"Big Data and the Missing Links",

Article *in* Statistical Analysis and Data Mining · August 2016

DOI: 10.1002/sam.11303

3 authors, including:

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

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

Big Data and the Missing Links

Richard De Veaux[1], Roger W. Hoerl[2] and Ronald D. Snee[3]•

[1]*Department of Statistics, Williams College, Williamstown, MA 01267, USA*

[2]*Department of Mathematics, Union College, Schenectady, NY 12308, USA*

[3]*Snee Associates, New Brunswick,• NJ, USA*

Received 13 December 2015; revised dd Month yyyy; accepted 14 December 2015
DOI:10.1002/sam.11303
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Although •Big Data can have the potential to help researchers in science and industry solve large and complex problems, basic statistical ideas are often ignored in the Big Data literature. It is not true that simply having massive amounts of data renders subject-matter models and experiments obsolete, alleviates the need to ensure data quality and no longer requires that variables accurately measure what they are supposed to. We refer to these fundamentals as missing links in the Big Data process. In this paper, we illustrate the challenges of making decisions from Big Data through a series of case studies. We offer some strategies to help ensure that projects based on Big Data analyses are successful. © 2016 Wiley Periodicals, Inc. Statistical Analysis and Data Mining: The ASA Data Science Journal 0: 000–000, 2016

Keywords: clustering; classification; mixture models; computational statistics; Bayesian analysis; ranking

1. INTRODUCTION

Big Data promises and often delivers big dividends. Analyzing large datasets can help the decision maker in almost any organization as well as the scientific investigator. However, finding the right data is not always easy, and even when found, data are often stored in disparate locations and databases. Merging information from different sources is always challenging, but when the different databases lack identifiers that can link them together, it can render the data useless. Missing links such as these, and others that we will discuss later, can spell the difference between a successful data-mining effort and a bust.

Companies expend much time and effort to collect data on various aspects of production for quality control, notably for use in Six Sigma programs. Once the product is shipped, warranty, usage and customer satisfaction data are used to monitor its success. Too often, however, those two databases do not communicate. A problem with the warranty that might be solved by knowing exactly under what conditions the parts were produced still exists because it is impossible to trace the components and their production history. In some industries, notably food and drug manufacturing, governmental regulations mandate that links between use and production be identified for safety reasons. If all manufacturing used the same discipline, the promise of Big Data could be realized much more broadly.

The missing link between production and the customer can have serious consequences both for the business as well as for the consumer, as illustrated by the Bon Vivant botulism case (New York Times October 5, 2007). On July 2, 1971, the U.S. Food and Drug Administration (FDA) released a public warning after learning that a Westchester County, New York man had died, and his wife had become seriously ill from botulism after partly eating a can of Bon Vivant vichyssoise soup. The company began a recall of the 6,444 cans of vichyssoise soup made in the same batch as the contaminated can. Bon Vivant did not have adequate records and controls of production lots and distribution in order to trace the products quickly.

The FDA discovered that the company's processing practices were questionable for all products packed by it and extended the recall to include all Bon Vivant products. The FDA ordered the shutdown of the company's Newark, New Jersey plant on July 7, 1971. The recall deeply damaged the

Correspondence to: Richard De Veaux (deveaux@williams.edu)

Bon Vivant brand. The company filed for bankruptcy within a month after the start of the recall. It changed its business name to Moore & Co., but its problems did not end there.

In this case, the lack of data linking production to distribution to the customer literally killed the company. Perhaps even more serious is the fact that this lack of data linkage put consumers' lives at risk as well. As noted earlier, traceability is a big issue in the food and pharmaceutical industries today. Not only must data be recorded properly but also records within the same database must be appropriately linked. Fortunately, the technology of relational databases provides the capability to do just this. The key question is whether such technology is effectively utilized.

The problems due to missing links between data bases are a practical issue that, with a little planning, can be avoided. However, another missing link of Big Data is the lack of the connection between the data being analyzed and the problem being solved. Rarely are the data in Big Data generated from designed experiments or surveys; rather, they are transactional data or other data compiled by observations without regard either to a specific purpose or design.

Unfortunately, conclusions based on happenstance data are as suspect today in 2014 as they were in 1950. While such data, which comprise most of the data being analyzed under the rubric of Big Data, can help researchers generate *hypotheses*, they can never be used to test or confirm hypotheses. In spite of all the hype about the quantity of data at hand, only carefully designed surveys and experiments justify the causal link between what is observed in the data and what will happen when those values are changed.

In this paper, we provide some fundamental strategies to ensure success when analyzing large datasets. We first illustrate the problem via a case study on the production of an inkjet printer. Then, we list four data analysis fundamentals that we find lacking in most of the Big Data literature. The next section provides a methodology for approaching Big Data problems sequentially, and we summarize our arguments in the final section.

2. INKJET PRINTER CASE STUDY

The backpack inkjet printer was a niche product by a major printer manufacturer for use in environmental sampling in the 1990s. An early version of the mobile printer, it enabled easy and convenient printing of bar codes in remote locations. Because it was battery operated, it could be taken into remote areas, next to streams or lakes where identifying bar codes could be printed and placed directly on the water sample containers. The product was a success (albeit in a niche market) for this company.

At some point after introduction, warranty issues began to surface. There seemed to be more problems with the inkjets than anyone had anticipated. Two statisticians, one in-house and the other a longtime consultant, were called in to help solve the problem. Unfortunately, like much warranty data, the information was thin. Customers rarely provide detailed information about the history or the circumstances of the product failure. In this case, the only variables the statisticians had to work with were the date of purchase, the location of the purchase, the price paid and some other information about the purchase channel.

The statisticians turned to the production team to find out more about the production process of the printers, hoping to find out if the defective printers were related in some way. Unfortunately, the production data was listed only by component number and was never assembled by product serial number, even though, naturally, each printer was shipped as a unit. All the data collected to help the Six Sigma effort in production were lost—at least they could not be linked to the product out in the field.

Fortunately, in this instance, the statisticians turned out to be lucky. In spite of the lack of usable data, they simply plotted the locations (from the zip codes) where the warranty claims were made. They then gathered a team of engineers, scientists and production workers to look at the map and generate some hypotheses about why certain regions might be problematic. They were fortunate that the problem was fairly simple and was pretty clearly due to extremely dry conditions—something that (after some thought) was fairly obvious from the regions with the most warranty claims.

This hypothesis—that the warranty problems were due to low humidity—now needed to be tested. Data mining generated the hypothesis, but only lab tests utilizing designed experiments could confirm that the desert conditions *caused* the problem. After a week of careful testing of various ink viscosities and jet diameters, it was confirmed that very low humidity was responsible for the poor performance of the current ink formulation. Moreover, the extremely small diameter seemed to exacerbate the problem by frequent clogging due to dust particles when the ink dried out. A combination of a new ink formulation and a slightly larger jet diameter seemed to perform best under a wide variety of conditions, and this was confirmed in the field over the next 6 months.

3. FOUR FUNDAMENTALS FOR SUCCESS IN THE BIG DATA WORLD

In our experience [1], the Big Data literature ignores four data analysis fundamentals, critical for success, that we summarize below and in Figure 1:

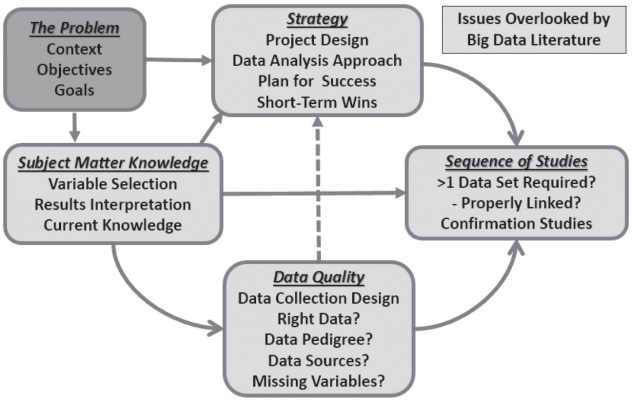

Fig. 1 Fundamentals critical to the success of big data projects. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

- Strategy—How We Will Win
- Value of subject matter knowledge
- Assessing data pedigree
- Sequential Studies

The inkjet printer case is an excellent example of the success that can come from paying close attention to these fundamentals. The problem was agreed to by all stakeholders, the production team, the design team and the sales and customer relation teams. Everyone understood that the objective was to understand the causes of the warranty claims and to adjust production based on the lessons learned. In this case, the key result of the data analysis was the formulation of the hypothesis that low humidity was the cause of the product failure. The analysts were extremely lucky. The data quality was poor, and there were missing variables. By a stroke of luck and a healthy amount of domain knowledge, the larger team was able to form a hypothesis from the location of the failures.

Of course, even that hypothesis was not the answer. A series of designed experiments, changing the ink viscosity and the nozzle diameter to work under a wider variety of environmental conditions, led to product improvement. However, the experiments followed directly from the knowledge gained by the data project. These modifications, in turn, were verified in the field after the new product design was shipped.

The important linkages highlighted above are discovered by adherence to the fundamentals. The following paragraphs explore these fundamentals more completely.

One needs to recognize at the outset that Big Data problems are different in more ways than just the size of the dataset. Big Data projects are generally not only large but complex and unstructured [2]. They generally involve datasets from different organizations and sources. One not only has to merge and integrate the data but also 'align' the agendas of the various groups involved with the objectives of the project. A different approach is clearly needed to deal with such problems. We refer to this approach as statistical engineering, the creative integration of multiple techniques to produce novel strategies for attacking such large, complex, unstructured problems [3].

3.1. Strategic Thinking—How Will We Make this Project Successful?

Early in the planning of a Big Data project, it is prudent to strategize how the project will be conducted and to produce a high-level plan for attacking the problem. The strategy defines how you will win. It defines the choices to be made—what you will and will not do. [4] admonishes us 'to begin with the end in mind'.

The 1998 Knowledge Discovery and Data Mining (KDD) conference included a now famous (or infamous) competition involving data from a direct mail campaign from a well-known philanthropic organization. The 21 teams that competed were given a training set of nearly 100,000 potential donor records that included almost 500 demographic and giving history variables. This training set included the results of the most recent solicitation, an indicator for whether the potential donor gave to the campaign and the amount received. The contestants also received a test set of 100,000 donors that included all the same variables as the training set, except for the two response variables that indicated the result of the solicitation.

The objective of the contest was to build a model that would predict who in the test set should have been sent the direct mail solicitation in order to maximize net donations. Using the model of the winner of the contest, by restricting mailing the solicitation only to a subset (about 55%) of the potential donors, one could have increased profit by about 41% over the 'mail to everyone' baseline strategy.

One of the authors (De Veaux) uses this dataset as the final project for a data-mining course. It simulates part of the Big Data paradigm, but of course, some of the pieces are already in place. The problem is well defined, the data are given and the objective is clear. However, students must still wrestle with variable selection, data cleaning and model building and assessment. One year, a student produced a model that gave results not only better than any student in previous years but one that would have won the 1998 contest (with a 45% profit increase). When De Veaux inquired what the student had done to develop such a great answer, he got the following response:

'I purposefully didn't start analyzing the data right away. Instead I took about three days to study the problem, understand the variables involved, re-express some of the variables and do some internet searches. From that research, I realized that although I had what seemed like plenty of variables, some combinations, including ratios of variables were missing, so I added a few more. An exploratory model supported my hypothesis that these variables were most likely to be informative, so I concentrated my data cleaning effort, because of the limited time, only on those key variables. The information and understanding gathered were very useful in developing the model and completing the analysis.'

Clearly, the student made understanding the subject matter part of her strategy to solve the problem. Although this is only an anecdote, it serves as a useful reminder of how important domain knowledge is to a successful Big Data effort. Unfortunately, some modelers do not incorporate their own business or scientific expertise into their data-mining explorations. Such knowledge is crucial to the success of the project because key variables are often not in the form that will ultimately lead to a successful model. In this sense, domain knowledge can be viewed as another missing link in many Big Data analyses. We will discuss the importance of subject matter knowledge in detail in the next section.

There are two aspects of strategy to consider when doing Big Data projects: project execution and data analysis and modeling. In project execution, we think about, among other things, the objective of the work (what success looks like), who to put on the team to bring the right skills to bear on the problem and who the key stakeholders for the project are. The strategy might also include creating short-term wins by dividing the work into pieces that can be completed in less than 3–6 months. People are impatient; they want to see progress. Short-term wins demonstrate that Big Data projects can be successful and return useful benefits.

The analysis strategy considers how the data pedigree will be assessed, how the data will be cleaned and integrated for effective and efficient analysis and what statistical modeling approach will be used. Some will jump straight to the analysis with no consideration of project execution or even analysis strategy. Such an approach generally has a 'low yield'.

3.2. Subject Matter Knowledge is an Important Ally

Subject matter knowledge is the next consideration that is underappreciated in current literature. 'Data have no meaning in themselves; they are meaningful only in relation a conceptual model of the phenomenon studied', wrote George E.P. Box, William G. Hunter and J. Stuart Hunter (1978) [5]. Without subject matter (or domain) knowledge, you have no context within which to define, analyze and interpret the results [6].

Some have taken the position that with vast amounts of data, subject matter knowledge and scientific theory are no longer needed (e.g. Anderson 2008 [7]) and claim that you can solve problems purely and empirically with data alone. Unfortunately, many analysis blunders have occurred because analysts did not understand the phenomenon under study, and they then misinterpreted the results [6].

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 (e.g., log, inverse and square root), selection of model form (e.g., linear, curvilinear, multiplicative and which variables should be considered x vs. y variables), interpretation of results and the ability to extrapolate findings. This principle was clearly illustrated by De Veaux's student.

One can do much more with data gathered and interpreted within the context of sound subject matter knowledge than can be done 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.

3.3. Assessing Data Pedigree

Understanding the pedigree of the data is particularly important because of the disparate sources of the Big Data. The assessment includes how the data were collected, the measurement process and the science, engineering and structure of the process. Assessing the pedigree of the data can help us avoid accepting poor quality data at face value and performing the wrong analysis of the data. Understanding the pedigree of the data results in a deep knowledge of the data and the associated process [8].

When assessing the pedigree of the data, a useful guiding principle is: do I really understand how the data were collected? Can I trace back and identify the origin of each data point? Such an assessment is enhanced by the use of graphics of data. The creation of process diagrams (schematics) is almost always helpful in assessing data pedigree and understanding the problem. The data pedigree should be assessed before, during and after the analysis:

- Before—understand the process, sampling procedure, data collection, analysis preparation and measurement system
- During—Constantly checking the data and results with the 'does this make sense' test aided with extensive use of graphical displays

- After—Evaluate the results to make sure results and conclusions make sense regarding what is known about the problem being investigated

It is also important to check the assumptions of the data collection process. When thinking about data quality per se, it is helpful to look for data that are clearly wrong (e.g., grossly atypical values, pregnant males, etc.), results and trends that do not make sense given the technical background of the problem and missing information and data critical to a useful analysis and making sound conclusions [9].

3.4. Sequential Approaches Are the Rule Not the Exception

While the mining of Big Data can potentially find interesting patterns and generate new hypotheses, it can never establish causality. Even a correlation based on large amounts of data is nothing more than a correlation. The only way to establish cause and effect is to use designed experiments to control some factors while purposefully manipulating others and then measuring the resulting response under a variety of randomized treatment conditions. One of the big challenges of designing experiments is the identification of a small number of factors to manipulate; this is where data mining comes in.

Obviously, one cannot follow the virtuous cycle between data mining and experimental design, a form of the iterative scientific method, if every problem is viewed as a 'one-shot study'. In our view, data competitions, such as the Netflix competition [6] or those on the website Kaggle.com, assume that the problem at hand is to develop the best possible model for a given set of data. Another underlying assumption is that additional data, such as from designed experiments, are not possible or needed. However, we must keep in mind that the data are not the problem but rather part of the solution. Statisticians and other analysts should never lose sight of the original problem they are trying to solve, such as improving the reliability of inkjet cartridges. Once the objective is clear, we are in a better position to think about how data might be part of the problem-solving process.

For example, we have found, in our collective experience, that in many cases the original problem cannot be solved with the initial data available. In other words, taking the time and effort to develop the best possible model for the original data set is often not a logical strategy but rather a waste of time and effort. A better approach is to consider the quality of the data and how analysis of this data would or would not be helpful. In some cases, the solution is in the original data, but in most cases, as in the inkjet case, it is not. In these situations, the key question to ask is not what is the best possible model for this data but rather what can we learn from this data to guide future data collection efforts in order to solve the original problem? This latter question may not require a model at all and often leads us to experimental design.

Of course, one experiment may not totally resolve the original problem either. We will no doubt learn from this experiment and answer some questions, but we may also raise new questions and suggest new theories. This is part of the iterative nature of the scientific method; breakthroughs in virtually all fields have come through iterative cycles of hypothesis generation, data collection, hypothesis testing and generation of new hypotheses based on the data analysis [5]. In our view, it is naive to expect that we can solve large, complex problems with the original data made available, regardless of how large the set may be. To solve real problems, we should rather expect that an iterative approach involving repeated cycles of data gathering and data analysis will be needed. Designed experiments can certainly accelerate the process and help us resolve the problem faster than analysis of happenstance data.

3.5. Use Big Data to Generate Hypotheses

Big Data projects typically involve diverse datasets collected in a number of different environments. Whenever data are collected by multiple sources, the opportunity for omissions and errors are always present. These result not only in missing values, missing variables, measurement variation and definition conflicts but also in missing identifying variables or links that connect the two datasets and allow the analysis to use the variables in both.

Even when a data mining-effort produces a model that seems to point to new insights, one should always proceed with caution. Big Data models are typically built from happenstance data. Even though they may point to associations between variables that can be manipulated to achieve a desired response, they cannot provide a causal link between the two. There may be other lurking variables that actually caused the response to change that either were not measured or were so correlated with other variables that their contribution was masked.

By cycling between discovery and experimentation, the knowledge process can progress. Even though data mining cannot answer the question, it can pose new questions, and insights gained from empirical models can be used to accelerate the learning from the experimental process. While Big Data has perhaps been given too much credit for proving hypotheses, we think it has not been given enough credit for generating new hypotheses. It is in this latter capacity that we think the true power of Big Data lies.

4. CONCLUSIONS

In this age of Big Data, we need to both capitalize on what these data can bring to our decisions and realize their limitations. By using Big Data analytics in conjunction with sound subject matter knowledge and designed experiments, we have a strategy that can both generate and verify hypotheses in an iterative fashion. Neither analytics or designed experiments alone will necessarily guarantee success, but the synergy resulting from the judicious linking of the two will help solve the problem of the missing link and result in the greatest probability of success.

REFERENCES

[1] R. D. Snee, R. R. DeVeaux, and R. W. Hoerl, Follow the fundamentals - four data analysis basics will help you do big data projects the right way, Quality Progress, January 2014 (2014), 24–28.

[2] A. DiBenedetto, R. W. Hoerl, and R. D. Snee, Solving Jigsaw Puzzles: Addressing Large, Complex and Unstructured Problems, Quality Progress, 2014, 50–53.

[3] R. W. Hoerl, and R. D. Snee, Closing the Gap; Statistical Engineering Can Bridge Statistical Thinking With Methods and Tools, Qual Progress, 2010, 52–53.

[4] S. R. Covey, Seven Habits of Highly Effective People, Franklin •Covey, 1989.

[5] G. E. P. Box, W. G. Hunter, and J. S. Hunter, Statistics for Experimenters, New York, John Wiley and Sons, 1978, 291 pp.

[6] R. W., Hoerl, R. D. Snee, and R.D., De Veaux. Applying Statistical Thinking to 'Big Data' Problems, Wiley Interdisciplinary Reviews: Computational Statistics, July/August (2014), 221–232. (doi: 10.1002/wics.1306).

[7] C. Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired Magazine*, Issue 16.07, 2008. URL http://www.wired.com/wired/issue/16-07.

[8] R. D. Snee and R. W. Hoerl, Inquiry on Pedigree—Do You Know the Quality and Origin of Your Data?, Quality Progress, December 2012 (2012), 66–68

[9] R. D. De Veaux and D. J. Hand, How to lie with Bad Data, Stat Sci, **3** (2005)**,** 231–238.

[10] National Research Council, *Frontiers• in Massive Data Analysis*. Washington, D.C, The National Academies Press, 2013.

AQ6

AQ7

QUERIES TO BE ANSWERED BY AUTHOR

IMPORTANT NOTE: Please mark your corrections and answers to these queries directly onto the proof at the relevant place. DO NOT mark your corrections on this query sheet.

Queries from the Copyeditor:

AQ1. Please confirm that given names (red) and surnames/family names (green) have been identified correctly

AQ2. Please provide the zipcode for affiliation 3.

AQ3. Please provide the revised date for the article.

AQ4. This capitalization has been retained across the file. Please confirm if this is fine.

AQ5. Please check as the Key Words section is missing.

AQ6. Please provide publisher location for ref. Covey (1989).

AQ7. Please note that ref. 10 has not been cited in text. Please provide it in the text or delete in the list.