Big Data, Neural Meshes, and Probabilistic Analysis: Are We Really Improving Decision Making?

J.D. Solomon, PE, CRE, CMRP

Key Words: Big Data, Algorithms, Decision Making, Forecasts, Monte Carlo, Uncertainty, Diagnostics, Prognostics

SUMMARY & CONCLUSIONS

All data has the potential to produce knowledge and improve decision making. This paper discusses several key aspects of big data, two key methods for its use in decision making, and two case examples.

"Black box" solutions driven by poorly understood algorithms should not be used in lieu of fundamental principles and sound statistical application. There are no short cuts when it comes to analytics used for quality decision making.

The author concludes that big data is improving our decision making in terms of problem diagnostics and short-term operational resilience. Decision making is much less improved in prognostics, prediction, and forecasting over the long term where uncertainty is a more significant factor. Uncertainty should be embraced, in the presence or absence of big data, and probabilistic methods offer the best opportunity for improving decision making in the face of uncertainty.

1 INTRODUCTION

At no time in our history has data been so readily available, affordable to collect, and cost-effective to analyze. Three common examples included equipment vibration, customer water use, and collection of water quality samples.

Vibration monitoring is an example of predictive technologies that improve proactive operations-related decision making by humans and provide a foundational component of machine-to-machine learning. In the past decade, the same vibration monitoring that once was only available with expensive fixed units can now be duplicated by hand-held devices for less than one thousand dollars. Moreover, the units can now be used as portable units that can be transmitted readily to pages and mobile devices in lieu of a plant's telemetry system. Data is abundantly more plentiful and easily accessible to each member of the operations and maintenance staff.

Water metering systems have now made similar advances as electricity metering systems. Fixed networks now easily provide water and power use for every customer on 15-second intervals. The amount of data created, some useful for decision making and much not useful, is not exponentially higher than when meter readings were performed just once per month.

Citizen testing of tap water and source water is much more common due to the affordability of the tests and sampling devices. While positive in terms of the amount of data that is now available, an underlying question regarding the quality of the sampling and laboratories must be incorporated. Whether to include "uncontrolled data" or how to include it is a key decision before blindly turning an algorithm lose on a data set.

Figure 1. Similar to modern geospatial positioning systems (GPS), hand-held vibration units such as the TPI 9070 has enabled equipment mechanics & operators to have an unprecedented opportunity for knowing where they are and where they are going.

2 TYPES OF DATA

All data has the potential to produce knowledge. However, not all data is of the same type and quality. Analysts of large data sets usually work to varying degrees with mixed, or heterogeneous, data. There are a number of approaches, methodologies and tools associated with mixed or heterogeneous data which goes beyond the scope of the paper; however, the underlying principles of data measurement, scales, and statistical approaches are briefly provided.

Measurement is a process of assigning numbers to objects in such a way that unique relationships of the objects are reflected as well as the unique properties of the numbers themselves. Two fundamental theories of measurement are representativeness, meaning that objects can be assigned to a number, and uniqueness, meaning that a number cannot be assigned to the same object by different measurers [1].

A central question in the analysis of opinion-based data, such as 1 to 5 ratings, and some types of classification data, such as geographic locations, is whether the measurements and their underlying scales are both representative and unique.

S.S. Stevens in his classic 1946 paper "On the Theory of

This paper was presented at the 2016 Utility Management Conference and subsequently presented with additional case examples at the 2017 RUC Operationalizing Resilience Symposium

Scales of Measurement" developed the predominate arguments for the measurement scale position [2]. He described four types of scales – nominal, ordinal, interval and ratio. Ratio scales, which are continuous, are the only type of scale that all types of statistical analysis are valid; for interval scales, a limited number of classic statistical analyses are valid. For ordinal scales, such as many 1 to 5 rating scales, Stevens states "ordinary statistics like the mean and standard deviation ought not be used"; the practical implication is that only non-parametric approaches are applicable to ordinal data. For nominal scales, such as cardinal numbers, Stevens discounts the use of any form of statistics.

Many big data sets contain a mixture of data types. The mixture produces a conundrum that is further complicated by several primary factors. One complicating factor related to the quality by which data, especially continuous and interval data, is collected and the significant figures to which it is reported. For all types of data, the computer has now made it too easy to errantly report data to decimals that are not significant. The second complicating factor is that often non-continuous data is reported to decimal places when in fact it should not be. A third complicating factor is that categorical data is indeed significant to understanding trends and estimating missing data but the underlying algorithms may not treat the data in a useful form. The analyst, whether man or machine, has many value decisions that must be made related to how to incorporate the data and what to leave in (or what to leave out).

3 DIAGNOSTICS

Diagnostics are the statements and conclusions drawn from an investigation or analysis of the cause or nature of a condition, situation, or problem. Diagnostics are fundamentally related to what has happened, or the past. Diagnostics are extremely important for both human and machine learning. This is important for operational resilience, or reactions to short-term tactical distresses and failures which are evaluated in terms of seconds, minutes, or hours.

3.1 Diagnostic Techniques

There are a number of techniques that can be used in a diagnostic procedure. In practice, several techniques are usually used. All require some form of data.

Most of techniques can be classified as: differential, or exclusion, diagnosis; comparative analysis; pattern recognition; criteria, or threshold, diagnosis; and analysis using computer-based algorithms.

3.2 Neural Networks

A neural network provides computer-based data analysis by using algorithms that imitate human brain cell functions. Modern neural networks are multi-dimensional arrays, which pass information, perform calculations, and iteratively close error messages. Neural meshes are described as the localized surfaces within the network that are created from the algorithms learning patterns and associations [3]. If this sounds like what is known as artificial intelligence or machine learning, it is.

Neural networks were conceptually framed before World War II. Frank Rosenbaltt is credited with being one of the primary pioneers in machine learning in the latter 1950s. However, computing power would limit both speed and the approach to single layers. By the mid-1980s, computer technology had advanced to a point where multi-layered approaches became common. The technology has now advanced to the point that some entrepreneurs are developing a neural lace, which theoretically can be used as a seed from the computer, inserted in the brain, and produce a neural mesh that allows the brain to process information with similar speed and clarity as a computer. The theory is that humans provided the logic for learning to the machines, and the machines will teach humans how to organize their brain neurons to process the information faster.

This dependent variable in the data set used software using a neural network can be numeric or alphabetic. If numeric, the dependent variable can be either continuous or categorical. Arguably the best use of a neural network is more for pattern recognition of categorical data and less reliable for closing gaps or performing projections of continuous data. However, the accuracy and precision of any neural network conclusion is dependent on the logic embedded in its algorithms.

Common applications today include speech recognition, facial recognition, machine translation, video games, filtering on social networks, recognizing political polling, and understanding of human purchasing preferences.

Luke Dormehl has documented some of the practical applications of neural meshes and their underlying algorithms [the Formula]. The applications include recommender systems such as those provided by Amazon and eHarmony, which map our preferences against those of other humans and suggest merchandise or people for us to experience. On-line media outlets highlight news that is "relevant" to us, and perhaps shape the news we obtain. Netflix gathered data from its 25 million users to discover the trends and correlations to find the nexus of viewer preferences for story theme, the actor Kevin Spacey, and director David Fincher to develop their blockbuster television series, *House of Cards*. Neural meshes and underlying algorithms have also been used to create movies with multiple endings (depending on the audience), and stocking decisions by Wal-Mart such as the correlation of severe weather warnings like hurricanes to trigger consumer purchases of strawberry Pop-Tarts.

However, big data mining and analytics are founded on a theory that is reflected by the biases of algorithm creators as to what is important and what is not. This is especially true when the algorithm is developed to answer a particular question. As Dormehl states, "a computational process that seeks to sort, classify and create hierarchies in and around people, places, objects and ideas carries considerable political connotations. [4]"

Neural networks have traditionally found more use in applied fields such as marketing and banking, and less in the

fields of statistics and sciences where understanding of causation is required rather than "black box' outputs that are just close enough. However, with each generation the confidence in all fields that the "black box" is a neutral and objective partner is growing.

3.3 Neural Network Case Example

The cited case example is related to the desire to improve energy efficiency by a private sector company at 1800 of its sites. Related decision making involved several alternatives including re-negotiating rates with energy providers, implementing new standard operating procedures to minimize energy use, upgrading existing equipment with more modern units, and changing technologies. Palisade's NeuralTools was selected as the software tool based on the project needs, data size, and the company's desire to have a tool that they could use independently in the future.

The data sets for the 1800 facilities contained more than a decade of data for 16 categorical variables and 40 numerical variables. The 1800 facilities had multiple functions (coded as Brick, steel, and other) and multiple scales of operation/production. The facilities were grouped into five management regions, and each region had different processes and standards for collecting, reviewing, and managing their data.

One major limitation that was noted early by the company was some of the gaps in the data. Although decision making related to energy strategies was desired as soon as possible, the company realized that some time may be needed to research the gaps in the current data or collect additional data. A key initial assessment was how much time and effort was needed to establish a meaningful set of data.

NeuralTools requires the same three basic steps as any neural tool. First, training of the neural network is performed on a set of known data looking for patterns. Second, testing of the neural network is performed for establishing its accuracy against a known data. Third, new predictions are made for incomplete data using the trained neural network. Each step generates margins of error which should be evaluated by the analyst [5].

NeuralTools provides a choice of several algorithms that the analyst may use. The software also allows for testing and training with different data sets or with subsets of the primary data set. The subsets may be generated by the software or by the analyst.

The analyst jumped into the evaluation by letting the computer generate its own subsets for training and testing. When margins of error were high, manual selection of data sets were selected. Margins of error were again higher than expected. Predictions were nonetheless attempted, but were unacceptable. A deeper analysis would be require more than the two days of effort that was initially anticipated.

Further analysis of the analysis indicated that the data from the East Region, the largest region, was problematic in terms of its consistency/quality. Some basic statistical analysis of the continuous data using Palisade's StatTools also led to the conclusion that the data was not just raw data, but may have been manipulated prior to being included in the final data set (further discussions with the company determined that this was indeed the case).

The statistical analysis also indicated that the continuous data was skewed by seasonal energy demands, which should have not been a surprise but was overlooked in the rush to expedite the initial analysis. Many neural network tools, StatTools included, use some form of linear or multi-regression algorithms that assume the underlying continuous data is predominately normally distributed. The reality is that continuous data associated with energy use is skewed and not independent. This should have been better accounted for in the analysis. However, the fact that some of the provided continuous data had either been "penciled in" or linearly estimated before being transmitted would have greatly comprised the results anyway.

The analysis did provide relatively good results related to categorical data such as locations, types of operations, regions, etc. The same cannot be said for the analysis of the continuous data such as energy consumption, costs, etc. The analysis did provide some important initial insights, but at the same time some of the important flaws may have been overlooked or discounted by less experienced analysts.

A number of best practices are underscored by this case example:

1. Take some time with data and do some detailed statistical analysis prior to plugging the data into a software tool. This is particularly important for providing an indication of data quality, data distributions, and data that should be considered representative, or not representative, for use in the model. As in many cases, time and/or budget often drives the analyst to skip or minimize this step and the consequences are dire.
2. Spend some time with the methods included in any software and developing an understanding of their underlying algorithms and their intended purposes. In this case, the software was well understood but the quick schedule and assumptions related to the data created some misalignment with the methodologies that could have been chosen.
3. Initially form some governing hypothesis and re-visit that hypothesis and underlying methods after the analysis is completed.
4. Many, maybe most, things in the real world are not linear and are not independent of one another. Be careful that your base algorithms related to continuous data do not errantly default to independent relationships, normal distributions, and linear behavior.
5. Avoid the temptation to dive into the analysis too quickly
6. Be careful not to draw too many sweeping conclusions too early and do not be afraid to conclude that either the data is poor or the selected method/tool is not appropriate. One component of this is an experienced analyst, but another part is being able to admit that the output does not generate a desired result.

In this case, the use of a neural network provided many key insights including filling data gaps in categorical data based on patterns found by the computer that were not apparent to the human eye. Decisions using the categorical data from the analysis were meaningful.

With respect to continuous data, it was much less effective for several key reasons. Making decisions based on the results of the continuous data diagnostics and prognostics could have had dire results had the analyst relied on the "black box" without further evaluation of the results. As in many cases associated with big data, the issue turned out to be not as straightforward as originally hoped.

4 FORECASTING AND PROGNOSTICS

Forecasting, prediction, prognosis, and projection are words that are used synonymously. However, each word has a slightly different meaning and, to make matters more confusing, their definitions often have different meanings depending on the field of practice.

In accounting, forecasting is a process that analyzes past information, primarily financial statements, and presents a belief of future performance. A financial projection also presents a future belief, but the projection is not based on a series of historical data.

In statistics, a forecast implies the use of times series data where a prediction does not. Both prediction and forecasts involve the use of one or more dependent variables in relation to the desired state of a future variable. An example of a prediction is the median or the mean taken from a given sample of items that estimates the same parameter for a whole population of the same items. An example of a forecast is the value of a corporate stock in 10 years, based on a collection of past variables collected over some period of historical time. In forecasts, it is important that the variables used in the analysis are variables that cause the future state to happen.

In medicine, a prognosis is a forecast of the future course of a disease or disorder based on medical knowledge. A forecast is an estimation of a future condition, but may be based on time series and non-time series data (mixed or heterogeneous data). Similar to the medical profession's relationship with human health, the field of operations and maintenance (O&M) has adopted a similar use of prognostics to describe the future predicted state of physical asset (equipment) health.

A key concept in forecasting an uncertain future is the difference between causation and correlation. Correlation is a statistical relationship between two or more variables. An example may be that when one variable goes up, the other variable also goes up. Correlation is typically established through the evaluation of past data. In the world of data analytics, technical analysts comb through past data to determine which variables are more strongly related to a variable of primary interest.

However, correlation between variables does not automatically mean that the change in one variable is the cause of the change in the values of the other variable. One common example that is often used in the literature is related to weight loss. My weight loss (or gain) may be correlated to a number of things, including the rainfall in a small country in the southern hemisphere or the number of oranges eaten by a foreign dictator. While these factors may be correlated over some period of past time, it does not mean that the change in one will necessarily cause a change in the other in the future. In other words, in the past they are statistically correlated but in the future one does not cause the other.

This fundamental concept is often overlooked in big data analytics. "Just as the alchemists were not successful in turning base metal into gold, the modern researcher cannot rely on the 'computer' to turn his data into meaningful and valuable scientific information [6]." Best practice requires both an analysis of past data, or *posteriori*, with the opinion of experts to determine based on experience in terms of causation, or *a priori*. [7]. The accuracy of forecasts based on a handful of well understood variables usually outperforms forecasts with tens of variables that fully explain historical correlation. Understanding causation is the key.

The literature also indicates that forecasts are, at best, only about 80 percent accurate [8]. In addition to the skill and data used to create the model by the forecaster, another variable is uncertainty concerning the future, and namely the future relationship of variables that are the source of causation. In other words, uncertainty means that shorter term forecasts are typically more accurate than longer term forecasts.

Big data analysts often take exception with some of these facts. Big data analysts often state that with enough data and better-quality data that predictions get progressively better. This is indeed true, but with enough high-quality data there would be no uncertainty in the future, which is never truly the case. And if so, at what costs in terms of data collection. If big data analytics can make us clairvoyants, then why are we at "work" rather than resting on a yacht in the Caribbean?

Sellers of "black box" predictions of the future also take exceptions with the limitations of forecasting. Quite simply, their promise of false hope of eliminating uncertainty in the future is simply unfounded. Spending more money on more historical analysis has diminishing returns. Simple models usually outperform more complex ones.

This is not to say that more data is not better than less or that computers are not better at picking up trends more quickly than humans. Both are true, especially in the short-term. And in some fields, such as marketing, correlation is perhaps more important than understanding the underlying causation – reaching 8 million people instead of 10 million is still reaching 10 million people, and in the case of Wal-Mart, understanding why people buy more strawberry Pop-Tarts when severe weather is in the forecast may be of less importance than they simply do. However, in fields like human health, safety, and reliability, being "80 percent" accurate is not good enough for decision making [8].

4.1 Prognostic Techniques

For the purpose of this discussion, the term prognosis and

forecast are used synonymously to describe making future predictions over time of intended outcomes. In human health terms, a prognosis is the prospect of recovery from the usual course of disease or peculiarities of condition. From an engineering perspective, a prognosis is focused on predicting the time at which a system or a component will no longer perform its intended function. For the purpose of big data used for asset health forecasts and for resilience, or the ability to bounce back from distress, the term prognostics is used to describe the ability.

Prognostics techniques can take s variety of qualitative and quantitative forms. Some of these include scoring (used in many human health applications), data driven using statistical analysis, computer algorithms, or machine learning, and model-based approaches.

4.2 Probabilistic Approaches

The Monte Carlo method is the most common probabilistic method. It uses thousands to hundreds of thousands of simulations to develop outputs based on a set of known inputs. The frequency of certain outputs divided by the number of simulations is a percentage. Probabilistic methods based on Monte Carlo simulations use probability distributions to describe the variation and uncertainty of the input while traditional deterministic methods require the analysts to choose a single point estimate for each input variable. Because no data set is perfect and every point estimate is to some degree a matter of opinion, probabilistic approaches such as the Monte Carlo method capture a much more complete range of possibilities. They are regarded by the international risk standard, ISO 31000, as one of the top methods for quantifying risk and uncertainty.

The primary criticism of the Monte Carlo method is that it does not provide a closed form mathematical solution. In layman's terms, this means that unlike calculus, the results of a series of simulations are slightly different depending on the number of iterations and the randomness of the input parameters chosen for each simulation. However, modern computer power has enabled simulations totally 10,000 to 100,000 or more, which as a practical matter provides convergence and repeatable results.

Similar to neural networks that were discussed earlier, analysis based on scenarios predates WWII but really began modern use with the advent of the computer in the 1950s. Monte Carlo methods were central to the simulations required for the Manhattan Project (nuclear bomb). It is there that the modern method received its name from scientist Stanislaw Ulam. However, the mass use of the Monte Carlo method was not realized until the late 1980s and 1990s when computing power began to grow exponentially.

Closed form, calculus-based methods dominate most engineering and technical education curriculums and most university undergraduate programs teach deterministic single input to single output problem solving methods. Probabilistic approaches such as the Monte Carlo method remain a bit of a niche practice among risk analysts and decision analysts as well as engineers and scientists who do not believe that the nature of things is always linear.

4.3 Probabilistic Case Example

The cited case example is from a division of local government in Florida. A master plan was independently developed based on the combination of community growth needs and the future needs of existing facilities. The master plan yielded a forecast over a 20-year time period based on single point estimates, or a deterministic methodology.

The unit of local government realized that there was risk and uncertainty associated with the forecasted funding needs that were not well incorporated by the deterministic analysis. They commissioned a probabilistic analysis of the same data to potentially improve the risk and uncertainties associated with other decision making.

A probabilistic analysis of this type breaks down major systems into thousands of underlying asset and subsystems. Years of asset performance data, including work order history, non-invasive testing such as vibration monitoring, and qualitative evaluation, are combined to predict asset health, when renewal/refurbishment will be needed, and when replacement will be needed. This aspect of the analysis for asset health is similar to the same type of prognosis performed to address human health.

The other part of the analysis addresses the financial costs associated with renewal/refurbishment and replacement. Many variables are included as part of this analysis for each physical asset, including cost escalation, the time value of money, historical costs, replacement costs, equipment obsolescence, and underlying maintenance approaches including insourcing versus outsourcing. This aspect of the analysis is similar to time series financial forecasting.

Large sets of data are used to develop probability distributions based on past performance. Those distributions are used to forecast future activity and costs for the large number of existing assets and anticipated new assets that will be added. A minimum of 10,000 simulations are typically run to predict future system needs.

In this example, the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentile results from the probabilistic analysis were overlaid with the deterministic forecast. The evaluation indicated:

1. the short-term prediction of the deterministic forecast was similar to that in the 50^{th} percentile range of the probabilistic forecast over the first 4 to 5 years;
2. the medium-term prediction of the deterministic forecast was less than the probabilistic forecast, and there was an approximate 30% probability that more money would be needed during this time period
3. the long-term prediction of the deterministic model corresponded to approximately the 85th percentile for the probabilistic forecast, meaning that funding required by deterministic forecast was likely overpredicted.

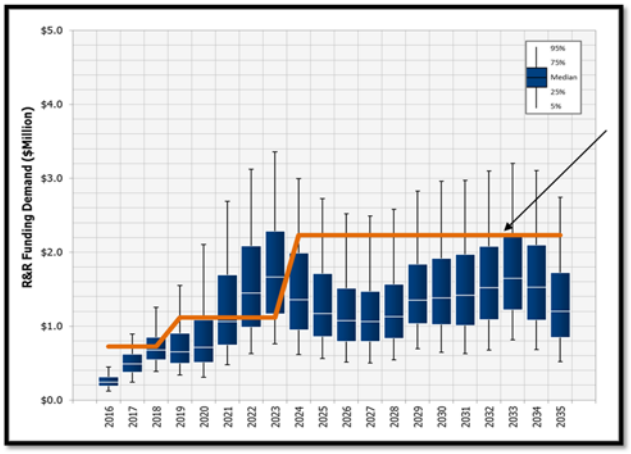

Figure 2. Deterministic forecast (orange line) overlaid on probabilistic forecast. Note potential underprediction in medium term and overprediction in long-term. Probabilistic analysis provides meaning for insights for complex decision making under uncertainty.

Many best practices are available for using Monte Carlo analyses. Like most big data, it involves data quality, data quantity, and applications based on analyst's knowledge and experience [9].

The probabilistic analysis yielded very meaning insights into large amounts of data. The probabilistic analysis created meaningful understanding of asset health prognosis and a meaningful understanding of associated time series associated financial forecast. It was considered essential for making decisions in the face of complexity and uncertainty.

5 CONCLUSIONS

All data has the potential to produce knowledge and improve decision making. This paper discusses several key aspects of big data, two key methods for its use in decision making, and two case examples.

We live in a time of unprecedented time of affordable access to vast quantities of data. Big data is a reality. Big data analytics will remain in vogue. However, poorly understood "black box" solutions driven by poorly understood algorithms should not be used in lieu of fundamental principles and sound statistical application. There are additional insights to be gained, but when it comes to quality decision making there are no short cuts.

Big data is improving our decision making in terms of problem diagnostics and short-term operational resilience. Decision making is much less improved in prognostics, prediction, and forecasting over the long term where uncertainty is a more significant factor. Uncertainty should be embraced, in the presence or absence of big data, and probabilistic methods offer the best opportunity for improving decision making in the face of uncertainty.

REFERENCES

1. JD Solomon, DA Vallero, K Benson (2017). "Evaluating Risk: A Revisit of the Scales, Measurement Theory, and Statistical Analysis Controversy." Proceedings of the 2017 Reliability and Maintainability Symposium (RAMS).
2. S.S. Stevens, "On the Theory of Scales of Measurement" Science, Vol. 103, No. 2684. June 7, 1946, pp. 677-680.
3. I. Ivrissimtzis, W.-K. Jeong, S. Lee, Y. Lee, and H.-P. Seidel (2004), Neural meshes: Surface reconstruction with a learning algorithm. *available at http://www.mpi-sb.mpg.de/ ivrissim/neural.pdf.*
4. Luke Dormehl (2014). The Formula, Penguin Books, New York, NY.
5. Palisade Corp., "@RISK and the DecisionTools Suite', accessed via http://www.palisade.com.
6. H.J. Einhorn (1972). Alchemy in the behavioral sciences. Public Opinion Quarterly, 36, 367-378.
7. Scott Armstrong (2012). Illusions in Regression Analysis. International Journal of Forecasting. 28. 689-694
8. JD Solomon (2017). Communicating Reliability, Risk and Resiliency to Decision Makers. Three Degrees of Learning.
9. JD Solomon and Adam Sharpe (2016). "Infrastructure Systems Renewal and Replacement Model Using Probabilistic Forecasting", Proceedings for the 2016 Reliability and Maintenance Symposium (RAMS 2016).

BIOGRAPHIES

J.D. Solomon, PE, CRE, CMRP
CH2M/Jacobs
3120 Highwoods Boulevard
Suite 214
Raleigh, NC 27604

E-mail: jd.solomon@ch2m.com

JD Solomon is an engineer, planner, facilitator, and consultant. He serves as a practice leader with Jacobs focusing on maintenance & reliability, asset management, financial management, strategic decision making, and master planning. He is the author of the book, "Communicating Reliability, Risk and Resiliency to Decision Makers." He is a Certified Reliability Engineer (CRE), Certified Maintenance and Reliability Professional (CMRP), is certified in Lean Management, and is a Six Sigma Black Belt. JD has a Professional Certificate in Strategic Decisions and Risk Management from Stanford, an MBA from the University of South Carolina, and a BS Civil Engineering from NC State University.