My facts are not your facts: data wrangling as a socially negotiated process, a case study in a multisite manufacturing company

Authors Names left out while in blind rev

ABSTRACT

The conditions under which the data wrangling process is undertaken has a profound impact on the quality of the results of the data wrangling and analysis. This paper presents the results of the analysis of the sociotechnical aspects of a data wrangling activity in a large, multi-site global manufacturer. This activity was technically demanding, as operational data from multiple sources and formats needed to be integrated, but also involved interaction with multiple stakeholders in different parts of the world with their own ways of collecting and structuring the data. The data had been captured previously for a different purpose. The clients were not aware that the data followed a different logic in the various sites and in some cases needed to be manually extracted and interpreted. The paper describes the data wrangling process and analyses the assumptions, goals and biases of the different stakeholders. The analysis raises questions and insights about how data can be trusted, and suggests that human intervention with data along the data wrangling process is often un-intentional, tacit and easily overlooked. It is suggested that contextual factors, such as data quality and assessment of consequences when acting/making decisions on the new data set is given higher attention during the specification of data wrangling assignments. The paper concludes with recommendations for data wrangling practitioners.

.

INTRODUCTION

Digitalization, and in particular the increasing reliance on data analysis, is one of the big trends in contemporary engineering practice [1]. Manufacturing companies have high expectations in that their access to data will give them previously unheard-of analytical abilities as the number of devices connected to internet continues to grow, having surpassed 35 billion in 2021 [2]. Rather than designing the data gathering process for each project, manufacturing companies often try to look for new insights in existing data sets. While statistics, data analysis and artificial intelligent (AI) are not new, they are rapidly gaining prominence in manufacturing companies. At present, companies see many opportunities, but don't fully understand the potential pitfalls yet. While effort is usually directed at obtaining the right data and finding suitable ways of analyzing it, the importance of preparing and analyzing data and the biases that can be introduced is less well understood, and manufacturing companies are less aware of the effort required to prepare data for analysis. Research on data wrangling is typically concerned with the technical aspects of handling and cleaning large data sets or establishing the quality and validity of data. The issues that are often overlooked are the sociotechnical processes surrounding it, i.e., in what way the organizational, social, and cultural aspects, impact the quality of the result. In particular, the quality of data used, imposes challenges for anyone who needs to use existing data in a different context [3], often referred to as re-purposing. Data analysis is a sociotechnical process, in which many decisions need to be taken by people based on the knowledge they have at the time [4]. Data scientist s are often called into manufacturing companies to compile and analyze distributed and heterogenous data sets and ensure that such compilations can be requested as standard

reports forming the basis for business decisions. This raises critical questions about the insights, responsibilities and impact data scientists have in relation to the 'customer', i.e., the decision makers who rarely have the competence necessary to judge the quality of the data wrangling effort made.

This paper reports on the data wrangling challenges existing in a global manufacturing company. A small case study of a data wrangling process is analyzed in detail to show the biases, goals and assumptions of the involved stakeholders and to illustrate the deeply sociotechnical nature of data wrangling. The paper analyzes a typical case where a data scientist has been called in to consolidate production data from multiple distributed sources to enable statistical analysis on improvements to the production process. The data had been collected for a different purpose that followed different recording logics. The cumulative effect of different biases and assumptions can have wide-reaching consequences that are only partially understood by the participants. For example, a manufacturing company might make a decision to shift production based on the production efficiency data, without ensuring that the different candidate sites submit data in a comparable structure and of a similar quality. A decision on the production site would affect the quality of the end products as well as the locations and communities that loose or gain the production. The paper argues that stakeholder goals, assumptions, and biases affect the data wrangling process and may impact the results. The ambition of this paper is (1) to identify situations where sociotechnical aspects impact the quality of the expected result and (2) to make suggestions based on the

analysis for how to bring attention to these factors, in data wrangling situations in industry.

Researchers developing decision support systems in engineering design, such as Hazelrigg [5] have successfully influenced engineering by providing consistent quantitative and data driven approaches to design, despite the difficulties in deriving quantified relations. In industry, statistical approaches to handle uncertainty and imprecision in data have also been successfully adopted, such as Design for Six Sigma (see for example [6]). The number of digital sources (software, sensors) generating and processing data increases the amount and variety of data to handle. This is largely seen as a technical problem and significant efforts have been made to better account for varying quality of data [7]. The recent ISO 8000 standard [8] offers shared interpretations of how to represent data quality.

Data is often reused beyond its original purpose and is deployed in other analyses, which leaves the people who were originally involved in capturing the data with little influence. Such repurposing [9] adds to the problems of 'ephemeral data use', i.e., when data is valid during a short period of time and the 'self-service data collection', where users interpret and add to the datasets when repurposed. As they lose visibility of the data they are associated with, the provision of data as well as the use of data become matters of interpretation and trust [3]. Data collected for different purposes often needs to be manipulated (restructured, partitioned, rescaled etc.) and combined with different data sets to be used in the new problems. This is usually done by people not involved with the original data collection process, who often misunderstand and

93 therefore unintentionally misuse the data [2]. Even conscientious documentation of
94 data collection might be insufficient, as later uses cannot be anticipated. This is, in
95 particular, an issue as many data projects hire in external expertise for these projects
96 and the knowledge is not maintained within the organization.
97 The problem addressed in this paper is that there is still a naivety about data and data
98 analysis, in that data is seen as facts and the results of data analysis as an expression of
99 "truth". As we have argued based on the same case study [10], data science (and, in
100 particular, data analysis) may appear as "magic" in at least two senses. Powerful
101 technology and science can "enchant" people [11], in this case, through the ability to
102 gain new knowledge [2]. At the same time the process of obtaining insights can seem
103 mysterious. Non-data scientists have overblown expectations. They use the results of
104 data analysis without consulting data science experts or understanding it themselves.
105 Following the introduction comes a brief introduction to data science as a sociotechnical
106 process of gathering and wrangling the data, forming hypotheses, analyzing the data,
107 and drawing conclusions from it. Our methodology section then describes the approach
108 taken to the analyzing the case study as well as the wider interviews in the same
109 organization. The case study itself is the presented descripting the problem that was
110 analyzed, the stakeholders and the steps that the process followed. A separate section
111 draws out some of the stakeholders goals, assumptions and biases. In the discussion
112 section, some of the implications for the company case studied and provides
113 suggestions for how to manage sociotechnical factors to the data wrangling community.
114 Conclusions are drawn at the end.

DATA SCIENCE IN ENGINEERING AS A SOCIAL PROCESS

Engineering has become a data intense activity, where the need to process data (e.g., translating, modelling, programming, analyzing, generating) is central. Engineering of industrial products, processes and systems has long been recognized as a sociotechnical process [12],[13]. Data analysis is beginning to be recognized as a sociotechnical process in its own right [14]. In data analysis, state-of-the-art techniques which are underpinned by social values encoded in the mathematical processes (e.g., [15]) needs to come together with a deep understanding of the problem context in which the data was gathered. This section will therefore look at the process of carrying out data science projects and set out some of the interfaces where different stakeholders need to interact and communicate their assumptions [2].

Depending on the aim, data science projects are typically either approached from a statistical point of view or a machine learning one and in consequence have slightly different processes, even though these can be conflated in the practice. At their core, both use statistical modelling techniques to estimate the mathematical relationship between the input variables and output variables. Classic statistics aims to facilitate a greater understanding of the underlying mechanisms by using statistical models to estimate the strength of the mathematical relationship between prior selected key variables. For useful insights to be gained from this abstract representation of the system of interest, two conditions must be fulfilled, firstly the model must be easily interpretable and secondly its underlying assumptions must be sufficiently fulfilled for

the theoretical guarantees assuring the validity of its results to hold. Big data / machine learning does not aim to provide insights, its focus is rather on finding patterns in the data in order to maximize the accuracy of predictions or classifications. Freed from the constraint of needing to be interpretable it can use a far wider range of model classes. Its algorithms usually automate the search for the best performing model, testing numerous models, often in complex combinations, frequently applying untransparent operations. What matters instead, is whether the model performs well on fresh data according to a predefined performance metric [17].

Data science as shared process

Data science is an emerging interdisciplinary field that uses technologies and methods to mine and analyze data in order to extract knowledge and value from them, thereby bringing together knowledge and techniques from statistics, data analysis, and machine learning. Data scientists may be involved at all stages of the data science life cycle. There is an increasing recognition that data analysis needs to be carried out by a team of experts with overlapping complementary skills. As the case study will illustrate, the data wrangling aspects are shared between data scientists and the domain experts. One of the key skills is interdisciplinary communication, [16] as data scientists need to understand the concerns and goals of different stakeholders and vice versa. They also need to negotiate access to data and negotiate the definition and scope of their tasks. Effective communication between engineers and other stakeholders is widely recognized as a successful factor in engineering (e.g., [32][33]). Communication is widely

seen as a constitutive part of any social system (e.g., [34]-[37]). The sociologist Luhmann [35] argues that systems are generated through communication, but can be challenging across disciplinary boundary and supply chain interfaces (e.g., [38]). Hales [39] argues, communication permeates all levels of design processes: the personal, the project, the corporate-, and the microeconomic- and macroeconomic levels. Engineers are used to working in their own object worlds [12], i.e., the experiences, domains knowledge and representation of their own field, and can find it difficult to collaborate with others who have different ways of thinking and use different vocabulary. This also applies to engineers and data scientists when collaborating. To work together they depend on boundary objects, which are understood by all the groups that are relevant [40]. The first step to enabling a successful collaboration is often the recognition that others do not understand what is required from them and providing them with clear and unambiguous explanations and instructions [41].

Ideally, the different stakeholders have a basic understanding of data analysis, to instruct and support the analysis process, form hypotheses and interpret the results. This requires a degree of data and information literacy, which is inextricably linked with statistical literacy [18]. Statistical literacy is a broad concept that encompasses a wide range of concrete abilities and contextual knowledge and is considered essential for any form of data-based decision making [19],[20]. As Gal [19] points out statistical literacy involves both basic statistical, mathematical and context knowledge but also the disposition to take a critical stance, which can interact in a dynamic manner. For many data science applications, big data literacy includes awareness of critical issues

associated with “ethics, epistemology, mathematical justification, and math washing” [21]. Data literacy affects all parts of the society, however professionals such as scientists require more in-depth knowledge [22]. However, to date, research on data literacy in engineering has focused on undergraduate teaching (e.g.,[23]).

The data analysis process

Historically data analysis was the domain of statisticians, who are provided with hypotheses and use analytical approaches to confirm or reject these hypotheses. Over the last 40 years computer science approaches have evolved to look for patterns in existing data, which are sometimes but not always guided by hypotheses (for example see [25] for a discussion of theory-guided data science). This data analysis is typically motivated by practical concerns and aims for classification or prediction, which can be useful without revealing the underlying rationale (see [26] for basic assumptions of different kinds of data science). Theoretically statistics and data analysis are different paradigms, but in practice it can be beneficial to combine the algorithms and quality measures of both [24].

Figure 1 gives an overview of the steps that a typical data analysis process goes through. The workflow is divided between the engineering application domain and the data science domain, such as statistics. The initial problem recognition comes from the engineering application domain, where the data needs to be generated or identified. Typically, the problem is abstracted and expressed as a small number of critical factors, such as process efficiency or quality. The data scientist frequently has to prepare the

data set before analysis, pulling it together from multiple sources and harmonizing vocabulary. After statistical analysis the results need to be interpreted and the resulting actions are based on this. During interpretation, follow up questions may arise, requiring the data analysis step and possibly also the data wrangling step to be revisited. The data wrangling process starts with the validation of the data, i.e., the removal of the unusable or implausible, and has three phases: extraction from multiple sources, integration of different sources and aggregation across the different data sources as the three steps of what the data warehousing community calls the ETL process (extract, transform, load) (see [28]).

However, this only describes the process partially. An often-overlooked aspect is that in the data wrangling phase, i.e., the “process of iterative data exploration and transformation that enables analysis” [27], significant effort needs to be put into representing the data in a suitable data schema, which structures the data as well as the generation of hypotheses which can be analyzed statistically.

This often involved combining data from different sources into common schemata. The schema contains the variables along which the data is structured and their relationships. Rahm and Do [28] classified problems with data quality into single source problems and multiple source problems. In both cases the problems occur at the level of the schema, i.e., the logic in which the data is presented and the instance level, which refer to errors and inconsistences in the actual data. They attribute schema problems to a "lack of appropriate model-specific or application-specific integrity constraints, e.g., due to data model limitations or poor schema design, or because only a few integrity constraints

were defined to limit the overhead for integrity control". Instance problems include misspellings, abbreviations, multiple entries, violated dependencies, and contradicting records. For example, in the case study one plant switched from recording initials to recording (occasionally misspelled) full names, which needed to be matched to analyze projects per person. Occasionally, a schema is modified to overcome limitations in the data set, e.g., aggregating the data to compensate for high levels of missing values. Negotiations over the structure of the data can have a huge impact on the overall project and its results.

In engineering, there are two motivations for using data [29] (1) to act on what is measured or (2) to gain knowledge to address other activities (decision making based on data impacting other decisions, such as prognosis, reuse etc.). This paper addresses the second point, as one needs to know more of the conditions from where the data comes from.

This paper argues that for effective data analysis, the domain experts owning the data, such as production engineers, and data science experts, need to establish clear shared activities to gain sufficient shared understanding. The data wrangling process is highly iterative in its character, since the task to prepare data for analysis can rarely be prescribed upfront. The application domain experts, e.g., production engineers in this case, expect data analysis to be useful for their decision-making process, and the data science experts need to overcome inconsistencies and problems with the data sets that may impact decisions. There is also often a need to iterate expectations by the manufacturing engineers with what is possible to provide from the data set available.

This paper in particular highlights the importance of different stakeholders sharing their assumptions and expectations. In particular, it shows that who communicates this and in which order can affect the outcome. The critical point is to define variables and have an understanding about how these might be related. While it is theoretically possible for a machine learning program to find these correlations, this can be slower, be computationally more expensive, require more data, and work less well [30].

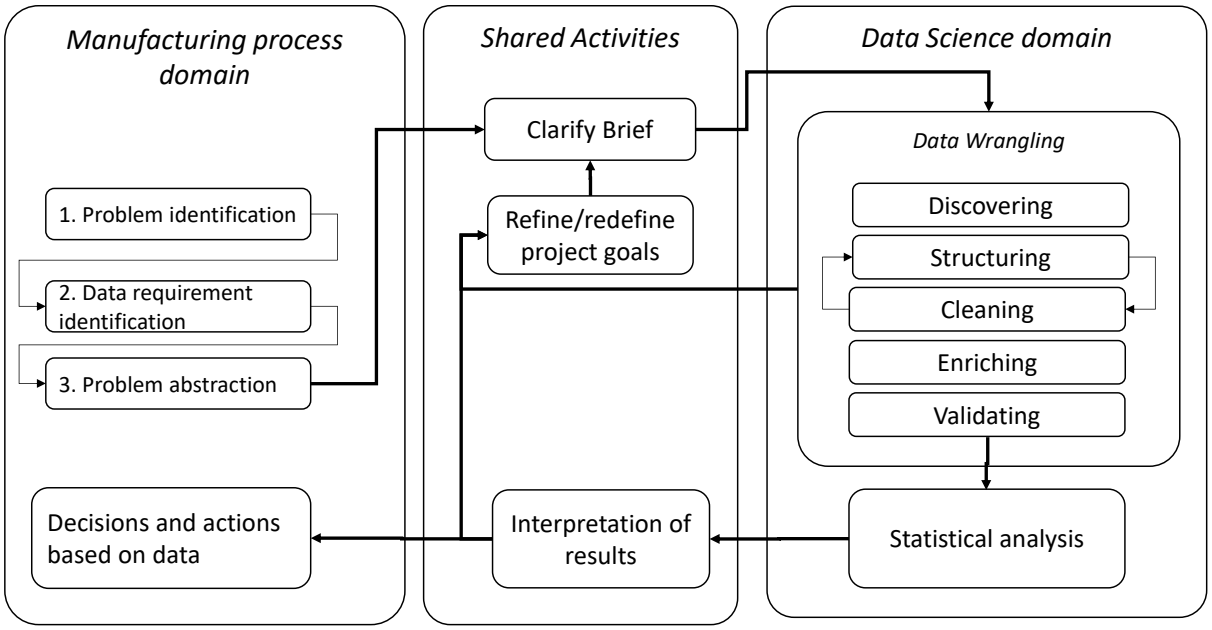

Figure 1 Data Analysis process connecting the manufacturing process and data science domains with key, shared activities.

Data analysis projects don't happen in isolation but are embedded in the context of the larger organization. If the organization has sorted out the broader data wrangling issues (see [31]) then individual projects can make use of well curated data.

METHODOLOGY

Data science is a rapidly evolving field and few empirical studies of the data science project in industry have taken place or focused on the benefits and challenges of deploying data analysis in particular tasks (e.g., [42] carried out case studies on the role of big data in agile manufacturing processes). However, practical issues of how to be a data scientist are widely discussed on websites and blogs aimed at novice or experienced data scientists. These have informed the previous sections. The research focus in the paper is on the sociotechnical aspects of data wrangling, rather than on the techniques and tools used by the computer scientist in wrangling data. The method has been chosen accordingly.

This paper is written in the spirit of participatory observation, which has been common in social sciences and anthropology since the 1920s [43],[44]. It enables research subjects to bring their own 'voices' to the research process [45] and yields insights into the everyday interactions of people**Error! Reference source not found.** The third and forth author are were active members of the case study, whereas the first and second authors advised them and listened to reports on the progress of the project. The issue of data wrangling as a prominent topic arose from analysis and reflection about the case study. As the third author is still in the organization, she could respond to issues arising directly from this paper.

The case study analyses the experiences of a young statistician (the fourth author of this paper), who has a background in statistics and philosophy. She undertook a one-month placement in the production organization of Volvo Powertrain (represented by the third author of this paper) at the company headquarters in Gothenburg in September 2018. This was an explorative pilot project and the team's first time working with a data scientist. While the team was extremely friendly and helpful to her on a personal level, they had no experience with what they needed to communicate and what they could expect from a data science project (see [10]).

The overall process following the same logic laid out in Figure 1. The details of the specific case study project are explained in Figure 3. After the end of the project the authors traced the impact of the project in the organization and had multiple discussion to analysis the dynamics of this project and the effect of the sociotechnical interactions. The project provided the client team with useful insights as well as data to support their qualitative understanding of the situation. However, the project did not create a quantifiable impact on the organization.

Since then several other larger scale data projects have taken place in the organization, which the third author of the paper has also been part of. The third author was selected to carry out the interviews based on the positive reception of the earlier paper in the organization.

THE CASE STUDY

The case study describes a student placement that took place in Volvo Group Trucks Operations, a large truck manufacturer, with many manufacturing sites globally. The pilot project was part a corporate learning effort into working with statistics and data analysis in an area where the findings had no implication on the quality of their products, but that was of genuine interested to them. It occurred at a time when the organization was hugely enthusiastic about data analysis but was in the process of discovering some of the challenges of going from opportunistic data collection to well-structured data (see also [31]).

The case selected represents a typical scenario where existing data needs to be collected from a dispersed data source and interpreted though analysis to deliver answers to specific questions.

The aim of this pilot project was to analyze the effectiveness of quality improvement measures through lean Kaizen [47] processes at the different plants. The centrally located improvement team wanted to understand how well the data for wastes and losses was captured and how well the Kaizen projects improved efficiency at the different plants. Both the data about the improvement measures and the data concerning the losses existed in five individual plants but was captured in different ways. Figure 2 illustrates the wide range of the routines, processes, and IT environments, that are locally influenced at the different manufacturing sites. Their work practices and routines with these data sources also differ. Therefore, information automatically obtained entirely straightforwardly at one site had to be gathered

manually at other sites. Any analysis across the sites required gathering and wrangling data into a consistent and relevant dataset.

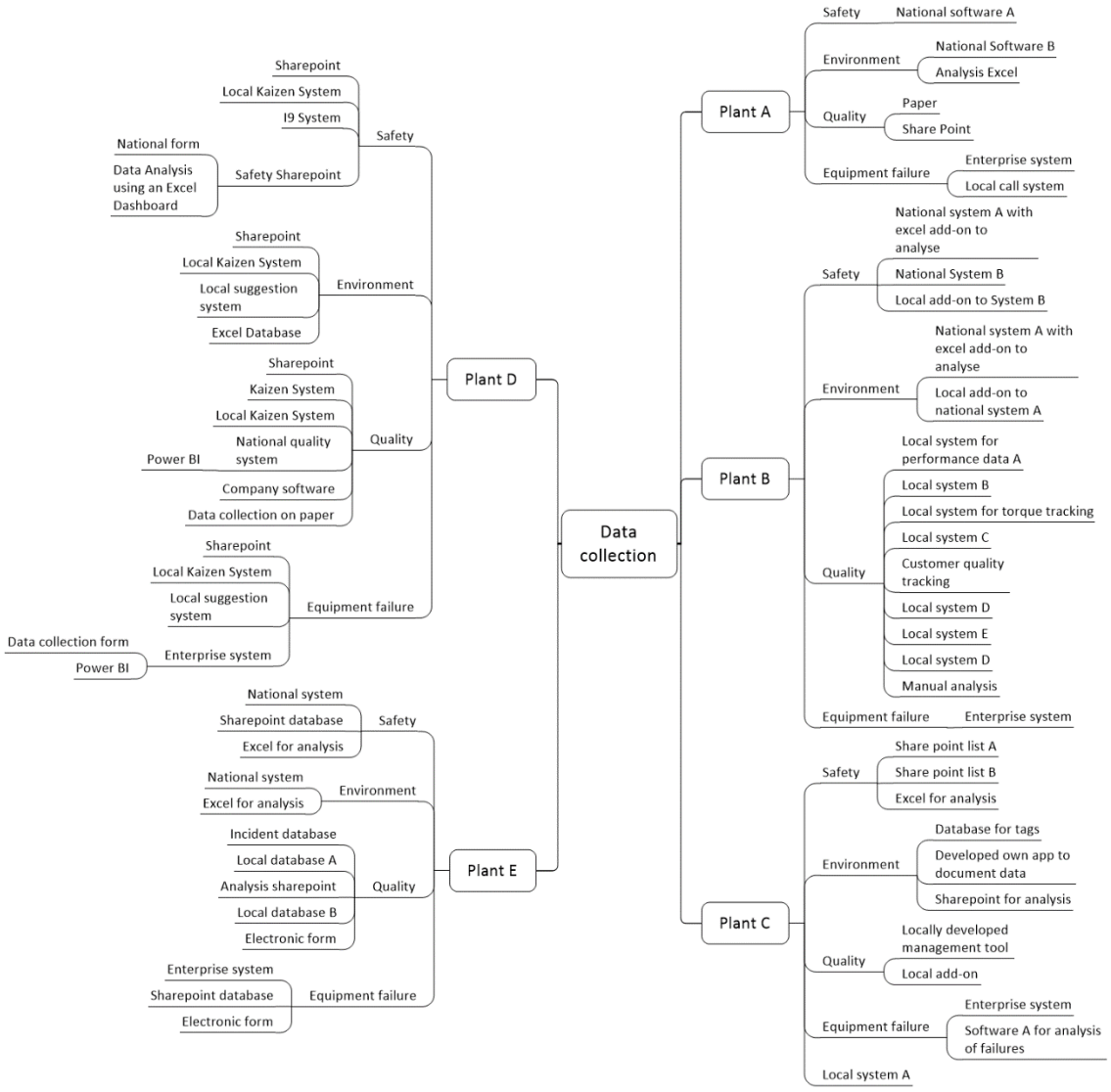

Figure 2 Overview of data sourced from within five different manufacturing sites in the case company

The challenge of this pilot project was to gather the data and bring it into a format that allowed comparisons of their Kaizen process across the organization. There was not enough data to carry out extensive statistical analyses. However, it was possible to develop an Excel tool which allowed the company to view the available data using

appropriate filters. Building the spreadsheet was technically straightforward; the challenge lay in understanding the problem and structuring the data in a way that allowed an, as unbiased as possible, analysis across the different plants. The case analyzed the process of data preparation and analyzed in detail to illustrate how different steps and activities are influenced by the assumptions, goals and biases of different stakeholders and their interpretation of their role and tasks.

The context of the case study

Volvo trucks has been building trucks in Gothenburg since 1928. The headquarters of the organization and most of the engineering activities are still carried out in Gothenburg, where the process improvement team is based. The main plants for the organization are established globally. All the plants collaborate closely with the production improvement team in Sweden. In addition to various backgrounds, the plants have different national cultures. This is particularly important in the context of reporting problems and improvements to them.

The company follows a lean production paradigm and engages in Kaizen initiatives [47]. The term Kaizen is used for improvement activities at different scales and levels of effort from simple and sporadic problems to longstanding and complex problems. A Quick Kaizen could be done by anyone - bottom-up as the company calls it - and without prioritization. It should take about one hour to implement the improvement action. This means that thousands of Quick Kaizens can take place every year in a plant. Standard, Major and Advanced Kaizens are carefully selected and take between one week and 3

months to complete. They require input from experts and are usually launched by management in response to major losses in operations. The more advanced the Kaizen, the more savings they are expected to generate. The effectiveness of Kaizens is measured by the ratio of the number of successful Kaizen implementation projects to the number and cost of the identified losses they have resolved.

The brief for the data scientist

The company was interested in finding out the return on investment for their Kaizen projects in different production plants. Their ultimate aim was to make the improvement processes themselves more effective and efficient.

The improvement team wanted to see how good each plant was at different aspects of process improvement. They also wanted to be able to carry out direct comparisons between plants on a number of parameters, such as the number projects started, proportion completed, savings generating, etc. They formulated 10 questions:

1. How much have we saved in total?
2. Actual savings vs. planned savings – how efficient is the problem-solving machine in total and per area?
3. Number of projects per loop per site over time evolution
4. Evolution of lead time per project
5. How has the loss map developed during the years?
6. How many projects / savings have been generated per focus area?

7. Compare the two above and see if they match
8. Evolution of direct, potential and cost avoidance over time
9. Number of project leaders that are involved
10. Analysis of delay in project starts

This needed to be represented in an accessible tool therefore they opted for Excel rather than R, which the data scientist would have preferred, because R was not a part of the standard toolset in their IT system.

The stakeholders

Like in other data science projects, the project had multiple stakeholders each with their own goals, perspectives and assumptions.

The client was the **improvement team,** which consists of one Vice President and three senior lean coaches, including the third author. It takes a strategic view at how to improve manufacturing processes. It coaches and evaluates the plants on their improvement effectiveness. They wanted to understand the effectiveness of the Kaizen projects in the different plants to target their support. With the case study projects, the team wanted to make more efficient use of the data while demonstrating that they are on the forefront of applying data analysis in their own processes.

They report to the **central management team,** which consists of support function heads and plant managers. It strives for world class performance in terms of safety, quality,

delivery, cost, sustainability and people. Its role is to support and challenge the plants by providing new technology and new ways of working to continuously improve.

The **management in the individual plants** needs to minimize the losses to run their plants effectively. **Financial teams** in the plant hold the main responsibility for measuring improvements in their plants. They have been collecting monthly data for more than ten years.

The **data analysist** was motivated by gaining experience in a large organization and working with a real data set while meeting the requirements of the improvement team as a client in the hope of a positive reference at the end of her project.

The data

The project was set up to make use of existing data, therefore the data scientist had no influence on the data generation or sampling process. The data from the Quick and some Standard Kaizens had been collected by each operator in a paper format and then transferred to data sheets, either by the team leader or the improvement leader, or reported into systems directly. The data from the major and advanced Kaizen initiatives is collected by engineers in spreadsheets. The design of the spreadsheets varies between locations, which contributed to the data wrangling challenge.

As Figure 2 illustrates, a wide variety of systems that are used to collect and/or store the data and each plant and even within each plant the departments have each their own

systems. In total 67 different systems were used to collect the data for the study discussed in this paper.

The plants were asked to quantify the financial benefit of the improvements that they have implemented. Many improvements are bundled, and the effects are multicausal, so that considerable judgement is involved in attributing financial gain to individual Kaizen projects. The improvement managers in the different plants appear to have taken a different line on what to include. However, they had neither requested clarification nor explained their rational for what they have included.

The data scientist was given two sets of data from Plant B while further data was requested from four other plants. The data arrived in individual spreadsheets. Some plants already had spreadsheets with data on all the improvement projects; others had to collate data from different spreadsheets recording projects in a 4- or 6-month time period. Some time periods were missing, which had to be requested individually. One plant never sent data and had to be excluded from the analysis. Some plants changed the way they recorded projects over time, so even on a plant level it was labor intensive to produce a uniform dataset.

While they were supposed to report largely the same variables, their interpretation of the data diverged in different plants. Some variables, despite having the same name, clearly referred to different things, indicating that the plants had different understandings of key concepts. With considerable effort it was possible to transform

the heterogeneous data sets into a uniform dataset that could enable a meaningful comparison. The rationale for the different interpretations of the definitions used in the source data was lost.

The uniform dataset was used to produce tables and plots that could give a visual overview over the data in each plant and compare the performance between plants. Each analysis could be filtered according to a range of factors to enable a more nuanced look at how they affect each plant in different ways. For example, some plants made greater use of major and advanced Kaizen while others focused mainly on quick Kaizen.

The overall process

Figure 1 showed a generic data analysis process. Figure 3 maps the process that the case study went through. It shows the timeline of activities in the form of a swim lane diagram, which shows the activities of each stakeholder group. The factories carried out the Kaizen projects before the case study began. The improvement team was aware that the data existed, and wanted to analyze it. They formulated a set to questions that they wanted to investigate through the data. When the data scientist arrived, she negotiated and clarified the questions. They requested the data and received the data from Plants B and E. The data analysist started with those data sets and got a brief explanation of the Kaizen processes and the nature of the data. It rapidly become clear that rather than being able to carry out a statistical analysis in a statistics package, the company was

interested in a visualization of the data in excel, which was widely available in the organization.

Originally the data scientist and the improvement team had assumed that a common data schema existed. Soon some of the major challenges of data wrangling become apparent. As the data from different plants arrived it become clear that the plants had a different understanding of the variables and the relationship between them. For example, one plant counted all quick Kaizens, even the ones that were not implemented, while others only counted the quick Kaizens that have been implemented.

Data wrangling took up the vast majority of the project time. This didn't surprise the data scientist as she had been warned that many experienced data scientists found themselves spending up to 80% of their time collecting, cleaning, organizing, preparing datasets for analysis. The process followed the usual data wrangling steps [60], although the process was highly iterative:

- Discovering: eyeballing the data and getting an overview over what data was there and how the questions in the brief could be addressed with it.
- Cleaning: There were many issues such as misspellings, abbreviations, multiple entries, violated dependencies and contradicting records. Some plants appeared to have captured the data on paper and introduced typos when entering the data into excel spreadsheets. For one data set it was discovered after translating the column names and consulting an engineer, that the plant had been using the wrong names for certain concepts. The individual data sheets also included

494 internal inconsistencies, e.g., different ways of referring to the same person,
495 which required a certain amount of guess work to straighten out.

496 • Structuring: This proved challenging as the plants chose to record different
497 information – some, for example, only listed the "loss area". The first step was to
498 understand the underlying logic by which each plant recorded their data. This
499 revealed discrepancies in the understanding of key concepts that the
500 improvement team had assumed were well understood – a finding they found
501 useful in and of itself. The example plants B and E understood "direct savings" as
502 the proportion of "potential savings" that were actually realized, Plants A and D
503 as an addition to "direct savings". The data sets were harmonized by subtracting
504 "direct savings" from the overall "potential savings" for the plants with a
505 proportional logic. The direction of the change could have been the other way
506 round enforcing instead a 'proportional' standard. It was largely arbitrary – albeit
507 made in consultation with the engineers – based on the logic of the data set that
508 arrived first.

509 • Enriching /augmenting: The original data provided had been incomplete.
510 Requests for missing data had been sent to the participants, but fields were left
511 blank when the data was not forthcoming. Most noticeably one plant never sent
512 the data.

- Validating: this was not possible in the given time frame. Checking the validity of the data with the companies that provided the data would have been a costly process and was suggested for continued efforts.

- Publishing: Having to work in Excel rather than a statistical language also made the documentation of the wrangling process difficult. The data from each of the sources had their own problems and idiosyncrasies that needed to be addressed before the data could be combined.

Data cleaning and structuring was an iterative process as she found during the data combining process that she had left important columns out and had to repeat parts of the combing process to include them. This came from the engineers not understanding the context of her questions and giving her explanations of concepts that while not inaccurate, failed to communicate their relevance to her and the research question she was answering. The most notable example was “Kaizen type” which she only, near the end of the project, realized classified the projects based on the time and resources needed to implement them – as it was very important to the engineers to be able to filter the analysis results based on this measure, she laboriously repeated large parts of the data combination process at the last minute to include it. Had the data from all plants arrived simultaneously, she may well have made different choices when designing the shared data.

As the spreadsheet was developed, the improvement team also saw the potential for other questions to be answered. While she waited for the last data set, she worked on the visualization of the data according to her understanding of the goals of the

improvement team. She presented the spreadsheet to the improvement teams as clients, which was favorably received, but also raised new questions. Before she left, she made some further changes to incorporate this feedback. The last data set from Plant C never arrived despite numerous requests by experts at different points of the company hierarchy. After the placement the findings were shared with decision makers in the organization. The main point discussed in the organization afterwards was the realization of the lack of understanding of the potential of data and the problems the company has with the available data today. The challenge is not only a technical matter or the quality of the data in itself, but also leadership, organizational and competence challenges.

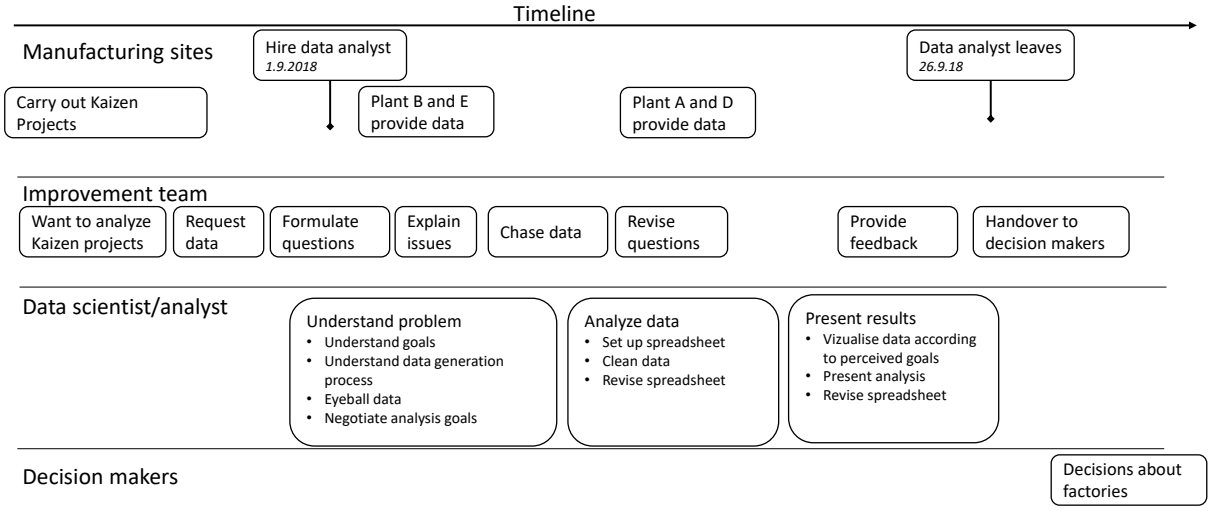

Figure 3 The specific data analysis process in the company

In the case study the data wrangling process was heavily biased by the order in which the data arrived. The data scientist used the data schema from plants B and E and fitted the other data sets to that with minor modifications. She may well have used another

plant's data structure or generated a hybrid structure, if the data had been available from the beginning. The members of the improvement team have been recruited from different plants and therefore answered questions about the data and its structure from the perspective of the plants they had come from.

The outcome of the project

The tool that the data analyst developed could visualize answers to questions 1 to 9. The improvement team gained useful insights and a comparable overview of the different plant's performance. For example, they could see which plant had the most accurate predictions of savings and the highest rate of closed projects, while another had 2.5 times more projects but the lowest rate of closed projects. On average actual net savings exceed their expectations. There is a strong positive correlation between the number of projects and adjusted total net savings by focus area, there is a strong positive correlation between leaders per project and completion rate across the plants.

The availability of financial data from each of the plants was generally very good from 2015 onwards with relatively few missing values among the projects completed. All plants already collect most of the important variables. The problem lied more in how the plants collect and store the data.

The improvement team understood that each plant is collecting and reporting differently. They also started to see the complexity involved when working with data, such as, that quality and standardization of data cannot be assumed. For example, there

were separate spreadsheets for each loop, the data structure has inconsistencies both between and within plants and the different plants data frames followed a different internal logic requiring complex transformations in the standardization process. Also, a simple factor like language; the datasets came in three languages and some variables e.g., Project leader, required laborious cleaning.

The potential consequences of the analysis

While this was a small project, without significant consequences in its own right, it contributed to the understanding of the Kaizen processes of different plants, that the central management and the improvement team built up. For the plants, this study did not have a significant strategic impact in itself. The improvement team could however identify that one plant significantly overestimated their improvement effectiveness and the team could also see that one plant had very low effectiveness which was not clear before. This was suspected before but could now be evidenced with data.

The project was a success for the improvement team as it enabled them to gain insights into the plants. The improvement team was also able to demonstrate that they were willing to test new ways of working. Through this project and other similar activities, the team is now considered to be in the forefront in digital literacy in the organization. The team was selected to be part of the wide-ranging data literacy study in the organization.

The data analyst achieved her aim of gaining insights into how a big organization worked and obtained the positive reference she was hoping for.

For the plants there was risk associated with providing the data as well as with not providing the data. Noncompliance with head quarter requests could be a problem as well as being seen as not being good or committed to Kaizen projects. Note in this project the plants had to actively supply the data, i.e., they could omit to do so as one plant did, in other circumstances the analysis might be carried out automatically on existing data. With the same consequences. This raises questions whether stakeholders need to be informed about ongoing analysis and potentially actively sign up to them.

In the case study, the effect on the consumer is very indirect because the analysis was aimed at measuring improvements to processes that already produce reliable products. However, improvements are a means to reduce cost, which could be passed on to the customer.

THE GOALS AND ASSUMPTIONS OF THE DIFFERENT STAKEHOLDERS

Throughout the progress of data wrangling the way different stakeholders approach the data needs to be considered. This requires an understanding of their goals, assumptions and biases.

The goals and assumptions of the stakeholders

Even a simple data science project like the case study, has multiple stakeholders. Table 1 gives an overview of the different stakeholders' understanding and goals associated

with the study. It shows that only the central improvement team and to a certain extent the plant finance teams understood the purpose of the study. For the plant finance teams this was one of many data requests. The study was not a priority for them and one plant that withheld the data despite multiple prods. The plant managers and the central management team who are impacted by the results of the study were not really aware that the study was going on. The data scientist wanted to gain experience and do a good job. She also wanted to support the organization, but only had a vague understanding what this entailed.

630 *Table 1: Stakeholder map drivers and goals*

Stakeholders	Overall drivers	Organisational dynamic	Purpose for the study	Goals for the study	Assumptions for the study	Consequences for the study
Central management	World class performance	All other stakeholders report to this function	Not aware	Not aware	Not aware	Not aware
Central improvement team	World class improvement system for achieving world class performance	To challenge and evaluate the plants on their improvement performance	- Improving the improvement system - Understand the challenges of working with available data	Understanding the effectiveness of the improvement system with the data already available	All plants are collecting and reporting data the same wat with good quality – why not collect all of it as it is already collected monthly	- Understood that each plant is collecting and reporting differently - Understood the complexity more when working with data - Could evaluate the plants' respective effectiveness
Plant management	World class operational performance	Is depending on the central improvement team to get the evaluation of the effectiveness of their improvement work	Not aware	Not aware	Not aware	Not aware
Plant finance	World class financial performance	Is responsible to lead and report the effectiveness of the improvement work	A part of a pilot to test a way to compare the effectiveness of the plants' improvement system	Providing already existing data which is normally never sent outside the plant	- The data is owned locally - One plant did not see this as a priority	Learned about the difference between the plants but do not see the need to harmonise
The data analyst	Gain experience, get good reference from happy client	Only relation to the improvement team in particular the third author	To gain experience	to get reference	- Data is internally consistent and she could do statistical analysis on it - She would get precise questions, that she could answer with the data	learned about massiness of data projects, obtained permanent job she wanted, learned what the asked and how to ask, need to understand the client holistically to understand what they could want.

631

Figure 4 illustrates the diversity of the goals of the different stakeholders and the organizational structure. This data analysis project was a small element of the assessment of the different plants and the allocation of work to them. This potentially made other employees in the plants also stakeholders in this analysis. Similarly, the quality of the results also reflected on other data analysis projects.

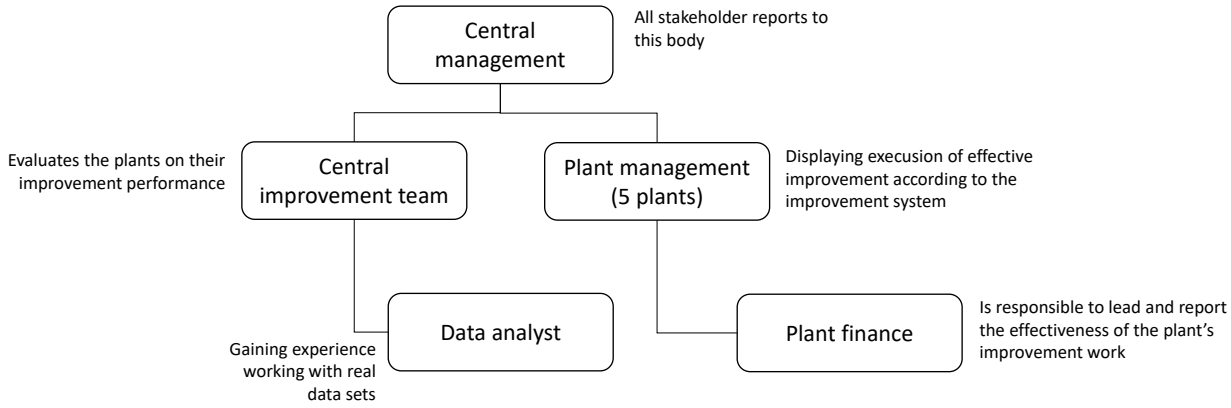

Figure 4: Alternative illustration of the stakeholders and their organizational dynamics

The improvement team set and managed the overall project goals. After the negotiation over the scope, the first nine of their ten questions listed in the brief (see section 0) have been answered. However, the analysis could, with little additional effort, be made more complex to enable a far more nuanced insight. They also benefitted from the project in terms of the kudos their proactive steps to data analysis gave them. Of all the goals that the different stakeholders had only the overall projects goals were explicitly stated in the request for data messages sent to the plants.

The process improvement team and the plant finance also had very different assumptions about the data project. The improvement teams assumed that the data

649 would be stored in the same way and that it would be entirely straight forwards and fast
650 for the plants to provide the data. They assumed that the challenge of project would
651 either be in the data analysis or the visualization. While the visualization in Excel was
652 fairly straight forward, the data wrangling was not. As they understood the underlying
653 processes and logic of the Kaizen projects, they assumed that this would be fairly easy
654 to understand and had little sense of what was obvious to the data scientist. For
655 example, they had delegated the initial explanation of the overall process to a departing
656 placement student, who explained the overall logic, but did not convey the Kaizen
657 classification as a fundamental driver of the way in which data was gathered and
658 structured.
659 Meanwhile the plant finance department saw the data under their devolved
660 responsibility. For them it was local data, which they monitored in the way they wanted.
661 In consequence providing the data was considerable effort. They assumed that head
662 office would appreciate that this project was a considerable amount of work for them,
663 that could not be provided at short notice as it could not be generated automatically.
664 They were also not aware that the data they provided would be fitted into the data
665 structure of Plants B and E.
666 The case study illustrates the importance for the data scientist of understanding the
667 different assumptions and goals of the stakeholders for the data wrangling activities.
668 Only then the data scientist can ask the right questions and frame them in a way that
669 the makes sense to the domain experts. For example, the data scientist at some point
670 realized that the different plants had different production volumes as well as country

dependent running costs and that therefore the monitory values of their saving would not be comparable. When she mentioned this to a Plant B expert, he dug out the companies preexisting conversion factor for international plant comparisons and sent it to her. This is one example where there is a need of a procedure to ensure a shared understanding.

Bias in the data analysis

Different goals and assumptions also bring a bias to how the data analysis process works. The composition of the group itself can already lead to bias. Frequently groups of developers and data scientists lack gender and ethnic diversity. For example, Cowgill et al. [49] found in a study on AI engineers that prediction errors were correlated with factors concerning the demographic of groups, in particular gender (groups are mostly male) and ethnicity (in the US mostly white and East Asian). In the case study, it was clear that the attitudes towards sharing data with management and wider in the global organizations differed. This may be due to the fact that adhering to group norms is a means of enabling group cooperation or simply that group members have an affective preference for each other. [50]. The data from Plants B and D was available much earlier than the other plants so it is what informed the data scientist's understanding of the data generating process and the "correct" way different variables relate to each other. It provided the template which all other data sources were modified to align with. Location can be a great source of bias. The data scientist visited Plant B and had a longer explanation of their logic.

Once composed, groups are then subject to all the biases introduced by the dynamics of any group decision making process. This could, in particular, be an issue, when the boundaries of groups are unclear. For example, in the case study plant B and E were more integrated with the improvement team, and might feel more part of the organization than the plants that have been acquired. In particular, Plant C, who did not send any data, viewed itself as a separate entity. The plants might also have been biased by their experience with previous requests for data. Plant A and C joint the Volvo group through mergers and had to change their processes to align with the parent company's processes. There was an element of competition between the plants; the long-term future of the plants and the jobs in them depends amongst other factors on the ability to improve their processes.

The way data is requested might set a bias. Like in a psychological experiment the phrasing could be very critical as it sets the expectations and sends out subtle messages (see [51]). In the case study the analysis was instigated by the improvement team, which consists of experts who are recruited from the different plants and were influenced by their plant's way of seeing things, which resulted in a sort of friendly rivalry. The data request was sent out by a member of the improvement team, who had come from Plant B and was still based there. Coming from him the request might well have been interpreted as an attempt to compare or benchmark other plants to his plant. The members of the improvement team also champion their own home plants to a certain extent and tried to persuade the data scientist that their home plants worked best and

had the best processes. The data scientist reflects that as she was strongly influenced by the enthusiasm of several members of the improvement team in favor of their respective plants. She became aware of their bias and later attempted to remain neutral. If she had been briefed from a single perspective, she might have not noticed this bias.

The analysis of the data was also subject to multiple influences. Data analysis is always, to a certain extent, a subjective process, as the data scientist is influenced by the tools and models that they are familiar with [48]. She was strongly motivated by doing a good job for the people who hired her. There is generally a confirmation bias of the hypothesis of the clients in statistical analyses. Typically, an analyst has a vested interest in finding results of interest to, or desirable for the client, therefore they are likely to put more effort into finding supporting evidence and/or interpret any evidence as more indicative of the desired result than they would most likely have done otherwise [53]. It is difficult, if not impossible, to avoid confirmation bias entirely, but awareness of its potential effects is the first step towards mitigating its distorting effects [52]. In the case study the data scientist wasn't given any hypothesis as such, but she was told which plants the improvement team considered to be most effective and so she was less likely to scrutinize their data for potential processing errors as closely, where it supported these a priori expectations.

Another source of bias is sampling. To a certain extent this is inevitable in any situation where it is not possible to include the entire population in an analysis. However, this can easily be accounted for in statistical analysis and valid inferences can still be drawn provided the following assumptions hold: (1) the sample is representative for the population as a whole and

(2) each instance has an equal probability of being included in the sample [54]. In the case study the original intension had been to cover all projects across in all power train plants from 2014 until 2018. However, the Plant A data couldn't be provided for certain months and the assumptions was made that this was coincidental. The total absence of any data from the Plant C makes inferences about the performance of power train improvement projects as a whole, impossible, which in turn affected assessments about typical behavior in other plants. Had Plant C been one of hundreds of other plants rather than one of five, the omission of one data set would have had much less effect.

Another source of the bias comes from the person who champions the data analysis [55] in the organization, since it is often their reputation that assures that others cooperate with the data analysis. This is less an issue when expensive teams of external data scientists are hired, but in the case study getting the data and getting information very much depended on the standing of the person who requested the data. As the case study project was carried out by an intern, more senior people were also less motivated to support the data gathering. For example, obtaining the Plant C data might have required intervention from a more senior level. If pulling rank is not an option, then communicating with all stakeholders becomes even more important.

The role of communication

It is widely recognized that communication is a vital success factor for data analysis [56]. The case study project was also a learning journey for the process improvement team as a client. They provided some briefing at the beginning, but many issues only come out through the process. In particular, when they saw the first demo, they were able to articulate more clearly what they wanted to know. As this was a short project, it was not documented in detail, so that a future data scientist would probably have to start from

765 scratch but would still benefit from the learning that had occurred in the improvement
766 team. This way to analyze the performance has not been continued but several
767 workshops have been held on the topic.
768
769 Data scientists are advised to understand the business problem and the priorities of the
770 stakeholders to be able to tailor explanations to the audience [56]. This can often be
771 stories rather than technical explanations for less technical audiences. Unless there is a
772 shared and established way of working that facilitates the necessary communication
773 between stakeholders, the data analysists are at risk of being biased.
774 Another important element is understanding what questions a data scientist can
775 reasonably answer. At the end of the case study project, the Excel spread sheet had
776 multiple useful filters and showed interesting relationships, which the data scientist
777 could point out. The team kept asking her, why she thought these phenomena would
778 occur. She could not answer these why questions, because she did not understand how
779 the plants had worked. On the other hand, the team did not ask for the data analysis,
780 that she could have easily provided. The members of the process improvement team
781 often discussed issues over lunch or coffee, where she could have provided evidence,
782 but they did not ask her. This points to the underlying problem that the team did not
783 have enough data literacy, see section 2. It is equally a typical case where the data
784 scientist is expected to interpret the impact of her analysis onto application contexts
785 where she doesn't have sufficient knowledge.
786

DISCUSSION

The case study revealed many situations and aspects where the sociotechnical issues impacted the quality of the results along the data wrangling process. This section makes suggestions how to elevate these issues. It was evident that the viewpoints, goals, assumptions and biases of the different stakeholders had a major impact on how the data was presented and analyzed and in consequence what conclusions could be drawn from it. While the consequences of these issues only manifested at the end of the process, they had originated from the way the data was requested and the data wrangling process. The clients, i.e., the improvement teams, only reviewed their questions and assumptions when they saw the visualizations and analysis of the data after analysis. They were part of the data wrangling process, but could not shape it. Equally, the data scientist could not ask the relevant questions until she had sufficient understanding of the applied context. In the case study both were relatively inexperienced, however an element of iteration between data wrangling and data analysis is probably inevitable as well as iteration between the business domain experts and the data analyst on setting sound objectives and goals that align with the data and analysis quality.

Lessons learned for the organization

The engineers in the manufacturing organization have traditionally focused on the product and production equipment. They increasingly need to understand the impact the systems they are designing have on the data that will be generated and collected from their designs. In the case study the Kaizen projects did not affect the product, but

the efficiency of the process. In the future they will also think harder about how they can measure and attribute improvements. Data is used by people who know very little of the context where the data is generated. In the same way, the people who generate the data do not have control over later uses of the data making. Until now, manufacturing engineers following a lean paradigm, they have mainly addressed problems connected to product or production issues, while the awareness of issues associated with the data itself are only slowly coming to the forefront. With the ongoing digitalization transformation and the effort to increase competence, industry is starting to build departments of data engineers. In the 1980s and 1990s companies had to learn to integrate manufacturing engineers into product development processes since it became evident that business disciplines such as production, development and marketing were too loosely aligned [61]. Now the data scientists need to be integrated in business processes such as manufacturing and development, as data quality becomes more important and the lean production teams require a competence shift to incorporate data in their problem solving.

The project improvement team in this study reflected on what could have been done differently. They realized that data is not something neutral, technical and un-biased, and that more attention needs to be given to the data in itself, before using it to reach any conclusions and decisions. This team also appreciated that they needed to learn about data gathering, data wrangling and data analysis. Several seminars on this topic have been held and workshops are planned with each plant management teams to also grow their knowledge in this aspect. Since the study, the lean team and the IT team is

working together to drive the digitalization strategy and roadmap together with the engineering teams. By doing deeper investigations in selected production flows, the technical quality of the data has been understood further, but there is still more to do when it comes to understanding the social and ethical aspects of data.

Lessons learnt for data wrangling

The overall process had a deep effect on all aspects of data wrangling in this project:

- **Data availability needs to be checked:** The improvement team made assumptions about what data would exist in different plans and in which format it existed. While most elements of data did exist, some had to be derived from other data, where again the assumptions made in the data capturing process were not entirely clear, while other data existed directly.

- **Data requests need to be phrased carefully:** The way the data was requested in terms of where the request comes from and how it was phrased had a huge effect on the willingness to provide the data. It must be clear what the data is used for and what is likely to result from the analysis, as well as the consequences of withholding data. As data requests can just be seen as work with little immediate benefit, they need to be issued through people with suitable organizational clout.

- **A sensible common data structure must be negotiated:** In the case study the data scientist used the logic of Plant B and E simply they were handed over first. Later discussion with everybody let to the conclusion that the Plant D's logic

would have been better and fairer to all the plants. The structuring of the data is fundamental to the analysis and therefore to the conclusions that can be drawn from the data. This process is a core activity that a domain expert needs to carry out together with the data scientists. If more different data sources are available, a proper sampling process would need to be put in place to get a sense of potential variability in the data structure. In this project a discussion of the logic of data capture in the different plants would probably have also helped with the overall aim of supporting improvement in the different plants.

- **Data needs to be harmonized:** The data needed to be adjusted to comparable values and time periods, for example currencies, were converted with simple conversion factors, which a more sophisticated understanding of the data could have adjusted.

- **Double counting needs to be avoided:** A big challenge was the multicausal nature of improvements, where the different plants had to make a judgement call how improvements are attributed to individual projects. This is linked to the structure of the data in different organizations.

- **Validation of data interpreted:** While formal validation can be practically difficult, a (self) critical perspective on data validity is advised. If the results from data analysis are perceived as strange or controversial by experts, the validity of the underlying data should be checked.

- **When repurposing data, original assumptions need to be revisited:** When requesting data, it is important also to request information on the original purpose of the data capture. Unless such information exists, the risk of using data that is no longer valid or relevant can have large consequences [9].

In the situations reported, the technical challenges of cleaning and consolidating diverse data sets were evident but were less of a problem for the data scientists. The structuring of data turned out to be an issue, even though the improvement team initially considered the data set to be rather complete and well structured. The quality of data, and the support tools for data wrangling was an expected issue, but these aspects are well discussed in the community already [27]. The sociotechnical aspects were however significant. One reason was the relative inexperience of the data analyst and the manufacturing organization to address data specific issues. When using larger data sets, the readiness and maturity, including literacy and systems support, are factors to include. Maturity models for big data analysis [58] have been proposed for similar situations.

Another lesson is that support is required to ensure shared understanding between the data analysts and the manufacturing engineers. To arrive at sensible expectations, we suggest planning for reviews and iteration of the brief, as in Figure 1, including revisiting goals and assumptions. In the case study, the dialogue between the improvement team and the analysist gained momentum once the first (preliminary) analysis results were visualized. To develop realistic and relevant objectives we propose to iterate the problem statement with preliminary analysis results between the client and the data

scientist, even if the underlying data set is still incomplete. As such, the “product”, i.e., the data analysis results, can be prototyped and relevant questions can be raised before larger efforts are being invested. This is also in line with best practices in ‘design thinking’ [59], that have gained popularity as a pragmatic strategy to bridge disciplinary boundaries in problem solving.

CONCLUSION

This paper has illustrated that the data wrangling process is a sociotechnical process, that is deeply influenced by individuals and their personal understanding, biases and assumptions. Human stakeholders affect the data throughout the collection and data wrangling process. They decide what to include and how to interpret parameters. This goes beyond the awareness of data validity and quality of data, and requires the data scientist to interact with the client and even data providers in an iterative manner to direct the effort of wrangling data effectively. The way and timing data scientists are introduced to the problem and the data further introduces biases. As data wrangling involves not only the data scientists, but also the clients and the providers of the source data, their time, resources, competences and abilities, affect the quality of the wrangling and therefore ultimately the results of the data analysis.

In conclusion, data wrangling is a process that needs to be actively managed and supported by all of the stakeholders. As the study has done in the case study organization, this paper aims to raise awareness of the sociotechnical issues and thereby minimize the biases introduced by individual stakeholders. In the future, processes need to be developed to engage all stakeholders, not just the data scientists, in the data wrangling process. These processes will include systematic iteration, that does not only address the properties of the data, but also the assumptions, biases and values of all stakeholders.

ACKNOWLEDGEMENTS

Left out while in review

REFERENCES

[1] Isaksson, O. and Eckert, C., 2020, "Product Development 2040: Technologies are just as good as the designer's ability to integrate them," Design Society Report DS107, https://doi.org/10.35199/report.pd2040.

[2] Statista Research Department, 2016, "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025", accessed April 5, 2022, https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/.

[3] Byabazaire, J., O'Hare, G. and Delaney, D., 2020, "Data quality and trust: Review of challenges and opportunities for data sharing," in iot. Electronics, 9(12), p.2083.

[4] Gregory, K.M., Cousijn, H., Groth, P., Scharnhorst, A. and Wyatt, S., 2020, "Understanding data search as a socio-technical practice," Journal of Information Science, 46(4), pp.459-475.

[5] Hazelrigg, G.A., 1998, "A Framework for Decision-Based Engineering Design," Journal of Mechanical Design, 120, pp. 653-658.

[6] Mader, D.P., 2002, "Design for six sigma," Quality progress, 35(7), pp.82-86.

[7] Cai, L. and Zhu, Y., 2015, "The Challenges of Data Quality and Data Quality Assessment in the Big Data," Era. Data Science Journal, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002

950 [8] ISO 8000-2:2020(en), “Data quality — Part 2: Vocabulary”

951 [9] Woodall, P., 2017, “The data repurposing challenge: new pressures from data
952 analytics,“ Journal of Data and Information Quality (JDIQ), 8(3-4), pp.1-4.

953 [10] Eckert, C., Isaksson, O., Eckert, C., Coeckelbergh, M. and Hagström, M.H., 2020,
954 “Data Fairy in Engineering Land: The Magic of Data Analysis as a Sociotechnical
955 Process in Engineering Companies,“ Journal of Mechanical Design, 142(12).

956 [11] Coeckelbergh, M., 2017, “New romantic cyborgs: Romanticism, information
957 technology, and the end of the machine,“ MIT Press.

958 [12] Bucciarelli, L.L.,1994,“ Designing engineers,“ MIT press.

959 [13] De Weck, O.L., Roos, D. and Magee, C.L., 2011. “Engineering systems: Meeting
960 human needs in a complex technological world”, MIT Press

961 [14] Elish, M.C. and Boyd, D., 2018,“ Situating methods in the magic of Big Data and AI,“
962 Communication monographs, 85(1), pp.57-80.

963 [15] O'Neil, C., 2016, “Weapons of math destruction: How big data increases inequality
964 and threatens democracy,“ Broadway Books.

965 [16] Rose, D. and Agile, C., 2016 ,“ Data science: Create teams that ask the right
966 questions and deliver real value,“ (pp. 3-251). New York: Apress.

967 [17] Domingos, P., 2012,“ A few useful things to know about machine learning,“
968 Communications of the ACM, 55(10), pp.78-87.

969 [18] Shields, M., 2005,“ Information literacy, statistical literacy, data literacy. IASSIST
970 quarterly,“ 28(2-3), pp.6-11

971 [19] Gal, I., 2002,“ Adults' statistical literacy: Meanings, components,
972 responsibilities,“ International statistical review, 70(1), pp.1-25.

973 [20] Wallman, K. K., 1993, “Enhancing statistical literacy: Enriching our society”, J. of the
974 American Statistical Association, 88(421), 1-8.

975 [21] Francois, K., Monteiro, C. and Allo, P., 2020,“Big-Data Literacy as a New Vocation for
976 Statistical Literacy,“ Statistics Education Research Journal, 19(1).

977 [22] Wolff, A., Gooch, D., Montaner, J.J.C., Rashid, U. and Kortuem, G., 2016,“ Creating
978 an understanding of data literacy for a data-driven society,“ The Journal of
979 Community Informatics, 12(3).

980 [23] Giese, T.G., Wende, M., Bulut, S. and Anderl, R., 2020,“ Introduction of Data Literacy
981 in the Undergraduate Engineering Curriculum,“ In 2020 IEEE Global Engineering
982 Education Conference (EDUCON) (pp. 1237-1245). IEEE.

983 [24] Shmueli, G., 2010,“ To explain or to predict?. Statistical science,“ 25(3), pp. 289-310

984 [25] Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A.,
985 Shekhar, S., Samatova, N. and Kumar, V., 2017,“Theory-guided data science: A new
986 paradigm for scientific discovery from data,“ IEEE Transactions on knowledge and
987 data engineering, 29(10), pp.2318-2331.

[26] Provost, F. and Fawcett, T., 2013,"Data science and its relationship to big data and data-driven decision making," Big data, 1(1), pp.51-59.

[27] Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N.H., Weaver, C., Lee, B., Brodbeck, D. and Buono, P., 2011," Research directions in data wrangling: Visualizations and transformations for usable and credible data," Information Visualization, 10(4), pp.271-288.

[28] Rahm, and Do, H. H, 2000, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., 23(4), 3-13

[29] Baxter, D., Gao, J., Case, K., Harding, J., Young, B., Cochrane, S. and Dani, S., 2007, "An engineering design knowledge reuse methodology using process modelling," Research in engineering design, 18(1), pp.37-48.

[30] Mendekar , V., "Machine Learning – it's all about assumptions", accessed April 5, 2022, https://www.kdnuggets.com/2021/02/machine-learning-assumptions.html.

[31] Terrizzano, I.G., Schwarz, P.M., Roth, M. and Colino, J.E., 2015, "Data Wrangling: The Challenging Yourney from the Wild to the Lake," In CIDR.

[32] Minneman, S.L., 1991, "The social construction of a technical reality: empirical studies of group engineering design practice," (Doctoral dissertation, Stanford University).

[33] Clark, K. B. and Fujimoto, T., 1991," Product Development Performance. Strategy, Organization, and Management in the World Auto Industry," Boston, Massachusetts: Harvard Business School Press.

[34] Luhmann, N., 1992,"What is communication?," Communication theory, 2(3), pp.251-259.

[35] Luhmann, N., 1995,"Social systems," Stanford University Press.

[36] Craig, R.T., 1999," Communication theory as a field," Communication theory, 9(2), pp.119-161.

[37] Krippendorff, K., 1971,"Communication and the genesis of structure," General Systems, 16, p.171.

[38] Bordonaba-Juste, V. and Cambra-Fierro, J.J., 2009, "Managing supply chain in the context of SMEs: a collaborative and customized partnership with the suppliers as the key for success," Supply Chain Management: An International Journal.

[39] Hales, C. and Gooch, S., 2004, "Managing engineering design," London: Springer.

[40] Star,L. S., 2010,"This is not a boundary object: Reflections on the origin of a concept.," Science, Technology, & Human Values, 35(5), pp.601-617.

[41] Stacey, M. and Eckert, C., 2003,"Against ambiguity," Computer Supported Cooperative Work (CSCW), 12(2), pp.153-183.

[42] Gunasekaran, A., Yusuf, Y.Y., Adeleye, E.O. and Papadopoulos, T., 2018, "Agile manufacturing practices: the role of big data and business analytics with multiple case studies," International Journal of Production Research, 56(1-2), pp.385-397.

[43] Malinowski, B. 1929, "The Sexual Life of Savages in North-Western Melanesia," New York: Halcyon House.

[44] Mead, M.,.1928, Coming of age in Samoa: A Psychological Study of Primitive Youth for Western Civilisation. New York: William Morrow & Co.

[45] Hickey, S., 2004, "Participation: from tyranny to transformation: exploring new approaches to participation in development," Zed books.

[46] Clark, A., Holland, C., Katz, J. and Peace, S., 2009,"Learning to see: lessons from a participatory observation research project in public spaces, "International journal of social research methodology," 12(4), pp.345-360.

[47] Liker, J. K., 1997, "Becoming lean: Inside stories of US manufacturers". CRC Press.

[48] https://towardsdatascience.com/my-favorite-machine-learning-models-all-data-scientists-should-know-2ab94b4db62d

[49] Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N. and Chaintreau, A., 2020, "Biased programmers? or biased data? a field experiment in operationalizing ai ethics," In Proceedings of the 21st ACM Conference on Economics and Computation (pp. 679-681).

[50] McAuliffe, K. and Dunham, Y., 2016, ”Group bias in cooperative norm enforcement, ”Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1686), p.20150073.

[51] Crocker, J., 1982,”Biased questions in judgment of covariation studies,” Personality and Social Psychology Bulletin, 8(2), pp.214-220.

[52] Interview with Pete Jones, 2019, “Unconscious bias – avoidable or inevitable?”, European Research Council Magazine, accessed April 5, 2022, https://erc.europa.eu/news-events/magazine/unconscious-bias-%E2%80%93-avoidable-or-inevitable

[53] Hogan, R. and Kaiser, R.B., 2005, ”What we know about leadership,” Review of general psychology, 9(2), pp.169-180.

[54] Kauermann, G. and Kuechenhoff, H., 2010,“Stichproben: Methoden und praktische Umsetzung mit R“. Springer-Verlag.

[55] Thomas, 2020, “10 reasons why data science projects fail,” accessed April 5, 2022, https://fastdatascience.com/why-do-data-science-projects-fail/.

[56] Laing, D. 2017, “Communication in data science: more than just the final report,” accessed April 5, 2022, https://ubc-mds.github.io/2017-11-10-DSCI-542-communication/.

[57] Day, R., 2020, “Communication Can Make or Break a Data Science Project - 4 Key Communication Skills to Make You a Better Data Scientist” accessed April 5, 2022,

https://towardsdatascience.com/communication-can-make-or-break-a-data-science-project-75ce3952de89.

[58] Comuzzi, M. and Patel, A. 2016, "How organisations leverage Big Data: a maturity model", Industrial Management & Data Systems, Vol. 116 No. 8, pp. 1468-1492. https://doi.org/10.1108/IMDS-12-2015-0495.

[59] Liedtka, J. 2018, “Why design thinking works,” Harvard Business Review, 96(5), pp. 72-79.

[60] Stobierski, T. 2021, “Data wrangling: what it is & why it’s important, ”HBR On Line, 19 JAN 2021, , Accessed on Aug 6 2022**,** https://online.hbs.edu/blog/post/data-wrangling?fbclid=IwAR05KkHdXfEnCyHa5QJugJ--ktIC_vA3SXqER8rSeXgt6CqnE1dCQ9O4ipY

[61] Wheelwright, S. C., & Clark, K. B. 1992, “Revolutionizing product development: quantum leaps in speed, efficiency, and quality,” Simon and Schuster.