Cost Estimating Maturity and a Vision for the Future

Patrick Malone
MCR Federal, LLC
550 Continental Blvd. Ste 185
El Segundo, CA 90245
pmalone@mcri.com

Henry Apgar
Retired Cost Engineer
Thousand Oaks, CA 91360
hapgar@froniter.com

William King
MCR Federal, LLC
550 Continental Blvd. Ste 185
El Segundo, CA 90245
william.king@mcri.com

***Abstract*— Through the years, the science of cost estimating has matured. From the Pre-Modern to Meta-Modern periods, cost analysis advances were progressive. Point estimates were developed, and cost estimating relationships were common. Probability distributions and estimating to the mean added risk insight. With the ongoing explosion of data being generated in the digital space, the Internet of Things, and related advancements, we can integrate robust historical methods and tools developed by philosophers, mathematicians, scientists and the clergy to process this ocean of information leading into the Meta-Modern environment of decision making in real-time. We look at Post-Modern tools and methods like data analytics, machine learning, artificial intelligence and natural language processing, and how to apply them to enhance forecasting, exploiting and organizing data. These additional capabilities help estimators become more effective at incorporating obtainable knowledge into high-fidelity cost models and illuminate new ways of implementing legacy methods with more recent applications for advancing the science of cost estimating and decision making in the Meta-Modern period and beyond. Last, we discuss future research and trends.**

TABLE OF CONTENTS

1. INTRODUCTION

The science of cost estimating and analysis has matured from pre-modern to Meta-Modern (present) and multiple industrial revolutions. Many estimating strategies were driven by one of two types of reasoning, deductive and inductive (Figure 1). At present, these strategies are still utilized to solve complex problems; however, with the ongoing explosion of data being generated in the digital space, the Internet of Things, and related advancements, we can integrate robust historical methods and tools developed by philosophers, mathematicians, scientists and the clergy to process this ocean of information leading into the Meta-Modern environment of decision making in real-time.

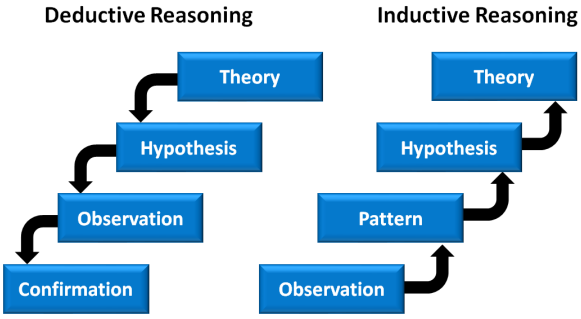

Figure 1 – Deductive and Inductive reasoning are still the basis of complex problem solving

We provide a brief history of method maturity, tool development and the people behind them through multiple eras and industrial revolutions, then introduce a vision for the future using advanced methods to embrace the exponential expansion of data and methods such as machine learning, artificial intelligence and natural language processing as they are applied in new ways within the Internet of Things (IoT)[a] framework and enhance the cost estimating community's capability to generate high quality and timely decision-making information.

The paper is organized into five main sections, 1) pre-modern, 2) early modern, 3) modern analysis, 4) Post-Modern and 5) Meta-Modern. Each period is also tied to technology, knowledge attainment, industrial revolutions,

[a] The Internet of things (IoT) describes physical objects (or groups of such objects) that are embedded with sensors, processing ability, software, and other technologies that connect and exchange data with other devices and systems over the Internet or other communications.

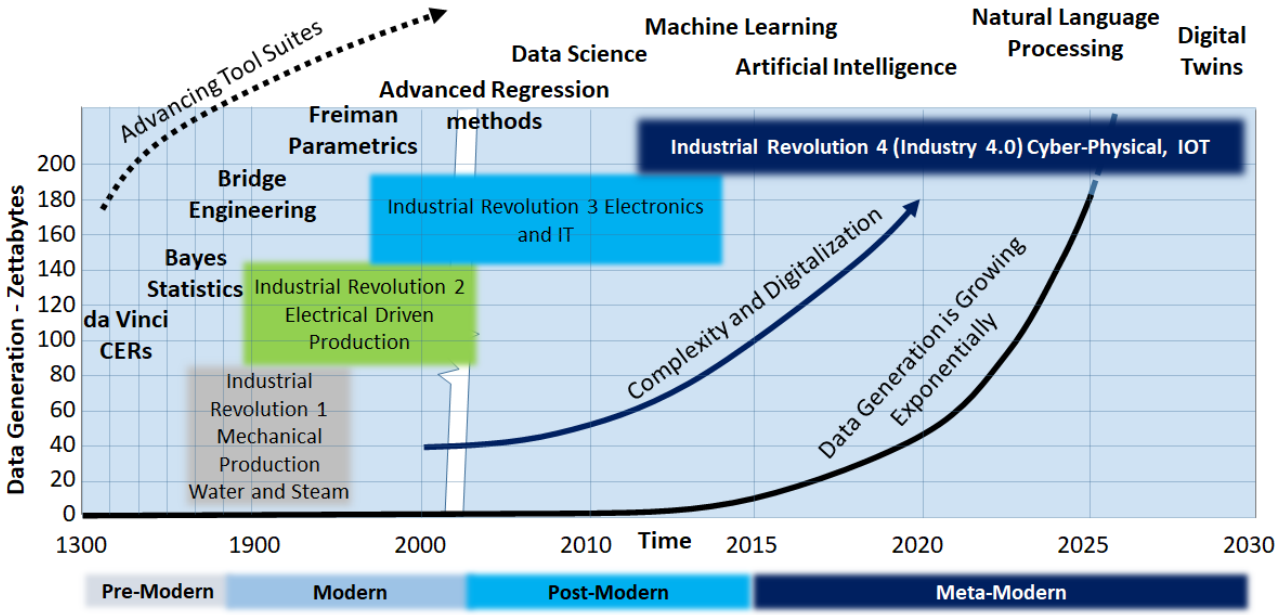

Figure 2 – Cost estimating maturity, method development, knowledge and data growth over the years

and data growth through the years. As time elapsed and industrial revolutions migrated forward implementing new technologies, new methods and tools have been developed out of need.

For example, early **cost estimating relationships** (CERs) were developed in sync with the first industrial revolution where entrepreneurs and war fighters needed to exploit the sudden availability of new technologies in ship building, bridge construction and railroad competition.

Figure 2 depicts the evolution of tools, technology and data across progressive periods and industrial revolutions. For another example, in 2020 Seagate and IDC stated that over 42 zettabytes of digital data were generated with a forecast of 175 zettabytes to be generated by 2025.[1]

2. Pre-Modern and Pre-Industrial Estimating

Overview

Early contributions to our estimating profession, while seemingly isolated solutions to immediate challenges, contributed to a unified and productive body of knowledge which is growing even today.

Early estimators and cost model builders weren't just accountants or pay masters; their occupations typically were in the military, politics, or civil engineering/logistics which thrived during the first industrial revolution, with widespread applications of mechanical production plus use of water and steam for power Some early estimators were entrepreneurs. [2] Most were motivated to develop focused estimating principles and methods because they had something more-immediate at stake – selling cargo ships, for example.[3]

Many medieval and later estimators typically had education and experience in mathematics and applied science (statistics, engineering, economics, et al). While they lacked our modern tools (computers) they used the tools at hand (mechanical or electric calculators, precise measuring devices, physical laboratories, etc.). They created a framework for similar techniques we apply today (learning curve, weight-driven cost relationships, mathematical simulation models, cost/benefit analyses, and probability analysis). Their approach, depending on their place in history, sometimes relied on the cost of standard units, standard ratios, or as derived from an algorithm typically based on size.

Note that early proposal estimating methods were driven by 1) outside civil and military requirements, including medieval church politics and modern US DoD standards, including the DoD "Green Book" of estimating standards, which introduced early discipline to estimating and 2) outside technological advantages from the first industrial revolution.[4]

The common goal was to satisfy immediate business applications – not scientific research; early cost models were typically unique to single products (ships, bridges, railroads), using mathematics and applied science. Estimators were likely classed as Logisticians (who typically were the project paymasters) and they were paid the same as an assistant architect. It is known that logisticians were educated at the Roman Lyceum.

Who Were the Pre-Modern Estimators?

Archimedes (287-212 BC) was a Greek mathematician, physicist, engineer, astronomer, and inventor from the ancient city of Syracuse in Sicily. He developed the concept of quantitative measurement, weight, and distance. Archimedes anticipated modern calculus and analysis by applying the concept of the infinitely small and the method of exhaustion to derive and rigorously prove a range of geometrical theorems, including: the area of a circle, the surface area and volume of a sphere, the area of an ellipse, the area under a parabola, and volume of an irregular shape; all were relevant to the later development of CERs based on size.

Diophantus (circ 275 BC), known as the father of algebra, published his Arithmetica comprised of numerical solutions of both determinate and indeterminate equations – the most prominent work on algebra in Greek mathematics. Algebra plays a critical role in the modern development of cost estimating models.

Leonardo da Vinci (1452-1519) exploited the science-driven inventions of quantitative analysis and early algebra to develop a sales-price CER in the first industrial revolution for contemporary Italian cargo ships based on their size, and capacity (Figure 3), and current methods of production. City-State Genoa derived its wealth from shipping; the citizens owned and operated shipyards – actually ship production assembly lines. They measured the actual labor and material costs at each station where planking, decking, deck houses, internal fittings, masts, and external rigging were added. da Vici, engineer and mathematician, established a process for developing a competitive **sales price** for each vessel. These early CERs are exhibited in the Vatican Museum.

Figure 3 – Italian cargo ships were built in Genoa

Isaac Newton (1642-1727) developed the binomial theorem, enhanced calculus, and laws of motion. The **binomial theorem** (or **binomial expansion**) describes the algebraic expansion of powers of a binomial, according to the theorem, it is possible to expand the polynomial $(x + y)^n$ into a linear combination involving terms of the form $x^b y^{n-b}$.

Calculus underpins optimization, which led to curve fitting and eventually artificial neural networks (which can utilize derivatives to update weights via backpropagation).

Thomas Bayes (1702-1761) (Figure 4) was founder of the Bayesian School of Statistics and had much in common with the estimators of today. He prepared a treatise for the Royal Academy on **parametric analysis**. An earlier manuscript read, in part, ", "*I am convinced that the Universe functions in obedience to some Divine Model…I am constrained to believe that among men of affairs, whose paths lie in fields of business and commerce, a means is needed whereby the units of monetary systems can be related to the products in the marketplace-…It should be possible to contrive a model by which money, skill, time spent in labor, and the fineness of the article crafted might be related."*

A key contribution of Bayes was to codify how to honestly comingle objective and subjective information. Under a Bayesian framework, Subject Matter Expert (SME) judgement probabilistically sets the distribution of prior beliefs, data is collected and evaluated against the likelihood of observing such a sample, and then prior beliefs are updated based on the data to form posterior beliefs. Bayesian approaches force you to translate subjectivity into a probability distribution (e.g., a prior mean and variance) and provides a principled framework to incorporate subjectivity into your analysis. Moreover, Bayes' contributions seep into Machine Learning through Bayesian Networks, Naïve Bayes Classifiers, and Bayesian Optimization, among others.

Figure 4 - Thomas Bayes

J. Carl Gauss (1777-1855) A German mathematician, developed the "least squares" or best fit method of CER development from cost and technical databases for economic models, in the form Y = A + Bx, in 1795. Adrien-Marie Legendre (1752-1833) expanded and published a similar work in 1805.

Charles Babbage (1792-1871) A mathematician, philosopher, inventor and mechanical engineer developed the first digital programmable computer which complex CERs could be developed and applied. Babbage originated the concept of the Analytical Engine, programmed using a principle openly borrowed from the Jacquard loom.

Isambard Kingdom Brunel (1806-1859), Bristol bridge builder (Clifton Bridge), ship builder (SS Great Britain), railroad builder (Great Western Railroad), and applied mathematician. Brunel was a pioneer parametrician, who evolved an elaborate series of **cost estimating relationships** (CERs) dealing with railroad car footprint, tractive power per unit of consumed fuel, and a custom parameter for rail "striction." Brunel ran a **cost/benefit analysis** based on cost per ton-mile and demonstrated a more cost-effective

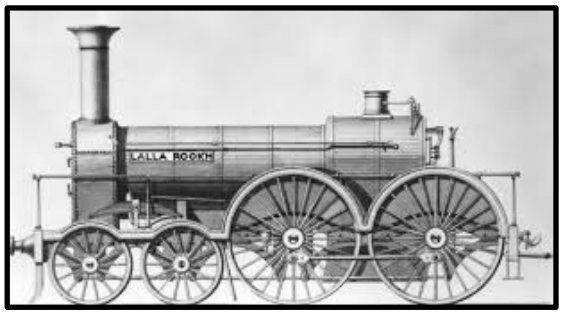

Figure 5 - Early Brunel locomotive for GWR and basis for early railroad CERs.

operation for the Great Western Railway (GWR) based on a new track width. Brunel declared that anything manufactured could be expressed in **monetary metrics per unit of weight** or size. (Figure 5) He later provided his railroad CERs to **Cyrus W. Field** (1819-1892) who was planning to lay the first submarine cable between Newfoundland and Ireland – a debatable idea at the time that CERs based on locomotive and railroad car weights from his limited railroad dataset from one product could be applied to another product altogether.

What Were Their Methods and Tools?

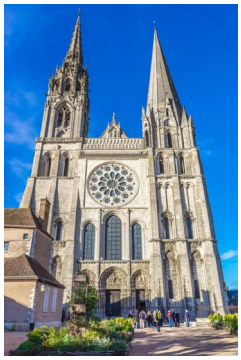

Figure 6 – Chartres Cathedral begun in 1145 estimated with standard units

French cathedral builders (1300s), developed a "standard unit" building method of estimating, defined by James [5] as a small section of a cathedral based on that portion which could be laid in ten man-days, including material. Chartres Cathedral, (Figure 6) for example, consists of 7,448 such units where one unit would cost $81,500 in 2018 US dollars. By this method Chartres Cathedral (begun in 1145 as a flamboyant gothic style cathedral with mismatched spires) would have cost **$607M** to build (in 2018 dollars) with limestone. By comparison, the 83,000 square foot Washington Cathedral (1907-1990) cost $65M in TY$. This included a $34M underground garage and two full sets of bells (carillons). Converting this then-year (TY) cost to fixed year (FY) 2018 cost would be **$665M**, or a cost today per square foot to approximately $8,010/sq ft.[6]

This early approach provides an example of analogy estimating for the building trade, using ratios and factors.

3. Early Modern Estimating (Industrial Revolutions 1 and 2)

Overview

The previous section described how early pioneers applied evolving math and statistical tools to meet specialized and immediate estimating requirements - how to predict production and deployment cost for *specific* applications (railroads, bridges, cathedrals, and ships) based on specialized data and experiences). These accomplishments were realizable as a consequence of the second industrial revolution which realized the benefits of electrical-driven production.

The following section describes what has been called the Golden Age of Estimating, with a modern focus on *generic* applications, implementing a wider range of data and experiences to deliver general-purpose models and databases which could then be calibrated to each user's specific products – during the fourth industrial revolution. Estimating concepts (based on small databases and logical relationships) were ripe for expansion with product complexity, digitalization of tools and techniques (computer), and the sudden availability of vast cost, schedule, and technical databases (probably as a result of WW2 technology).

This section describes generic applications in parametric estimating methods, the application of statistical estimating, general acceptance of parametric estimating methods, and the role played by professional societies in the development and acceptance of modern estimating methods.

Who were the early estimating pioneers? What were their goals?

Frank Freiman an RCA cost estimator, statistician, industrial engineer, and WW2 logistics officer was seeking a generic (but disciplined) parametric approach to engineering estimating of a universe of defense and commercial products **(1962-present)**. He developed a set of general-purpose, non-

Figure 7 - Frank Freiman receives Honorary Director Certificate from Bryant Barnes, first ISPA President, at close of Charter Meeting in Washington DC (1979).

proprietary CERs (typically relating development, production, and logistics cost to size, complexity, percent (%) new/reuse, production quantity, and elapsed time) which could then be calibrated to a specific user's own products and processes. Originally developed for RCA internal use, but later applied by DoD, PRICE H for hardware products (1975) and PRICE S for software products (1977) were released commercially. Frank encouraged the expanded and disciplined development of recently-available cost, schedule and technical user databases for calibrating the generic models for special-purpose applications. Other commercial general-purpose models were developed by Dan Galorath (SEER), Capers Jones (software productivity), Randy Jensen (SEM), Larry Putnam (SLIM), David Novick (Aircraft production), and Barry Boehm/USC (COCOMO). At this time new parameters came into use including: function points and equivalent software lines of code (ESLOC).[7]

RCA PRICE Systems unified international parametricians with their PRICE Model Users Group, which morphed into the first independent professional society (1979) devoted to the acceptance of parametric estimating methods by DoD - the International Society of Parametric Analysts (ISPA). For his active role in ISPA's development, **Frank Freiman** was honored by ISPA as its Honorary Director, as shown in Figure 7.[8]

David Novick (1930s-1960s) of RAND Corp, used statistical estimating post WWII, along with NASA and USAF where they studied multiple scenarios concerning how the US should proceed into the age of jet aircraft, missiles, and rockets. The military saw a need for a stable, highly skilled, cadre of operations researchers (OR).[9]

In 1950, **the RAND Corporation** established its Cost Analysis Department for the purpose of analyzing weapons system costs using OR methods developed during WW2. An early challenge was to identify the "elements of cost," later to become the DoD standard work breakdown structure (WBS).[10]

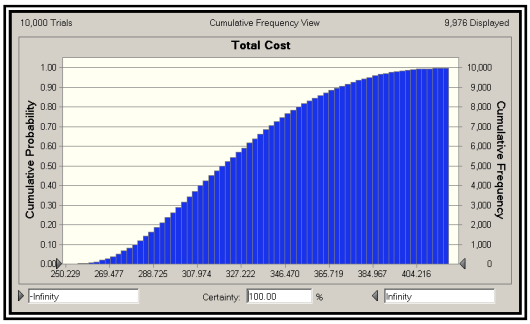

Figure 8 – OR methods support cost risk analysis based on historical experience. This curve explains the cumulative probability (estimating confidence) or each possible cost estimate.

Once cost elements became common across multiple models, cost drivers could be defined, and cost risk methods were accepted, (Figure 8) the RAND analysts focused on developing mostly-aircraft generic cost estimating relationships (CERs); the term parametric cost estimating became common by 1952. Novick developed the concept of "cost considerations in systems analysis." And cost estimators, searching for identity and status in an increasingly complex technical world, instantly became systems engineers. Subsequently, RAND found it necessary to

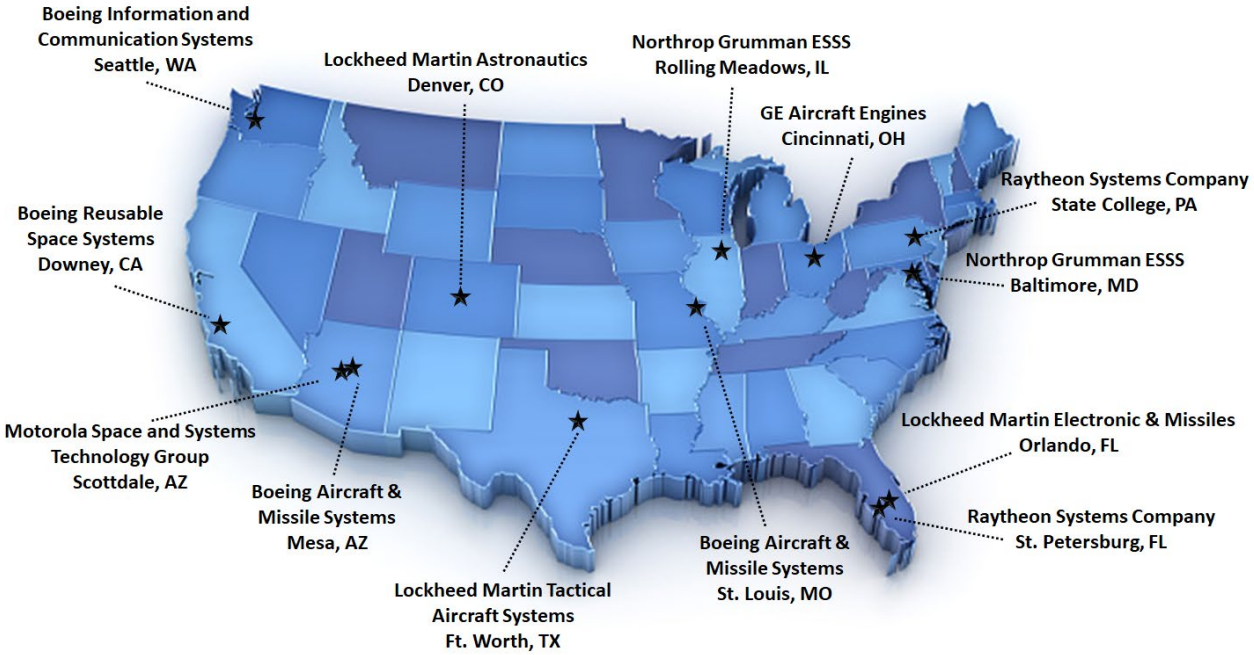

Figure 9 - Parametric Reinvention Laboratory Teams were located across the United States (1994).

differentiate between one-time outlays - non-recurring expenses for investments (development), recurring expenses for procurement (manufacturing) and maintenance expenses for logistics (operations). David Novick retired in 1971; Their work was expanded under other estimating scholars, including Steve Book of The Aerospace Corporation and later of MCR.

Monte Carlo Statistical Estimating Predictions were originated by Los Alamos National Laboratory (LANL) while estimating nuclear bomb yields. Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. It is a technique used to understand the impact of risk and uncertainty in prediction and forecasting cost estimating models.

The first known reference for using Monte Carlo methods to solve particle transport problems on computers was reported by LANL in 1947.

Parametric Cost Estimating Initiative (PCEI - 1994): a joint government-industry endeavor to sanction and encourage parametric (top-down) estimating methods for federal proposal estimating, the Defense Contract Audit agency (DCAA) issued new top-down estimating guidelines based on already-accepted bottoms-up government estimating rules. Figure 9 is a mapping of the United States Parametric Reinvention Laboratory Teams (circa 1994).

Professional Estimating Societies (1960 – present): were dedicated to the development of cost estimating tools and methods, as well as the practice of free and open discourse (papers and workshops) by learned government and commercial estimators. Some professional societies created discussion platforms for development of focused WBSs which were eventually adopted and integrated as government standards, i.e., MIL-STD-881. Figure 10 illustrates the expansion and consolidation of professional societies over the recent decades as cost analysis changed to meet industry and government needs.[b]

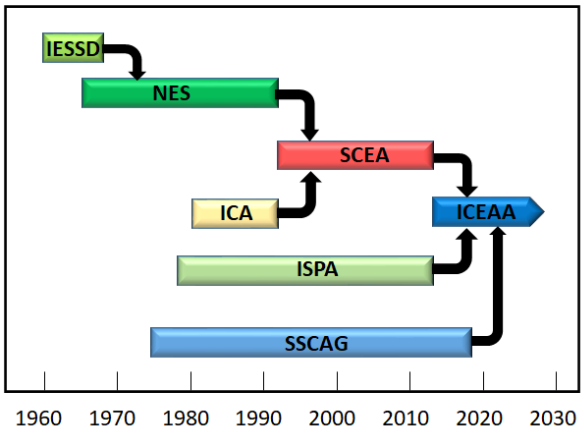

Figure 10 - Consolidation of estimating societies in America reflects the evolving methods employed by cost estimators.

What Were Their Methods and Tools?

Government cost estimating centers (1980s to present): The Air Force Center for Cost Analysis (AFCCA), Naval Center for Cost Analysis (NCCA), Army Cost and Economic Analysis Center (CEAC), DoD Cost Assessment and Program Evaluation (CAPE), and their role in standardizing the government-environment has enhanced cost estimating discipline. These centers have also developed and provided non-proprietary versions of very large cost/technical databases such as the Unmanned Spacecraft Cost Model (USCM), the NASA/Air Force Cost Model (NAFCOM), and Air Force Space and Missile Systems Center (SMC) Software Database which have proven useful in applying general purpose models.

Spreadsheet tools (1980-present): they were developed to satisfy the immediate need for a mathematical framework for estimating the labor hours of project development for the newly available personal computer and the expansion of general-purpose estimating models and widely available cost estimating relationships (CERs). Spreadsheet programs such as VisiCalc, SuperCalc, Multiplan and Lotus 1-2-3 were advanced enough; they just lacked the user-friendly nature like many operating systems at the time. But, they promoted growth in general purpose estimating tools.

4. Modern Analysis (Industrial Revolution 3)

Overview

Over 100 years ago H.G. Wells observed that "The great body of physical science, a great deal of the essential fact of financial science, and endless social and political problems are only accessible and only thinkable to those who have had a sound training in mathematical analysis, and the time may not be very remote when…for [the] complete initiation as an efficient citizen…it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write." While not explicitly mentioning statistics, Wells' often misquoted observation clearly stresses the importance of statistical thinking and procedures in various financial and business situations, and in fact, many of the emerging techniques during this period are routinely used now.

[b] Society abbreviations:
IESSD – Industrial Estimating Soc of San Diego
NES – National Estimating Society
SCEA – Society of Cost Estimating and Analysis
ICA – Institute of Cost Analysis
ICEAA – Institute of Cost Estimating and Analysis Association
ISPA – International Society of Parametric Analysts
SSCAG – Space Systems Cost Analysis Group (working group)

Descriptive Statistics and Statistical Inference

Descriptive statistics provides a concise and efficient means to summarize and communicate data. Commonly providing a measure of central tendency (e.g., a mean of median) and dispersion (e.g., a standard deviation), descriptive statistics often serve as the foundation for inferential statistics. Whereas descriptive statistics seek to merely describe a sample, inferential statistics seek to describe an underlying population from a sample. Drawing from a pre-existing body of knowledge (e.g., prior research, literature, etc.) or subjective beliefs, a null hypothesis can be formulated to test its compatibility with newly observed data. After finding sufficient evidence of incompatibility between the data and our null hypothesis, we can reject our status quo assumption in favor of its logical complement. Univariate prediction intervals provide a basic example of applying inferential statistics to determine ranges of likely outcomes for future events.

Regression Models

Expanding on descriptive statistics, regression models use organized raw data to develop relationships between two or more variables using linear or nonlinear equations.[11] With a pedigree tracing back to Da Vinci CERs, regression models explicitly quantify the relationship between variables and provide a statistical framework to assess the statistical significance the observed relationship (i.e., whether the observed relationship would be expected to occur by random chance, or not). Historically, linear least squares has served as the starting point to investigate the relationships between historical data. This is primarily due to the fact that there is a closed form solution to the linear least squares estimator (meaning we do not need a computer to numerically minimize the sum of square errors) and the widely held notion that linear models provide a reasonable approximation (at least locally) to complex or non-linear phenomena. A rich statistical literature (e.g., the Gauss-Markov theorem) has firmly established the assumptions and conditions (e.g., additive errors) for linear least squares regression models to serve with minimal sampling variance and how to probabilistically model predictive uncertainty. In many cost estimating problems, errors follow a multiplicative pattern, wherein the errors are proportional to the magnitude of the predictive variables. In such cases the variables are transformed (typically with natural logarithms) to convert multiplicative errors into additive errors and preserve the benefits of the linear least squares estimator; however, such transformations introduce bias when converting the transformed equations into unit-space.

More recently, alternative regression techniques have been developed to eliminate the bias introduced from variable transformations, which include the Minimum Unbiased Percentage Error (MUPE) and Minimum Percentage Error under Zero Percentage Bias (ZMPE) techniques. Whereas MUPE eliminates bias from applying iteratively reweighted least squares, ZMPE eliminates bias through a constrained optimization process. All regression models are driven using existing data support a deductive reasoning approach and are commonly utilized to solve forecasting problems in business applications.[12]

Factor Analysis

Factor analysis is a multivariate statistical method that utilizes the covariance structure of observed variables to model the data in terms of a smaller set of unobserved (i.e., latent) variables called factors that originated with "early-20th-century attempts of Karl Pearson, Charles Spearman, and others to define and measure intelligence."[13] The basic idea behind factor analysis is that variables can be grouped in a manner such that correlation is high between variables within a group and low between variables in another group; in such cases, there is a latent "factor" that causes the observed values. Factor analysis is closely related to Principal Components Analysis (PCA). Whereas PCA seeks to find linear combinations of the observed variables that maximize variation while remaining orthogonal, factor analysis seeks to find uncorrelated factors whose linear combinations best explain the observed data. In both cases, the dimensionality of the data is reduced; however, factor analysis provides rotation techniques that help elucidate latent factor interpretation. Notwithstanding factor rotation, factor analysis is generally an exploratory or descriptive procedure that requires subjectivity in the interpretation of factors.

For example, El-Choum developed a Factor Analysis Model (FAM) to identify key factors that influence a cost estimate at a project's bid phase.[14] The first step in the development of a model was identification of all possible parameters that contribute to cost overruns. By examining the correlation matrix, 8 variables were identified that contributed to cost overruns. The intent was to enhance and improve the quality of budgeting for cost estimates in general and is to facilitate the decision-making process in providing a more realistic and practical representation of the construction industry. Based on model results, it was found there are key parameters that significantly affect cost overruns. They are: 1) lack of supervision; 2) scope changes; 3) inexperienced management; 4) design changes; 5) improper supervision; and 6) feedback procedure. The model can be used to investigate the effect of potential variables prior to award, so that corrective action can be taken to adjust the cost estimates for a particular project.

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is used to quantify the degree of association between two sets of multivariate data (not necessarily with the same dimensionality) and originated with Harold Hotelling's desire to relate measures for arithmetic and reading abilities in the first half of the 20th century.[15] CCA identifies linear combinations from each set of data that achieve the largest degree of correlation to each other. From another perspective, CCA represents a means for supervised dimensionality reduction; effectively, we compress the multivariate data into linear combinations while

ensuring a high degree of association with the outcome variables of interest.

For example, CCA lets us relate a set of technical requirements (e.g., cost drivers) to measures of performance (e.g., cost and schedule). The linear combinations of technical requirements derived from CCA can be thought of as complexity measures, since they are designed to scale linearly with linear combinations of performance outcomes.

Discriminant analysis

Discriminant analysis, or linear discriminant analysis, (LDA) is a statistical technique used to classify observations into non-overlapping groups, based on scores of one or more quantitative predictor variables.

This technique has been here for quite a long time. First, in 1936 Fisher formulated linear discriminant for two classes, and later on, in 1948 C.R Rao generalized it for multiple classes. LDA finds linear combinations of the quantitative predictor variables that best separate observations into pre-defined groups or classes. In this sense, LDA can also provide supervised dimensionality reduction that results in a set of variables that best separate between classes.

Discriminant analysis is a valuable tool and has gained widespread popularity in areas from marketing and finance to facial recognition. There several reasons for this; when its underlying assumptions are met, LDA is more accurate than logistic regression[16] and can operate well with small sample sizes.[17]

Similar to it is archetypical application in predicting firm bankruptcy, LDA can be applied to predict whether commonly experienced risks are likely to be realized and impact cost or schedule. While classification techniques such as LDA have not been commonly utilized in the cost estimating field, we see ripe application in risk analysis.

Time Series

Time series has also been around for many years and is still a popular analytical method. It is a specific way of analyzing a sequence of data points collected at regular time intervals that are serially correlated. The use of time series analysis provides analysts consistent data over a set period of time[18]. The benefit to using this method is that the periodic data can show how key elements change over time; it shows the direction data is trending and supports statistically rigorous forecasting. Although time series analysis requires a significant number of observations, the large set of observations provide additional insights including: patterns, trends, seasonality and unusual trending. Classically. variants to the simple Autoregressive Integrated Moving Average (ARIMA) or Vector Autoregression (VARs) models are applied to account for the correlation between observations across time and the interaction between endogenous and exogenous predictors. For example, using time series data for a project using earned value management provides performance trending, early warning, and areas for corrective action.

Decision Trees

Decision trees represent a series of splitting rules (i.e., logical if-then statements) organized into a flow-chart. Beginning with the single parent node (or root) and following the logical splitting rules (e.g., if TRUE go Left, else go Right), will results in a single terminal node (or leaf) that captures the predicted value or class for an observation. (Figure 11) To learn these splitting rules, we find the optimal segmenting criteria that partitions the predictor space into regions and simply forecast the average (or mode) of the response variable for the predictors within each region.

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. This is very popular and highly interpretable learning algorithm. While it often underperforms against other machine learning methods it is still useful for obtaining first order decisions.

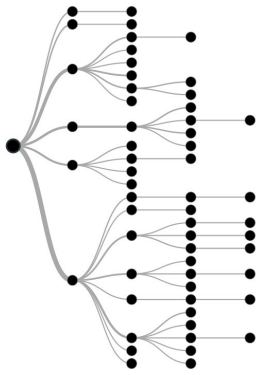

Figure 11 – Decision trees provide a straightforward method of decision making

Other benefits of using decision trees are: domain knowledge is not required; they are easy to comprehend; the classification steps of a decision tree are very simple and fast. However, without rigorous cross-validation to prune away unnecessary splitting rules, decision trees will tend to overfit their training data will not generalize well. A well-pruned (i.e., simple) decision tree model will generally under-perform relative to other machine learning or regression approaches but is makes up for the lack of accuracy in terms of a higher degree of interpretability.

5. POST-MODERN (INDUSTRIAL REVOLUTION 3)

Overview

In this era of processing and memory intensive methods, data science begins to flourish. With cheap memory and computing costs and an explosion of data in the digital environment, we apply more greedy, computationally intensive algorithms that have been out of reach such as K-Nearest Neighbors, K-Means, Neural Networks, Fuzzy Logic, and Evolutionary Programming. Historically, with limited computing power, we were forced into simplified

solutions. Now however, with advanced computers and processing power, almost any type of analyses is possible.

Cross-Validation

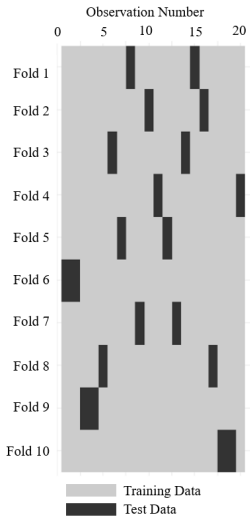

Figure 12 – Clustering, K-means and KNN provide flexibility in machine learning and AI

With advances in computing power, comes the ability to apply resampling methods to assess out-of-sample predictive performance. While there are many variants of Cross-Validation (CV), k-fold CV remains one of the most popular. k-Fold CV begins with randomly partitioning the training data into k non-overlapping sets (called "folds"). Sequentially, each fold is held back as an unobserved test set, while the remaining (k-1) folds are used to train a model and predict the held back set. Performance metrics are then calculated for each held back set. (Figure 12) The CV performance of a model against each of the k-folds represents a how the model would have predicted those data points had they been unobserved at the time the model was generated. Comparing CV performance statistics (e.g., Root Mean Square Error (RMSE)) to those calculated on a model learned from the entire training data (i.e., the classical approach to model development) yields insight into overfitting and give an indication to how well the model will generalize to predict new data. As an added benefit, CV can be utilized to perform principled feature selection and hyperparameter tuning for machine learning models.

Clustering

Cluster analysis is an unsupervised statistical method for processing data. It is unsupervised in the sense that there is no response variable (e.g., cost) that needs to be specified or collected. It works by organizing items into groups, or clusters, on the basis of how closely associated they are under a variety of (dis)similarity criteria. The objective of cluster analysis is to find similar groups of objects, where "similarity" between each pair of objects is defined by some global measure over the whole set of characteristics. This is a key analytical method for data mining.

While the number of clusters is not usually known, *a priori*, there are a number of techniques to determine the optimal number of clusters for a given dataset (e.g., the Elbow Method, Silhouette Distance, or Cross-Validation). Once the analysis is performed, it provides information about where associations and patterns in data exist, but not what those might be or what they mean.[19] As a result it is an initial step in the analytical chain. For example, we want to know what transponder to choose for an upcoming spacecraft design. Comparing the data from a list of manufacturers, we want to know the best reliability versus cost. We might compare the number of space flights, quantity of parts within the unit, historical failure modes, etc. These are some of the clusters that could be configured to perform the trade.

One of the most commonly used algorithms in machine learning is k-Means clustering. While sometimes confused with k-Nearest Neighbors (kNN) due to the presence of the k letter, they are different methods k-Means clustering is a centroid based unsupervised algorithm where distances between points in a specific cluster(s) are used to determine the multivariate "center" or centroid mean of the cluster. In contrast, KNN is a supervised learning algorithm used for classification or regression of data which needs labelled data to train on. Figure 13 shows examples of these methods. kNN makes predictions based on the most common class membership or average of the response variable of the k nearest neighbors to a new data point. Similar to unsupervised clustering algorithms, kNN, defines nearness with a variety of (dis)similarity measures (e.g., Euclidean, Manhattan or Minkowski distance).

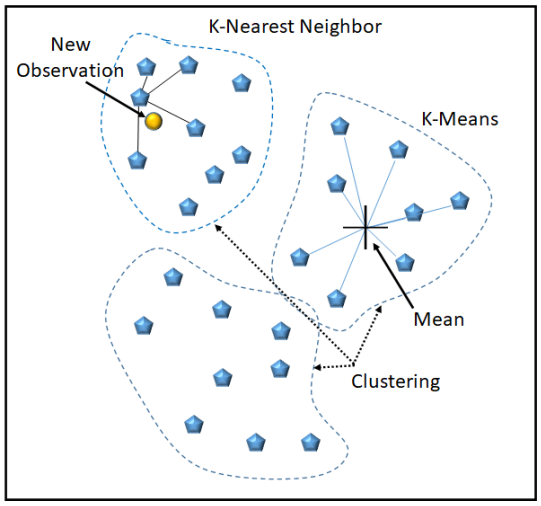

Figure 13 - Clustering, K-means and KNN provide flexibility in machine learning and AI

Neural Networks

Artificial Neural Networks (ANNs), often just called a "neural network" (NN), are a subset of machine learning and present a brain metaphor for information processing. These models are biologically inspired computational models (algorithms) and consist of interconnected groups of artificial neurons. As a result, they quickly process information and have the unique ability to extract meaning from imprecise or complex data to find patterns and detect trends that are too convoluted for the human brain or for other computer techniques (Figure 14).[20,21]

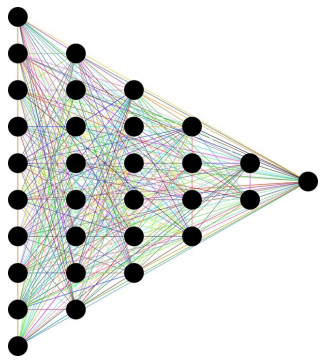

Figure 14 – Neural networks provide an efficient approach to pattern recognition

ANNs can be made to be arbitrarily flexible and serve as a class of universal function approximators. To achieve this flexibility, they require learning a large number of weights, which in turn, requires a large amount of data. The example in Figure 14 begins with 10 input variables and maps them to an initial hidden layer of 8 neurons that requires learning 80 weights. These 8 neurons then map into a second hidden layer of 6 neurons (requiring 48 weights) which maps into a third layer of 4 neuron (which requires 24 weights) which maps into a fourth layer of 2 neurons (which require 8 weights) which finally map into a single layer (requiring 2 weights). In total, this ANN needs to learn 162 weights which requires substantially more than 162 data points for meaningful results. While there are no clear-cut choices for the number of hidden layers, nor the number of neurons per hidden layer, cross-validation can assess whether ANNs overfit their data and provide a principled mechanism for selecting the ANN structure from a grid of plausible choices.

The application of neural networks in data mining is very broad. With sufficiently large data sets, they have a high acceptance ability for noisy data while delivering a high degree of accuracy. Neural networks according to Ramos, have been shown to be very promising systems in many forecasting and business classification applications.[22]

Fuzzy Logic

Fuzzy logic is related to probability theory, it is a technique for representing and manipulating select information by attaching numeric values between 0 and 1 to each proposition in order to represent uncertainty. Once assigned, the data is manipulated in a way to measure how likely a proposition is to be correct and fuzzy logic measures the degree to which the proposition is correct. For example, the proposition 'President Obama is young' may have a degree of correctness 0.8 [23] Fuzzy logic methods are a key ingredient of artificial intelligence. Another example where, Chiang and Jyh-Shing implemented fuzzy logic algorithms on the Cassini Spacecraft attitude control system in the 1990's the result was an optimized solution to minimize fuel use.[24]

Evolutionary Programming

Evolutionary programming is a stochastic optimization strategy similar to genetic algorithms with an emphasis on behavioral linkages.[25] It was first used by Dr. Lawrence J. Fogel in 1960 in order to use simulated evolution as a learning process aiming to generate artificial intelligence.

This method is typically used to provide good approximate solutions to problems that cannot be solved easily using other techniques (as previously discussed). Many optimization problems fall into this category. For instance, it may be too computationally-intensive to find an exact solution but sometimes a near-optimal solution is sufficient.

NASA is using evolutionary programming to conceptualize and optimize communication solutions. For example, they may want to know how to get the best signal return from an antenna on Mars back to earth. Using evolutionary programming, the spacecraft lander will search to optimize the signal.

In another example, Moraglio set up a programming problem to using polygons and wheels to run on a terrain. Using evolutionary programming, Figure 15 provides the evolutionary solution.[26] In these situations, evolutionary techniques can be effective. Due to their random nature, evolutionary algorithms are never guaranteed to find an optimal solution for any problem, but they will often find a good solution if one exists.

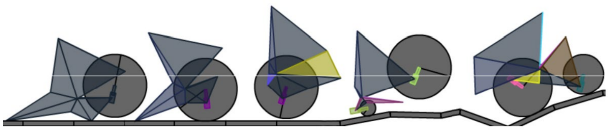

Figure 15 – Evolutionary programming accelerates solutions to complex problems using abstract constructs

6. META-MODERNISM, (INDUSTRIAL REVOLUTION 4 AND BEYOND)

Overview

Meta-Modernism and the fourth industrial revolution are centered in a digital transformation shaping how researchers, technologists and business operations are transforming and adapting to the environment. It is the era of connectivity (of everything) using high fidelity sensors that get fed back to users through the data ocean. Big data analytics is growing in lock step with data growth. This section provides maturing tools and methods to leap into the future. Under many names, data science had been around for decades. Within it is artificial intelligence, machine learning, natural language processing as well as the classical descriptive statistics and regression. The use of deductive and inductive reasoning support complex problem solving. The difference now is the rate or velocity data is being generated and analyzed across sectors and industries. The methods in this section are the steppingstones into the future and the next step for cost estimating and analysis.

Table 1 - Two Fundamental Machine Learning Methods and key attributes.

Types	Supervised	Unsupervised
What does it do?	Learn a mapping from inputs to outputs (i.e., how to predicts outcomes)	Identify clusters, association or anomalous data and reduce dimensionality
What kinds of data does it require?	Labeled training data (i.e., a response variable in addition to predictors)	Training data (a response variable is not required)
What kinds of Algorithms can be applied?	Linear, non-linear and logistic regression, LDA, Decision Trees, Random Forests and Gradient Boosted Trees, Support Vector Machines, and ANNs	K-Means and Hierarchical Clustering, Gaussian Mixture Models, Principal Components Analysis (PCA) and Autoencoders
What are the key limitations?	Needing to balance the Bias-Variance tradeoff (Underfitting and Overfitting)	Interpreting clusters requires human intervention and no guarantee of meaning

Data Science

Data science is the continuous process of collecting large data sets to observe, form and test models (or assumptions). This mix of programs, statistics, domain expertise, calculations and visualizations help identify hidden patterns and trends in data to support informed decision making and predictive modeling activities.[27]

It is "The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it. That's going to be a hugely important skill in the next decades…"[28]

Artificial Intelligence and Machine Learning

Artificial intelligence (AI) and AI systems are used to perform complex tasks in a way that is similar to how humans solve problems. Machine learning (ML) is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior.

AI is a collection of programs and algorithms that uses ML and deep learning to simulate human intelligence. AI algorithms can learn, reason and self-correct to perform autonomous actions to solve problems as they continue to capture more data. AI can translate, understand speech, and make decisions self-sufficiently to automate repetitive tasks or strategically support product and service advancements.

ML is a method consisting of algorithms to look at data sets and identify patterns, then, using those insights better understand and complete its assigned task. While there are two major types of ML, supervised and unsupervised, that have particular relevance to cost estimation, reinforcement learning offers many applications for real-time decision support (e.g., self-driving cars). Supervised learning involves learning a mapping between training data and labeled outputs, whereas unsupervised learning discovers relationships based solely on the data itself. With reinforcement learning, algorithms learn what actions to take within an environment that maximize a reward function over time and that alter its state within the environment (which potentially change what actions are available).

Table 1 provides some key attributes of the primary ML methods.

Supervised learning is performed with a labeled set of data (labeled meaning it includes a dependent, or response variable). It can be used to address regression (i.e., where the output is a numeric value – e.g., cost) and classification (i.e., where the output is a category – e.g., "successful") problems. When applying supervised learning algorithms, the primary goal is to infer a mapping from the predictor variables to the response variable that will perform well (i.e., generalize) for new, unobserved data. The key to achieving this goal is finding a balance between the bias-variance (i.e., complexity vs accuracy) tradeoff. As ML models become more complex, they tend to overfit (i.e., essentially memorize) the training data and are unable to predict well against new observations (e.g., fitting a high-degree polynomial to a small dataset); conversely, insufficiently complex ML models will tend to underfit the training data (e.g., fitting a simple average to data that clearly follows a linear trend). Luckily, cross-validation offers a principled and effective means to find the optimal tradeoff between complexity and accuracy. During CV, overly complex model will tend to overfit the training data in each fold, and poorly predict with the held back test set; similarly, insufficiently complex models will poorly predict both the training and held back test set in each fold. Simply selecting the ML model (or parameterization) with the best CV performance is usually effective at optimizing this tradeoff. Figure 16 shows a number of methods trading accuracy with interpretability. For example, Decision trees are typically easier to interpret while neural nets are more accurate.

As our community considers adopting highly accurate, yet uninterpretable methods, we can temper our historical distrust of "black-box" models with techniques that explain

the predictions and provide human intelligible reasons for what drives individual predictions. For example, Locally Interpretable Model-Agnostics Explanations (LIME) make small perturbations to the inputs for a single data point and records how those perturbations change predicted output. Applying linear regression to the set of perturbated inputs and corresponding outputs yields a highly interpretable summary of how the inputs variables influence a particular prediction.

Unsupervised learning is performed with unlabeled data. The algorithms extract patterns and relationships. While there are numerous ways to implement unsupervised learning, there are several common methods; 1) Clustering, in which the algorithms are programmed to find similar data points within a data set and group them accordingly (as previously discussed, 2) density estimation, this method uses algorithms to look at how a data set is distributed and any associated patterns, 3) anomaly detection, where the algorithms search for data points within a data set that are significantly different from the rest of the data, and 4) principal component analysis (PCA), which can be utilized to reduce the dimensionality of the data.[29]

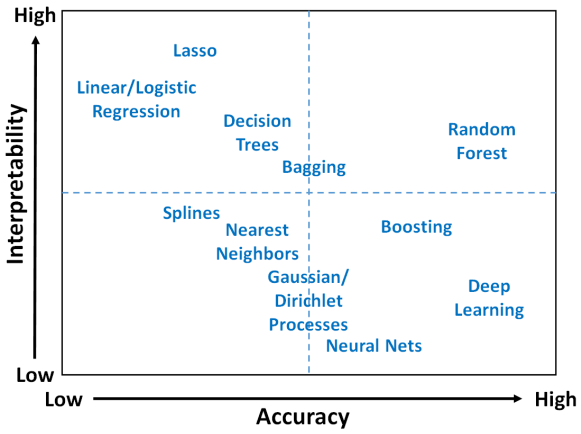

Figure 16 – Notional Machine Learning Algorithmic Trade-Offs using different statistical methods

Natural Language Processing

Natural Language Processing (NLP) is a branch of AI whose purpose is to understand human generated text and spoken words. It is a set of techniques and algorithms that combine approaches from statistics, ML and others to enable models to process and find meaning within unstructured text. A possible future tool for cost estimating is one that can translate written requirements into a set of WBS -oriented cost drivers that ultimately can support a complete cost estimate.

NLP techniques can be used to extract and exploit volumes of narrative data associated with cost estimation. For example, applying NLP techniques for tokenization (converting every word or character in a sentence into "tokens" or variables) and part-of-speech tagging (labeling nouns, verbs, etc.), analysts could extract action verbs from a set of functional requirements for a software development effort and then apply simple function point counting to automatically generate an early cost estimate for the effort. Moreover, utilizing other NLP techniques for creating document term matrices or "word2vec", an analyst could quickly identify similar or redundant requirements.

Digital Twins

A digital twin is a virtual model designed to accurately reflect a physical object. The object being is outfitted with various sensors related key areas of functionality to provide input to a virtual model that can then be used to run simulations for studying actual performance and synthesize possible improvements or troubleshoot issues, leading directly back to the physical specimen.[30] The idea of digital twins was conceptualized in the early 1990's and further developed coined by NASA in 2010[31]

The benefits are boundless and support better research and development, higher efficiency products and product disposal strategies. Today the use of digital twins is expensive with limited use on large scale products or projects. Industries currently engaged in this technology include: systems engineering, auto manufacturing, production aircraft and others.

However, the future is limitless for this technology. Researchers foresees a digital reinvention that will disrupt operating models with artificial intelligence and other enabling technology to increase the cognitive ability of the models and then implement it into the physical asset.

7. A Vision for the Future

Key Points

The data ocean is growing at an exponential rate. As a result, evolutionary and advanced methods, tools and techniques will continue to advance and mature to keep pace to allow for sorting, characterization and pattern recognition. The resulting data lakes can then be evaluated based on industry specific criteria allowing analysts and subject matter experts (SME) to select the appropriate analytical frameworks to set up models and perform complex analyses that will result in providing actionable information for stakeholders.

Methods Summary

We discussed numerous methods to perform data analysis. Today many of these methods are still used, some for general applications others for specific purposes. As the newer methods are implemented in the big digital data ocean, more capable models will be developed to provide disruptive methods to problem solving.

Key drivers to this disruption are the five tenets of big data to support value added solutions[32]:

Volume – The amount of data produced is increasing (Figure 2), At smaller scales it is now measured on a petabyte scale

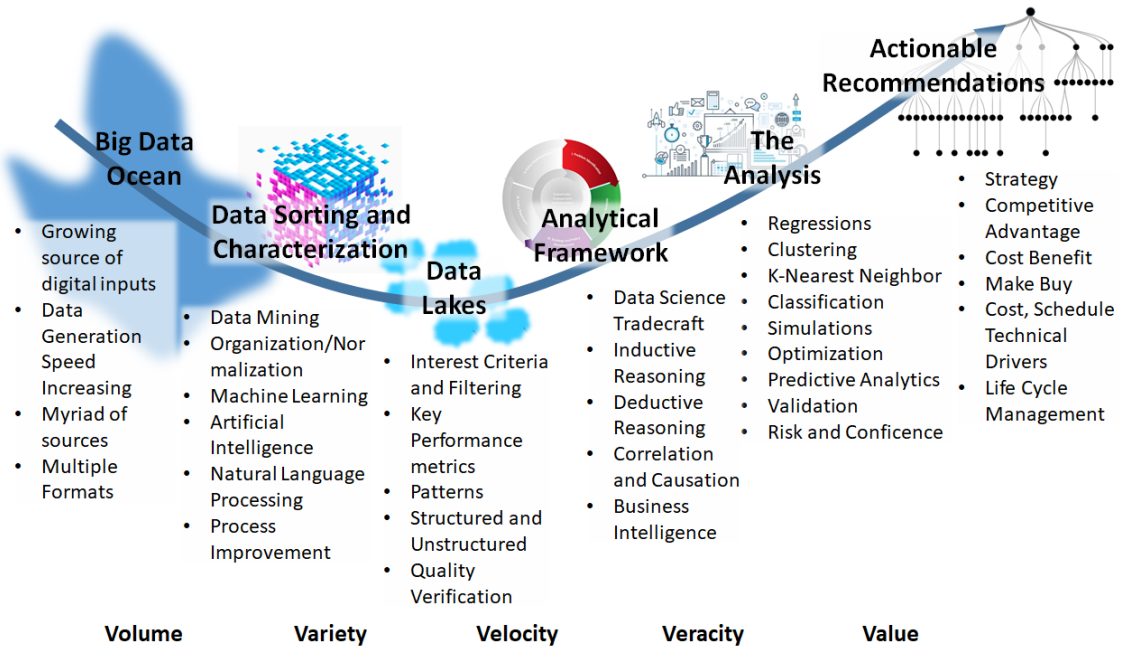

Figure 17 -Big data tenets provide the foundation for robust analytical solutions

(1,000,000 gigabytes) for data lakes. Whereas "big Data" is on a Zettabyte scale (1,000,000 petabytes) for data oceans. Refer to Appendix B for data volume measurements.

Variety – Is the form in which data can be stored or processed (e.g., structured or unstructured data). Variety is useful for inductive reasoning and unearthing previously hidden insights (Figure 1). Most often structured data is used in cost estimating and analysis due to many specialty data models with their own unique data characteristics and specific analyses.

Velocity –As the speed at which data is created increases, consideration should be to how it is captured, organized and stored for quick recall and use. This is likely a game changer in the future as it will enable real-time decision support.

Veracity –Good quality is the cornerstone of reliable analysis and results. Using effective data cleansing and validation is needed early. It often requires a disproportionate amount of time to get it right. Typical issues include input errors, fragmented or missing data labels, missing values and consistent units.

Value – Providing data to preserve history and support forecasting and trending to add value for decision makers. Its intrinsic characteristics include solving problems, identification of cost drivers (e.g., cutting costs, increasing revenue, etc.), providing transparency and insight into resources and efficiency, and others.

These five tenets provide a basis of effective analysis from pulling data out of the data ocean, performing data sorting and characterization into application specific data lakes that can then be used to define the modeling framework for performing insightful analysis resulting in robust, defendable recommendations. Figure 17 provide an illustration of the data science process lifecycle.

Impact to Cost Estimating

The cost estimating community must prepare for the emerging and critical disruptive change in the way cost analysis and forecasting is performed. As more cost related data is fused with technical and programmatic information the ability to perform real-time analysis with many organizations is becoming a reality.

The advanced tools and analysis must keep pace. In the future as AI, ML and NLP matures further, the day will come to apply experience, utilize existing tools in new ways, even speak to a computer to generate real time a cost estimating relationship. For example, someday a person may be able to just speak into a microphone and say: "*Computer – generate - a - cost estimating - relationship - for - a – satellite – system – with a weather payload – based – on – mass – power – IR -frequency – and LEO – polar - orbit*". Almost like asking "Alexa" to tell you a joke of the day. The data will then be generated and displayed on the tablet you are holding via the cloud. This may be a little farfetched today.

A more realistic step is the ability to apply NLP to directly cost requirements without intermediary measures (e.g., weight, ESLOC, function points, etc.). We imagine applying NLP techniques to the requirements of the system (e.g., direct inject GEO at a zero-degree inclination, with a given delta-V requirement for on-orbit maneuvers, that performs remote sensing in these band ranges, etc.) and then learning a

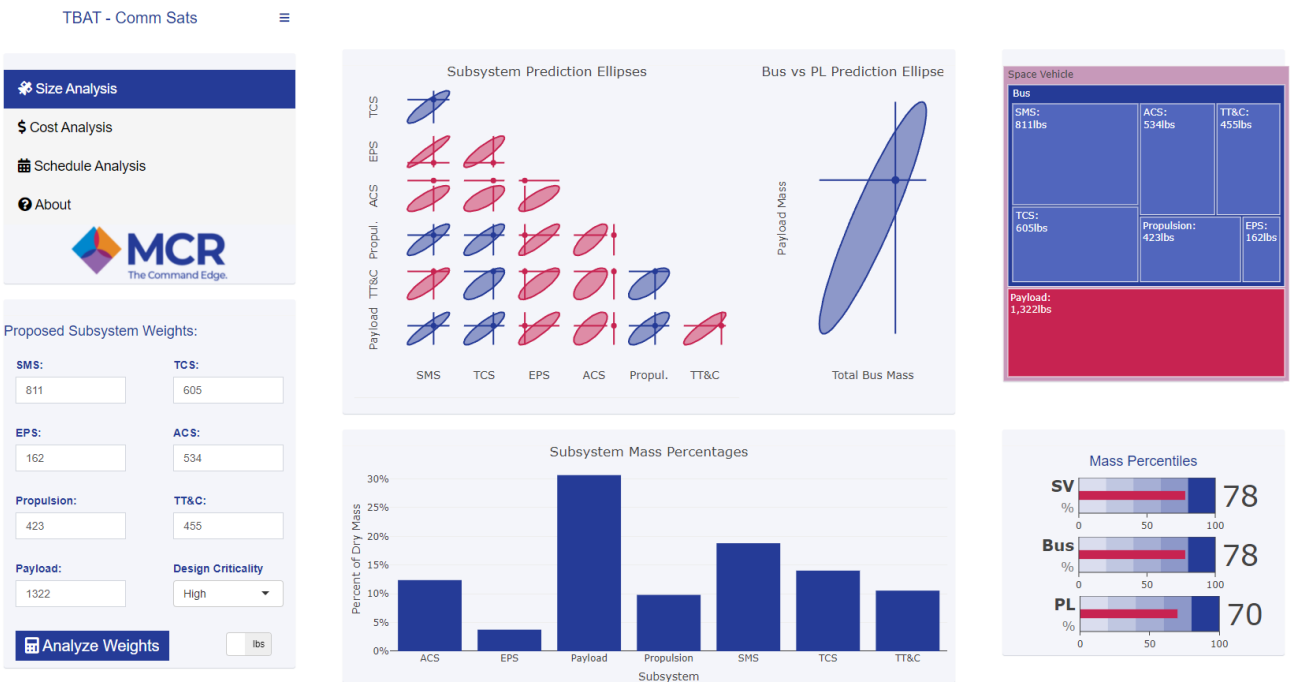

Figure 18 – The visionary Technical Baseline Assessment Tool provides optimal solutions to complex problems.

mapping to between these requirements and cost or schedule. This would require access to lots of data in addition to the computing power to perform the machine learning and natural language processing tasks. Moving forward, we expect this kind of shift in the "inputs" to our cost models (i.e., going from measures that have historically been correlated with cost but don't have a clear causal relationship (e.g., weight or lines of code) to the requirement sets that have clear causal impacts).

To enable this, as a community, we need to embrace, teach and advocate for data science-based approaches to cost estimation (e.g., cross-validation). While there is still a way to go, enhancing estimator familiarity and proficiency with free and open-source computational tools (e.g., R or Python) could help a great deal with advancing the state of our community.

An immediate step is illustrated in the Technical Baseline Assessment Tool (TBAT) with a snapshot in Figure 18 showing a baseline set of parameters and charts showing optimal solutions across various sub systems and class spacecraft with error bands. This illustrates future possibilities to quickly provide optimal solutions to spacecraft design and development. The tool applies multivariate statistical techniques and sequential quadratic programming to determine optimal sub systems that work with the mission requirements.

Last Word

Disruptive change is coming, current tools and methods are maturing. Preparing for the future of analysis will provide estimators a significant capability for timely decision making.

8. Future Research and Trends

We have provided a brief survey of tools and methods showing how the big data ocean is changing the way decision making is done and how it might change in the future. Future research will include delving deeper into AI, ML and NLP specific to the cost estimating community to provide advanced tools that enable robust analysis at light speed to decision makers.

Thinking about the future of our profession, based on a better understanding of the evolutionary development of estimating tools and estimates, makes us better prepared for that future.

Theodore Roosevelt once said, **"I believe that the more you know about the past, the better you are prepared for the future."**

We have provided a history and maturity of the cost estimating methods and tools that are used today by estimators and those in the community. Next, we discussed more advanced techniques with advanced tools like computers. Using these advancements will cause changes in how solutions are provided to decision makers in the future.

To facilitate this more training in information design (à la Edward Tufte)[c] can enable engineering-based graphics (using software tools such as those discussed in this paper (e.g., R and Python) that are beautiful publication level visualizations.[33] A near term recommendation will be to add a ICEAA track/award for information design. This will provide forward looking opportunities and get people thinking about using different plot types (e.g., https://datavizproject.com/) and understanding which is best to communicate information in what contexts.

[c] Edward Tufte was a pioneer in data visualization and representation methods. Professor Emeritus at Yale University

APPENDICES

A. ACRONYMS

AACEI	American Association of Cost Engineering International
AFCAA	Air Force Cost Analysis Agency
AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
CAPE	Cost Assessment and Program Evaluation
CCA	Canonical Correlation Analysis
CEAC	Cost and Economic Analysis Center (Army)
CER	Cost Estimating Relationship
COCOMO	Constructive Cost Model
CV	Cross-Validation
DCAA	Defense Contract Audit Agency
DoD	Department of Defense
EP	Evolutionary Programming
ESLOC	Equivalent Software Lines of Code
ICA	Institute of Cost Analysis
ICEAA	Institute of Cost Estimating and Analysis Association
IESSD	Industrial Estimating Society of San Diego
IoT	Internet of Things
IQR	Interquartile Range
ISPA	International Society of Parametric Analysts
KNN	K-Nearest Neighbor
LDA	Linear Discernment Analysis
LIME	Locally Interpretable Model-Agnostics Explanations
MAD	Median Absolute Deviation
ML	Machine Learning
MUPE	Minimum Unbiased Percent Error
NAFCOM	NASA/Air Force Cost Model
NASA	National Aeronautics and Space Administration
NCCA	Navy Center for Cost Analysis
NES	National Estimating Society
NLP	Natural Language Processing
NN	Neural Network
OR	Operations Research
PCA	Principal Component Analysis
PCEI	Parametric Cost Estimating Initiative
PRICE	Parametric Review of Information for Costing and Evaluation (PRICE Systems)
PRL	Parametric Reinvention Laboratory
RMSE	Root Mean Square Error
SCEA	Society of Cost Estimating and Analysis
SEER	Software Evaluation and Estimation of Resources (Galorath)
SEM	Software Estimating Model
SLIM	Software Lifecycle Management
SMC	Space and Missile Systems Center (now SSC)
SME	Subject Matter Expert
SSC	Space Systems Command (formerly SMC)
SSCAG	Space Systems Cost Analysis Group
TBAT	Technical Baseline Assessment Tool
US	United States
USAF	United States Air Force
USCM	Unmanned Space Vehicle Cost Model
VAR	Vector Autoregression
WBS	Work Breakdown Structure
ZMPE	Zero Percentage Bias Methods

B. DATA VOLUMES

Data Volumes: The volume of data in a single file or file system can be described by a unit called a byte. However, data volumes can become very large when dealing with Earth satellite data. Below is a table to explain data volume units (credit Roy Williams, Center for Advanced Computing Research at the California Institute of Technology). Kilo - means 1,000; a Kilobyte is one thousand bytes.

Mega- means 1,000,000; a Megabyte is a million bytes.

Giga- means 1,000,000,000; a Gigabyte is a billion bytes.

Tera- means 1,000,000,000,000; a Terabyte is a trillion bytes.

Peta- means 1,000,000,000,000,000; a Petabyte is 1,000 Terabytes.

Exa- means 1,000,000,000,000,000,000; an Exabyte is 1,000 Petabytes.

Zetta- means 1,000,000,000,000,000,000,000; a Zettabyte is 1,000 Exabytes.

Yotta- means 1,000,000,000,000,000,000,000,000; a Yottabyte is 1,000 Zettabytes.

BIOGRAPHY

***Patrick Malone**, P.E., is Certified Cost Estimator/Analyst (CCE/A), and senior analyst at MCR, LLC, where he performs in-depth cost, earned value and risk assessments for space, air and marine programs of all types and sizes. His analytical insight provides creative advancements in science and technology. An experienced engineer and analyst, Pat has researched methods such as technology and system readiness levels, agile applied to program management and system engineering that provides additional insight to managing programmatic processes and risks to support positive project outcomes.*

***Henry Apgar** is a retired cost engineer. He has been employed by MCR and Aerospace Hank has more than 50 years' experience developing parametric models and statistically based parametric cost estimates. He is a lifetime ICEAA member and a Certified Cost Estimator/Analyst (CCEA), with a degree in electrical engineering and an MBA. Hank authored the Cost Estimating Chapter for the Space Mission Engineering Handbook. He is co-founder and past president of ISPA and recipient of their lifetime achievement Freiman Award.*

***William King** is a Data Scientist for MCR who specializes in statistical analysis and cost model development. Formerly, he has worked for the Air Force Cost Analysis Agency (AFCAA) and supported cost estimating for numerous programs at the Space Systems Command. He holds a Master of Applied Statistics degree from Pennsylvania State University, a M.A. Economics from University of California, Irvine, and a B.S. in Mathematics from the University of Redlands.*

REFERENCES

[1] https://www.i-scoop.eu/big-data-action-value-context/data-age-2025-datasphere/

[2] Burbridge, Keith; "A Touch of History" 1984, self-published vignettes of parametric cost estimating applications from ancient Greece through the industrial revolution.

[3] Apgar, Henry; "The Legacy of Parametric Estimating," April 2019, presented at ICEAA Conference (Tampa, FL)

[4] https://comptroller.defense.gov

[5] James, John; "Funding the Early Gothic Churches of the Paris Basis;" Parergon; 1997; also, his unpublished work on "Medieval Units of Measure."

[6] Kraus, Henry; "Gold was the Mortar – the Economics of Cathedral Building;" Barnes & Noble; 2012.

[7] Apgar, Henry; "Who Was Frank Freiman;" ISPA Parametric World; 2009 (3 issues).

[8]Freiman, Frank; "History of Parametric Developments," personal diary of how Frank developed the PRICE H Model, March 8, 1995. Frank was the only recipient of the ISPA Honorary Director Award in 1980 and is the namesake of the ICEAA Freiman Lifetime Achievement Award.

[9] Novick, David; "Rand: A History of Cost Analysis;" Journal of the National Estimating Society; 1981.

[10] Fisher, Gene; "Rand: Cost Considerations in Systems Analysis;" Elsevier; 1975.

[11] Mason, R. and Linde, D. "Basic Statistics for Business and Economics 2nd Ed., McGraw-Hill, 1993

[12] Joint Agency Cost Estimating Relationship Handbook https://www.dau.edu/tools/Lists/DAUTools/Attachments/387/CER_Dev_Handbook_Feb2018_Final.pdf, 2018, pg. 91.

[13] Johnson, R. & Wichern, D. Applied Multivariate Statistical Analysis (6th Edition). 2019. Pg. 481.

[14] El-Choum, M. "An Integrated Cost Control Model", AACE International Transactions, 2000

[15] Johnson, R. & Wichern, D. Applied Multivariate Statistical Analysis (6th Edition). 2019. Pg. 539.

[16] Trevor Hastie; Robert Tibshirani; Jerome Friedman. The Elements of Statistical Learning. Data Mining, Inference, and Prediction (second ed.). Springer. Pg. 128.

[17] BÖKEOĞLU ÇOKLUK, Ö, & BÜYÜKÖZTÜRK, Ş. (2008). Discriminant function analysis: Concept and application

[18] https://www.tableau.com/learn/

[19] https://www.qualtrics.com/experience-management/

[20] https://www.ibm.com/cloud/learn/neural-networks

[21] https://www.smartsheet.com/neural-network-applications

[22] Ramos, Diana, "Real-Life and Business Applications of Neural Networks", Smartsheet.com, Oct 17, 2018.

[23] https://www.scientificamerican.com/article/what-is-fuzzy-logic-are-t/

[24] Chiang, R and Jyh-Shing, J. "Fuzzy Logic Attitude Control for Cassini Spacecraft", IEEE World Congress on Computational Intelligence, Orlando, FL, 1994.

[25] https://www.cs.cmu.edu/Groups/AI/l

[26] https://www.cs.bham.ac.uk/internal/courses/ f

[27] https://us.nttdata.com/

[28] Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

[29] Delua, J. IBM Analytics SME, https://www.ibm.com/

[30] https://www.ibm.com/topics/what-is-a-digital-twin

[31] Gelernter, D. "Mirror World: Or the Day Software Puts the Universe in a Shoebox…How it will Happen and What it Will Mean", Oxford University Press, 1991

[32] Arrow, J., Markhus, A., Singh, V., Ramos, C. Bakker, S., "Project Controls & Data Analytics in the Era of Industry 4.0", AACEI Risk.2926 Technical Paper, 2018.

[33] Tufte, E. "Visual Display of Quantitative Information", 1985.