Clearly, these reported numbers cannot all be correct – if the "90% within two years" was valid in both 2013 and 2018, the daily rate must have increased by a few orders of magnitude! The variety is also astounding, ranging from social media, business communications, and digital images – in both structured and unstructured formats – and increasingly, the explosive growth of the Internet of Things

(IoT) composed of diverse "smart" devices that are regularly (and in some cases, continually) collecting data and communicating with each other or with us (Burgess 2018).

Some have advocated a look at more V's (including veracity, validity, and volatility, viability, value, and victory), but others argue that these are derived from context-specific analytics while the original 3 V's capture the "intrinsic, definitional big data properties essence of big data" (Grimes 2013). Regardless, the nature of the V's determine the analysis tools that can be used and, in turn, the insights that can be gained.

The 3 (or more) V's have a different flavor in simulation than in other big data situations. Velocity and volume are partially controlled by the analyst, who determines how to run the simulation (e.g., on a single core or on a high-performance computing cluster), how much data to output (e.g., aggregate end-of-run statistics, batch statistics, or full time-series output) for each performance measure, and the number of performance measures to study. Generated output can have a variety of types, but the variety does not include many of the problems that we find with observational data (such as incompatible data formats or inconsistent data semantics). In the near future, many simulators may make use of big data tools for tying their models to real-time or near real-time data sets for model input (Elmegreen et al. 2014). With regard to some of the "wanna V's," verification and validation have long been cornerstones of effective simulation practice. A structured V&V process means that the simulation output should not suffer from lack of veracity. Of course, the question of validation – whether the simulation model's behavior is sufficiently close to that of the real-world problem of interest for decision-making purposes – is always with us.

2.2 The 3 F's of Data Farming

Moving beyond the 3 V's of observational big data, in the arena of large-scale simulation experiments we can focus on the 3 F's of inferential big data: *factors, features*, and *flexibility*. All these should be "big" when viewed with a data farming mindset.

Factors refers to a broad view of the inputs (or functions of inputs) that, if varied, affect the simulation and should be manipulated to increase our understanding of the simulation responses. A "big factor" view includes many aspects. Clearly, a large number of factors may be of interest: in fact, one can argue that if we did not feel an input was important, then we would not have included it in our simulation model. The big factor view also means that factors may vary over wide ranges, rather than limited ranges. They may be of different types – qualitative, discrete, or continuous – rather than homogenous. Factors can be further broken down into decision factors that can be controlled in the real-world settings; noise factors that are difficult or impossible to control in reality, but can be controlled during the simulation experiment; and artificial (simulation-specific) factors – such as run length, warm-up period, batch size, or random number seed – that may not have an analogy in the real world, but can influence the way we conduct our simulation experiments and the results we obtain.

Features refers to the simulation responses. A "big feature" view includes many aspects. We may be interested in multiple responses, rather than limiting ourselves to a single one. For stochastic simulation models, our responses may have complex variance and covariance structures. It is worth mentioning that this characteristic, as pervasive and accepted as it is in simulation, is still the exception rather than the rule for classical experiment designs. As with the factors, the responses may be of different types. We may be interested in short-term, transient behavior for one response, while at the same time we are interested in the long-run quantiles of another response. We may also be searching for different types of features in the response surface landscapes. For example, in trade-off analyses, maxima and minima are often much less interesting than so-called "knees in the curve," where we find that further changes in one or more factors lead to diminishing (or increasing) returns. Other interesting features include thresholds where responses change suddenly, such as a model entity experiencing an abrupt shift from usually losing to usually winning; broad, flat regions that indicate we have a solution that is robust to uncertainties in the underlying factors over certain ranges; and Pareto sets containing those alternatives that cannot be completely dominated (in terms of desirable multivariate responses) by any other alternatives in this set.

Flexibility represents the fact that we should be able to answer a variety of questions from our experiments, without needing to specify these questions *a priori*. A "big flexibility" view affects the type of design we use: by choosing a design that facilitates a broad variety of metamodeling, data mining, and graphical analysis tools, we are less likely to find that the data we grow are not sufficient for our needs than if we pick a restrictive design that is intended for specific (but limited) types of analysis. One way of gaining flexibility is to have a space-filling design: this term has typically been used for designs involving continuous-valued factors. The analogy for discrete-valued or qualitative factors is to have balanced or nearly-balanced designs. Note that in simulation, we can have additional flexibility built into how we store and retrieve our data. As long as we store the design points and random number seeds, we have the option of "growing" new data from our original experiment by rerunning our simulation and printing out additional output, if needed. For example, if we initially print out only end-of-run summaries and then find out that strange things have happened in a handful of the thousands of runs we have conducted, we can rerun that handful and inspect (or animate) the entire sequence of events to better understand what has transpired.

3 REVISITING CORRELATION AND CAUSATION

The vast majority of big data is currently observational in nature. This means that patterns found in big data (using data mining, machine learning, and artificial intelligence) are correlative, and do not necessarily reflect causal relationships in the underlying systems. The focus on correlation rather than causality has been extolled by some and lamented by others. For example, in their book about the big data revolution, Mayer-Schönberger and Cukier (2013) claim that "the need for sampling is an artifact of a period of information scarcity" and that big data "leads society to abandon its time-honored preference for causality, and in many instances tap the benefits of correlation" (pp. 13 and 18, respectively).

There are certainly times when uncovering correlation is very useful. As Horgan (2014) points out, "Epidemiological studies demonstrated more than a half century ago a strong correlation between smoking and cancer. We still do not understand exactly *how* smoking causes cancer. The discovery of the correlation nonetheless led to anti-smoking campaigns, which have arguably done more to reduce cancer rates over the past few decades than all our advances in testing and treatment." Yet there are also many more times when identifying correlation is not nearly good enough. It is well known that trolling through mountains of data can yield spurious correlations: an excellent discussion of how this affects published research results appears in The Economist. (2013). Common sense can help eliminate some spurious correlations (Vigen 2015), but one of the tenets of big data is that it is not repeatable. Correlations that appear strong at one point in time and weak at another might mean that the underlying system has changed.

If observational data reveals a correlation between two variables X and Y, the ground truth can be one of four basic situations: (i) changes in X cause changes in Y; (ii) changes in Y cause changes in X; (iii) changes in X and Y are both the result of changes in other, potentially unknown or unobservable factors; or (iv) this is a spurious correlation. One drawback of observational data is that we have no real way of determining which of these situations is true.

The simplest way of establishing cause-and-effect is via an experiment. Design of experiments (DOE) is a well established field (see, e.g., Montgomery 2017 or Ryan 2007). Three important concepts in DOE are *control*, *randomization*, and *replication*. For real-world experiments, we exercise control over the situation by deciding which values of X are of interest. We also decide how to control for everything else that is not of interest, perhaps by holding it constant or using a control group for comparison purposes. Randomization is used to guard against hidden or uncontrollable sources of bias. For example, if we are measuring the miles per gallon of different vehicles by using a single driver, randomizing the order in which they are driven will remove any systematic bias due to fatigue. With replication we collect multiple observations to assess the magnitude of the variability associated with Y, so we can construct confidence intervals or conduct significance tests.

In simulation experiments, the analyst has total control, and we use these concepts in different ways. Potential factors in simulation experiments include the inputs and distributional parameters of a simulation

model, whether or not they are controllable in the real world. The analyst also has control over the random number seeds and streams. This means that, unlike physical experiments where uncontrollable noise occurs, the results from a simulation experiment are perfectly repeatable, and randomization is not needed to guard against hidden or uncontrollable sources of bias. Replication means we get multiple experimental units (runs or batches) to gain a sense of the magnitude of the variability associated with Y. While homogeneous variance is commonly assumed for physical experiments, heterogeneous variance is pervasive in stochastic simulation. Consequently, we should not view response variability as merely a nuisance for estimating means and other output statistics, but as an important characteristic of the simulation's behavior.

We just reviewed correlation and causation in its simplest terms as the linear relationship between two variables. In a "better big data" context, there are many X's and many Y's, and their interrelationships can be much more complicated.

In short, observational and inferential big data are not the same. Observational big data washes over us like a force of nature, but big data from large-scale simulation experiments puts the analyst in control. If the goal of analysis is to yield better outcomes via controlling or influencing the inputs, what is more important than establishing causality?

4 STEPPING INTO DATA FARMING

4.1 Setting Appropriate Goals

For a decade and a half, we have advocated three basic goals of large-scale simulation experiments: (i) *developing a basic understanding* of a particular simulation model or system; (ii) *finding robust* decisions or policies; and (iii) *comparing the merits* of various decisions or policies (see, e.g., Sanchez and Lucas 2002; Kleijnen et al. 2005). These are broader than other goals that are often mentioned for deterministic computer experiments, namely: (i) constructing accurate *predictions*; (ii) *calibrating* the simulation to real-world data, or (iii) *optimizing* a system (see, e.g., Santner et al. 2003). Casting our results in a big data framework helps to clarify why we choose different goals. We believe that simulation is the proper choice for modeling complex problems. In these circumstances, it is unlikely that decision makers will be interested only in the answer to a single, narrow question. Developing an understanding – including identifying important factors and interesting features – allows us to address a much richer range of questions. Similarly, we believe that robust (resilient) alternatives found by seeking configurations that perform well over a variety of noise factor settings may be much more useful in practice than alternatives found by optimizing a single performance measure while implicitly holding noise factors constant. Finally, in trade-off analyses it may be much more useful to identify the existence of "knees in the curve," where we start to see increasing or decreasing rates of return, than to numerically predict exactly where these inflection points fall.

4.2 Portfolio of Potential Designs for Large-scale Simulation Experiments

How do we run simulation experiments that meet the 3 F's of factors, features, flexibility in a way that is sufficiently fast? First and foremost, we use a designed experiment capable of meeting these requirements given our computing resources and the time frame available for making the decision. Many of the classic experimental designs can be (and have been) used in simulation studies – but do not actually meet the needs of a "big data" view. The context for real-world experiments can be much more constrained than for simulations in terms of costs, number of factors, time required, ability to replicate, ability to automate, etc., so a framework specifically oriented toward simulation experiments is beneficial.

There are many considerations in selecting a design, including but not limited to:

- the goals of the experimenter;
- the ease of conducting a designed experiment, either within the simulation modeling platform, via external scripts, or via a customized data-farming wrapper;

- the length of time required for the simulation runs, which may vary greatly depending on the design point as well as due to random variation;
- the availability of parallel computing resources; and
- the timeline available for the study.

Elsewhere in these proceedings is a tutorial on designing and conducting large-scale simulation experiments, with examples of different classes of designs useful for data farming studies (Sanchez et al. 2018). A "consumer reports" chart that provides guidance to those interested in conducting large-scale simulation experiments is kept on the SEED Center web pages, along with software and spreadsheets that can be freely downloaded (https://harvest.nps.edu). The chart characterizes designs in terms of their factors, features, and flexibility; gives notes with additional guidance; provides citations for the source papers; and highlights designs that we have found to be good starting points. More on this "better big data" philosophy appears in Sanchez and Sanchez (2017). Books that discuss DOE specifically for computer or simulation models include Santner et al. (2003), Law (2014), and Kleijnen (2015).

While we do not discuss specific designs in this tutorial, efficient designs are the key enabler of large-scale simulation experiments, because they allow us to break the curse of dimensionality that arises from a brute-force approach. Suppose we have a simulation model that runs as fast as a single machine instruction and has 100 inputs, each varied at two levels (low and high). The world's most powerful supercomputer is the Summit, with an astounding 200 petaflop capacity (Simonite 2018). However, brute-force computation of all 2^{100} combinations for our simulation model would keep the Summit running at full capacity for over 178 millenia. Efficient design of experiments can break this curse of dimensionality at a tiny fraction of the hardware cost. Recent breakthroughs provide designs, such as those in Hernandez et al. (2012) or Vieira et al. (2013), that can be used to explore multiple replications of 100 or more factors in days to weeks on a single machine, or hours to days on a computing cluster – even for simulations that take minutes or hours to complete a single run.

4.3 Fast Data Farming Environments

If inferential simulation data sets are already on hand, we can identify whether they are characterized by enough factors, features, and flexibility to address our questions. But if we are preparing to grow our data, then the timeliness of getting the results is also important.

Efficient DOE is *absolutely required* for large-scale simulation experiments. So, if you use designed experiments, is that enough to say that your turnaround time is fast? Not necessarily. For example, if changing the factor levels in your simulation model can only be accomplished through a graphical user interface (GUI), then the bottleneck is often the analyst's time, rather than the computational time. Manually changing all the factor settings is a time-consuming and error-prone process. If you are facing a very short deadline, your only alternative may be to use a small design and double-check or triple-check your input settings. In contrast, setting up a data farming environment is extremely worthwhile. Automating the run generation process immediately expands your capability for growing data that does not suffer from input errors, and paves the way for running your model on a cluster.

While efficiency is needed, is the small design always preferable? For simulation experiments, the answer is a resounding 'No!' If your cluster has 1000 cores and your simulation takes a fixed amount of CPU time to complete, it will take no longer to conduct 1000 runs than to conduct a single run; if so, calling a design with 10 design points "faster" than one with 1000 designs points is both wrong and counterproductive! It is better to gather enough data, via larger designs and multiple replications, to be able to explore the simulation's performance without resorting to lots of simplifying assumptions. The bottom line is that a large-scale simulation is fast if you get the kind of data you need in time to act on it.

Even using good designs, the data farming approach includes automating the process of data collection whenever possible. Once you have chosen the design (or design algorithm), you should use a computing script to automatically run the experiments, allocate individual runs to distributed computing assets, and

consolidate the output in a form suitable for analysis. This requires some programming expertise and may take a bit more time initially, but the payoff is worthwhile. Open source software is available to help you get started. For example, HTCondor is open source software that controls the distribution of runs across computing nodes, either on a single machine or a computing cluster (HTCondor 2018). Xstudy and OldMcData, available from https:\harvest.nps.edu, facilitate the conversion of a design file into the associated set of input files for instantiating the simulation runs, for modeling platforms with scenario inputs stored in XML formats. SESSL, available from sessl.org, is a domain-specific language with a declarative core developed to facilitate the process of setting up and conducting simulation experiments Warnke and Uhrmacher (2018). Sanchez and Sanchez (2017) and Warnke and Uhrmacher (2018) provide additional guidance related to the "nuts and bolts" of automating large-scale experiments.

Another issue to consider is the number of replications. For run-based experiments with long individual run times, it makes sense to iterate through replications of the entire design, rather than iterate through a large number of replications of design point 1, then design point 2, etc. The former makes it more likely that you will have useful information if you stop the experiment early, and it allows you to halt once you have seen enough.

4.4 Analysis Approaches

Clearly, the design you choose will impact the types of analyses you can conduct. Goals (i) and (ii) of Section 4.2 are often addressed by developing a metamodel (i.e., a statistical model of the simulation model's I/O behavior) that can be used to convey key features of the model; metamodels also facilitate predictions of the model's performance at untried design points before rerunning the simulation, and can aid in the search for so-called optimal or robust solutions. In this section, we describe a few different classes of metamodels that can be useful. We then provide a few graphs and examples of how visualization can provide additional insights when we take a big data view.

Polynomial metamodels, built using regression techniques, are a class of metamodels often used in both physical and simulation experiments. We find that main effects model are too restrictive, but a good starting point for a single response Y is often a model that can include second-order effects, such as quadratic terms and two-way interactions. For some designs, such as central composite designs, it is possible to estimate simultaneously all first-order and second-order terms.

Alternatively, space-filling designs based on orthogonal or nearly-orthogonal Latin hypercubes are much more efficient, growing as $O(k)$ or $O(k^2)$, rather than the $O(2^k)$ required to simultaneously estimate all 2^k potential effects. A variety of polynomial metamodels can be fit, from first-order models, models with higher-order terms involving a subset of factors, up to a ridiculously large-order polynomial involving a single factor. Stepwise regression or some other automated method can help to determine the subset of terms that are most important. Note that in a big data world, many statistically significant terms may, nonetheless, be eliminated from the metamodel because they are not deemed to be of practical importance, i.e, they are dominated by other terms with much larger effects.

A second approach we find to be quite useful is the data mining approach of partition trees, also called classification and regression trees (see, e.g., James et al. 2013, ch.8). Partition trees employ a binning and averaging process to successively split a large group of heterogeneous data into two smaller groups of data, where each leaf in a split is individually more homogeneous, while the difference between the leaves is large. Partition trees have several advantages. They are non-parametric, they can be used for both qualitative and quantitative factors and responses, and they can quickly identify important factors and convey the main results in a form easy to explain to both technical and non-technical audiences. Large-scale designed experiments can be very useful aids to model verification and validation during the model development process. For example, a simple partition tree (two splits from 27 potential factors) identified conditions for which a preliminary model failed to run to completion. In short, conducting experiments during the model development process can allow modelers to identify and eliminate coding bugs, and fix conceptual or documentation errors – all while decreasing the model development time.

Two other metamodeling approaches of interest are kriging (also known as Gaussian process modeling) and stochastic kriging. Kriging has been heavily used for deterministic computer experiments, and also adapted for stochastic simulation experiments (see, e.g., van Beers and Kleijnen 2003; Ankenman et al. 2010; Kleijnen 2017). However, the analysis approach is computationally quite intensive, so kriging model-fitting is typically conducted for experiments involving relatively small numbers of factors and design points. These metamodeling approaches are quite flexible in terms of the model form and consequently are useful for prediction, although they can be harder to interpret and the results can differ for different software packages (Erickson et al. 2018).

Any metamodeling approach – multiple regression, logistic regression, partition trees, kriging – can be used to identify the most important factors and interactions for a single response. Yet this does not mean we have learned all we can from the data. Many other statistical and graphical approaches can provide additional insights, as part of our *big features* view of data farming.

For example, we might be interested in assessing trade-offs among multiple performance measures. Cheang (2016) explores a simulation inspired by a training scenario in which an aircraft carrier traveling through the Straits of Hormuz must fend off an attack from small boat swarms, while retaining enough weapons for its primary mission. Figure 1, adapted from Cheang (2016), is a parallel plot of 12 different measures of effectiveness (MOEs) for this scenario. Each line traces the average MOE values, over 50 replications, for one of 513 design points in a designed experiment. The trace highlighted with a red dashed line corresponds to the baseline performance where the carrier must rely on its own weapons to fend off the enemy attack. The traces highlighted in bold blue solid lines lines correspond to design points where there are 3 littoral combat ships (LCSs) and 5–6 tactical unmanned aerial vehicles (UAVs) that can also be employed. The scales differ for different columns, so numeric values are provided for the highlighted alternatives. Here, the five left-most measures all decrease dramatically from the baseline scenario (dashed red line) to the alternatives of interest (bold blue lines). These include the average numbers of enemy missiles and small boats that leak through the carrier's defensive ring, as well as the average percentage of different carrier weapons used. MOEs 6–8 correspond to the usages of UAVs of different types, MOEs 9–10 correspond to the fuel used by the LCSs and UAVs, and MOEs 11–12 represent two types of costs. This parallel plot displays a wealth of information that is not easily conveyed in a non-graphical format.

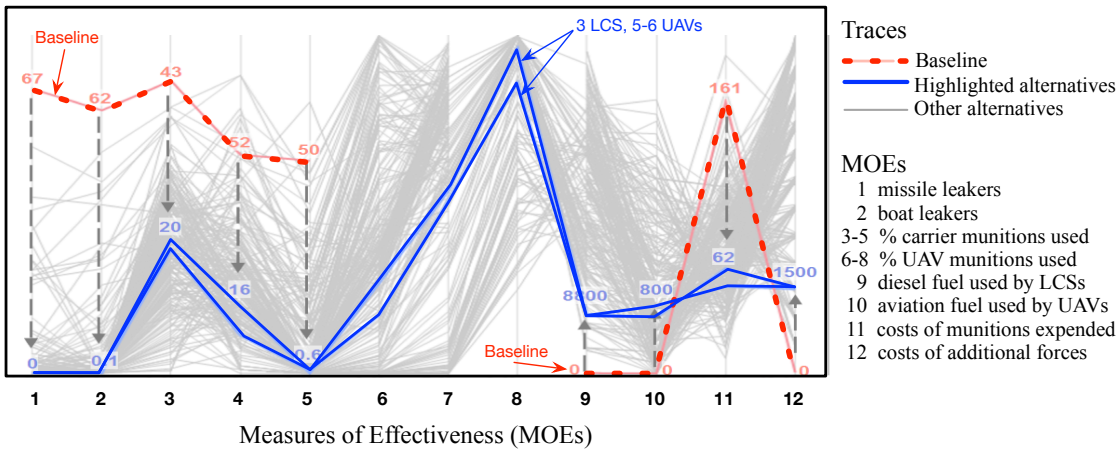

Figure 1: Parallel plot for 12 performance measures from a naval convoy protection simulation. Each of the 513 traces the average of 50 replications for the performance measures shown (adapted from Cheang 2016.)

We now give a few graphics based on dynamic behavior within simulation runs. Figure 2(a), adapted from a presentation based on Marlow and Novak (2013) shows the results of a single replication from a simulation that tracks the state of each of 24 naval helicopters in a fleet over their 30 year life cycle.

The colors represent ten different states of the helicopters – including embarked and ashore flight hours; scheduled, unscheduled, and deep maintenance – when a particular heuristic is used for scheduling flights and rotations. Despite how compressed the information is, we can see some diagonal stripes corresponding to specific states. We circle one such pattern that reveals a swath of operational but idle aircraft. Figure 2(b) is a resource heatmap, one of many visualization and analysis tools implemented in the STORMMiner tool for exploring campaign analysis scenarios in the Synthetic Theater Operations Research Model (STORM). This resource heatmap displays how often a particular resource was above or below a desired threshold for 10 different surface naval assets over time, from 30 replications of a single design point. The results clearly indicate that two of the assets (C and J) rarely, if ever, meet the threshold after the first day of the naval campaign. Assets F and G have similar bands of low inventory potential at certain points of time, while several assets (A, D, E, and H) have resource levels above the threshold throughout the campaign. Graphs such as those in Figure 2(a–b) have the potential to reveal useful information that may otherwise be difficult to identify without prior knowledge (see Morgan et al. 2018 for further examples of STORMMiner displays and tools for this unclassified naval campaign-level simulation). Similar scripts could be developed to automatically generate these types of plots for other simulation modeling platforms.

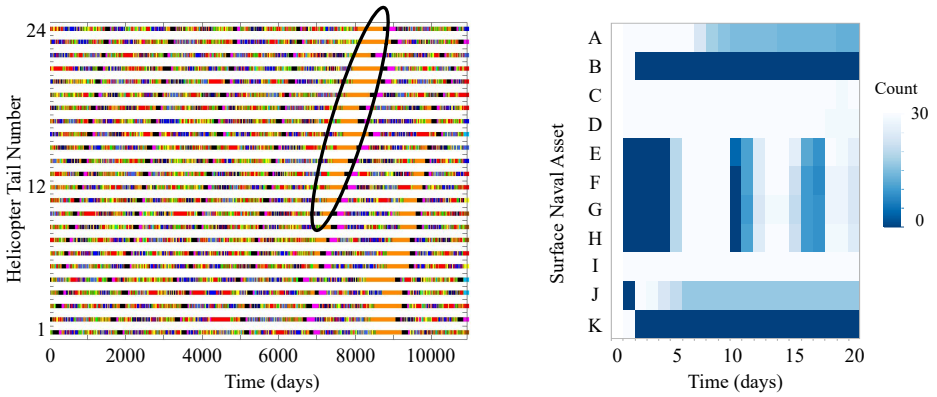

Figure 2: Two graphs (best viewed in color) revealing several dynamic outputs for a single design point. (a) States of each of 24 naval helicopters over a 30 year life cycle for a single replication (adapted from Marlow and Novak 2013). (b) Likelihood of a resource dropping below a critical inventory level of 50% capacity over time, for surface naval assets A–K based on 30 replications (adapted from Morgan et al. 2017).

Lin and Nelson (2016) coined the term "virtual statistics" to refer to responses that are conditional on an event occurrence and are estimated post-experimentation by retrieving sample path information from simulation runs. Lin et al. (2017) expand on this idea to estimate derivatives and variances for virtual statistics, in addition to means. Graphs like those in Figure 2(b) might help the analyst identify situations where virtual statistics would be particularly informative. In this way, computing virtual statistics for quick "what if?" analyses might be able to provide insights without requiring additional runs. In addition, if virtual statistics about mission success change dramatically based on resource levels just below or just above a threshold at a particular time period, this new information could be incorporated into future experiments or suggest policy changes that might help mitigate the problem.

Time-related graphs are also of interest when we show the results of either individual replications, or of summary statistics (e.g., averages, variances, or quantiles) computed over time from different replications. Olabode (2016) performs several data farming experiments on an optimization model for time-shiftable load scheduling for temporary shelters in a desert environment. There are nine shelters in this scenario, served by a mix of fossil fuel and renewable energy sources. The goal of the model is to keep the shelter temperatures within allowable boundaries during the daytime hours, while turning on and off the

various energy sources based on either perfect knowledge or predictions of the upcoming (random) cooling requirements. Figure 3 shows two graphs adapted from one of his experiments. On the left, the lines represent the shelter temperatures (in °C) over time, averaged over all design points, when perfect future information is available. One interesting feature of the plot is the consistent differences in average shelter temperatures during most of the time horizon; further investigation confirmed that this occurs because of different thermodynamic characteristics of the shelters, and the fact that some energy sources provide cooling to multiple shelters at a time. Another interesting feature is the uptick near the end of the time horizon, as all the power sources are turned off early whenever possible to save energy. The graph on the right shows the relative efficiency, over time, of operating with uncertain future demand versus perfect future knowledge. Interesting features of this graph are the high degree of variability early in the day (showing that perfect future knowledge might or might not be helpful), the relative consistency of the ratio with a value near 1.0 for most of the design points (showing that perfect future knowledge has minimal effect as time progresses), and the single design point whose sample path is quite different from the others.

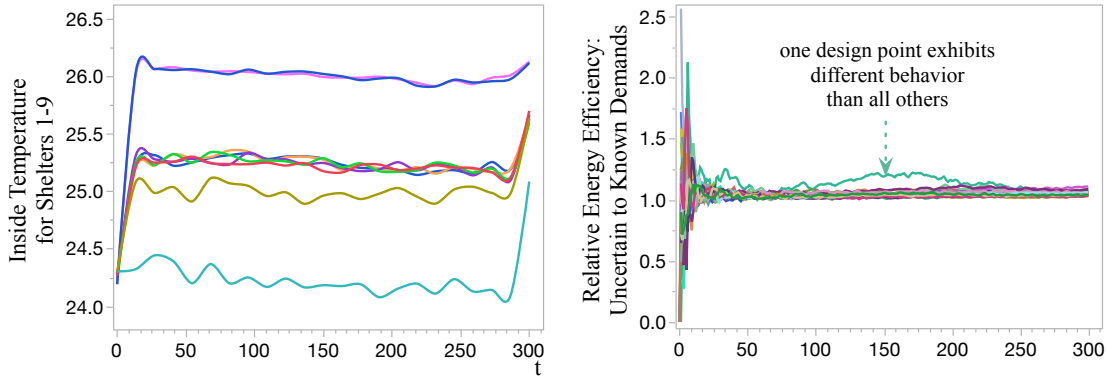

Figure 3: Performance measures related to energy usage in temporary desert shelters (adapted from Olabode 2016). (a) Average temperature inside each of nine shelters over time. (b) Relative efficiency of operating with uncertain versus known future energy demands, over time.

One other capability that is extremely powerful for analysis, but can only partially be displayed using static plots, is the ability to easily explore the data interactively. *Visual analytics* is another term for semi-automated processes that combine the visualization skills of people interacting with the data-processing strengths of computers (see, e.g., Keim et al. 2010). This interactivity can be achieved in different ways. Some statistics packages, such as JMP (https://www.jmp.com) are designed with interaction in mind. Those using R (https://www.r-project.org/) may wish to create templates, reusable scripts, or R Shiny applications that can run in web browsers (https://shiny.rstudio.com/) to facilitate this type of exploratory analysis. For those who foresee ongoing needs for conducting data farming experiments on multiple scenarios instantiated using a particular modeling platform, more effort may be needed to set up reusable data farming wrappers and post-processing tools – but the benefits can transform the way simulation studies are conducted.

4.5 Testing the Results: Another Reason Simulators Have Better Data

In large-scale simulation experiments, if we identify that some factor X belongs in a metamodel for some response Y, the ground truth can be one of four different situations: (i) changes in X (either individually, or in conjunction with other factors) cause changes in Y; (ii) changes in X are partially confounded with changes in high-order interactions of other factors; (iii) changes in X are fully confounded with changes in high-order interactions of other factors; or (iv) we have a false positive effect. In the simulation world, it is possible to explicitly test the results and determine whether or not our effects are truly important.

Our choice of design determines the degree of confounding we are willing to accept *a priori*. This, in conjunction with model confirmation runs or secondary experiments, lets us test to see if the metamodel is sufficiently accurate at previously-untested regions in the factor space. If so, we can feel comfortable using the metamodels, in place of running additional simulations, for quick-turn analysis. If not, we can use new information and additional designs to enhance the metamodels until their performance is satisfactory.

4.6 Reality, and Back Again

All the techniques we have described allow us to gain a better understanding of our simulation models. We have found them very useful throughout the simulation process: during model development and model verification, as well as for "production runs." Yet it is important to remember that our experiments are providing us with information about the simulation model's behavior, not necessarily the behavior of the real-world system they attempt to emulate. In cases where it is possible to get some real-world data for validation purposes, we can use our simulation experiments to suggest real-world test cases that may be most informative or interesting. In other situations, where we do not have (and may hope never to have) real-world data, the ability to perform large-scale designed experiments allows us to examine a plethora of "what if?" and "what matters?" analyses in an efficient and effective manner.

5 THOUGHTS FOR THE FUTURE

Several aspects of the intersection between big data, simulation, and decision making will be of increasing interest in the near future. Here are a few that we hope will resonate with our simulation community.

5.1 Future Simulation Clients

All too often, the elegance and rigor of having a closed-form or mathematically tractable solution have been touted as advantageous over a simulation modeling approach. This ignores the introduction of "type III errors" (Mitroff and Featheringham 1974) defined as "solving the wrong problems." This bias towards stylized analytical models becomes harder to justify in the face of readily available big data. For example, when it is not necessary to task someone to go and collect a lot of information because that information is already available, it is harder to justify assuming i.i.d. exponential random variates. Increasingly, the lack of a closed-form solution is not an issue when our software is capable of computing results to a desired level of precision in a small amount of time; this is called *computational tractability* (Lucas et al. 2015). Climate change, economics, transportation, combat, and social dynamics are just a few of the areas where closed-form analytic models will not suffice – computational models are better at capturing the complexity of the underlying systems, and so are better choices for investigating these types of problems.

When stakeholders have complex problems and are studying complex systems, they are not likely to be interested in answers to simple questions. Just as "having" big data from the internet meant that companies found new and exciting things to do with it, we have seen that having big data from simulation experiments offers the opportunity for new and interesting ways of looking at the results. "How should I set up my transportation network?" and "What are the impacts of the affordable care act on health costs and health outcomes?" are much more complicated (and interesting) questions than "What is the expected time-in-system for a customer in an M/M/1 queue with no balking and unlimited buffering?"

The current fascination with big data has several secondary effects. The rapid evolution of data science means that a greater number of simulation and non-simulation professionals will be becoming more adept at scripting, modeling, graphical displays, and statistical analyses. Decision makers may, similarly, be less likely to shy away from using observational or model-driven data to inform their decisions. At the same time, enhancements to methods for rapidly creating, merging, searching, displaying, and analyzing data from large repositories may prove to be useful new tools for the simulation community. Comfort with computerized and computer-based decisions is also increasing in other ways. If we trust a well-written

computer program to drive a car (Jaffe 2014), why not trust a well-written computer program for other types of decisions?

5.2 Future Simulation Methods

Most often, we see examples of simulation studies defined to address a specific question. It is time to view simulation-based decision making as an ongoing process, not an end state. Why do we churn up the CPU cycles when we are in the midst of an analysis activity, and then let our computer sit idle for the rest of the time? One intriguing idea is to go back and forth between models of different types or different fidelities, as we seek to learn more about these systems: this is being done in some scientific computing communities, such as computational physics, and it may have interesting parallels in the discrete-event simulation community. Another approach, even if we begin with a specific question, is to generate more output (in a structured manner) so that we are prepared for the next round of questions from our model, and we have been able to identify interesting features in the response surface metamodels that might not have initially been apparent.

There has been a great deal of good work in our community on simulation optimization, ranking and selection, and response surface modeling. But these presuppose that the decision maker knows what questions to ask. More research is needed on multi-objective procedures, exploitation of parallel computing, adaptive methods, and the design and analysis of large-scale simulation experiments. At the same time, there appear to be new opportunities for some established research areas. Importance sampling becomes potentially more of interest as we pull in real-time data. Can we easily update the state of our simulation, identify branching opportunities, and move forward quickly in parallel? Regarding simulation optimization and other adaptive search techniques, it may be that we should be doing optimization on metamodels, rather than on the simulations themselves – and that we need automated ways of reoptimizing as these metamodels evolve over time.

5.3 Future Simulation Software

Data capture is being automated at an incredible rate from real-world systems, ranging from satellite imagery, to web site navigation, to social network analysis, to engine systems. In the future, the process of creating linkages between real-world data and simulation modeling environments may be facilitated as the Internet of Things keeps expanding. This increases the potential for using simulation as a real-time decision support and control system, as it has in recent biopharmeceutical applications (Johnston et al. 2008). Major simulation packages may adopt an "app" approach, and use common data exchange protocols and interface protocols to link simulation models with external data sets. If so, the same protocols might also allow the practitioner to easily find or create suitable apps for analysis (e.g., simulation experiments, simulation optimization, ranking & selection, importance sampling) as well as for big data visualization. We anticipate further increases in the use of adaptive, automated analysis methods.

Simulation software developers should start taking more advantage of cloud computing, coupled with the ability to run models remotely via a Web interface. Software developers might consider whether there is an analogy to a subscription service for running simulation models, rather than licensing software for individual machines or users. At the server side, intelligent resource allocation ("automated data farming") can take advantage of parallel processor capabilities in stand-alone clusters or in clouds.

The use of smarter computational agents is an area that is ripe for improvement. The medical field has a few applications where intelligent software agents search through large data sets and find correlations. These have led to theories (e.g., on environmental or genealogical links to certain diseases later in life) that can then be examined more thoroughly and tested by medical researchers. Can we do the same with simulation? One way is to construct intelligent agents to search through model-driven data sets, identifying important factors and interesting features in the responses. Another way is to embed some of this capability into our models themselves; for example, rather than relying on calls to random variate generators with

fixed parameters, we might allow intelligent agents within our simulation model to access near-real-time big data and assess whether or not these distributional models still appear to be valid.

5.4 But Wait, There's More!

The thoughts in this section represent a few changes that I feel are on the horizon, but if there is one thing that the past few decades have taught us, it is that we never truly know what the future will hold. When the internet got started, we viewed e-mail as a faster alternative to letters, and word-processing as a potential way of cutting down on waste paper – in other words, incremental change instead of revolutionary change. Similarly, when the Web got started, we did not envision how this connectedness would change our society. So whatever the future holds, the simulation community should be poised to identify, respond to, and ideally blaze a trail that leverages emerging technologies.

Our simulation community was interested in many of these ideas before they captured the public's attention. Because we have been wrestling with them for years, we already have a rich literature of effective ways to deal with complex problems. If we take steps to push both the fundamentals and state-of-the-art simulation techniques out to broader communities, we will help them to avoid reinventing the wheel – or worse, repeating the mistakes of the past. More importantly, we will help jump-start the process of improving decisions that may affect our businesses, our lives, and our planet.

As we look to the future, our simulation community has an important role to play. A data farming approach – where we design experiments that handle big sets of factors, big sets of features, yet are flexible and fast – allows simulation researchers and practitioners to stake out the area of inferential big data for addressing complex and important problems. Another benefit – namely, that simulation can be used to make investigate systems and situations that do not yet exist, or that are too dangerous or costly for real-world experiments to be viable – is a cornerstone of simulation modeling, but may not be obvious to those with an observational big data worldview. Our simulation community has the chance to become recognized as the gold standard for model-based decision making within the big data analytics community, and we should seize this opportunity.

ACKNOWLEDGMENTS

This tutorial is an update of Sanchez (2015). This work was supported in part by the USMC Expeditionary Energy Office and the ONR/NPS CRUSER initiative. Portions are adapted from Elmegreen et al. (2014). DoD Distribution Statement: Approved for public release; distribution is unlimited. The views expressed in this document are those of the author and do not necessarily reflect the official policy or position of the DoD or the U.S. Government.

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. C. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58:371–382.

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

Burgess, M. 2018. "What is the Internet of Things? WIRED Explains". *WIRED*, February 16th, www.wired.co.uk/article/internet-of-things-what-is-explained-iot, accessed July 10th, 2018.

Cheang, M. 2016. *Operational Energy Capability Portfolio Analysis for Protection of Maritime Forces Against Small Boat Swarms*. Master's thesis, Naval Postgraduate School.

Dodson, S. 2014. "Big Data, Big Hype?". *WIRED*. www.wired.com/insights/2014/04/big-data-big-hype/, accessed July 10th, 2018.

Elmegreen, B. E., S. M. Sanchez, and A. S. Szalay. 2014. "The Future of Computerized Decision Making". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk et al., 943–949. Piscataway, New Jersey: IEEE.

Erickson, C., B. E. Ankenman, and S. M. Sanchez. 2018. "Comparison of Gaussian Process Modeling Software". *European Journal of Operational Research* 266(1):179–192.

Grimes, S. 2013. "Big Data: Avoid 'Wanna V' Confusion". *Information Week*, August 7th.

Hernandez, A. S., T. W. Lucas, and M. Carlyle. 2012. "Enabling Nearly Orthogonal Latin Hypercube Construction for any Non-saturated Run-variable Combination". *ACM Transactions on Modeling and Computer Simulation* 22(4):20:1–20:17.

Horgan, J. 2014. "So Far, Big Data is Small Potatoes". *Scientific American Blog Network*, June 9th, blogs.scientificamerican.com/cross-check/2014/06/09/so-far-big-data-is-small-potatoes/, accessed July 10th, 2018.

HTCondor 2018. research.cs.wisc.edu/htcondor/, accessed July 10th, 2018.

Jaffe, E. 2014. "The First Look at How Google's Self-Driving Car Handles City Streets". *The Atlantic Citylab*, April 28th.

James, G., D. Witten, H. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Johnston, L., L. Schruben, A. Yang, and D. Zhang. 2008. "Establishing the Credibility of a Biotech Simulation Model". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason et. al., 822–826. Piscataway, New Jersey: IEEE.

Keim, D., J. Kohlhammer, G. Ellis, and F. Mansmann. (Eds.) 2010. *Mastering the Information Age: Solving Problems With Visual Analytics*. Goslar, Germany: Eurographics Association.

Kleijnen, J. P. C. 2015. *Design and Analysis of Simulation Experiments*. 2nd ed. New York: Springer.

Kleijnen, J. P. C. 2017. "Regression and Kriging Metamodels With Their Experimental Designs in Simulation: A Review". *European Journal of Operational Research* 256:1–16.

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. 2005. "A User's Guide to the Brave New World of Designing Simulation Experiments". *INFORMS Journal on Computing* 17(3): 263–289.

Laney, D. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety". In *Application Delivery Strategies*, Number 949. Stamford, CT: META Group Inc.

Law, A. M. 2014. *Simulation Modeling and Analysis*. 5th ed. New York: McGraw-Hill.

Lin, Y., and B. L. Nelson. 2016. "Simulation Analytics for Virtual Statistics via *k* Nearest Neighbors". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder et al., 448–459. Piscataway, NJ: IEEE.

Lin, Y., L. Pei, and B. L. Nelson. 2017. *Virtual Statistics in Simulation via k Nearest Neighbors*. Technical report, Northwestern University.

Lucas, T. W., W. D. Kelton, P. J. Sánchez, S. M. Sanchez, and B. L. Anderson. 2015. "Changing the Paradigm: Simulation, Now a Method of First Resort". *Naval Research Logistics* 62(4):293–303.

Marlow, D., and A. Novak. 2013. "Using Discrete-Event Simulation to Predict the Size of a Fleet of Naval Combat Helicopters". In *Proceedings of MODSIM2013*, December 1st–6th, Adelaide, South Australia, 2506–2512. MSSANZ.

Marr, B. 2018. "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read". *Forbes*, May 21st, www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/, accessed July 10th, 2018.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt Publishing Company.

Mitroff, I. I., and T. R. Featheringham. 1974. "On Systemic Problem Solving and the Error of the Third Kind". *Behavioral Science* 19(6):383–393.

Montgomery, D. C. 2017. *Design and Analysis of Experiments*. 9th ed. Hoboken, New Jersey: Wiley.

Morgan, B. L., H. C. Schramm, J. R. Smith, T. W. Lucas, M. L. McDonald, P. J. Sanchez, S. M. Sanchez, and S. C. Upton. 2018. "Improving U.S. Navy Campaign Analysis Using Big Data". *Interfaces* 48(2):130–146.

NATO 2014. *Data Farming in Support of NATO*. Technical Report TR-MSG-088, NATO Science & Technology Organization.

Olabode, J. 2016. *Analysis of the Performance of an Optimization Model for Time-shiftable Electrical Load Scheduling Under Uncertainty*. Master's thesis, Naval Postgraduate School.

Press, G. 2013. "A Very Short History of Big Data". *Forbes*, May 21st, www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/, accessed July 10th, 2018.

Ryan, T. P. 2007. *Modern Experimental Design*. Hoboken, New Jersey: Wiley.

Sanchez, S. M. 2015. "Simulation Experiments: Better Data, Not Just Big Data". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al., 800–811. Piscataway, New Jersey: IEEE.

Sanchez, S. M., and T. W. Lucas. 2002. "Exploring the World of Agent-Based Simulation: Simple Models, Complex Analyses". In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yucësan et al., 116–126. Piscataway, New Jersey: IEEE.

Sanchez, S. M., and P. J. Sanchez. 2017. "Better Big Data via Data Farming Experiments". In *Advances in Modeling and Simulation*, edited by A. Tolk et al., Chapter 9, 159–179. Cham, Switzerland: Springer International Publishing AG.

Sanchez, S. M., P. J. Sánchez, and H. Wan. 2018. "Work Smarter, Not Harder: A Tutorial on Designing and Conducting Simulation Experiments". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe et al. Piscataway, New Jersey: IEEE.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

SEED Center for Data Farming. 2018. harvest.nps.edu, accessed July 10th, 2018.

Simonite, T. 2018. "The US Again Has the World's Most Powerful Supercomputer". *WIRED*, June 8th, www.wired.com/story/the-us-again-has-worlds-most-powerful-supercomputer/, accessed July 10th, 2018.

SINTEF 2013. "Big Data, for Better or Worse: 90% of World's Data Generated Over Last Two Years". *ScienceDaily*, May 22nd, www.sciencedaily.com/releases/2013/05/130522085217.htm, accessed July 10th, 2018.

The Economist. 2013. "Unreliable Research: Trouble at the Lab". *The Economist*, October 19th.

van Beers, W. C. M., and J. P. C. Kleijnen. 2003. "Kriging for Interpolation in Random Simulation". *Journal of the Operational Research Society* 54:255–262.

Vieira, H., S. M. Sanchez, K. H. K. Kienitz, and M. C. N. Belderrain. 2013. "Efficient, Nearly Orthogonal-and-Balanced, Mixed Designs: An Effective Way to Conduct Trade-off Analyses via Simulation". *Journal of Simulation* 7(4):264–275.

Vigen, T. 2015. *Spurious Correlations*. New York, NY: Hachette Book Group.

Warnke, T., and A. Uhrmacher. 2018. "Complex Simulation Experiments Made Easy". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe et al. Piscataway, New Jersey: IEEE.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2017. *Data Mining: Practical Machine Learning Tools and Techniques*. 4th ed. Cambridge, Massachusetts: Morgan Kaufmann.

AUTHOR BIOGRAPHY

SUSAN M. SANCHEZ is a Professor of Operations Research at the Naval Postgraduate School, and Co-Director of the Simulation Experiments & Efficient Design (SEED) Center for Data Farming. She also holds a joint appointment in the Graduate School of Business & Public Policy. She has been an active member of the simulation community for many years, and has been recognized as a Titan of Simulation and an INFORMS Fellow. Her web page is http://faculty.nps.edu/smsanche/ and her email is ssanchez@nps.edu.