Rejoinder - Experiences with Big Data: Accounts from a Data Scientist's Perspective

Kulahci, Murat; Frumosu, Flavia Dalia; Khan, Abdul Rauf; Rønsch, Georg Ørnskov; Spooner, Max Peter

Published in:
Quality Engineering

Link to article, DOI:
10.1080/08982112.2020.1808223

Publication date:
2020

Document Version
Peer reviewed version

REJOINDER: Experiences with Big Data: Accounts from a Data Scientist's Perspective

Murat Kulahci[1,2], Flavia Dalia Frumosu[1], Abdul Rauf Khan[1], Georg Ørnskov Rønsch[1], Max Peter Spooner[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

We would like to thank all discussants for their valuable comments and additional remarks on our views on Big Data applications in industry. Many of the discussants expressed their agreement with our experiences and lessons learned along the way. As noted in a disclaimer in the paper, our views have been shaped and somewhat constrained by our own experiences. Therefore, it was very comforting to read confirmatory remarks by a group of esteemed scholars who have been actively involved in industry projects and arrived at similar conclusions. We also value greatly their further comments on issues that were missing in our own accounts. We would like to address some of the general themes that emerged in the points raised by the discussants.

In many of our projects, the analysis of the existing data suggested the need for more data from the same source and at times from new sources as well. The latter usually comes at a higher acquisition cost as it often includes expenses like the installation of new sensors and revising the IT infrastructure. Hence, prior to such an undertaking, a cost/benefit analysis is needed to balance the potential information content of the data against the cost of acquiring it. In one of our current projects, we worked on the effective use of different data sources in increasing quality but acquired in increasing cost as well. This is done initially on a pilot study to provide the necessary evidence for the decision makers in establishing a more elaborate data acquisition scheme. Our final conclusion was that only a slight modification to the current data acquisition system that would provide more detailed process data was sufficient in delivering good models to predict the product quality. We encourage such studies (whenever

possible) to be performed before any hasty decision on collecting more data from the process.

An issue raised by some of the discussants is about the traceability of the data from raw material to the final product and maybe even beyond. As we indicated in the paper, this is often very hard to establish and usually missing in many manufacturing applications. We are for example currently working on a collaboration on digital fingerprinting technologies that can be embedded on products to be read in real time through optical means. Particularly in high-volume manufacturing, this will greatly facilitate synchronizing process data with product characteristics data for process improvement and optimization purposes. Moreover, we expect this to have an impact on warranty and reliability studies as the product can be traced back even after it has been in use for a while. In that sense, such technologies will not only ensure the quality of the product during manufacturing but also help establish "the quality over time" of the product as well.

The most common discussion point was on the differences in the characteristics of the data. This was brought up under different comparative scenarios such as experimental vs. observational, structured vs. unstructured or quality vs. quantity. We do certainly agree with the discussants pointing out that Big Data is usually observational and unstructured. We do also share the majority view that experimental data often available in small quantities can be of extreme quality and address many production related issues more efficiently. We are however equally aware of the cost of obtaining such data. It has always been a struggle to convince the management to run production experiments as they are seen as, at best, a hindrance to everyday operations and at worst, a waste of time and money. That has certainly contributed to the allure of the observational data as it has become more abundant. In recent years, we have in various occasions heard researchers casually reading the obituary of experimental design as we know it. Our experience has been somewhat different. In academia particularly, the experimental work is as strong as ever. In industry, even in the presence of abundant observational data, we have had some opportunities to run experiments. This could very well be an unintended consequence of understanding and appreciating the value of high-quality data when observational data fails to match up to the hype. Furthermore, in some of our projects the goal is to develop a digital

twin of the production environment. This is going to create many opportunities for virtual experiments in which some production concerns can be addressed. Similarly, we expect our future research activities in experimentation using discrete event simulation, computational fluid dynamics and finite element analysis to expand with various applications in manufacturing.

Over the years, we have also had some opportunities for combining both experimental and observational data. In some of our projects, the processes in question would operate under very stable conditions based on tight controls, years of experience or due to the intended consequence of quality improvement schemes such as Six Sigma. Resulting lack of variation in product characteristics or in key performance indicators hampers the attempts of building predictive models. This has been of particular importance when using machine learning and deep learning methods that demand large amounts of data to train the models. We have successfully used designed experiments to systematically go beyond the range of operating conditions to generate more variation in the process data and consequently in the output characteristics. We suggest however such attempts to be performed with care due to the danger of selecting a range that can have detrimental consequences.

Another example illustrates how the observational data can guide the experimental efforts. The aim of the project is to figure out the important factors such as corrosion causing product failures in electronic components. The industry partners have international market presence and some failure modes are known to be a function of the weather conditions on the location of use. In the first phase of the project, we developed a scheme to cluster the world into zones based on any relevant weather characteristics such as temperature and humidity. Subsequently, the profiles of these characteristics typical to each cluster were obtained. These were then given to experimenters in order to run experimental plans reflecting realistic use conditions for different markets the industry partner may wish to enter. The tool comes in the form of an easy to use and flexible dashboard for the project partners.

Some discussants also mentioned educating the new breed of data analysts/scientists. We have had similar discussions in our latest PhD hires. The candidates for a data analytics position tend to fall into two categories: statisticians or computer scientists.

However, in most of our projects, we need to use a mixed set of tools from basic data exploration techniques to sophisticated machine/deep learning algorithms. Then the question becomes which educational background is more malleable to the needs of the project and learning new methods. We believe many universities are in the process of looking into this gap. Yet we fear when it comes to the ultimate outlet, the priority will be given to the analytics needs in areas like social media, image analysis, healthcare and finance. Therefore, a more concerted effort needs to be made for manufacturing. There is certainly a rapidly growing need for scientists skilled in production analytics methods and the thirst for such talent will not be satiated anytime soon.

Another common theme in the discussions is about the hurdles surrounding the implementation of data analytics methods in industry. Data readiness as a subset of digitalization readiness is a major concern when it comes to effective and efficient use of production analytics methods. It certainly does not help that these efforts may sometimes look like an attempt in “Keeping up Appearances”. It is not a new phenomenon for upper management to cling on to a management fad with unfounded or even unrealistic hopes. We can sit on the analytics side and criticize this all we want. The reality is that this is not temporary and the need/demand for data analytics in production is going to grow further. We have often had our best luck with engineers in terms of managing expectations even though (or most likely because) they carry a healthy dose of skepticism towards data analytics methods. They will remain our best allies in this pursuit. In any rate, the road to success starts with actually attempting to get on that road and winning the industry over with what data analytics can offer. We should indeed not shy away from what we would call a consultant’s quick fix if that gets your foot in the door. With properly managed expectations and a growing number of success stories under our belt, we should be able to pave the way towards better utilization of production data and contribute to our industries accordingly.