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:
[10.1080/08982112.2019.1686641](https://doi.org/10.1080/08982112.2019.1686641)

Publication date:
2020

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Kulahci, M., Frumosu, F. D., Khan, A. R., Rønsch, G. Ø., & Spooner, M. P. (2020). Experiences with big data: Accounts from a data scientist's perspective. *Quality Engineering*, 32(4), 529-542.
<https://doi.org/10.1080/08982112.2019.1686641>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

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

Murat Kulahci^{1,2}, Flavia Dalia Frumosu¹, Abdul Rauf Khan¹, Georg Ørnskov Rønsch¹, Max Peter Spooner¹

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

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

Abstract

Manufacturing has been rejuvenated by automation and digitalization. This has brought forth the new industrial era also called Industry 4.0. During the last few years we have collaborated with companies from various industries that have all been going through this transformation. Through these collaborations, we have collected numerous examples of (sometimes troublesome) experiences with Big Data applications of production analytics. These experiences reflect the current state of production data and the challenges it poses. Our goal in this paper is to share those experiences and lessons learned in dealing with practical issues from data acquisition to data management and finally to data analytics.

Keywords: Industry 4.0, Manufacturing, Digitalization, Big Data, Production Analytics

1. Introduction

A new industrial revolution is upon us. Aptly named Industry 4.0 (Lasi et al. (2014)), this fourth revolution is mainly driven by automation and information exchange with heavy reliance on digitalization, see Figure 1. Accordingly, many companies from various industries are all going through this transformation where the goal is to digitalize the information content in production and hence enhance knowledge about their processes. Usually, the expectation is that the abundant production data should contain information that can be used to further the goals of the company (e.g., reduce operating costs and/or increase profits). Moreover, the edge some companies have obtained through past movements such as Lean Production and Six Sigma may be winding down resulting in a search for new approaches.

Figure 1. Industrial revolutions from the 18th century until the present

In their pursuit towards digitalization, many companies from a wide range of industries are now frantically scrambling to collect “more data” from their processes under the wishful thinking that it contains the “necessary information.” It has however been our unfortunate and yet repeated observation that this is done in a frenzy with little consideration to key questions such as:

- What kind of problems do we actually want to solve?
- What kind of data is in fact needed and how can we collect such data efficiently?
- How can we handle such data properly and how can we make sure the extracted information is used most effectively?

Instead, this haste in data collection often generates more problems than it solves. As a response, academics and practitioners alike rush into yet another frenzy of dealing with these problems, which could very well be simple artifacts of poor planning in data collection or in problem solving. Over the years, we have had the privilege of working on many projects in collaboration with industry partners where the analysis of (big) production data has been the focus. These projects involved various industries ranging from high volume parts manufacturing as in injection molding and electronics as well as batch process manufacturing in food, pharmaceuticals and bio-based production. Granted our collaborations may not constitute a “big enough” sample for overreaching generalizations, but the small amount of variation in the common hurdles we needed to overcome leads us to believe that at least some of these issues are quite widespread. It should also be noted that throughout the paper, we restrict our attention to data analytics in manufacturing and hence use the term *production analytics* to summarize our efforts in product development and improvement as well as in process improvement, optimization, monitoring and control. Some of our current work in predictive maintenance as presented in Section 5.4, *New Opportunities*, would also fall under this umbrella. Since the experiences cited in this paper are mainly in manufacturing, we encourage readers from other fields of data analytics applications to draw their own parallels to our conclusions with caution and as they see fit. In the rest of the paper, we present our observations regarding general expectations in the Big Data movement followed by a discussion of issues related to data management, content and analysis. We conclude our paper with the concept of digitalization readiness, as we have perceived it through our interactions with our industrial partners.

2. A pre-requisite: Matching expectations

Many of our production analytics collaborations involve two distinct parties: problem owner (industry) and problem solver (consultant/academia). Needless to say, both of these parties need to work together for a successful completion of any production analytics project. However, the increased complexity of production challenges has resulted in an ever-widening gap between academic output and industrial needs. The examples included in textbooks remain too sterile and new theoretical solutions developed by the academic research tend to get buried in journals that are not on the

industry radar. Moreover, as in any other relationship, the success in a production analytics collaboration is closely tied to matching the respective expectations. That is, while close collaboration and commitment are preliminary conditions, we need to go further and establish early on a realistic mutual understanding of what is to be achieved.

The current inclination in industry, as they face an increasing pressure in highly competitive markets, is whenever possible to simply hand over the production data to the data analyst for the discovery of any hidden information. Another expectation from the industry is to obtain a fast and simple solution. We should at this point re-emphasize that this paper refers to our collaboration with industry as an external agent. There are companies with expanding in-house data science teams established towards achieving similar goals. Yet their journey may be plagued by problems that would be different from the ones we have encountered and therefore we refrain from elaborating on them further.

The seemingly insurmountable amount of production data and increased complexity of the problems combined with the reassuring promises from software companies and academics to deliver “the” answer through empirical evidence may lead the industry to assume a rather passive role. This should however, not be taken as a criticism since it is a natural reaction to ever evolving targets and the progression of digitalization and Internet of Things (IoT) by the production engineers who are not necessarily trained to address these new challenges. When we couple this with the increasing concerns of upper management believing that their competitors have a much better handle on digitalization, it is no wonder that we often hear companies proclaiming “We need AI!”. This however is seldom accompanied with a clear vision of its ultimate use and purpose. In our limited view, we may certainly be overlooking a potential “If you build it, he will come!” moment as in the movie “Field of Dreams” but we do nonetheless strongly believe that there are still plenty of opportunities for the use of real intelligence when it comes to manufacturing. This however, requires the industry to actively participate in data analytics efforts and not simply remain as an outside spectator expecting a data machinery to spit out an outcome that will alleviate all their problems. This includes not only providing useful input and feedback based on their process knowledge but also understanding and appreciating the capabilities and more importantly the shortcomings of the methods used therein.

The data analysts, on the other hand, tend to have little to no knowledge about the processes they work on. Yet they are usually conditioned to start any data analysis with defining the problem. This should not come as a surprise as most data analysts, similar to engineers, are trained in the basic scientific method in which setting up the hypothesis constitutes a crucial step. In the past, data analysts would often rely on the industry partner's expertise to help with this important juncture. As challenging as this may be at times, most engineers would have a reasonable handle on their process and could readily come up with suggestions. In case of predictive modeling, for example, it would not be too far-fetched to think that the engineers could provide a list of output(s) and potential inputs based on past experience and process knowledge. This would often be possible as that list would have usually consisted of a relatively limited number of process variables which were deemed important enough to collect the data for in the first place. In modern production however, the list of potential inputs and outputs has for almost any process expanded with the advances in sensor, data acquisition and storage technologies. It should be noted that this expansion has not necessarily been a direct result of a well thought out process in data gathering but rather a side-effect of a haste in collecting, often indiscriminately, everything we can get from the process. Unfortunately this results in stretching the process knowledge of the industry partners beyond their comfort zone and often prevents them from offering more concrete help in defining the problem to tackle. Hence to avoid an impasse, the data analysts have to adjust their expectations about the help they would traditionally require in defining the problem and seek to be part of that process based on empirical evidence. For example, if historical data is available, a preliminary analysis involving all potential variables, both as outputs and inputs, should be performed and observed correlations should be examined with the process experts for relevance and practicality. That in turn would help to define the problem and lead to a more focused study. In that sense, historical data can serve in setting up hypotheses. However, if the historical data does not reveal relevant information due to the many reasons we will discuss in the following sections, this initial study would still serve as the basis for what to do next in terms of what type of data is needed and how to collect it. In the next sections, we discuss many data related issues we have faced and provide, when possible, solutions we have adopted to overcome these issues.

3. Data Management

The promise of Big Data is one of the main characteristics of digitalization of modern production. Yet it often comes at a price. Before indulging in the content and the analysis of data, one has to deal with data management issues. Below, we discuss some of those issues. In many ways, our experiences and conclusions presented in this and subsequent sections follow the same path of data readiness levels suggested by Lawrence (2017), where the author proposes three data readiness levels concerning data accessibility, cleanliness and relevance.

3.1 Accessing the Data

Until recently, access to data would be achieved primarily through floppy disks, compact disks (CD's), memory sticks and attachments to emails. The sheer volume of production data from modern manufacturing processes renders these data transfer media inadequate. Nowadays, we often resort to accessing the data directly at its source or at the very least at a virtual data warehouse, e.g., cloud. This however has proven to be quite challenging as well. Connection to cloud storage or to various databases with varying protocols requires different expertise than most data analysts have been trained for. Furthermore, the collaboration with industry is usually established through engineers and operators who themselves have varying levels of understanding of data acquisition systems. Therefore, a close collaboration with IT support in these companies becomes crucial for the success of the project. As IT departments often have other priorities, such as maintaining the security of production data, granting access to an external user usually creates more concerns on their part. Therefore, rather than granting full access to the data, the data analyst is given the data on a need-to-have basis. In this case, the engineer in the collaboration plays the role of an intermediary in providing access to the data. Hence, in a typical transaction, the data analyst would need to go through the engineer to receive the data.

This two-step process often delays the data analysis greatly due to communication issues and prioritization mismatches. Ultimately, to avoid delays in data retrieval and analysis, the data analyst has to take the initiative to gain direct access to data. This brings back the original obstacle of the data analyst lacking the necessary background to cope with modern database systems and protocols for data transfer. Even in the case of a data analyst being well versed in these fields, security

concerns remain an issue, especially when handling sensitive production data. These concerns are also valid in the use of cloud applications or setting up remote access for the data analyst. Establishing a proper and convenient data access protocol is a must before any serious attempt at data analysis.

In our applications so far, the most reliable option has been gaining access by Virtual Private Network (VPN) to the company's database through a portable computer, which is specifically configured by the company's IT support. This brings additional challenges such as the speed in data processing and security of the device. Yet it still remains to be the preferred solution in many circumstances.

It should also be noted that the IT department and the engineers have different expectations from production data as they have different aims and priorities. Engineers would like to make the best use of the data to achieve important production goals. The IT department, on the other hand, is mainly concerned with data management, which includes acquisition, transfer, storage, security and access. In that regard, what is convenient for one does not necessarily provide the same convenience for the other. One inadvertent outcome of rapid digitalization in most companies has been that the IT department has become the owner of the data with the above-mentioned priorities. Hence, engineers much like data analysts will need to adapt to this new situation by making the utmost effort in learning more about data management to establish a common language with the IT department.

As for data analysts working with production data, the days of sitting in front of the computer waiting for the data to arrive are over. The distinction has been made between data analyst and data scientist in that the latter is expected to be more computer savvy and knowledgeable in data management. This transition needs to be made to remain a relevant player in production analytics efforts.

3.2 Merging Databases

As engineers and data analysts working with processes, we tend to focus on process and product data with little to no focus on more managerial issues such as operational and planning concerns. Hence, we often see our work as limited to making sense of production data. As relevant as this is, other types of data are also available and should be used in conjunction with process data to provide relevant and timely information for decision makers. A complete dataset consists of information from the company's Enterprise Resource Planning (ERP) system (Meer (2005)),

Manufacturing Execution System (MES) (Meyer et al. (2009)) and historians containing process data (plant data), see Figure 2 (Liu et al. (2002)). These IT systems are often used as separate tools (and therefore not necessarily linked) and in many cases not designed for extraction of large amounts of data, which is mainly the case for ERP and MES. This makes it even more challenging to construct complete datasets encompassing all these systems. This is because of the current limited capacity for data extraction but also because the level of aggregation often differs in these systems. We may retract back to our old modus operandi and isolate our efforts to production. But our prospective solution as great as it may very well be, will most likely fall on deaf ears if it cannot be planned for properly or is not implementable at all.

Figure 2. Illustration of a typical data hierarchy within a manufacturing company

Furthermore, besides the main data platforms described above, companies often use specialized systems for lab and quality measurements (e.g., Lab Information Management System (LIMS)) and local systems in R&D functions. This makes it even more challenging to construct the needed dataset. Therefore, with all its challenges, the urgency of connecting various datasets reflecting different components of daily operations is becoming more and more relevant.

3.3 Traceability

Many processes consist of several stages where raw material goes through a series of unit operations before the final product is obtained. When the aim is a comprehensive data analysis solution, which considers all these unit operations at once, traceability becomes a crucial issue. However, true cradle to grave traceability is hard to obtain, particularly in chemical processes. Often different production streams are combined and/or split to make the traceability almost impossible.

In one of our projects, the raw material was going through a series of operations in batches before the final product is packaged and shipped. Each of these operations is very tightly monitored and controlled through a series of sensors and controllers. This naturally gave the impression of the existence of traceability and availability of abundant process data. Upon further analysis, we realized that we actually had data for only a few batches from the beginning to the end. This was not because the data was not available. The data for all batches was simply collected locally for each operation without necessarily considering the preceding or following operations. The batches were tagged accordingly and it required an intimate knowledge of the process to match these tags for different operations. The data analyst working remotely lacked this knowledge and hence ended up with a significantly low number of batches with traceability, which clearly did not represent the reality. This problem was resolved in an accelerated manner by a second data analyst who was more familiar with the process and had direct access to key process engineers and IT personnel.

Another example we have encountered involved a company producing high precision metal parts in a milling process. The metal part undergoes several steps in different milling machines using different milling tools. The goal of the analysis was to develop a model for predicting tool wear and when tools should be changed. The dataset consisted of part information extracted from CAD, task information (type of tool, milling steps, duration, speed, etc.), machine tool measurements (length and diameter), and tool wear evaluations. When the first obstacles with data extraction were overcome, including consolidation of data from different local databases and the analysis started, it became clear that there were some major issues with traceability. The final evaluation of the tool wear was only done on a subset of the tools reducing the dataset dramatically. Also there were no ID numbers on the individual tools, making it impossible to link the results back to the tool usage. The key learning from

the project was that there were missing tags to combine different data sources, key process variables were often not collected and the data structure was not suitable for the required analysis.

Traceability is of essence if the overall process is to be scrutinized. While many companies make great strides towards achieving it, the impression is that in many cases we are still a long way from tracking down a single product all the way back to the raw material. When starting a new project involving the entire production steps, we highly recommend establishing a flow diagram indicating each process step and availability of the data therein. Then the next step is to understand and establish (if none existing) the connections among these local data sources for assuring traceability. As self-evident as this recommendation may be, it has been our experience that this was overlooked in the mist of talks of Big Data, cloud computing, and newfangled data analytics methods.

4. Data Content

4.1 Historical Data

It is not uncommon to hear in our early meetings with the companies that they have a vast amount of historical production data.

However, in many cases this claim can be quite misleading. First of all, operating conditions in any unit operation let alone in the entire production can drastically change over time to make parts of the historical data incompatible with the rest. It should be noted that what we mean with drastic changes are changes beyond the unavoidable regime changes that processes usually go through, for example, a recent renovation of the plant, getting brand-new production equipment or permanently changing raw material with completely different characteristics as in the case of eco-friendly products. This may very well render previously obtained data using for instance the old production equipment incompatible with data obtained after the renovation.

Data collection schemes and measurement systems are also prime examples of such changes. In a particular study on railways, the geometric measurements of the tracks were regularly collected by a so-called measuring car. These measurements were then used to quantify the wear and tear that the tracks were experiencing in time and a degradation model was to be constructed for better scheduling of maintenance activities such as tamping of the tracks. These measurements would sometimes be

taken by subcontractors that may change the measuring system in time. Furthermore, different subcontractors from different countries may have used different equipment with varying tolerances. This can once again render different parts of the historical data being incompatible for further analysis.

Even in the rare circumstances when the above-mentioned changes do not occur, there could still be issues rendering parts of the historical data irreconcilable. Historical data was often collected for documentation and control purposes, and not necessarily for analysis toward better understanding of the process or for its optimization. That is, old data may simply be used for example in case of a recall or to identify and isolate potential batches of products affected by a production problem. Since the process improvement would not necessarily be the primary goal in the collection and storage of historical data, it may correspondingly not contain valuable information to address serious production questions of today. This should by no means be a deterrent from the use of historical data but it should simply be taken as a word of caution when it comes to claims of “plenty of historical data” being available at the start of a project.

4.2 Multiple Production Sites

Over the last few years, one of the main activities of the IT departments has been building data historians. These are practically data warehouses where data from different unit operations and production sites is gathered, see Figure 3. The goal has been not only to provide easy access to data collected in multiple production sites but also to potentially generate opportunities for real time applications such as process surveillance and predictive maintenance. However, often those production sites are subjected to different operating conditions that may severely impair the possibility of combining data from different sites.

In one particular application, we worked on a fermentation process from three production sites. The idea was to combine the data to see the differences in key performance indicators, identify the causes for these differences and also perform process surveillance. However, these processes were tightly controlled with engineering control schemes, which differ from site to site; analogue and digital control schemes. Since the former is considered to be an old technology, we were then asked to focus only on the sites for which the digital controller is used. Furthermore, for one of the remaining two sites, it was deemed that the local crew was acting too

differently from the main production site when it came to day-to-day operations and hence the large variation in key performance indicators could simply be attributed to this “cultural” difference. Considering the fact that the fermentation process was exposed to not only engineering control but also to manual control performed at the discretion of the skilled operators to avoid wasting any batch, it made more sense to initially focus on a single site with more relevant and somewhat more stable production characteristics.

In this case, one site was using a soon to be obsolete technology and it did not make sense to include the data coming from that process. For another, manual interference that was hard to predict with the available data limited our focus to production from one site only. In fact, manual control of the operations by experienced operators is not an uncommon occurrence even in modern production processes. The true digitalization should aim to minimize this and incorporate that knowledge gained based on years of experience through automated systems. Ultimately though, site-to-site differences are of course extremely valuable to extract in order to eventually minimize these differences and learn the impact of different operating conditions on the final product. In that regard, merging the datasets from different production sites will indeed be of importance but needs to be done with care.

Figure 3. Establishing data historians (Computer clip arts are obtained from <http://clipart-library.com/cloud-server-cliparts.html>)

4.3 Multi-stage Processes

In many industrial processes consisting of multiple stages, the focus has historically been on unit operations sometimes taken care of by different, semi-autonomous groups within production. Each group would be responsible for its own unit operation and optimization would be performed locally with data collected and labeled accordingly. As a result, the data collection is seldom standardized, and connecting all these data becomes extremely challenging. This is also related to the traceability issue we discussed earlier. Another concern is that optimized individual unit operations do not necessarily translate into an optimized overall production. In the same vein, we have seen great interest from industry to pursue a general approach to data analysis and consider all stages of production at once. The aim in such a study is usually to discover the relationship of process variables from different stages of production with a final key performance indicator such as product quality. This could be achieved through observational or experimental data. The former often requires methods beyond standard statistical methods due to characteristics of data discussed below. However, it should be noted that the observational data could be used to obtain predictive models as in the case of machine learning methods (Hastie et al. (2009)). These can in turn be used for risk management, i.e., if you can predict accurately what is going to happen, you can take actions to hedge the risk in anticipation accordingly. However, if the goal is to inflict a desired change in key performance indicators through adjustments to the process variables, we need to uncover causal relationships. The time-tested approach to achieve this is through controlled experiments. The design and analysis of experiments for multi-stage processes can be elaborate and we currently need more methodological and applied work in this field (Tyssedal et al. (2011), Tyssedal and Kulahci (2015), Kulahci and Tyssedal (2016)).

Another concern related to considering all production stages at once is that the data collected in these different stages comes in great variety; numerical, categorical, text, etc. This so-called mixed data represents a significant hurdle for many analysis methods to overcome.

A somewhat similar circumstance happens when production occurs in multiple parallel lines with varying production equipment. We have had experience with processes involving injection molding where tens if not hundreds of molding machines in one production site is used on a continuous basis. The need to combine data from all these machines could for example, compensate for the scarcity of rare

events that are to be predicted. We have been encountering this issue more and more frequently with the proliferation of Six Sigma processes for example, where low defect rates create unintended consequences in modeling as in classification (Khan et al. (2017a, 2017b), Khan et al. (2018)) as well as in statistical process control (Agresti and Coull (1998), Montgomery (2013), Wang (2009)). We will discuss this issue further in Section 5.2.

4.4 Data Transformation

From the time, data is measured by the sensors in production to the time that it reaches the data analyst, it would usually undergo several transformations. Firstly, the sensor itself will transform its input signals, depending on the sensor configuration, to produce the initial measurement, which may be taken by the sensor at unevenly spaced time intervals. Measurements from the sensor may be further manipulated upon storage in a data historian. Typically, the data is compressed in order to save storage space and increase retrieval speed (Thornhill et al. (2004)). A common compression method used by data historians is to only retain those measurements from a sensor that differ by some minimum amount compared to the previously stored measurement from that sensor. Finally, upon retrieval of data from the data historian, further manipulation can be made where the data historian itself can often by default fill in the previously discarded measurements according to the method selected by the user, and provide the data in a convenient format. This final manipulation is perhaps the most insidious, as interpolated data can be mistaken for actual observations measured during production. As a result, of these transformations, the data takes on an ephemeral quality and it can be a challenge for the data analyst to know what version of the data has been acquired.

We have found that the best approach in these situations is to insist on obtaining data from as close to the source as possible – which usually is in the form in which it is stored in the data historian. As tempting as it may be, it is misguided to rely on the retrieval modes of the data historian to provide a ready-to-use dataset, as important features of the data may be masked. For example, we have experienced cases where one variable was stored at the rate of less than 10 observations per hour, whilst another variable was stored at the rate of several thousand observations per hour. If we had not insisted on the data in its rawest format, but relied on the data historian's reconstructed version of the data where all variables were evenly spaced in

time, we would have been unaware of this huge difference in sampling frequency, and unable to advise the company to reexamine the data storage parameters. Many analysis approaches to serially dependent data assume regularly spaced sampling and cannot handle irregular sampling frequencies, so some interpolation/imputation is required. By insisting on acquiring the raw data, the data analyst has at least a chance of understanding what is truly being measured, and has control over the interpolation method being used. Some multiresolution/multi-granularity methods have been proposed with the aim of dealing directly with data containing different granularities (Li et al. (2001), Reis (2019)) but the problem has otherwise been somewhat overlooked.

One of our projects involved injection molding of plastic components. These processes often operate at high production rates in a matter of a few seconds per “shot.” They also show characteristics of batch processes for which each injection molding shot is a batch and process variables such as temperature and pressure in the mold follow expected profiles during the batch. However, due to physical constraints (e.g., difficulty of mounting sensors in the injection molding machine) and the high production rate, these profiles are seldom available. Instead, pre-determined features of the profiles as in maximum or steady state values may be available. Sometimes, the profiles are shown for each shot for the operator’s instantaneous review but never to be stored afterwards. It was our first instinct to inquire about those profiles and ask to have access to that data accordingly. The company uses different equipment suppliers for the same product and only some allowed access to that data. Even then, a lot of IT work was needed to actually access those profiles. In the end, insisting in obtaining the entire profile data for the overall production or generating solutions based on the assumption of easy access to those profiles would ultimately require replacement and/or adjustment of production equipment. Hence, before venturing down that road, we needed to properly investigate the features that were readily accessible. Even if these features did not contain the right information, the costly adjustment of production equipment needed to be justified against the potential benefit of full access to production data. For this, we would like to once again refer back to the question we started with: how much and what type of data do we actually need? The systematic collection of data as we have advocated would suggest to tone down our inhibitions about the amount of data or lack thereof for that matter, gradually work with what is

available and determine the next steps.

5. Data Analysis

5.1. Combining Process and Product Data

One of the key issues in process understanding, optimization and control is to connect process data with product characteristics. The process data is usually collected with the help of sensors and through automated data collection schemes. When companies claim to possess large amounts of data, they often refer to process data. Such data can be used in process monitoring, for example. Modeling of such data is done through so-called unsupervised learning methods where no response, e.g., product characteristics, is available.

Except for the case of 100% inspection, which has been an exception rather than the norm in quality assurance, product quality data is often more scarcely collected. This is due to issues such as cost of inspection, physical constraints and production rate. Fast production rates make it particularly difficult to obtain product quality for each product. When that happens, only a fraction of process data has the corresponding product quality characteristics. Depending on the magnitude of that fraction, this can result in cases where a majority of process data cannot be used for modeling the product characteristics, see Figure 4. In this case, semi-supervised methods where supervised and unsupervised data are combined in the same modeling approach can provide some relief (Frumosu and Kulahci (2018, 2019)). In these methods, the goal is to make use of unlabeled (unsupervised) data for which no output is available to improve, for example, the predictive model, which will otherwise be based solely on the labeled (supervised) data. In real life applications, the main concern is the similarity of the unlabeled and labeled process data to allow for merging these two for predictive modeling. Also, for fast production rates, the issue of serial dependence (autocorrelation) in process data becomes more evident and needs to be taken into consideration in many production analytics applications (Bisgaard and Kulahci (2011), De Ketelaere et al. (2015), Vanhatalo and Kulahci (2015a, 2015b), Vanhatalo et al. (2017)).

Eventually, we can aim for all labeled data, i.e., an established, direct connection between the product characteristics and corresponding process data. More sensorics applications are certainly needed to accomplish that. The advances in barcodes and QR codes for example will allow for tracing each product throughout

production and even beyond. These codes can provide dynamic information, i.e., product specific information at the time of production or static information, e.g., ingredients in all pills with the same contents in tablet production in pharmaceutical industry. If the traceable product with encoded dynamic information can then be inspected automatically through for example, image analysis, the connection between the process data and the product characteristics can be established. We should once again emphasize that this will require a multi-disciplinary approach where engineers and IT personnel will work towards developing the hardware and software for tags based on the physical limitations of the process and products, sensorics experts will develop the right inspection scheme of an individual product and finally the data analyst will not only analyze the data but also support the efforts to encode the needed process information in the tags. In that sense, it is of utmost importance for the data analyst to be involved early in these efforts to avoid delays further down the line.

Figure 4. Semi-supervised learning where the original data (X) is split into supervised (X₁) and unsupervised (X₂) components, and a predictive model is established using both.

5.2 Lack of Specialty Data

Often the data available to the data analyst originates from routine production, where the quality characteristics of the end product must lie within narrow specification limits. Many manufacturing companies have gone through quality management schemes like Six Sigma and Lean Production and therefore, they are usually capable in producing products with desired specifications. Consequently, it is quite common

to end up with a dataset containing very little variation in the measured production settings and quality characteristics. It has been on the other hand noted by data analysts that more interesting discoveries can be made when things go wrongly (Borne, 2014), so a lack of process and/or product failures can be an obstacle to new insights.

Similarly, a common starting goal in many projects is to develop a model that can predict end-product quality based on the measured production conditions at some point earlier in the process. A lack of variation in the observed product quality in the available data poses a major challenge in the identification of the factors that lead to good or bad quality. This challenge cannot necessarily be addressed by acquiring more data, if the new data is just “more of the same”. One approach to consider is to revise the goal and look for features in the data that do display promising variation. For example, in one project we ended up focusing on the production duration of each product as the variable to be predicted, after our prior attempts to predict product quality were unsuccessful due to lack of variation. Of course, it is vital to discuss the revision of goals with the industry partners to ensure the new goal is one that will provide value to the company.

Even in the case when enough variation is found, there can still be a lack of “bad processes” due to very tight control of the process particularly in industries such as in pharmaceutical and biomedical industries, where batch processes are being extensively used (Croughan et al. (2015)). Each batch may involve a huge volume of raw material and take hours or even days to complete. It is, therefore, extremely costly to have a “bad” batch. Hence, these processes are very carefully controlled through both automatic control schemes and manual interventions by the operators. If in a process surveillance study the aim is not only to detect an out-of-control situation but also diagnose of the root cause, data capturing various fault scenarios is needed.

In terms of modeling, this translates into a highly imbalanced division between “bad process” versus “good process” classes. Several methods have been proposed to alleviate the imbalance problem at both data and algorithm level (Haixiang et al. (2017)). At the data level, data sampling methods such as Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. (2002)), under-sampling and over-sampling have been used with some success. For algorithms, ensemble methods, or algorithm modifications of traditional classifiers have been suggested (Haixiang et al. (2017)). Also, cost-sensitive learning methods have been proposed,

which usually assume higher costs for the misclassification of the “bad process” samples (Haixiang et al. (2017)).

5.3 Large Data Processing

With increased accumulation of production data, one of the biggest challenges has become the allocation of enough computational resources to process it. Although new technologies, such as parallel computing and quantum computing, have revolutionized the whole field, memory capabilities are still limited. Most of the well-known data analytics methods worked on the principal of in-memory processing. Computing frameworks such as Hadoop and Spark (Zaharia et al. (2010)) enable in-memory computation of large data streams and provide solutions to the problems prompted by the continuous streams of data (Agneeswaran (2014)). In terms of data storage, there is currently a transition towards NoSQL (“non SQL” or “non-relational”) databases (Leavitt (2010)) as opposed to the traditionally structured relational databases. One of the key advantages of NoSQL databases is that they can handle large unstructured data efficiently.

This issue certainly ventures into the domain of IT folks whilst data analysts (or even data scientists) will most likely not be well-versed in this issue. But, the voracious appetite of digitalization is not likely to be satiated anytime soon and data management (or data wrangling for analytics purposes) will only get more complicated in time. “Not my problem” attitude to data management could no longer be a response for the data analyst/scientist. We should strive to be part of the solution for the reasons we discussed earlier. This requires at the very least being able to communicate with the IT folks and establish a common language early on in the process. Otherwise, we will risk falling out of the loop of data management process altogether.

5.4 New Opportunities

In some industries, such as in electronic component manufacturing and aerospace industry, inspection of each product is a legal requirement. But, now due to the affordable sensor technologies, it is becoming more common in other industries as well to inspect and report the quality characteristics of each product as we mentioned in Section 5.1. This information along with process data reinvigorates the possibilities to investigate new opportunities to optimize the production processes. In process

monitoring, for example, the focus may shift to product characteristics as they contain process information anyway. This may simplify the monitoring schemes significantly. The process data and its connection to product characteristics can then be used for predictive purposes or when possible, to create a desired change in the process as in optimization and robustness studies. Furthermore, the reaction time to a fault can be significantly improved when we shift from sampling to 100% inspection.

As we mentioned in the previous section, it has been our experience that despite all efforts in automation, manual interventions by the operators are alarmingly high. This is generally accepted as good practice relying on the years of experience of these operators. In modern manufacturing, this process knowledge needs to be digitalized and fed back to the process when needed. This will prevent unintended variation and lead to more standardized operations. Currently, not all these interventions are recorded making them impossible to digitalize. Action needs to be taken for proper collection of these manual interventions so that empirical models can be trained and put to use with this enhanced process understanding.

Another emerging field for using Big Data is in predictive maintenance (Alaswad and Xiang (2017), Peng et al. (2010), Wang (2002)). There have been primarily two approaches to maintenance employed by the companies: Time-based preventive maintenance and corrective maintenance. The former is very intuitive and often effective as planning and scheduling of the maintenance activities can be done on a regular basis. Yet the downside of it is that this may cause unnecessary or delayed maintenance if the scheduling does not reflect the actual wear and tear in the equipment/system. Similarly for the corrective maintenance, the damage would be done by the time maintenance is performed. In predictive maintenance (sometimes also referred to as condition based maintenance), maintenance is performed when deemed necessary by a predictive model reflecting the condition of the production equipment. While this is expected to reduce the maintenance cost, it also brings more challenges due to uncertainties in planning and scheduling of maintenance activities caused by prediction. We are currently involved in multiple projects both on delivering predictive models but also on planning and scheduling of maintenance activities. Production companies see tremendous value in predictive maintenance, particularly in cases where high throughput is the main goal and stoppages are of major concern. Similarly many engineering companies that produce the production equipment are interested in selling not only the equipment but also a service contract

to provide the maintenance. The use of production data is the key in all of these efforts.

Advances in computer technology also had a very encouraging effect on obtaining simulation models of the production processes. These could be the key in one of the very important concerns in data analysis: correlation vs. causality. As we mentioned earlier, controlled experiments are the classical tool for unearthing causal relationships. We also indicated that there are many challenges when experimental work is to be performed on the process as a whole. But if a simulation model of the process exists, experiments can be run at a relatively lower cost without any disruption to the actual process. Once the crucial relationships are uncovered or the right settings of the process variables are established using the simulation model, confirmatory experiments can be run on the actual process. It should however be noted that simulation experiments differ in characteristics from their physical counterparts (Santner et al. (2003), Dehlendorff et al. (2011)). Nonetheless, the simulation models of the processes offer a great platform for process understanding and improvement.

6. Data Utilization Maturity – Digitalization Readiness

We have observed that many companies were not at the required level of maturity when it came to utilizing data within the manufacturing environment whereas the same companies were fairly good at utilizing data in other parts of the business, e.g., finance, marketing and R&D functions. There have been different illustrations and descriptions of data maturity, which have inspired the development of our interpretation of data maturity within a manufacturing environment as depicted in Figure 5.

Figure 5. Evolution in Data Maturity and Data Utilization. The flags illustrate an estimated time of incubation and implementation.

Our assessment is that the jump many companies are trying to make from unexplored terrain, e.g., from using Business Intelligence techniques in manufacturing, to advanced analytics as in AI is considerably bigger than any company initially anticipates. In this attempt, there will be lessons to be learned related to data connectivity, data accessibility, data management, traceability, etc. In the worst case, these lessons will be so costly that they will hinder future data related activities. Our recommendation is, therefore, to follow a sequential approach as in the Data Maturity Road illustrated in Figure 5.

A natural first step is to connect to the manufacturing equipment and initiate data collection. Some of this data can be transformed into valuable information by just visualizing the data, e.g., in the form of simple time series plots for the relevant users (operators, line managers, specialists). This will also help to confirm the pre-existing understanding of the process to a large extent. A natural next step is to establish the variation in the process through process monitoring. If this exceeds the

expected level of variation, changes to the process are in order. This will be done through the relationships between process variables and key performance indicators such as product characteristics. Predictive models based on correlations can be initially obtained for further experimental work to discover/confirm causal relationships through which desired changes to the process will be made.

It is important to acknowledge that going further on the data utilization journey will require establishing new competencies within the company. As one moves further on the data utilization road and starts relying on empirical models for insights, a new set of challenges will arise. Many companies immediately pursue AI for being able to optimize production. Often this seems more appealing than going for a sequential approach starting with understanding the process and hence understanding what the real needs are. However, we should always keep in mind that data and data analytics are simply means to an end and not the end itself. Therefore a sequential approach to digitalization and use of data will most likely be more fruitful.

7. Conclusion

With the proliferation of automated data collection schemes, the amount of production data that is being generated is going to increase at an accelerated pace. However, currently available data often lacks the necessary information to understand, improve and optimize production. While some useful information can be obtained from historical data, its analysis usually reveals that new data should be collected. In that sense, Big Data activities should be forward looking and performed in a sequential manner. Moreover, the aim should not necessarily be to collect as much data as possible from production but rather to collect the data systematically through appropriate sampling strategies. These efforts can then be targeted towards solving specific production problems. We should reiterate that this could be a treacherous journey with plenty of deterrents along the way. Industry should have the patience and commitment to go through the processes while acquiring and developing the right skills and expertise. Data analyst/scientists should be open to new challenges and learning opportunities, and expand on their communication skills as today's problems are truly multi-disciplinary and isolation will easily translate into irrelevance.

Acknowledgements

The authors would like to thank the industrial partners and MADE SPIR, MADE DIGITAL, BioPro2 and Innovation Fund Denmark for making this work possible.

References

- Agneeswaran, V. S. (2014). "Big data analytics beyond Hadoop: real-time applications with storm, spark, and more Hadoop alternatives." FT Press.
- Agresti, A. and B. A. Coull (1998). "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, 52, 119–126
- Alaswad, S. and Y. Xiang (2017). "A review on condition-based maintenance optimization models for stochastically deteriorating system," *Reliability Engineering and System Safety*, 157, 54-63
- Bisgaard, S. and M. Kulahci (2011). *Time Series Analysis and Forecasting by Example*, New York: Wiley
- Borne, K. (2014). "Outlier Detection Gets a Makeover - Surprise Discovery in Scientific Big Data," *Statistics Views*
- Capaci, F., E. Vanhatalo, M. Kulahci, and B. Bergquist (2019). "The Revised Tennessee Eastman Process Simulator as Testbed for SPM and DoE Methods," *Quality Engineering*, 31, 212-229
- Chawla N.V., K. W. Bowyer, L.O. Hall and W.P. Kegelmeyer (2002). "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 16, 321-357
- Croughan, M. S., K. B. Konstantinov and C. Cooney (2015). "The future of industrial bioprocessing: Batch or continuous?" *Biotechnology and bioengineering*, 112, 648-651.
- De Ketelaere, B., M. Hubert and E. Schmitt (2015). "Overview of PCA-Based Statistical Process-Monitoring Methods for Time Dependent, High Dimensional Data", *Journal of Quality Technology*, 47, 318-335
- Dehlendorff, C., M. Kulahci and K. K. Andersen (2011). "Designing Simulation Experiments with Controllable and Uncontrollable Factors for Applications in Health Care," *Journal of Royal Statistical Society, Series C*, 60, 31-49
- Frumosu, F. D. and M. Kulahci (2018). "Big Data Analytics using Semi-supervised Learning Methods," *Quality and Reliability Engineering International*, 34, 1413-1423

- Frumosu, F. D. and M. Kulahci (2019). “Outliers Detection Using an Iterative Strategy for Semi-Supervised Learning,” *Quality and Reliability Engineering International*, 35, 1408-1423
- Gajjar, S., M. Kulahci and A. Palazoglu (2016). “Use of Sparse Principal Component Analysis (SPCA) for Fault Detection,” *Proceedings of the 11th IFAC Symposium on Dynamics and Control Process Systems*
- Gajjar, S., M. Kulahci and A. Palazoglu (2017a). “Real Time Fault Detection and Diagnosis Using Sparse Principal Component Analysis,” *Journal of Process Control*, 67, 112-128
- Gajjar, S., M. Kulahci and A. Palazoglu (2017b). “Selection of Non-zero Loadings in Sparse Principal Component Analysis,” *Chemometrics and Intelligent Laboratory Systems*, 162, 160-171
- Gao, H, S. Gajjar, M. Kulahci, Q. Zhu and A. Palazoglu (2016). “Process Knowledge Discovery Using Sparse Principal Component Analysis,” *Industrial and Engineering Chemistry Research*, 55, 12046-12059
- Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue and G. Bing (2017). “Learning from class-imbalanced data: Review of methods and applications,” *Expert Systems With Applications*, 73, 220–239
- Hastie, T., R. Tibshirani and J. Friedman (2009). *The Elements of Statistical Learning*, 2nd Ed., New York: Springer
- Khan, A. R., H. Schiøler, M. Kulahci and T. S. Knudsen (2017). “Selection of Objective Function for Imbalanced Classification: An Industrial Case Study,” *Proceedings to 22nd Emerging Technologies and Factory Automation*, Limassol, Cyprus
- Khan, A. R., H. Schiøler, T. Knudsen and M. Kulahci (2017). “Big Data Analytics for Industrial Process Control,” *Proceedings to 22nd Emerging Technologies and Factory Automation*, Limassol, Cyprus
- Khan, A.R., H. Schiøler, M. Zaki and M. Kulahci (2018). “Rare-events classification: An approach based on Genetic Algorithm and Voronoi Tessellation,” *Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Melbourne, Australia
- Kulahci, M. and J. Tyssedal (2016). “Split Plot Designs for Multistage Experimentation,” *Journal of Applied Statistics*, 44(3), pp. 493-510
- Lasi, H., H-G. Kemper, P. Fetteke and T. Feld (2014) “Industry 4.0,” *Business and Information Systems Engineering*, 6, 239-242
- Lawrence, N. D. (2017) Data Readiness Levels, <https://arxiv.org/abs/1705.02245>

- Leavitt, N. (2010). "Will NoSQL Databases Live Up to Their Promise?" *IEEE Computer*, 43, 12-14
- Li D., S.L. Shah and T. Chen (2001). "Identification of fast-rate models from multirate data," *International Journal of Control*, 74, 680-689
- Liu, W., Chua, T. J., Larn, J., Wang, F. Y., & Yin, X. F. (2002). "APS, ERP and MES systems integration for semiconductor backend assembly," *7th International Conference on Control, Automation, Robotics and Vision*, 3, 1403-1408
- Meer, K.H. (2005). *Best Practices in ERP Software Applications*. iUniverse
- Meyer, H., F. Fuchs and K. Thiel (2009). *Manufacturing Execution Systems: Optimal Design, Planning, and Deployment*. New York: McGraw Hill
- Montgomery, D. C. (2013). *Statistical Quality Control - A Modern Introduction*, 7th edition. Wiley: New York
- Peng, Y., M. Dong and M. J. Zuo, (2010). "Current status of machine prognostics in condition-based maintenance: a review," *International Journal of Advanced Manufacturing Technology*, 50, 297-313.
- Reis, M.S. (2019). "Multiscale and Multi-Granularity Process Analytics: A Review," *Processes*, 7, 61-82
- Santner, T. J., B. J. Williams and W. I. Notz (2003). *The Design and Analysis of Computer Experiments*. Springer: New York
- Thornhill, N. F., M. S. Choudhury and S. L. Shah (2004). "The impact of compression on data-driven process analyses," *Journal of Process Control*, 14, 389-398.
- Tyssedal, J. and M. Kulahci (2015). "Experiments for Multi-Stage Processes," *Quality Technology and Quantitative Management*, 12, 13-28
- Tyssedal, J., M. Kulahci and S. Bisgaard (2011). "Split-Plot Designs with Mirror Image Pairs as Subplots," *Journal of Statistical Planning and Inference*, 141, 3686-3696
- Vanhatalo, E. and M. Kulahci (2015a). "The Effect of Autocorrelation on the Hotelling T^2 Control Chart," *Quality and Reliability Engineering International*, 31, 1779-1796
- Vanhatalo, E. and M. Kulahci (2015b). "Impact of Autocorrelation on Principal Components and their Use in Statistical Process Control," *Quality and Reliability Engineering International*, 32, 1483-1500
- Vanhatalo, E., M. Kulahci and B. Bergquist (2017). "On the Structure of Dynamic Principal Component Analysis Used in Statistical Process Monitoring," *Chemometrics and Intelligent Laboratory Systems*, 167, 1-11

Wang, H. (2002). “A survey of maintenance policies of deteriorating systems,” *European Journal of Operational Research*, 139, 469-489

Wang, H. (2009). “Comparison of p-control charts for low defective rate,” *Computational Statistics and Data Analysis*, 53, 4210–4220

Zaharia, M., M. Chowdhury, M.J. Franklin, S. Shenker and I. Stoica (2010), “Spark: cluster computing with working sets,” *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, Boston, USA