Follow the fundamentals: Four data analysis basics will help you do big data projects the right way

Article *in* Quality Progress · January 2014

Ron Snee
Snee Associates, LLC
306 PUBLICATIONS **3,409** CITATIONS

Roger Hoerl
Union College
64 PUBLICATIONS **1,262** CITATIONS

FOLLOW the Fundamentals

Four data analysis basics will help you **do big data projects the right way**

by Ronald D. Snee, Richard D. DeVeaux and Roger W. Hoerl

In 50 Words Or Less

- Big data projects are becoming more prevalent in today's business world.
- To ensure project success: use high-quality data, leverage subject matter knowledge to put things in perspective, realize more than one data set or experiment is needed to solve the problem, and develop a strategy for conducting the project.

NEW TECHNOLOGY FOR acquiring, storing and processing data is being introduced at an ever-increasing pace. In 2012, the White House launched a national "Big Data Initiative."[1] According to IBM, 1.6 zettabytes (10^{21} bytes) of digital data are now available. That is a lot of data—enough to watch high-definition TV for 47,000 years.[2]

Real-time data acquisition is becoming the norm. These developments are having and will continue to have major effects on how quality professionals and statisticians conduct product and process design and quality improvement projects.

So what should we do? Is new thinking required, or will the usual approaches to data analysis work? A little thought suggests any change this big will require a new way of thinking.

Big data are a collection of data sets that are too big and too complex to be processed using traditional database and data processing tools.[3] Big data projects, as generally practiced today, seem to overlook four fundamental ideas that ensure successful outcomes:

1. Data quality.
2. Subject matter knowledge.
3. Sequential approaches.
4. Strategy.

1. Data quality

Data quality is a major issue because big data projects typically involve diverse data sets collected in different environments. Whenever data are collected by multiple sources, the opportunity for omissions and errors is always present. The result is missing values, missing variables, measurement variation and definition conflicts that can halt successful analysis.

Unfortunately, much study and literature assume large data sets contain high-quality data measuring the right variables at the right frequency, and are devoid of missing values, missing variables or outliers. If only this were the case. Typically, a great deal of time and effort is needed to produce high-quality analysis.

Consider a pharmaceutical company's quality improvement problem that involved a major product manufactured at five different locations. One of the first steps was to find the vital few variables related to the quality issue. Unfortunately, three of the most critical variables identified by the quality improvement team were not being measured on the process in one location. Missing variables in a portion of the data set is not an uncommon occurrence and complicates the analysis.

Missing data within recorded variables presents another common problem. In many data sets, unfortunately, it is unclear whether values recorded as zeros are actually zeros, or whether they represent missing values. Consider this online survey question: "If you are experiencing technical issues, click here." Treating a nonclick as missing would result in assuming 100% are having technical problems.

Another common data quality problem with big data is multicollinearity: Data are often collected without regard to survey or experimental design. This redundancy of a predictor results in a high correlation among them. This can make it difficult to separate the effects of the correlated variables on the response. Multicollinearity also can have effects on the performance of regression algorithms, such as the stepwise method and decision trees. Analysts of big data always should look for multicollinearity and take steps to mitigate its effects.

It's important to understand the pedigree of the data—how the data were generated and collected, including the process producing the product or service and the process used to measure.[4] Clear understanding of the pedigree of the data is critical to the accurate assessment of the data quality. Even sophisticated algorithms cannot extract information that is not in the data.

2. Subject matter knowledge

Subject matter knowledge is the next consideration that is underappreciated in current literature. "Data have no meaning in themselves; they only have meaning within the context of a conceptual model of the phenomenon under study," wrote George E.P. Box, William G. Hunter and J. Stuart Hunter.[5] Without subject matter (or domain) knowledge, you have no context within which to define, analyze and interpret the results.

Some have taken the position that with vast amounts of data, subject matter knowledge and scientific theory are no longer needed; you can solve problems purely and empirically with data alone.[6] Unfortunately, many analysis blunders have occurred because analysts did not understand the phenomenon under study, and then they misinterpreted the results.

Amazon.com, for example, used an automated algorithm to set an initial price on a new biology textbook, *The Making of a Fly*, at $1,730,045 in 2011. Even more shockingly, competitive websites benchmarked this price electronically and started raising their prices, resulting in Amazon.com doing the same. Shortly after, the book's price on Amazon.com reached $23,698,656.[7] Surely, a little common sense and understanding of how textbooks sell would have ruled out such prices.[8]

Subject matter knowledge—the theory underlying the process and the data itself—can be used fruitfully in many ways, including selection of variables and appropriate scales (for example, log, inverse and square root), selection of model form (for example, linear, curvilinear, multiplicative and which variables should be considered x versus y variables), interpretation of results and the ability to extrapolate findings.

You can do much more with data gathered and interpreted within the context of sound subject matter knowledge than you can with either data or theory alone.

From a practical point of view, understanding the process that produced the data provides context that helps frame the problem, create the plan for analysis, understand the results of the analysis, and guide the creation and implementation of the solution.

3. Sequential approaches

Experience has taught us important problems are rarely solved using a single data set or experiment. The Wright Brothers, Thomas Edison and, more recently, Bill Gates and Steve Jobs all achieved success after many rounds of trial and error. Fortunately, they learned with each attempt, gradually building up their understanding (subject matter theory) of the problem they were addressing.

This is a key aspect of the scientific method: becoming a little smarter with each round of experimentation, gathering better and more relevant data until you eventually solve the problem. Obviously, organizations such as Apple and Google never stop learning: They indefinitely continue this sequential approach, leading to continuous learning, improvement and innovation in products and services.

Unfortunately, there seems to be an implicit assumption in much of the big data literature that all problems can be solved with one data set and one analysis. Online websites, for example, such as www.kaggle.com are now hosting online data analysis competitions. The original problem, however, is reduced to: "What is the best model I can create from this data set?"

The data set, of course, should not be the focus of your efforts—the problem you are solving should be the focus. A better question would be: "What can I learn from this data set that would help me collect even better data in the future so I can solve the original problem and continuously learn?"

A good analysis of a specific data set typically will answer some questions, but will almost always lead to more questions requiring the collection and analysis of additional data sets. It should not be overlooked that the final—or near final—model also should be used to predict an independent data set to confirm and measure the prediction accuracy of the model.

Note that an additional data set, collected at a different time and under different circumstances from the original data set, goes well beyond holding out some of the original data as a prediction set, sometimes referred to as cross validation.

Your strategy must move beyond the one-shot study mindset if you are to be effective over the long term. It may be prudent to consider the current model as a hypothesis generator for the next investigation rather than confirmation of an existing hypothesis.

4. Strategy

Early in the planning of a big data project, it's prudent to strategize how the project will be conducted. A strategy is an overall high-level approach to attacking a problem. The strategy defines how you will win. It defines the choices to be made—what you will and will not do.

During World War II, for example, the Allies' strategy was to "win in Europe first." Thus, when a choice had to be made regarding resources, Europe got the resources before the Pacific war theater.

The following aspects of strategy are based on the recommendations of John P. Kotter.[9]

First, it is important to develop a sense of urgency: Why it is important for the organization to address this problem at this time and how interested is the organization's leadership in the problem's solution? The organization will be more interested in providing needed support and resources when the sense of urgency is well understood.

Next, it is critical to have a guiding coalition of influential leaders that believe in the initiative and will help ensure its success. The support and involvement of the guiding coalition will be needed throughout the project's life.

The guiding coalition also can help develop the vision for the project: What things will be in place and what it will look like when the project is complete? The construction of the vision is similar to Steven R. Covey's recommendation of "beginning with the end in mind."[10] The vision helps the affected people see themselves in the creation and implementation of the solution to the problem.

Clear understanding of the **pedigree of the data** is critical to the accurate assessment of the **data quality.**

It is also important to create a communication plan and process for the project. The people affected by the project must be continuously and clearly reminded of the purpose of the project and informed of progress made to date, including the schedule for project completion. Because people take in and process information differently, a variety of media—such as email, video, one-on-one meetings and small discussion groups—should be used, depending on the nature of the project.

The affected people should be empowered as needed to perform effectively in their new roles in the creation and implementation of the project solution. This will include training and education, and perhaps new equipment.

Generating short-term wins means dividing the work into pieces that can be completed in less than three to six months. People are impatient—they want to see progress. Short-term wins demonstrate that big data projects can be successful and return useful benefits.

When projects run longer than three to six months, people move to different assignments, business conditions change affecting the project and the organization loses interest.[11] Naturally, some projects cannot be completed in less than three to six months. An effective approach for such situations is to create subprojects that can be completed quickly, thus satisfying the organization's need to see progress.

Framework

Over the years, we have learned that problem solving, particularly in a team environment, is most effective when guided by a framework that provides the team a step-by-step process to follow. As a result, the team knows what work is to be done (team alignment), in what sequence and where the team is in the problem-solving process at any point in time. Two well-known examples of such problem-solving frameworks are plan-do-check-act and define, measure, analyze, improve and control.

Big data projects are typically related to large, complex and unstructured problems. Data come from many different sources. Several groups are involved, each with its own agenda and its own ideas about what the problem is and how to find the solution. As a result, the problem is typically ill-defined, requiring work to decide what problem to attack and who should be on the team. After the team is formed and the work begins, the strategy for how the problem will be addressed, and the tactics of the analysis must be worked out.

Such a situation is similar to the problems addressed with statistical engineering, which has an associated framework to guide the work.[12-15] While there is no magic number of easy steps to take when conducting such projects, the work typically follows the following phases:

- **Identify** high-impact problems that are doable and that management will support.
- **Create** structure for the problem. Big data problems are typically unstructured initially.
- **Understand** the context of the problem, such as the process in which the problem resides, data types and sources, historical background and agendas of stakeholder groups.
- **Develop** an overall strategy, as discussed earlier.
- **Establish** tactics on how to implement strategy.

Paying attention to these phases will significantly increase the success rate of big data projects.[16]

Not replacing the scientific method

The glass is half full: Big data offer a unique opportunity, but there can be many pitfalls along the way. Using powerful statistical software packages incorrectly can prevent effective analysis of big data sets. Ignoring statistical engineering fundamentals can lead to ineffective and perhaps wrong solutions.

Not understanding "data pedigree," use of sequential approaches, integration of subject matter knowledge in all aspects of the problem-solving process—problem identification and formulation to solution implementation—and the development of an overall strategy or plan of attack can decrease the probability of success.

Big data analytics are here to stay, and that's good news. Careful attention to the fundamentals of data analysis can lead to very useful and profitable problem solutions and process improvements. The secret to success is to understand and use the fundamentals. QP

REFERENCES AND NOTES

1. Tom Kalil and Fen Zhao, "Unleashing the Power of Big Data," White House Office of Science and Technology, Office of Science and Technology blog, April 18, 2013, www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data.
2. Mike Ebbers, "5 Things to Know About Big Data in Motion," IBM DeveloperWorks blog, IBM, June 12, 2013, www.ibm.com/developerworks/community/blogs/5things/entry/5_things_to_know_about_big_data_in_motion?lang=en.
3. Wikipedia, "Big Data," http://en.wikipedia.org/wiki/big_data.
4. Ronald D. Snee and Roger W. Hoerl, "Inquiry on Pedigree," *Quality Progress,* December 2012, pp. 66-68
5. George E.P. Box, William G. Hunter and J. Stuart Hunter, *Statistics for Experimenters,* John Wiley and Sons, 1978, p. 291.
6. Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired Magazine,* June 23, 2008, www.wired.com/science/discoveries/magazine/16-07/pb_theory.
7. Kevin Slavin, "How Algorithms Shape Our World," TED Conference presentation, July 21, 2011, www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world.html.
8. Richard D. DeVeaux and David J. Hand, "How to Lie With Bad Data," *Statistical Science,* Vol. 20, No. 3, 2005, pp. 231-38; discusses several other examples of bad data.
9. John P. Kotter, *Leading Change,* Harvard Business School Press, 1996.
10. Steven R. Covey, *Seven Habits of Highly Effective People,* Franklin Covey, 1989.
11. William H. Gates III, *Business @ the Speed of Thought,* Warner Books, 1999.
12. Roger W. Hoerl and Ronald D. Snee, "Closing the Gap," *Quality Progress,* May 2010, pp. 52-53.
13. Ronald D. Snee and Roger W. Hoerl, "Engineering an Advantage," *Six Sigma Forum Magazine,* February 2011, pp. 6-7.
14. Ronald D. Snee and Roger W. Hoerl, "Proper Blending," *Quality Progress,* June 2011, pp. 46-49.
15. Ronald D. Snee and Roger W. Hoerl, "Further Explanation," *Quality Progress,* December 2010, pp. 68-72.
16. The success rate of big data projects will be the subject of a future *Quality Progress* Statistics Roundtable column or feature article.

RONALD D. SNEE is president of Snee Associates LLC in Newark, DE. He has a doctorate in applied and mathematical statistics from Rutgers University in New Brunswick, NJ. Snee has received ASQ's Shewhart and Grant medals. He is an ASQ fellow and an academician in the International Academy for Quality.

RICHARD D. DeVEAUX is the C. Carlisle and Margaret Tippit professor of statistics at Williams College in Williamstown, MA. He has received ASQ's Shewell and Frank Wilcoxon awards and is a fellow of the American Statistical Association.

ROGER W. HOERL is the Brate-Peschel assistant professor of statistics at Union College in Schenectady, NY. He has a doctorate in applied statistics from the University of Delaware in Newark. Hoerl is an ASQ fellow, a recipient of the ASQ's Shewhart Medal and Brumbaugh Award, and an academician in the International Academy for Quality.