*Proceedings of the 2021 Winter Simulation Conference*
*S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo and M. Loper, eds.*

# INCREASED NEED FOR DATA ANALYTICS EDUCATION IN SUPPORT OF VERIFICATION AND VALIDATION

Christopher J. Lynch
Ross Gore

Virginia Modeling, Analysis, and Simulation Center
Old Dominion University
1030 University Blvd.
Suffolk, VA, USA

Andrew J. Collins
T. Steven Cotter
Gayane Grigoryan

Department of Engineering Management and Systems Engineering
Old Dominion University
2100 Engineering Systems Building
Norfolk, VA, USA

James F. Leathrum, Jr.

Department of Computation Modeling and Simulation Engineering Department
Old Dominion University
1300 Engineering and Computational Sciences Building
Norfolk, VA, USA

## ABSTRACT

Computational simulation studies utilize data to assist in developing models, including conducting verification and validation (V&V). Input modeling and V&V are, historically, difficult topics to teach and they are often only offered as a cursory introduction, leaving the practitioner to pick up the skills on the job. This problem of teaching is often a result of the inability to introduce realistic datasets into class examples because "real world" data tends to come in poorly formed datasets. In this paper, a case is made that teaching data analytics can help ease this problem. Data analytics is an approach that includes data wrangling, data mining, and exploratory analyses through visualization and machine learning. We provide a brief discussion on how data analytics has been be applied to computational modeling and simulation in the context of verification and validation, but also for input modeling as that gives a basis for validation.

## 1 INTRODUCTION

Modeling and Simulation (M&S) strives to gain an understanding of a system's structure and behavior, be that system real or imagined. This occurs through the development of models that abstract the system down to its critical elements and the implementation of computational simulations that can be executed to explore the model of the system. The simulation model then can undergo verification and validation (V&V) to determine if the implemented simulation matches the intended model design and if its results match the expected behaviors from the real system. This second part requires data to be collected from real-world sources, even if this data is speculative and hard to obtain.

M&S education has classically glossed over much of the difficulties relating to data. Educational examples of input modeling use clean sets of data that translate easily into pre-determined data distributions for inputs into the simulation. V&V is considered too challenging to be given proper coverage; instead it is presented as a required step in the process with appropriate methodologies defined, without the required

**978-1-6654-3311-2/21/$31.00 ©2021 IEEE**

practices to make a student sufficiently skilled in its application. Applying V&V to academic models and simulations, which are used in M&S education, would not be of great benefit to the student due to the simplicity of these clean data sets. While model design and simulation development can scale from academic exercises to the real world, V&V practices cannot. We propose that a more formal education in data analytics helps prepare students to become competent practitioners of V&V. By expanding students' understandings of the importance of cleaning, prepping, describing and sampling, and managing data, the student is better equipped to understand and properly interpret and communicate the significance, both statistical and practical, of simulation testing outcomes. Understanding data analytics can address shortcomings in transitioning from verifying and validating simulations within academic settings to applying and interpreting V&V outcomes to practical, realworld scenarios.

In the 2020 Winter Simulation Conference (WSC), the authors made a case for the importance of data analytics in the M&S education process (Leathrum et al. 2020). This paper carries that concept further by focusing on the necessity to enable a proper introduction to V&V and input modeling. A primary difficulty in presenting more realistic problems in which to address V&V is the required effort of the student to become intimate with the data, a step that data analytics allows. The work presented here is the result of the development of a series of short courses to enable an organization to move into data analytics and M&S. Being that the courses train real-world practitioners, a cursory introduction to V&V and input modeling is inappropriate. Taking advantage of the already developed background in data analytics enables the courses to delve deeper into the concepts than previously possible.

## 2 BACKGROUND

### 2.1 Verification and Validation Education in Practice

The lack of fully integrated V&V training in the M&S curriculum impedes the training of future M&S professionals. Sargent and Balci (2017) performed an informal survey at WSC 2016 of vendors, and they found that while it was recognized that the Department of Defense and some large companies require V&V, outside of that venue there was a surprising lack of the practice. They estimated that 50% of simulation models receive adequate V&V, 25% some V&V, but inadequate, and 25% receive no V&V. One cause of this could be insufficient education of professionals on the topic. Education of the practical application of V&V requires data, but handling data requires an intense effort for discrete event simulation projects (Skoogh and Johansson 2007). Thus, it is difficult to dedicate sufficient time, in an M&S course curriculum, to present real-world test cases necessary for proper skill development. This gives the impetus for making students more comfortable with the data through modern data analytics techniques.

Data is used throughout the M&S process. Model formation can be heavily data reliant as it assists in forming initial hypotheses about the system and its expected behaviors. These behaviors form the benchmarks that can later support how well the simulation matches the model (verification) and how well the model matches the real system (validation) (Whitner and Balci 1989, Sokolowski and Banks 2010). Data helps inform model development, helps conduct V&V, helps explore model outcomes to gain insight, and helps effectively communicate findings. Data commonly needs to be cleaned, filtered, formatted, or structured into usable formats to fill these needs. Data Analytics (DA) contains numerous methods that support these M&S activities.

The world is becoming more data-driven. Gantz and Reinsel (2012) describe the measure of data created, replicated, and consumed in a single year as a "digital universe." The growth and emphasis of data for decision-making contributed to the popularity of the concept of data analytics early in the 2000s (Tyagi 2003). DA is the application of computer systems to analyze large data sets to support decisions (Runkler 2012). Runkler describes DA as an interdisciplinary field with close ties to disciplines such as statistics, machine learning, pattern recognition, system theory, operations research, or artificial intelligence. Several DA terms are often used interchangeably, including data science, data analysis, and data mining.

Data science is the use of the scientific method with data analytics; however, it tends to be used as an umbrella term for data analysis, analytics, and mining as well as domain-specific research problems where

data is thrust to the forefront (Van Der Aalst 2016). Data analysis uses data analytics to study the parts of the whole and help form hypotheses (Tukey 1962). Data mining uses computational algorithms to illuminate meaning, relationships, and patterns; these findings generate input for the decision-maker and support further evaluation of the data. Many DA activities support various activities within the M&S process. Figure 1 provides a simplified view of the M&S process and illuminates areas where intersections exist between DA and M&S. This paper focuses on the input modeling and data modeling as applied to V&V.

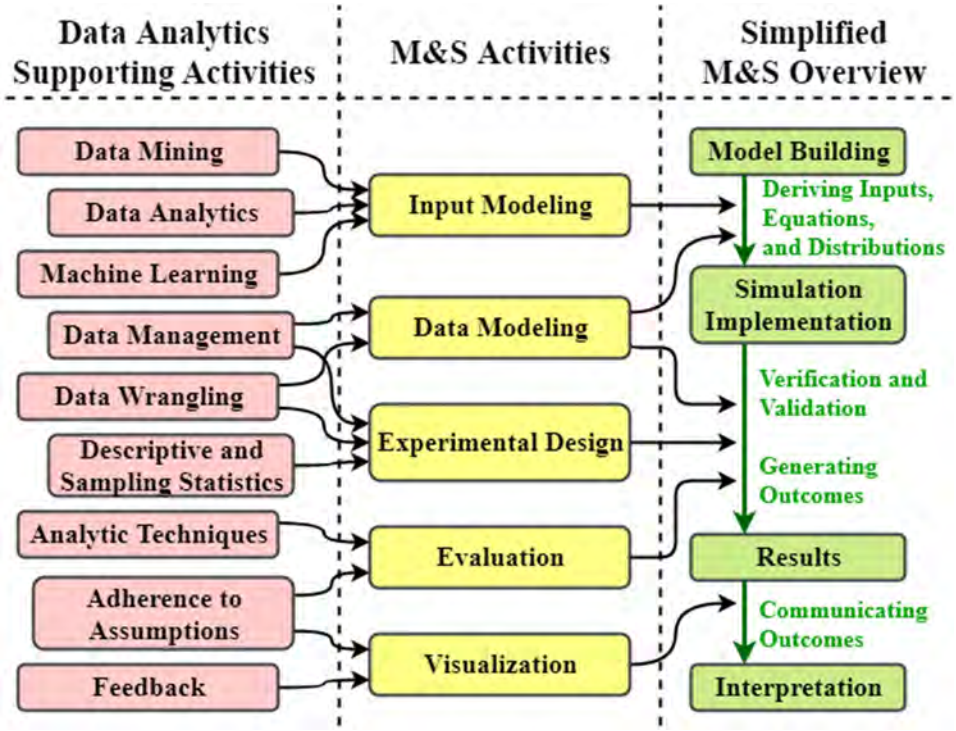


Figure 1: Connections between Data Analytics activities and Modeling and Simulation activities.

## 2.2 Barriers and Challenges to Verification and Validation

Numerous techniques exist for verifying and validating simulations with vast differences in requirements in mathematical formality (Sargent 2013, Balci 1998). Law (2008) presents techniques for building valid and credible simulation models and underlines the role of data and the importance of data analysis. Informal methods have been found to be more commonly utilized over formal methods in practice (Padilla et al. 2018). Informal techniques generally have lower learning curves, do not require rigorous mathematical background, and rely on intuitive feedback mechanisms for obtaining results. The type of medium used to convey V&V outcomes impacts the associated credibility of the test result (Lynch 2019). Subjective techniques rely on subject matter expert review, graphical inspections, or animation to help identify the presence of potential errors. However, statistically supported tests may better aid in identifying the cause of the problem while covering larger portions of the simulation space. Statistical tests provide mathematical indicators, such as correlation and coverage, in support of searching for unexpected simulation occurrences (Gore, Lynch, and Kavak 2017). These techniques generally utilize more cumbersome experimental design considerations.

V&V applications are commonly re-engineered based on examining a simulation's outputs with respect to its intended use (Balci 1998, Sargent 2013), and the burden falls on the model builder to evaluate correctness. As a compounding effect of focusing on intended use and the development of ad hoc V&V

solutions, few generic tools exist for facilitating V&V. Subject matter experts or domain experts may be barred from using or trusting simulation results as a result of high learning curves associated with conducting and interpreting V&V (Sargent 1981), for acquiring the programming and statistical backgrounds necessary (Whitner and Balci 1989), or for obtaining the M&S knowledge needed for tracing simulation outcomes back to their specifications (Gore, Lynch, and Kavak 2017).

Challenges exist for facilitating the selection and application of specific V&V techniques (Padilla et al. 2018, Lynch et al. 2020). Several studies conducted of software testers across industry and academia (Andersson and Runeson 2002, Ng et al. 2004, Wojcicki and Strooper 2006, Eek et al. 2015) have enumerated many challenges in adopting, applying, and accepting techniques. Some of the main challenges to the incorporation of V&V identified from these studies include: tradition of use, resource requirements, ease of use, and the practice of applying techniques in an *ad hoc* manner to their simulations. Table 1 summarizes the questions and perceptions driving these challenges.

Table 1: Challenges contributing to perceived barriers for adopting V&V techniques.

| Challenge | Questions | Perceived Barriers |
|---|---|---|
| *Resource requirements* | What resources are required to evaluate if new techniques are needed? How are any of the proposed techniques beneficial? How much time may be needed for training and application? | (1) High costs to purchase commercial products; high time requirements; and cost-effectiveness. (2) Time to evaluate whether new techniques fill a need and to evaluate how long it may take to learn and incorporate them existing toolkits (Ng et al. 2004) |
| *Tradition of use* | What has worked before (individually or within the organization)? How familiar are the modelers with the techniques? Are tools available to assist testing? | (1) New techniques may look less promising based on guidance from senior researchers and in deference to successfully applied techniques in the past (Andersson and Runeson 2002, Arthur and Nance 1996) |
| *Ease of use* | How complicated is the theory underlying the technique? How long will this take to learn? What are the assumptions and requirements to apply the technique? | (1) Level of understanding required to retrieve the right type of data, to post-process or format output data, or to understand simulation results. (2) Lack of expertise or high difficulty in using a technique can prevent its adoption (Shannon 1998) and informal techniques are most often utilized in practice (Padilla et al. 2018) |
| *Ad-hoc solutions* | How to apply the technique? Does the technique contribute something new? Is the contribution relevant? | (1) Solutions are rebuilt on a case-by-case basis. (2) No universal solutions exists and a group of techniques may often be needed to explore the simulation correctness (Glasow and Pace 1999) |

Considering these questions from the perspective of Data Analytics can aid understanding of the importance and practical significance of V&V for students, teachers, researchers, and professionals, as presented in Figure 1. The growing availability of statistical programming languages and platforms for handling larger data sets, aiding in the formatting and cleaning of data, and providing more intuitive means for representing and conveying information can help to reduce these barriers. The role of Data Analytics within this frame should be understood with respect to how it augments, not replaces, the M&S processes that it supports. To support model building and V&V practices, education in Data Analytics should reflect the specific components of M&S activities that are supported while reinforcing how these processes should be further utilized or expanded upon within the life cycle of the M&S process.

## 3 MANAGING INPUT AND OUTPUT DATA

Data, from a curriculum perspective, provides a transparent link between the simulation model and the real system. Data collected from the real system can be used to inform the types of distributions utilized within the model as well as forming an empirical baseline for conducting V&V comparisons on the correctness of

simulation outcomes. Input modeling provides the connections between data and its probabilistic mechanisms within the real system and generally manifests in the forms of probability distributions and uncertainty (Wagner and Wilson 1995, Leemis 1999, Cheng 2017). Input modeling and V&V can both be supported through cleaned and well-structured input and output data. Input data aids in deriving structure (such as by applying machine learning), deriving input parameters, and identifying underlying model assumptions. Output data allows for more easily applying statistical tests, assessing sample sizes, examining simulation adherence to underlying model assumptions (i.e. minimum sample sizes or boundary values). Identifying an analysis plan and corresponding sample size requirement *a priori* can greatly facilitate the analysis processes for V&V. Model credibility can be supported through adequate sample sizes in the experimentation phase (Hariharan et al. 2017, Lynch and Gore 2021). For instance, statistical debugging has been used to delve into complex simulation occurrences and behaviors without requiring formal mathematical specifications to help isolate sources contributing to the errors (Gore et al. 2015, Gore, Lynch, and Kavak 2017).

Large amounts of data about human activities, engineered systems, and smart devices can be utilized to inform model structure (Kibira et al. 2015, Jain et al. 2017) and the representation of human behaviors (Kavak et al. 2018). If a well-defined experimental input space cannot be specified for a set of predictor variables, the analyst must resort to gathering available unstructured data as generated by or as available from other studies or analyses of the phenomenon of interest. Unstructured messy data can induce bias structures and unequal variance estimates in differing regions of the sample space when fitting an assumed linear model with the least-squares or maximum likelihood algorithms. Sources of messy data include: (1) data with differing linear scales or mixed linear-nonlinear scales; (2) sample discontinuities along a continuous dimension or from multiple studies; (3) nonrandom or semi-random missing structures; (4) cross-classified data with unequal sampling replication; (5) semi-structured data or completely unstructured data containing combinations of the prior problems; (6) any multivariate model of less than or greater than full rank given the information contained in the predictors; (7) strong collinearity among predictor variables; or (8) a multivariate model of full rank from consistently structured multivariate data in the same linear dimensions with leverage points or outliers.

The ability to convey understanding of how to effectively conduct data collection and how to properly utilize the collected data is critical for input modeling and supports the foundation of V&V activities. Data identification, collection, and assessment informs model development and supports evaluation (Leemis 1999, Onggo and Hill 2014, Lynch and Gore 2021). Input models generate the random numbers used as input to the simulation model and lead to the eventual creation of the output data (Schmeiser 2001). Data collected on the modeled system provide a comparison point to assess how well the simulation performs at recreating the known behaviors of the modeled system. Establishing a measure of credibility for the simulation model depends on the quality of the collected data. Issues such as biased and missing data, correlated and autocorrelated variables, or the presence of messy data sets can greatly impact the reliability of any assessments of simulation correctness and contribute to type I and type II errors. Curriculum should teach methods for proper collection and storage of data, techniques for cleaning and preparing data, and provide transparent relationships between the collected data and any developed metrics for incorporating empirical data (where possible) for verifying or validating the simulation.

### 3.1 Regression with Messy Data

Regression analysis of messy data is usually approached by assuming a feasible linear model and conducting appropriate residuals diagnostics on the fitted model. Tackling messy data problem for regression analysis is a simple example to highlight the impact of messy data on the model performance. A regression model is considered feasible if it accounts for the structural variance components observed in the phenomenon under study while remaining small enough to be considered as parsimonious (Wooldridge 2015). In the balanced data, full rank case, interpretation of the mapping of the assumed linear model to the phenomenon's structure and behavior yields understanding or new knowledge that is most representative of the phenomenon. In the messy unbalanced data, less than full rank case, this mapping to the

phenomenon's structure and behavior yields, at best, subjective interpretations and less optimal understanding or misunderstanding of the phenomenon. Further, it affects the model validity and prediction accuracy. More about regression model validation can be found in Snee et al. (1977).

Managing messy data is the most critical step in the data analytics modeling process. Before data acquisition, the analyst must specify the analysis objective: screening, modeling, prediction, or optimization. The objective of screening is to identify the predictor variables whose linear combination in a regression model adequately represents the structure or behavior of the phenomenon. The questions to address here are: (1) how the combined range of the variables limit the modeling space of the phenomenon's structure or behavior; and (2) how the non-completeness of the available data affects the statistical significance in variable selection relative to practical significance in phenomenon structure or behavior. Given a candidate set of predictors, the objective of modeling is selection of the set of linear models that isolate the types of effects (categorical, main, interactions, polynomial, nonlinear) most predictive of the phenomenon's structure or behavior. The selected model set should not be statistically different in information content as indicated by a model selection metric such as R-square, R-square adjusted, Mallow's Cp, Akaike Information Criteria (AIC), or Bayesian information criterion (BIC). The selected model set should be arrived at through triangulation comparison of the selection rankings by two or three of the metrics in order to mitigate weaknesses in each. The question that must be addressed is how the non-completeness of the data may induce bias, unequal variances, and instability in the sample space. Induced bias manifests as incorrect predicted or optimized structure or behavior when compared to observed phenomenon. Unequal variances filter through as incorrect confidence and prediction interval estimates in models that assume equal variance. Instability filters through as non-constant coefficient estimates and differing model sets across the sample space.

After identifying the analysis objective, data wrangling (1) sources the data; (2) reduces the set; (3) normalizes for consistency; and (4) parses data into required structure. Data must be sampled from the phenomenon or from sources, such that: (1) analysis objective can be met; (2) data includes all key predictor variables necessary to capture model structure; (3) the time period represents all relevant phenomenon behavior; and (4) sufficient observations exist to achieve required confidence and power of fitted models. Next, data are transformed into the correct, ordered, and simplified forms necessary for model building. The objective is to produce a compact data structure which uses memory that is as close as possible to its information-theoretic lower bound and that permits the most efficient modeling computation. Data normalization contributes efficient computation by scaling the predictor variables to common ranges. Scaling methods include min-max, mean, unit-length, and z-score normalization. Centering the data may be necessary for modeling additive bias variance components. Parsing the data structure simply means translating the compacted, scaled data into vector, matrix, or array formats required regression software.
Missing data may introduce bias and compromise predictions with fitted regression models. Potential bias from missing data depends on the missingness mechanism and the estimation method applied to estimate any missing values, such as: (1) Missing Not at Random (MNAR) - a relationship exists between the missing values and their pattern in the data set; (2) Missing at Random (MAR) - a systematic relationship exists between the missing and observed values, but not the missingness pattern in the data set; and (3) Missing Completely at Random (MCAR) –a relationship does not exist between the missing values and any other values in the data set. Widely applied methods for missing values include:

1. Completely Recorded Units – complete case analysis removing all cases with missing data, This approach is simple to apply and is satisfactory with small quantities of missing data. It can induce significant bias and is not statistically efficient with estimating missing values for subpopulations.
2. Weighting – randomized inferences from sample data with missing values weighted by sampling design weights inversely proportional to their probabilities of selection.
3. Imputation – (a) mean estimation –estimate the mean from observed values; (b) substitution – impute missing values from samples not included in the data-wrangled set; (c) hot-deck imputation

– randomly select values from samples included in the data-wrangled set; (c) cold-deck imputation – systematically select values from other samples with similar values for other predictors; or (d) local least squares regression imputation. To achieve unbiased estimates, standard analyses must be modified to account for differences between observed and imputed values. Regression models with imputed values may understate model variance, confidence, and prediction intervals.

4. Model-Based – (a) multiple imputation; and (b) the expectation-maximization algorithm. Mixture weights are constrained to nonnegative and sum to one. Convergence requires the gradient of the objective function and possibly the Hessian depending on the method. Convergence may be slow and attain local optima. Forward and backward probabilities are required. Advantages include: flexibility; avoiding more ad hoc imputation methods; ability to evaluate model assumptions underlying the imputations; and estimates of variance that account for the missing data structure.

Input modeling, verification, and validation depend on the quality of data utilized to generate a model's components, such as input distributions, as well as for establishing empirical comparison points to aid in V&V, such as data utilized for statistical hypothesis testing. Therefore, understanding how messy data was addressed, when applicable, aids in how to properly interpret and communicate model results and the assumptions under which the results are valid. This knowledge supports the application of V&V by strengthening the evidence for how testing was performed and how the assumptions of the data and the assumptions of the applied tests relate to each other.

### 3.2 Data Management

Given the volume of data required and the analytic objects generated to support development of just a single model let alone an organization's modeling information needs, it is essential to follow a data management protocol and the organization implement a formal data management program to preserve and secure original data sets, intermediate structured and cleaned data sets, and final experimental or regression designs, models, and graphical objects. The individual data management protocol must address: (1) acceptable data set formats for the selected experimental design and regression analysis software; (2) raw and structured data set file structures and metadata definitions; (3) experimental design and regression definition objects file structures and metadata; and (4) model specification and related graphical objects outputs file structures and metadata. The Data Management Association DMBOK2 provides a framework to plan and coordinate the use, archive, retrieval, control, and purge of data sets and model objects.

Metadata provides information about data sets, experimental and regression definition objects, model objects, and graphical objects to facilitate storage, access, and retrieval as a structured data set. There are three general types of metadata: descriptive, structural, and administrative

Existing open source and proprietary data management software facilitates organizational data management. The R statistical computing software provides functions such as ***attr()*** to add metadata to variables, ***contents()*** to generate metadata about data frames, and ***metadata()*** to attach metadata to raster objects. R package metadata can be used to initialize, install, and read metadata packages about collections of R objects. Python provides library metadata for recording and retrieving metadata about python objects.

## 4 UNDERSTANDING DATA

Within the perview of V&V, the importance of understanding data is the relationship between how well the simulation matches its intended design and how well it matches the referent system. The aim is to increase confidence in the implemented simulation by trying to identify errors and to show their absence. Such as establishing the truth of a simulation against its requirements (Fishwick 2007) by accepting or rejecting the presence of errors based on the provided evidence. Additionally, the need to understand data is present during the exploration of simulation outcomes when evaluating results, analyzing interactions, tracing behaviors, and gaining numerous other forms of insight. Simulation data can be one-dimensional, multi-

dimensional, and specialized (Feldkamp, Bergmann, and Strassburger 2020). Statistical methods can help deal with difficulties in experimental error and the complexities of the effects studied (Box, Hunter, and Hunter 1978). To reinforce understanding of the importance of properly structured data and to support verification and validation, we suggest that data analytics education focus on topics of data wrangling, descriptive and sampling statistics, and machine learning. These analytic activities strengthen the M&S activities provided in Figure 1 by explicitly communicating data properties alongside model development decisions as well as model outcomes.

### 4.1 Data Wrangling

Data cleaning is an important problem, but it is an uncommon subject of study in M&S. This is somewhat unexpected as a nontrivial amount of time is spent on the process of cleaning and preparing data during the process of constructing a model and the ensuing simulation (Dasu and Johnson 2003). Data preparation is not just a first step but must be repeated many times over the course of analysis as new problems come to light or new data is collected (Duggan 2018). Despite the amount of time it takes, there has been surprisingly little research on how to clean data well. Part of the challenge is the breadth of activities it encompasses from outlier checking, to date parsing, to missing values. Structuring datasets helps to facilitate analysis.

A dataset is a collection of values, usually either numbers (if quantitative) or strings (if qualitative). A data set is clean if the values are organized so that every value belongs to a variable (i.e., a column) and an observation (a row). A variable (i.e., a column) contains all values that measure the same underlying attribute (like height, temperature, duration) across units. An observation (i.e., a row) contains all values measured on the same unit (like a person, or a day, or a race) across attributes. Real datasets almost always do not adhere to these definitions. For example, some combination of the following frequently occur: (1) column headers are values, not variable names, (2) multiple variables are stored in one column, (3) variables are stored in both rows and columns, (4) multiple types of observational units are stored in the same dataset and (5) a single observational unit is stored in multiple tables (Wickham 2014).

Clean datasets simplify many M&S activities, such as: (1) simulation calibration; (2) combining outputs from multiple sources; (3) visualization; (4) computing descriptive statistics; and (5) computing deviations from trusted dataset. This also simplifies data manipulation within simulations. Clean data sets' structure enables straightforward variable filtering and transformation, aggregating multiple values, and changing the order of observations. Ultimately, these capabilities result in more explainable models with more effective insights. Research and domain experience show that structured input data reduces the need for complexity and reduces the errors made during the model development process (Dasu and Johnson 2003, Kavak et al. 2018). Likewise, structured output data allows for the consistent and replicatable (versus *ad hoc*) application of V&V tests (Gore et al. 2015, Diallo et al. 2016, Duggan 2018).

### 4.2 Descriptive and Sampling Statistics

Descriptive statistics convey expected values for sets of observed outcomes with the data being treated as though it constitutes the whole (Navidi 2008). Sampling statistics investigate data based on samples from the full population to infer conclusions about the population. Statistical significance is evaluated for these statistics to determine the likelihood that the observed occurrence could have occurred by random chance. To this end, statistical indicators are common indicators into simulation behaviors (Chen et al. 2008, Lynch et al. 2020). Most simulation platforms provide the ability to visually display at least one value pertaining to their visual constructs during runtime, such as queue size, stock value, or transition probabilities (Sargent and Balci 2017). Numerical indicators aid in easily comparing and communicating structural issues, such as finding bottlenecks or comparing transition probabilities against specifications (Lynch 2019) and standard error measures are commonly applied to assess sampling error (Schmeiser 2001).

Understanding the role of descriptive and sampling statistics can yield greater evidence in support of the characteristics of the simulated data with respect the the real system's data. Additionally, this knowledge helps to communicate a test's coverage over the simulation space and can aid the interpretation of testing

outcomes with respect to the generalizability of findings across populations. For input modeling and verification and validation, increased understanding of descriptive and sampling statistics can support the validity of model behaviors and outcomes as well as the veracity of the utilized input destributions and probabilistic mechanisms.

### 4.3 Machine Learning

Examining whether an output of a black-box machine learning model "makes sense" gives a user confidence that the model is trustworthy (i.e., valid). Using machine learning, explanations can be created for a given single prediction from a black-box model. A prediction's explanation is based on a local surrogate model. The surrogate model is an interpretable model (i.e. a decision tree) that is learned on the predictions of the original black-box model (Guidotti et al. 2018, Hu et al. 2018). However, portions of a machine learning model are not designed to be transparent in terms of how the model makes its prediction. With this design decision, machine learning models can achieve a greater power of expressivity by adding more parameters and nonlinearity computation. However, many users will ask: "Can I trust this model?" (Castelvecchi 2016).

As an example, consider a neural network text classifier for differentiating the subject of bills from the United States Congress. The model achieves 94% on a held-out testing set, but the explanation provided by surrogate models show that the decisions result from arbitrary reasons, such as counting the words "read" and "host" which have no direct connection with a bill's subject areas (Collingwood et al. 2013). Even though this model achieves nice accuracy, it cannot be trusted without refinement. Furthermore, the surrogate model provides insight on how to refine neural network, such as better preprocessing on the text.

A tradeoff of complexity and nonlinearity at the expense of explain-ability is found in agent-based models (ABMs) which explore systems of spatially situated agents interacting over time. While agents' rules are transparent, the complex patterns that emerge from those rules are not (Epstein and Axtell 1996). A similar machine learning strategy can be applied to an ABM's outputs as in the development of surrogate models (e.g., linear/logistic models, decision trees). ABMs can be instrumented with predicates to highlight dynamics during a model run and which can serve as features within the surrogate models. Previous research has developed meta-models of ABMs in the form of logistic regression models for validation (Gore et al. 2017) and derived agent behaviors through machine learning (Kavak et al. 2018). Future research pursues developments in machine learning as a means of explaining emergent ABM behaviors.

Machine learning models are often developed as black boxes, utilized for predictions, and that utilize diverse ranges of validity measures (Guidotti et al. 2018). Understanding the properties behind each of these machine learning characteristics is required to inform input modeling and V&V activities for models derived from machine learning. This will aid the proper conveyance of machine learning model results as well as the proper conveyance of the assumptions surrounding their development and validation.

## 5 CONCLUSIONS

The expanding field of data analytics provides the opportunity to educate future M&S professionals such that they can efficiently become comfortable with the data and data models necessary in an M&S project. This further allows for the introduction of real-world V&V test cases in the educational process; these test cases allow students to explore systems complex enough to highlight the difficulties of V&V. Understanding the role of data analytics in support of M&S activities can:

- facilitate understanding of the many strengths provided to the M&S process,
- assist in conveying proper interpretations of a simulation's input and output data by properly connecting with the assumptions and cleaning methods applied to the real data,
- mitigate the perceived barriers for performing V&V, and
- aid in the establishment of trust and credibility in a developed simulation.

Effective application of DA techniques can provide empirical support for data-driven development decisions for model design while simultaneously serving as sources of requirements or specifications that

can be used for comparison during the cyclical V&V testing phases of the modeling process. Understanding how the data used to create the model design was collected, cleaned, and transformed (i.e. converted into input distributions) provides numerous benefits for properly interpreting outcomes, evaluating validity, and assessing generalizability. Focusing on how to deal with messy data, managing data, and understanding data through the lens of data analytics helps to build a solid foundation for properly interpreting and conveying simulation results. This provides the advantage of being able to clearly connect model components, such as input distributions, with the real data from which the distributions are derived. In cases where real world data exists on expected system behaviors, knowledge of these topics aids in communicating simulation outcomes alongside the assumptions of the data to build additional confidence in the validity of the model and the correct implementation of the simulation.

To exemplify this benefit, the authors developed a curriculum consisting of a series of short courses exploring data analytics, predictive analytics, and data modeling. This has allowed the authors to be more aggressive in the development of the M&S portion of the curriculum to allow students to consider everyday problems they encounter dealing with real data. This would not have been possible if students could not get the data into an appropriate form and develop the data models in support of a simulation, in particular the V&V process. Students can carry this work forward through the courses, building an understanding of the data to enable V&V without a large overhead.

## ACKNOWLEDGMENTS

This work was supported, in part, by the United States of America's Naval Sea System Command [grant number GS-10F-097CA].

## REFERENCES

Andersson, C., and P. Runeson. 2002. "Verification and Validation in Industry - A Qualitative Survey on the State of Practice". In *Proceedings of the 2002 International Symposium on Empirical Software Engineering*. Los Alamitos, California: IEEE Computer Society.

Arthur, J. D., and R. E. Nance. 1996. "Independent Verification and Validation: A Missing Link in Simulation Methodology?". In *Proceedings of the 1996 Winter Simulation Conference*, edited by J. M. Charles, D. J. Morrice, D. T. Brunner, and J. J. Swain, 230-236. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Balci, O. 1998. "Verification, Calibration, and Testing". In *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, edited by J. Banks, 335-393. New York: John Wiley and Sons, Inc.

Box, G., W. G. Hunter, and J. S. Hunter. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. *Wiley Series in Probability and Mathematical Statistics*. New York: John Wiley & Sons.

Castelvecchi, D.. 2016. "Can We Open the Black Box of AI?" *Nature News* 538(7623):20.

Chen, M., D. Ebert, H. Hagen, R. S. Laramee, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. 2008. "Data, Information, and Knowledge in Visualization." *IEEE Computer Graphics and Applications* 29(1):12-19.

Cheng, R. 2017. "History of Input Modeling". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page. Piscataway, 181-201. New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Collingwood, L., T. Jurka, A. E. Boydstun, E. Grossman, and W. H. van Atteveldt. 2013. "RTextTools: A Supervised Learning Package for Text Classification". *The R Journal* 5(1):6-13.

Dasu, T., and T. Johnson. 2003. *Exploratory Data Mining and Data Cleaning*. Vol. 479, Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons.

Diallo, S. Y., Gore, R., Lynch, C. J., and Padilla, J. J. 2016. "Formal Methods, Statistical Debugging and Exploratory Analysis in Support of System Development: Towards a Verification and Validation Calculator Tool". *International Journal of Modeling, Simulation, and Scientific Computing*, 7(01):1641001.

Duggan, J. 2018. "Input and Output Data Analysis for System Dynamics Modelling using the Tidyverse Libraries of R". *System Dynamics Review* 34(3):438-461.

Eek, M., S. Kharrazi, H. Gavel, and J. Ölvander. 2015. "Study of Industrially Applied Methods for Verification, Validation and Uncertainty Quantification of Simulator Models". *International Journal of Modeling, Simulation, and Scientific Computing* 6(02):1-29.

Epstein, J. M, and R. Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, Massachusetts: The MIT Press.

Feldkamp, N., S. Bergmann, and S. Strassburger. 2020. "Visualization and Interaction for Knowledge Discovery in Simulation Data". In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Honolulu, Hawaii, 1340-1349.

Fishwick, P. A. 2007. "The Languages of Dynamic System Modeling". In *Handbook of Dynamic System Modeling*, edited by P. A. Fishwick, 1-12. New York: Chapman & Hall/CRC.

Gantz, J., and D. Reinsel. 2012. "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East". *IDC iView: IDC Analyze the Future.* Framingham, Massachusetts: IDC Go-to-Market Services.

Glasow, P., and D. K. Pace. 1999. "Simulation Validation (SIMVAL) 1999, Making VV&A Effective and Affordable Mini-Symposium and Workshop". Alexandria, Virginia: DTIC Document.

Gore, R., S. Y. Diallo, C. J. Lynch, and J. J. Padilla. 2017. "Augmenting Bottom-up Metamodels with Predicates". *Journal of Artificial Societies and Social Simulation* 20(1):1-20.

Gore, R., C. J. Lynch, and H. Kavak. 2017. "Applying Statistical Debugging for Enhanced Trace Validation of Agent-Based Models". *Simulation: Transactions of the Society for Modeling and Simulation International - Special Issue on Modeling and Simulation in the Era of Big Data and Cloud Computing: Theory, Framework, and Tools* 93(4):273-284.

Gore, R., P. F. Reynolds Jr, D. Kamensky, S. Y. Diallo, and J. J. Padilla. 2015. "Statistical Debugging for Simulations". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 25(3):1-26.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2018. "A Survey of Methods for Explaining Black Box Models". *ACM Computing Surveys (CSUR)* 51(5):1-42.

Hariharan, P., G. A. D'Souza, M. Horner, T. M. Morrison, R. A. Malinauskas, and M. R. Myers. 2017. "Use of the FDA Nozzle Model to Illustrate Validation Techniques in Computational Fluid Dynamics (CFD) Simulations". *PLoS One* 12(6):1-25.

Hu, L., J. Chen, V. N. Nair, and A. Sudjianto. 2018. "Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP)". *arXiv preprint arXiv:1806.00663*:1-15.

Jain, S., D. Lechevalier, and A. Narayanan. 2017. "Towards Smart Manufacturing with Virtual Factory and Data Analytics". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 3018-3029. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kavak, H., J. J. Padilla, C. J. Lynch, and S. Y. Diallo. 2018. "Big Data, Agents, and Machine Learning: Towards a Data-Driven Agent-Based Modeling Approach". In *Proceedings of the 2018 Spring Simulation Multi-Conference*. San Diego, California: Society for Modeling and Simulation International.

Kibira, D., Q. Hatim, S. Kumara, and G. Shao. 2015. "Integrating Data Analytics and Simulation Methods to Support Manufacturing Decision Making". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, V. W. K. Chan, I.-C. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 2100-2111. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Law, A. M. 2008. "How to Build Valid and Credible Simulation Models". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, 1402-1414. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Leathrum, J., A. J. Collins, T. Cotter, R. Gore, and C. J. Lynch (2020). "Education in Analytics Needed for the M&S Process". In *Proceedings of the 2020 Winter Simulation Conference*, edited by N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 3236-3247. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Leemis, L. 1999. "Simulation Input Modeling." In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 14-23. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lynch, C. J., S. Y. Diallo, H. Kavak, and J. J. Padilla. 2020. "A Content Analysis-based Approach to Explore Simulation Verification and Identify its Current Challenges". *PLOS ONE* 15 (5):e0232929.

Lynch, C. J. 2019. *A Lightweight, Feedback-Driven Runtime Verification Methodology* Doctor of Philosophy Dissertation, Computational Modeling and Simulation Engineering (CMSE), Old Dominion University (22619686).

Lynch, C. J., and Gore, R. 2021. "Application of One-, Three-, and Seven-day Forecasts During Early Onset on the COVID-19 Epidemic Dataset using Moving Average, Autoregressive, Autoregressive Moving Average, Autoregressive Integrated Moving Average, and Naïve Forecasting Methods". *Data in Brief* 35:106759.

Navidi, W. C.. 2008. *Statistics for Engineers and Scientists.* New York: McGraw-Hill Higher Education.

Ng, S. P., T. Murnane, K. Reed, D. Grant, and T. Y. Chen. 2004. "A Preliminary Survey on Software Testing Practices in Australia". In *Proceedings of the 2004 Australian Software Engineering Conference*, 1-10. IEEE Computer Society.

Onggo, B. S. S., and J. Hill. 2014. "Data Identification and Data Collection Methods in Simulation: A Case Study at ORH Ltd". *Journal of Simulation* 8(3):195-205.

Padilla, J. J., S. Y. Diallo, C. J. Lynch, and R. Gore. 2018. "Observations on the Practice and Profession of Modelling and Simulation: A Survey Approach". *Simulation: Transactions of the Society for Modeling and Simulation International* 94(6):493-506.

Runkler, T. A. 2012. "Data Analytics". *Wiesbaden: Springer.*

Sargent, R. G. 1981. "An Assessment Procedure and a Set of Criteria for Use in the Evaluation of Computerized Models and Computer-Based Modelling Tools." Griffiss Air Force Base, New York: DTIC Document.

Sargent, R. G. 2013. "Verification and Validation of Simulation Models". *Journal of Simulation* 7(1):12-24.

Sargent, R. G., and O. Balci. 2017. "History of Verification and Validation of Simulation Models". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 292-307. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Schmeiser, B. W. 2001. "Some Myths and Common Errors in Simulation Experiments". In *Proceeding of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 39-46. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Shannon, R. E. 1998. "Introduction to the Art and Science of Simulation". In *Proceedings of the 1998 Winter Simulation Conference*, edite by D.J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 7-14. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Skoogh, A., and B. Johansson. 2007. "Time-Consumption Analysis of Input Data Activities in Discrete Event Simulation Projects". In *Proceedings of the 2007 Swedish Production Symposium*. 1-8.

Snee, R. D. 1977. "Validation of Regression Models: Methods and Examples." *Technometrics*, 19(4), 415-428.

Sokolowski, J. A., and C. M. Banks. 2010. *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*. Hoboken; New Jersey: John Wiley & Sons.

Tukey, J. W. 1962. "The Future of Data Analysis". *The Annals of Mathematical Statistics* 33(1):1-67.

Tyagi, S.. 2003. "Using Data Analytics for Greater Profits". *Journal of Business Strategy* 24(3):12-14.

Van Der Aalst, W. 2016. "Data Science in Action". In *Process Mining*, edited by W. Van Der Aalst, 3-23. Eindhoven, The Netherlands: Springer.

Wagner, M. A. F., and Wilson, J. R. 1995. "Graphical Interactive Simulation Input Modeling with Bivariate Bézier Distributions". *ACM Transactions on Modeling and Computer Simulation* 5(3):163-189.

Whitner, R. B., and O. Balci. 1989. "Guidelines for Selecting and Using Simulation Model Verification Techniques". In *Proceedings of the 1989 Winter Simulation Conference*, edited by E. MacNair, K. J. Musselman, and P. Heidelberger, 559-568. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Wickham, H.. 2014. "Tidy Data". *Journal of Statistical Software* 59(10):1-23.

Wojcicki, M. A., and P. Strooper. 2006. "A State-of-Practice Questionnaire on Verification and Validation for Concurrent Programs". In *Proceedings of the 2006 International Symposium on Software Testing and Analysis*. New York, New York: Association for Computing Machinery.

Wooldridge, J. M. 2015. *Introductory Econometrics: A Modern Approach*. Boston, Massachusetts: Cengage Learning.

## AUTHOR BIOGRAPHIES

**CHRISTOPHER J. LYNCH** is a Lead Project Scientist at the Virginia Modeling, Analysis, and Simulation Center (VMASC) at Old Dominion University (ODU). He holds a Ph.D. and MS in M&S from ODU. His email address is cjlynch@odu.edu.

**ROSS GORE** is a Research Assistant Professor at VMASC at ODU. He holds a Ph.D. and MS in computer science from the University of Virginia. His email address is ross.gore@gmail.com.

**ANDREW J. COLLINS** is an assistant professor at ODU in the department of Engineering Management and Systems Engineering (EMSE). He has a Ph.D. in Operations Research from the University of Southampton. His email address is ajcollin@odu.edu.

**T. STEVEN COTTER** is a Senior Lecturer in the department of EMSE at ODU. He holds a Ph.D. in Systems Engineering from ODU. He holds certifications as a SQL Server Application Developer and Database Administrator. His email address is tcotter@odu.edu.

**GAYANE GRIGORYAN** is a Ph.D. student and graduate research assistant in the EMSE Department at ODU. She received her MS in Economics from ODU. Her email address is ggrig002@odu.edu.

**JAMES F. LEATHRUM, JR.** is an Associate Professor and Chief Departmental Advisor of the Department of Computational Modeling and Simulation Engineering at ODU. He holds a Ph.D. in electrical engineering from Duke University. His email address is jleathru@odu.edu.