**How is Statistical Engineering Different from Data Science?**

Allison Jones-Farmer, Miami University

Roger Hoerl, Union College

January 2019

## Introduction

The emerging discipline of data science has clearly revolutionized the world, and changed our lives dramatically (Friedman 2016). It is very hard for most people today to remember a world without email, smart phones, or being able to "Google" a person or a phrase. Further, as pointed out in an even earlier book by Friedman (2005), this technology revolution has to a large degree "flattened" much of the world, in terms of providing people in developing countries unprecedented access to knowledge and information.

The ability to acquire, store, process, and analyze data has grown roughly exponentially for decades ("Moore's Law"), resulting in larger and larger data sets, and more and more sophisticated methods to analyze this data. There are numerous methodologies involved in such analysis, such as statistics, analytics, artificial intelligence (AI), management science, data mining, operations research, and machine learning. The term data science is often used to represent a combination of the above analysis approaches with modern computer science and information technology for data acquisition, processing, and analysis.

Given the existence of such an established set of analytics approaches, what exactly is the value added by statistical engineering? We attempt to answer this question below. First, we provide a high-level explanation of what statistical engineering is, and how it is additive to applied statistics. To understand how statistical engineering is different from data science, it is important to first understand how it is different from applied statistics.

## Statistical Engineering and Applied Statistics

The field of statistical engineering attempts to provide the thought leadership for developing sound, evidence-based approaches to producing practical, sustainable solutions to complex problems. The International Statistical Engineering Association (ISEA - https://isea-

change.org/) defines Statistical Engineering as *the systematic integration of statistical concepts, methods, and tools, often with other relevant disciplines, to solve important problems sustainably*. The focus of statistical engineering is on integrating methods to solve large, complex, unstructured problems in a way that provides a solution that works within the broader context of an organization.

Real problems that require the application of statistical methods range from simple to complex. Hoerl and Snee (2017) suggest that many problems can be solved through the application of one "correct" statistical method, often one that can often be looked up in a textbook. These would be examples of applied statistics problems, but not statistical engineering problems, in that no integration is required. An example of a straightforward applied statistics application may be the use of a linear model to estimate the effect of neighborhood blight removal on the market value of a home. Provided the data are given to the analyst in a nicely formatted spreadsheet with a well-defined dependent variable (market value) and predictor (a measure of neighborhood blight removal), along with suitable control variables, this should be a straightforward – not necessarily easy – applied statistics problem.

Few real problems of any merit, however, reach the analyst in such a clear-cut form. There is often a tremendous amount of work behind the scenes before the analyst can begin the process of fitting a statistical model. In order for the analyst to properly fit a model and deliver a sustainable solution, he or she should ideally be involved in the problem solving from the beginning and see it through to the end.

The problem of understanding the relationship between neighborhood blight removal and property value is a real, messy problem on which the first author of this white paper actually worked. Our solutions were based on work with a local county government and a team of student researchers. This is an example of a statistical engineering problem for several reasons. First of all, neighborhood renewal is a high-impact problem. Blight removal refers to the demolition of blighted properties. In this case, residential properties that are dilapidated or are beyond repair are often demolished in order to reduce crime and improve the value of the existing properties in the neighborhood. The goal of this study was to understand if there is a

relationship between blight removal and property values. The technical and political aspects of blight removal, property value, and home sales over time are complex issues.

The data for the blight removal study were curated from local government agencies. While most data are available from the county auditor, the data exist in multiple data tables that must be reformatted for consistency and carefully merged. Many discussions among researchers, government officials and colleagues in analytics, information systems, economics and political science were held. Key discussion points included definitions of blight, proximity, market value, and what variables should be used as controls. The inclusion/exclusion criteria for properties are complex and politicized. There are no right answers for how to proceed with most of the issues with this study. Although there is a small academic literature on this type of problem, most of the solutions are limited and there is plenty of room for improvement. Throughout the process, decisions must be made, justified, and carefully documented. Clearly, this is not a textbook applied statistics problem. But how should such problems be approached?

Below is a high-level overview of the steps completed to frame the problem and prepare the data for analysis.

1. Work with the client to understand the goal of the study. In this case, the client consists of county government officials who are funding neighborhood blight removal.
2. Work with the local government entities to obtain the data. Standardize and reformat the data for consistency. Define the key measures such as property value and blight removal. Define the control variables.
3. Use geographic information systems to geocode the data and develop a system to summarize the spatial and neighborhood characteristics of each property in the county.

The descriptions of each of the three steps above are an oversimplification of the work that was involved. Each step required several months of work. Once these steps were completed, the team was ready to model the data. The following steps were taken.

4. Research appropriate statistical methods to model the spatial and temporal aspects of the data while estimating the relationship between blight removal and property value.
5. Develop a statistical model to estimate the relationship between blight removal and market value.
6. Develop materials and whitepapers to be delivered to the client. Work with client to use the results of the statistical model to direct future neighborhood renewal initiatives.

As the steps above illustrate, statistical engineering is about solving big problems sustainably using statistical methods in thoughtful and responsible ways – typically by integrating multiple methods and even multiple disciplines. It is about engineering solutions. Statistical engineering often requires bringing together people from several different disciplines at several different levels, often in multiple organizations, to work on a problem together. This study included participants with expertise in statistics, information systems, geographic information systems, and economic modeling. All were required to find a solution that worked for this problem.

**Statistical Engineering and Data Science**

A challenge in clearly articulating the difference between statistical engineering and data science is that there are multiple definitions of data science in use today. Virtually all of them include a mixture of statistics, machine learning (neural networks, support vector machines, etc.), and methods from computer science and information technology, such as cloud computing, data processing and storage, and coding.

Regardless of which definition used, however, many data science problems are statistical engineering problems. The data may be larger, more complex, and may require distributed computing. The methods used may be machine learning or statistical learning methods in addition to statistical modeling. For data science problems, the underlying approach used to fit the model is often predictive rather than explanatory (see, e.g. Brieman, 2001; and Shmueli, 2010). In the data science paradigm, data are typically plentiful, and models are developed and

validated empirically rather than theoretically, using what can be described as a common task framework (Donoho 2017) that employs benchmark data sets or cross-validation methods.

Although the goal in predictive modeling is often focused on selecting an optimal model based on some measure of optimality, such as lift, gain, or root mean squared error, it is still imperative to select a model that is sustainable and works within the context of a broader organization. Therefore, the problem-solving skills needed to engineer a solution for large, complex, unstructured problems are the same. In this sense, statistical engineering and data science are synergistic.

Donoho (2017) presents a broad framework for the field of data science that he refers to as Greater Data Science (GDS). GDS includes six activities: (1) data exploration and preparation; (2) data representation and transformation; (3) computing with data; (4) data modeling; (5) data visualization and presentation; (6) and science about data science. The sixth domain, the science about data science, refers to applied methodological research, including meta-analysis and studying the validity of methodologies applied in practice. In Donoho's opinion, the data science literature, at least up to 2017, was more focused on activities 3, 4, and 5, and less so on 1, 2, and 6. If one takes Donoho's broader view, then greater data science (GDS) covers much of statistical engineering.

Still missing from this broad definition of the field of GDS, however, is any discussion or guidance as to how one would go about using these methodologies to solve a high-impact, large, complex, unstructured problem. Statistical Engineering, while technically old, is a new discipline that has emerged out of a collaboration of industrial and academic statisticians and engineers, working together to build a *unified approach* to solving such problems. *This is the main aspect of statistical engineering that we find value adding to data science, including the broader view advocated by Donoho.*

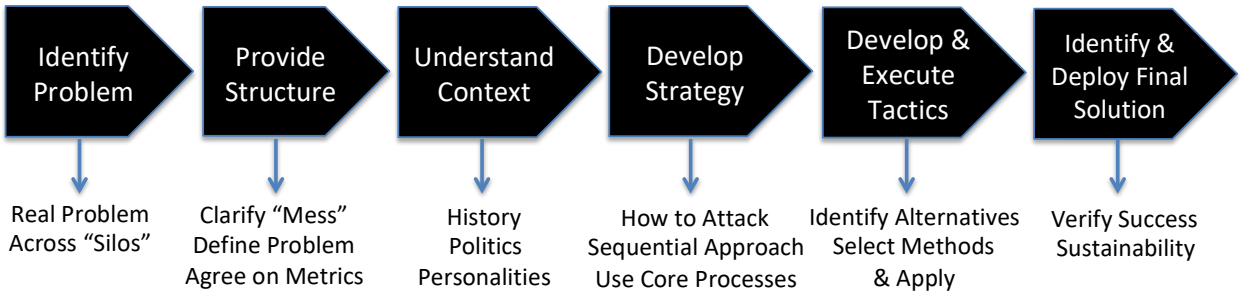
**The Unifying Goal of Statistical Engineering**

Whether using statistical or data science methods, high-impact, complex, large, unstructured problems abound. These problems need to be solved, and we need a unified strategy to solve them. Hoerl and Snee (2017) describe the phases of a statistical engineering

project using a model similar to the one given in Figure 1. Although not intended as a "cookbook" for problem solving, a generic model such as this can serve as the general guide to help practitioners develop a tailored approach for many complex problems.

Hoerl and Snee (2017) further point out that smart statisticians and engineers have been solving large, complex, unstructured problems for a long time, people such as Fisher, Gosset, Box, Tukey, and so on. What has been missing, however, is documentation of exactly *how* these brilliant people actually approached their problems and found solutions. Figure 1 is therefore an attempt to do just this, at least at a high level, based on the existing literature in statistical and engineering problem solving, and the authors' own experience. Ideally, others who are not Box or Tukey could use this approach to increase their own chances of success. Further, this model could be taught to statisticians and engineers in academia, speeding up the learning curve to solve large, complex, unstructured problems.

Figure 1. Typical phases of a Statistical Engineering Project




Of course, many other methods exist for solving data-related problems including Knowledge Discovery and Data Mining (KDD, Fayyad et al., 1996), CRISP-DM from data mining (Chapman et al., 2000), and the Data Analytics Life Cycle (EMC Educational Services, 2015). Many practitioners have also used structured approaches to problem solving such as those provided in Lean Six Sigma (LSS). Jones-Farmer and Krehbiel (2016) discussed LSS practices and compared these to the use of Agile methods, and in particular, the Scrum methodology as a way to break complex problems into smaller tasks in order to rapidly iterate towards a solution. For some problems, they found Agile and Scrum to work better than LSS. Clearly there is no

universally best method for solving problems, as each has its own unique aspects and challenges.

Although many problem-solving methodologies have been suggested previously, we note that most are *tactical* rather than *strategic*, in that a reasonably well-formulated problem is typically assumed. We find the comprehensive approach given in Figure 1, when augmented with the additional detail on how to actually deploy this approach in practice (available on the ISEA webpage) to be much more strategic. There exists no evidence-based agreement on the best strategic approach for solving high-impact, large, complex, unstructured, problems, but the approach given in Figure 1 is fairly generic, and should provide at least a starting point to a broad array of problems. As noted, there is further detail on how to succeed in each phase on the ISEA website.

**Summary**

*Developing evidence-based strategic approaches for solving complex problems is the primary focus of the field of statistical engineering, and in our view is its most important value-add to both applied statistics and data science.* These strategies can serve as an umbrella for solving data-related problems that are of a statistical (explanatory) nature and those of a data science (predictive) nature for both large and small data scenarios. Developing continuing education curricula for existing data practitioners in statistics, engineering, analytics, and data science is also a key initiative of the ISEA. In addition, ISEA collaborates with universities to develop curricula to in order to educate the next generation of problem solvers. https://isea-change.org/ for further details.

**References**

Breiman, Leo (2001). "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3), 199-231.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0. *CRISP-DM Consortium*, *76*, 3.

Donoho, D. (2017) “50 Years of Data Science”, *Journal of Computational and Graphical Statistics*, 26, 4, 745-766.

EMC Educational Services (2015). *Data Science and Big Data Analytics: Discovering, Analyzing Visualizing and Presenting Data*. John Wiley & Sons: Indianapolis, IN.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases." *AI Magazine* 17(3), 37.

Friedman, T.L. (2005) *The World is Flat*, Ferrar, Straus, and Giroux, NY.

Friedman, T.L. (2016) *Thank You for Being Late*, Ferrar, Straus, and Giroux, NY.

Hoerl, R. W., and Snee, R. D. (2017). "Statistical Engineering: An Idea Whose Time Has Come?" *The American Statistician*, *71*(3), 209-219.

Jones-Farmer, L.A. and Krehbiel, T.C. (2016). "Agile Teams: A Look at Agile Project Management Methods"*, Statistics Digest* 35(3), 30-35.

Shmueli, Galit (2010). "To explain or to predict?" *Statistical Science* 25(3) 289-310.