Data Fairy in Engineering Land: The Magic of Data Analysis as a Sociotechnical Process in Engineering Companies

Claudia Eckert
Faculty of Science, Technology, Engineering & Mathematics
The Open University, UK
claudia.eckert@open.ac.uk

Ola Isaksson[1]
Division of Product Development
Department of Industrial and Materials Science
Chalmers University of Technology
412 96 Gothenburg, SWEDEN
ola.isaksson@chalmers.se

Calandra Eckert
Department of Statistics
Ludwig-Maximilians Universität München
Geschwister-Scholl-Platz 1
80539 Munich, GERMANY
calandra.eckert@hotmail.com

Mark Coeckelbergh
Department of Philosophy
University of Vienna
Universitätsstraße 7 (NIG)
1010 Wien, AUSTRIA
mark.coeckelbergh@univie.ac.at

Malin Hane Hagström
Division of Product Development
Department of Industrial and Materials Science
Chalmers University of Technology
412 96 Gothenburg, SWEDEN

and
Volvo Powertrain
SE-405 08 Gothenburg, SWEDEN
malin.hane.hagstrom@volvo.com

[1]Corresponding Author

Abstract

In the era of digitalization, manufacturing companies expect their growing access to data to lead to improvements and innovations. Manufacturing engineers will have to collaborate with data scientists to analyse the ever-increasing volume of data. This process of adopting data science techniques into an engineering organisation is a sociotechnical process fraught with challenges. This paper uses a participant observation case study to to investigate and discuss the sociotechnical nature of the adoption data science technology into an engineering organisation. In the case study, a young data scientist / statistician interacted with experienced production engineers in a global automotive organisation to mutual satisfaction. However, the case study highlights the mis-aligned expectations between engineers and data scientists and knowledge in what is necessary to successfully benefit from manufacturing process data.

The results reveal that the engineers had an initially romantic and idealistic view on how data scientists can bring value out of dispersed and complex information residing in the multi-site manufacturing organisation's datasets in a "magic" way. Conversely, the data scientist had not enough engineering and contextual understanding to ask the right questions. The case reveals important shortcomings in the sociotechnical processes that undergo changes as digitalisation is brought into mature engineering organisations and points to a lack of knowledge on multiple levels of the data analysis process and the ethical implications this could have.

1 Introduction

"Come, faeries, take me out of this dull house!
Let me have all the freedom I have lost—
Work when I will and idle when I will!
Faeries, came take me out of this dull world,
For I would ride with you upon the wind,
Run on the top of the dishevelled tide,
And dance upon the mountains like a flame!" W.B. Yeats [1], The Land of the Heart's Desire

Data generated in manufacturing industry is increasing rapidly [2]. Expected business benefits from this data are high, and companies launch initiatives, hire consultants and invest to come out as winners in the ongoing digital transformation. Data scientists are seen as magicians, bringing their expertise in Big Data analytics, machine learning (ML), artificial intelligence (AI) or statistics to help companies to reveal the golden insights residing in their data on product quality and process efficiency. This paper uses a case study to discuss the expectations that engineers and managers in the manufacturing industry have about data scientists, and report on problems and practical challenges when engineers and data scientists work together to analyse the data sets.

The opportunity for this paper arose from a production systems development engineer and her team in a global manufacturing unit within the truck manufacturer Volvo (the fifth author of this paper) decided to engage a young statistician (the third author of this paper) for a short term student placement. They asked the statistician to investigate their data set on production quality at different production sites using statistical analysis. They recognised a clear potential in their data set and knew that it needed to be analysed further. Volvo also saw the placement also as a way to learn from engaging with statisticians. This type of engagement is typically awarded to professional consultants and thus rarely reported in literature. This was a unique opportunity to observe and reflect with the participants on the sociotechnical nature of the process. It illustrates issues that a data scientist can meet when entering into the engineering land of a manufacturing company such as Volvo and how the magic of data analytics actually happens.

The objective of the paper is to discuss the sociotechnical aspects when introducing new data science related skillsets necessary into a manufacturing engineering industry context. In particular, the mis-aligned expectations of what can be achieved, and what is needed to succeed is addressed. The short placement is used as a case that give empirical examples for the discussion.

The engineering of industrial products, processes and systems has as long been recognised as a sociotechnical process [3], where the successful delivery of manufactured products largely depends on the process by which they are designed and manufactured. Manufacturing industries invest continuously in improving and renewing these processes. Mature businesses such as manufacturing in the automotive sector have for decades stayed competitive by a combination of economy of scale and continuous improvement together with offering more advanced products. Management approaches such as Lean Manufacturing [4] and Six Sigma [5] build on rational and data driven prediction, analysis and decision-making [6]. Manufacturing companies are often large, global and complex organisations where practices, norms and tools for how data is processed used typically vary.

Although engineering in general has been at the forefront of using – and even advancing – computing and data analysis technologies for many decades, the situations that arise from a broader digitalisation transition are new. Artificial intelligence, statistics and big data analysis have typically been applied in a more limited and technical context, typically where the engineers have first hand access to the sets. Therefore, the sociotechnical aspects of gathering, handling and analysing data have been less pronounced when it was kept within one technical competency. Data is now gathered from multiple sources, often generated for different purposes, and represented in a large variety of formats and forms. We are now at a point of change where engineers working on all aspects of the product life cycle have to engage with data to an unprecedented extent.

Ongoing improvements now include taking advantage of the rapidly increased ability to generate and process large quantities of data within all the activities along a product life cycle. The ability to collect and store data increases and the sheer volume of data gathered is steadily growing. It provides the potential for greater predictability of human behaviour [6]. Data from manufacturing processes is seen as a credible and objective way to gain wider insight and knowledge and informed decisions on e.g. quality improvement and process performance. This knowledge is vital to improve and optimize current production processes. It is also needed to gain insights for decisions onto how to design next generation products and processes. However, the journey from data to resulting decision is not a deterministic process, but one where humans have to take multiple decisions on the way, where data analytics can help to prioritise human attention [7].

European Union [8] update is pushing the “digital transformation” as a key strategy to for both societal development and business success. The expectations of society at large and engineers that data can resolve current problems are currently extremely high. Regarding AI, Elish and Boyd [9] put it as “AI has the potential to collapse under the weight of the hype that surrounds it, even though the analytic contributions it has to offer have significant potential”. AI is currently discussed in the public debate through three interwoven strands of discourse: AI as portrayed in science fiction, high-profile real-world development in AI, and futurology pointing to the risk of machine intelligence outstripping human intelligence [10].

Gartner [11],[12] describes hype cycles of technology with the following phases: Technology Trigger, Peak of Inflated Expectations, Trough of Disillusionment, Slope of Enlightenment and Plateau of Productivity. These are based on classical S-curve models [13], but distinguish between the hype around a technology and the actual development. While not being scientifically rigorous, (see [14]), the hype phases describe the current romantic enthusiasm for data science and the disillusionment that is likely to befall it unless the engineering community engages with some of the technical and ethical challenges that lie ahead and of which this paper gives a flavour. For example, the notion of data is

ambiguous, as the data sources deliver data of different composition and quality, whereas expectations about comparative analysis following statistical processing raises unexpected issues.

This implies that engineers either acquire data management or analysis skills or interact directly with data scientists such as statisticians. Like all cross disciplinary collaborations, this brings its own challenges. This paper highlights some of the different perspectives of engineers and statisticians in approaching data sets and the difficulties they have in understanding each other's and their own information needs, and their divergent interpretations of key concepts, like what constitutes a model statistical analysis needs hypotheses in terms of the relationships that are being investigated. While it is possible to discover correlations from data sets, it is important to do so only where it makes sense in the context of the specific situation. Otherwise spurious correlations could lead to time consuming or harmful actions.

Manufacturing industries typically generate data and deploy it in decision making and efficiency improvement in three different life cycle phases, as illustrated in Figure 1. Based on specifications from business units (marketing, sales etc) engineers initially develop new and improved technologies and products that can be offered to the market. In the second phase, the products are being manufactured and assembled; and finally the products need to be sustained and maintained. Each phase uses and generates an increasing amount of data. In the Technology and Product Development phase, typically through modelling, simulation and testing of technologies and forthcoming products to ensure that requirements and expectations are met. In the second phase – Production – data is being generated to with the aim of efficiently realising products. Measuring and controlling activities and material flows generate data that is used for quality control and prognosis of process performance in general. In the In-Service phase, data is generated from the product in use, both to operate it, e.g. for autonomous cars, and to monitor it, e.g. for predicting and scheduling maintenance. Data is captured both within each phase, but also shared between the phases. In particular data from production and in-service provides a foundation for design improvements.

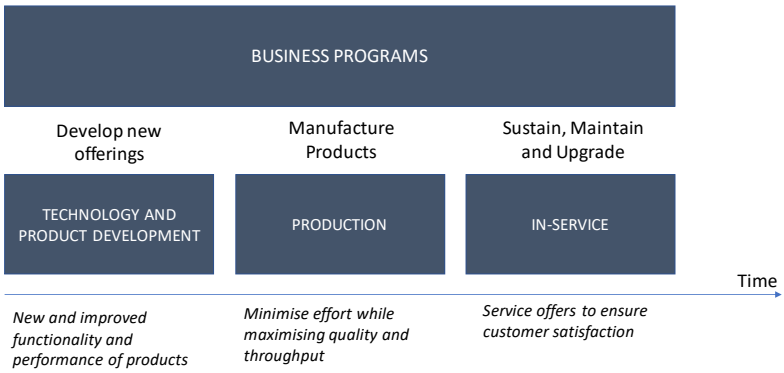

Figure 1 Main engineering intensive life cycle phases of a manufacturer.

Managers and engineers alike see an underutilized source of insight hidden in the increasingly rich data sets that are being captured. It is recognised that such data need to be analytically studied to extract these insights. Big data approaches and – in absence of a clear understanding – statistics become an almost magical way to finding links and gaining insight. In this paper we will discuss this in the context of discussions of the magic of technology. However, Domingos [15] summarised it nicely when he said: “In retrospect, the need for knowledge in learning should not be surprising. Machine learning is not magic; it can’t get something from nothing.

This paper puts data analysis in the context of working across disciplines and discusses the rapidly changing nature of data analysis. Section 2 sets the background of working with data scientists as a sociotechnical problem. Section 3 illustrates the sociotechnical nature of data analysis by discussing a case study of a data analysis problem in industry, whereas section 4 focuses on issues arising from the pivotal role of data models. Section 5 puts the fascination with data in the context of the long running debate on the magic of technology and AI and points to the layers of misunderstanding that can exist in a data analysis process. The paper draws conclusions about the wider implications of our study for the practice of manufacturing engineering.

2 Working with data experts

Data analysis, Big data and AI are wildly recognised trend areas in engineering, [6],[16],[17] and engineering companies are bracing themselves to respond and work with data experts.

2.1 The challenges of cross disciplinary working

Realising the full potential of new technologies often requires greater collaboration across disciplinary boundaries, which can be challenging for engineers from different backgrounds. Engineers are used to working in their own object worlds [18], i.e. the concepts, terminology and domain knowledge of their own field, and can find it difficult to collaborate with others who have different ways of thinking and use a different vocabulary. To work together they depend on boundary objects, which are understood by all the groups that are relevant [19]. The first step of 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 [20]. This however requires a sufficient understanding of the information needs of the group they are working with. In addition, as engineers work across supply chains or companies merge, the engineers also have to overcome cross-cultural differences in procedures and methods as well as concepts and terminology. In the case study data had to be obtained from several recently acquired companies, who were still operating some of their old systems. The companies often use different data base and file formats, so this presents additional challenges when consolidating data before analysis. Further, the work processes and IT tools used also differ within large organisations.

In industrial processes, cross-disciplinary collaboration is necessary, for example between manufacturing engineers, design engineers and operational research and logistics experts. These groups have been working together for a long time and frequently run into problems that emerge in integration due to communication problems [21]even though an increasing number of engineers are trained at least to a certain extent in both areas. When data is gathered within the In-Service phase, the primary use of data is to ensure availability of equipment and product through careful maintenance and overhaul processes, yet the data is found increasingly useful for the design of new products, business programs and business models. Data needs to be correctly interpreted consistently by different domain specialists and roles in the organisation.

2.2 The rising prominence of data in engineering

Data is prominent in all phases of the product life cycle in engineering. For a very long time, engineering has been at the forefront on taking advantage of – and even leading – advancements in data analytics and use of IT. Analytics have traditionally focussed on physics-based phenomena, but have started to address social and behavioural aspects of real world variation and problems. Alignment and compliance with physical testing is still seen as the main source of validation. Measurement-based approaches such as Lean Production [22] and Six Sigma [23] introduced systematic procedures to understand and improve production processes; and have paved the way for increased level of automation in industrial processes. Industry 4.0 aims to achieve greater autonomy in production though introducing sensor-based technologies. It enables companies to link data across the life cycle from production and use of a product from a multitude of sources to monitor and manage their products in the so called IOT (Internet of Things). "Digital Twins" create digital clones

of products and processes, experiments, prognoses, deviations, change impacts etc, that can be simulated to provide useful input for maintenance planning, production flow optimisation and effects of product, process changes and so on [24]. This data can then be combined with data from modelling and simulation activities to inform the development of the next generation of the product and assure a robust design.

Expectations are high. The European Union expected an annual efficiency gain in European manufacturing of 6 to 8 % through the introduction of industry 4.0 [25]. The EU measures progress with digitisation through "digital transformation scoreboards"[26]. Production is priority area and big data analytics is seen as a driver of digital adoption The lack competence in digitalisation has been raised as a bottleneck. As an example, Soban et al. [27] emphasised the role of visual analytics for decision makers as a means to interact with datasets, incomprehensible unless these are processed and prepared.

However, not all data will be generated automatically and directly from sensors. Instead some of the critical data will need to be gathered manually or derived from other data sources. Analysis of data in industry dependents on the quality of data. This has led to the development of new standards for data quality to handle the ever-increasing richness and complexity of data sources. Within the ISO/TC 184/SC 4 Industrial data organisation, a new ISO 8000 series on data quality [28] has been developed with the view that the value of data is the information derived from the data, and that such information depends on how data can be trusted, the relevance of data, its representation and its timeliness. The ISO8000 standard is a basis for certifying Master Data Quality Managers. The standard provides a structure to represent and manage the value of data, as well as the figures.

2.3 Skills of a data scientist – the myth of the data science unicorn

Data science is a quickly evolving field, which is still debating the skills it requires, as illustrated by the debate on professional websites cited in this section.

Engineers often hope that they can hand data over to a data scientist, who then reports back with useful insights with little further input [29]. This is unrealistic. Unless the data scientists understand the context in which they are working, such as how the model will be used, where the data comes from and what the priorities of all key stakeholders are, they cannot ensure that their model or analysis actually provides the required value for the engineers. Likewise, the engineers must understand the results of the data science project to use them, for example to make decisions or integrate them into their process. For the results to be trusted enough to be used the data scientists need to be able to answer technical questions to a level that satisfies the engineers in a language that they can understand. This understanding can only be gained though open communication on both sides [30]. A lack of effective communication skills is widely regarded to be a key reason why data science projects are often considered failures[31]-[34]. The amount of effort to correctly understand, interpret and act in a new context is underestimated.

Interdisciplinary communication is just one of a wide range of skills and prior knowledge needed in data science projects. Other required skills include: extracting data from different sources; pre-processing data so it can be used; finding patterns in complex data structures that could be of business interest; building effective and efficient models; conducting rigorous statistical testing; visualise and present results in a compelling and accessible fashion; communicating effectively with shareholders from different backgrounds and building infrastructure and develop software [35]. Finding a single individual capable of doing even a large fraction of those jobs is so rare that they are often referred to as unicorns [36][37]While "Many organizations are seeking unicorn data scientists, that rarest of breeds that can do it all" [38] Baškarada and Koronios [38]established through an interview study, that they rarely do. The skills that the individuals bring with them depend on their own background and skillsets. While an increasing number of universities are offering designated data science degrees, most data scientists working today have crossed over from a wide range of different fields such as

computer science, mathematics / statistics, natural or social sciences, engineering or economics or business (source) and taken supplementary courses[39].

Most data science projects are carried out by teams, who collectively ought to hold the skills needed to carry out all steps in the project. Wrong team composition is a frequently cited reason for data science projects failing [30]. Data science team roles are often ambiguous[40]. Common job titles and associated roles include:

- *Data analysts* search for patterns in data [41]
- *Business analysts* have the domain knowledge to determine which patterns could be of potential business interest and often facilitate communication between the project partners and the data science team [42],
- *Data engineers* are usually involved in building the infrastructure to extract and use large amounts of data; they need to be very good at programming [35]
- *Machine learning engineers* also build predictive models using different modern machine learning [43]
- *Data scientists* are expected pre-process, analyse and visualise data, carry out statistical tests and build models using both classic statistical methods as well as modern machine learning techniques, and communicate results [43].

In practice these roles can be shift, for example the data analyst could also be the team's "data storyteller" responsible for constructing a compelling narrative around the results[35].

2.4 Data science as a sociotechnical process.

De Weck et al [3] point out that today "Today, working in an engineering system, that same engineer has to interact with a host of socioeconomic complexities and "externalities" – impacts, either positive or negative, that are not a direct part of the artefact or even a self-contained system or process under consideration." For the engineers the skillset of the data scientists is one of these externalities. They need to have a close look at the skills the data scientists bring, as well as the skills available on the side of the engineers to support them, to see whether these are likely to be sufficient to cover all stages of the task at hand. However, it is not only the technical skills of the engineers and data scientists that determine whether they can successful communicate, but also the personal interaction and personality. The views of the engineers that engage with the data scientists are those that are privileged and reflected in the models and analysis.

Communication between engineers and data scientists is a vital aspect of the social process. The data scientists need understand the domain context. They not only have to negotiate the access to data [44] but also negotiate the definition and scope of their tasks. On the other hand, the engineers need to instruct the data scientists and interpret their results. The ability to understand and critically evaluate statistical results is a major aspect of statistical literacy [45] -[49]), which is essential for any form of data-based decision making. As the rise of easy to use machine learning tools makes statistical results more widely accessible, the ability to interpret results becomes all the more important. Statistical literacy is a broad concept that encompasses a wide range of concrete abilities and contextual knowledge [45].

Statistical literacy in teams is often collective ability, as one person rarely holds all the relevant knowledge and abilities. Even then the number of equivalent terms for the same concepts can pose a problem [50]. Which terminology data scientists and those aiming to collaborate with them are most comfortable with depends largely on their educational background or object world [18]. Statisticians or those who took statistics courses at university would most likely speak of 'fitting' a model to 'describe' the relationship between the 'independent variables' and the 'dependent variables' using a data set containing a number of 'observations' [51]. While people who came to data science from a machine learning angle would describe the same task as 'training' a model to 'learn' the relationship

between ‘features’ and ‘labels’ using a ‘labelled set of examples’ [52]. This distinction is particularly important in this context, because many engineers have encountered statistics in their degree, whereas the majority of data scientists come from a machine learning background.

3 Empirical example of sociotechnical aspects of data analytics

The previous section establishes data science as a sociotechnical activity. This section highlights some of the practical challenges a data analysis activity can face. It shows up some of the issues that can arise if engineers approach a data analysis tasks without much prior preparation.

3.1 The case study methodology

The specific example is based on a short-term placement of a young statistician (the third author of this paper), who has a background in statistics and philosophy, in the production organisation within Volvo Powertrain (represented by the fifth author of this paper). The company often brings in students and gives them freedom to explore to get different views on the system and the application of lean improvement methods. This is a conscious learning strategy of the company, when learning to utilize techniques new to the organisation. On the personal level the team was extremely friendly and helpful and made the statistician feel part of the team.

This paper is written in the spirit of participatory observation, which has a long history in social sciences and anthropology going back to the pioneering studies of Malinowski [53]and Mead [54] in the 1920s. It has risen to prominence in a wider research community to enable more socially relevant research where research subjects ″can bring their own ‘voices’” to the research process [55]. It enables researches to “produce knowledge about the everyday interactions of people”[56]. Unlike in ethnography study, two of the authors were active participants in the process, however they did not enter the study with specific hypotheses or questions [57], [58]. Rather the paper arose from their telling of their experiences to the other authors. The arguments put forward in this paper are themselves a sociotechnical artefact negotiated by all the authors. This enables the authors to get the perspectives of both the engineers and the data scientists and to discuss the sociotechnical issues freely. The willingness of the participants to share their experiences was a large part of the motivation to write this paper.

A powertrain of a truck is a part of the product platform and is used in vehicles of multiple brands as well as sold independently. Volvo Powertrain has five plants worldwide building powertrains in Volvo Group. Volvo Group has grown by acquisition which means that some plants had their own quality procedures stemming from the period when they had operated as independent companies. The aim of the placement in the production organisation at the main site in Sweden was to analyse how effective quality improvement measures were at the different plants by understanding how well the data for wastes and losses are captured and how effective the different plants are in addressing these losses. At the same time Volvo Group saw this as an opportunity to learn how to work with a data scientist.

There was not enough data to carry out extensive statistical analyses and the statistician developed 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 unbiased analysis across the different plants. Even though the team was extremely welcoming and approachable, they did not know what they needed to explain and the statistician did not know what to ask.

The company was happy with the results of the placement and can now analyse how effectively the plants carry out projects to reduce loss. The study has been presented in the knowledge networks within the company and is now analysed on global level as input to take the next steps in the improvement journey.

3.2 The task and the available data

The data was not collected deliberately for this project, but part of their established process for documenting Kaisen Projects

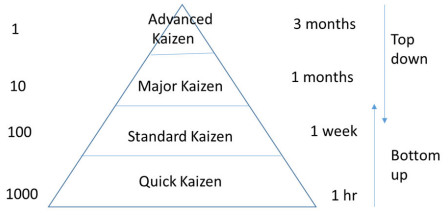

Figure 2 Kaizen hierarchy

The company follows a lean production paradigm and engages in Kaizen initiatives [22]. It uses the term for improvement activities in their manufacturing processes at different scales and levels of effort from simple and sporadic problems to chronic and complex problems as illustrated in Figure 2. A Quick Kaizen could be done by anyone, bottom-up as the company calls it, and without prioritisation and should take about one hour to implement the improvement action. This means that thousands can take place every year in a plant. The Standard, Major and Advanced Kaizens are all carefully prioritised and triggered in a top-down manner, as they are more expensive to perform. They take longer to complete, require experts and are usually triggered from the top-down perspective. A Standard Kaizen should be implemented within one week, a Major Kaizen within one month, and the Advanced Kaizen can take up to three months to implement. The more advanced the Kaizen, the more savings are expected to be generated.

The data from the bottom-up Kaizens is collected by each operator in a paper format and then transferred to data sheets, either by the team leader or the improvement leader. The data from the top-down initiatives is collected by engineers into spreadsheets. The design of the spreadsheets varies between locations, which contributes to the data cleaning challenge.

The statistician was given two sets of data from the Goteborg plant while further data was requested from plants in Sweden, the US, Brazil and France. The plants maintained an independent work culture. While they were supposed to report largely the same variables, their interpretation of the data diverged in different plants. The data arrived in individual spreadsheets. Some plants already had spreadsheets with the data on all the improvement projects; others had to collate data from different spreadsheets recording projects in a 4 or 6month time period. Some time periods were missing, which had to be requested individually. Some plants changed the way they recorded projects over time, so even on a plant level it was labour intensive to produce a uniform dataset. Between plants the structure of the datasets varied considerably, in terms of what was recorded and how the variables relate to one another. With considerable effort it was possible to transform the heterogeneous data sets into a uniform dataset that could enable a meaningful comparison.

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

3.3 Challenges for carrying out the data analysis

The case study highlights some factors that could affect wider collaboration between engineers and data scientists. In the case study the statistician was located with the engineering team and had the opportunity to ask questions during the day, unlike others who are in data analysis departments or

work as consultants, which might exacerbate problems. The team did not know what to expect from a trained statistician.

Timely availability of data. The greatest challenge was to get sufficient data to carry out the analysis. While data from the Swedish plants was available at the beginning, it was much more difficult to obtain the data from the other plants. The data was repeatedly requested by various members of the engineering teams. Most of the data eventually arrived, but one data set was never provided. It clearly had not been possible to communicate the benefit of the data and the analysis of the data to the other plants.

Visualisation versus Correlations. The team wanted to have a convenient way of viewing details of the data in an Excel spreadsheet, that they can click through. They wanted filters put into the data set, e.g. filtering the data by year to view different element of the data together, so that they could directly see relations in the data. In addition, statisticians want to find correlations or patterns in the data through statistical analysis. Correlations can indicate causal connections and therefore be used to identify where to look for the root cause of problems.

Using suitable statistical package. The engineers were familiar with using Excel and had Excel as standard software on their machines. R, the freely available statistical analysis language, would have been better suited to analysing the data, because it would have enabled easier changes when the new data with a different structure became available. However, getting permission to load R onto the company machines could have taken an infeasibly long time.

Understanding the importance of the structure of the data. The statistician started with one data set and got an initial explanation how the process in the plant worked and what different data entry points meant. From hindsight this turned out to be the partial view of one person on that data set, which was difficult for a new person to challenge. It later emerged that the underlying logic of the data set was different in terms of what the individual plants measured, where sometimes variables of the same name had different meanings. For example, the different plants had different time cycles over which they captured some of the data. The excel spreadsheet was therefore prepared following the logic of the first data set that was made available and the other data sets are adapted to follow this logic. This has a profound effect on the analysis that can be carried based on the data and ultimately the actions that can be derived from this. The order in which data sets are provided and explanations are given therefore has a profound effect on the outcome.

3.4 Challenges in making use of the potential of data analysis

Knowing what the statistician can know. The data analysis revealed that one of the plants had better results, i.e., fewer mistakes than they had predicted themselves. This would point to either very conservative estimates in the first place or underreporting of faults that actually occurred. These results surprised and concerned the engineers. They therefore asked the statistician for reasons, which cannot be provided just by looking at the numbers. In practice the engineers have to make the step from correlation to potential causality themselves.

Understanding how statistics or data could support arguments. During lunch time the engineers often discussed their on-going work, in particular needing to write reports or requests for changes to the manufacturing process. The statistician tried to extract and test hypotheses based on these informal conversations but the data set would have yielded many more supporting graphs or figures, if it had occurred to the engineering team to ask the statistician for these graphs directly.

Recognising what was useful when they saw it. When the engineers saw the visualisations and the ability to "click through" the data they recognised that some of graphs would be really useful for them. They saw the potential of their data when it was prepared for them rather than providing input into the decision-making process. This makes data analysis serendipitous, rather than goal directed.

Pushing the envelope of the analysis. The engineers largely looked for quantitative confirmation or contradiction of the impressions they formed based on anecdotal evidence. The engineers did not ask for additional analysis nor tried to push the scope of the analysis beyond what they had at that moment. The data analysis would have allowed them to get interesting results out; however nobody asked the statistician additional questions that would have prompted her to generate this analysis. The people who understood the process well did not think in terms of the criteria and measures to improve the process. Instead this left the statistician with anticipating what would be of interest to them and incorporating that in her analysis.

Communicating the benefits of the data analysis of the different plants. The plants might also have felt a certain unease about the impending analysis, which might "find them out". Rather than seeing the analysis as a help to identify their own problems, they might have seen it as a thread to their established way of working. The engineering team was not surprised at some of the reluctance to hand over data and saw that as a part of an ongoing power struggle between the plants.

The different plants had their own systems of capturing and reporting data, where data was recorded in different systems or spreadsheets with a different structure. Therefore, somebody had to manually prepare the data for the analysis by copying particular rows or columns into a dedicated file. This became evident in data cleaning as some entries were clearly in the wrong place.

3.5 Sociotechnical aspects of the case study

Analysing data and drawing conclusion from it is complex sociotechnical process in its own right [9], where a deep understanding of the problem from which the data arose needs to come together with state-of-the-art analysis techniques. Social scientists have pointed out social values are underpinning the techniques and practices and are encoded in the mathematical processes (e.g. [59]).

The case illustrates that in data analysis the challenge is often not technical, but social or rather that the social and human challenges need to be overcome before it is possible to engage with the difficult technical challenges. The engineers were very happy with the Excel tools they were given because they wanted to see their data and draw their own conclusions from the data. In this sense the data was not analysed. If the data had been available at the beginning and arrived in a suitable structure, it would have been possible to apply statistical methods to look for significant correlations.

The team had little awareness of the different aspects of a data analysis task and roles as outlined in section 2.3. They expected that the data scientist would yield useful results, but underestimated their own role in guiding this process. It was obvious to them that without data no results would be forthcoming, so they helped obtaining the data sets. However, for the engineers it was much less clear what they needed to explain to the data scientist. Equally, the data scientist was unsure what questions to ask. This is typical for cross disciplinary collaboration and shows the importance of a basic level of statistical literacy (see section 2.3). The participants have learned from this experience and future interactions are likely to be smoother. In established collaboration, different groups know what inputs the other groups require and what results to expect. This was not the case here, the engineers still needed to learn what guidance they needed to provide and what information was needed to direct the process. From hindsight they would have benefitted, if they had explained the context more clearly upfront and voiced some of the concerns or suspicions, they had about the different manufacturing plants, which could have been turned into testable hypotheses.

The remainder of the paper will pick up on issues that were very prominent in the case and have a wider implication for current data science: the central role of a data model as a boundary object between the data scientists and the users; and the belief in the almost magical properties of data analysis for understanding and addressing problems, which otherwise would be difficult to tackle.

4 The central role of data models

This section argues that one reason was that the engineers did not recognise the preparation of the data as a modelling exercise. Underlying this are different notions of what a model is, held by engineers and data scientists.

Models are abstractions of reality for a specific purpose which can described in different ways [60]. Models typically represent an aspect of reality [61]; however the relationship between models and target systems, i.e. the piece of reality they are modelling, is not always unproblematic.

Statistical modelling is the generation of mathematical relationships between the input variables and output variables that reflects an abstracted image of their real world connection in order to gain a greater understanding of the relationship between the input and output variables [62]. In order to do this a model needs to be found for which the assumptions are sufficiently fulfilled to enable confidence in the possibility of generalising beyond the given sample. Typically, statisticians take the data as given and do not engage actively in the generation of the data. Machine learning follows a different paradigm. It aims to build models that can reliably predict outputs variables using the input variables. Therefore, underlying assumptions are not relevant [15]. The two approaches to data analysis have the same mathematical foundations and share many of the same methods (even if they have different names for them) but they are used to achieve different goals.

Models are a fundamental way for engineers to interact with the emerging product and the process by which it is generated. Engineering models have been researched much less than models in science, but have a more complex relationship between the models and their target system as many models are generated before a target system exists. Engineering models can be divided into generative models, who describe the future product and evaluation models, which assess whether the actual or intended design have the intended properties [63]. In spite of the enormous importance of models in engineering, engineers rarely reflect about the nature of models, and typically adopt the view that a model represents its target system. In most engineering contexts the modelling conventions are fairly clear and standardised (e.g.[64]), so that engineers do not have to think about what is included in a model.

The performance of manufacturing processes can be modelled in different ways to design, monitor and improve the production process. In the targeted organisation within Volvo this also depends on the perspective of the engineer: The perspective of the global process owner or the perspective of a specific plant. The models capture different elements of the target system. Some are based on physical flows, while others also capture more invisible information flows. For example the case study used a model of the physical order of the stations in the production line, a loss data model based on the platform of eleven key knowledge areas, and the Kaizen classification of project effort based on problem complexity (see Figure 2).

The data model in the case study is similar to the schemata developed by the data scientists, albeit on a smaller scale. The engineers left the modelling process to the statistician with a relatively loose brief. The process of generating a coherent data set is itself a social process, as the articulation of categories and deploying of classification is a historically determined social process [65].

4.1 Challenges arising from the analysis brief

Deciding on the purpose or the purposes of the models. The purpose determines the selection of the relevant information. While this does not preclude unintended uses, a clear purpose also assures consistency in the selection. The engineers did not provide a clear list of all the aspects they were interested in.

Missing assumptions. The engineers had made multiple assumptions about what improvements would be included by the data reported from the plants in their data sets and what would constitute good or bad practise in the plants. Improvements to Kaizen events were attributed differently.

Developing a coherent and agreed vocabulary for the variables used in the model. The terminology of the data model was based on the data set from Swedish plants, because the other data sets where not yet available. The other plants were theoretically expected to use the same data capturing template as their Swedish parent company, but deviated considerably. Some variables were included, because it would be useful for further analysis even though not every plant had selected them, such as the classification of the loss which the project aimed to address. The intent was also that by including it in the standardised merged dataset the other plants could be encouraged to collect data on these variables in future allowing analysis at a later date. In the case study company, the data was not even in the same language and some measures were taken in different currencies. Particularly for key variables like "focus area" the abbreviations had to be standardised across all datasets.

4.2 Challenges from using data schema

Significant effort was needed to represent the data in suitable data schemata. Harmonising data from different sources with different structures required making assumptions about the 'right' way terms and concepts relate to each other. These assumptions needed to be explained. Here again the aspect of data science as a sociotechnical process can be seen, as the perspective and priorities of whichever individual engineers communicate with the data science team will be privileged in the model as they shaped the data scientists' understanding of the data generating process and project goals. In the case study, the way the Swedish plants were run, was expressed in the data model.

A schema contains the variables which structure the data and their relationships. The issues of data structuring or data warehousing is of greatest concern to the data science community. Rahm and Do [66] 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 at the instance level, where errors and inconsistences in the actual data occur. 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 contradictory records. Cleaning the data took a very large fraction of the effort, in particular as the data was in different languages and had different abbreviation conventions.

4.3 Challenges from the data sets

Misunderstanding the purpose of the modelling for the data providers. The different plants needed to pull data out of existing spreadsheets. The differences in the provided data sets arose from misunderstandings about what the metrics meant and how they related to each other; and what the rationale for the metrics was. There were clear differences in the quality and validity of the provided data.

Deciding on the logic by which the data is structured. There are often choices about how certain processes are measured and which variables need to be connected. Data can be bundled, which might camouflage certain relations, or individual points might be included which gives them undue importance in the analysis. For example, only two of the plants explicitly measured the size of the losses avoided in the future as a result of the implementation of the project. For the other plants only savings were recorded. Another interesting difference could be found in the relationship between direct and potential savings. In some plants one project could have direct savings, potential savings and cost avoidance and these were added together to make total savings, while at other plants a project could only have one of either direct savings, potential savings or loss avoidance. At the fourth

plant direct loss was given as percentage of potential loss, which shows that these concepts were fundamentally understood in different ways at different plants.

5 The magic of data

As outlined in the introduction engineering is placing high hopes in data science to understand their products, markets or problems better. Some of these hopes arise from the general hype around data science, others from the knowledge that good data can provide insights. The expectation that data reveals its meaning by itself, is even more prevalent when it comes to the analysis of big data. However, big data approaches are based on the assumption that the data itself is coherent and valid and that the quality of the data is roughly similar across the data set. In the example of the case study, a big data approach would be not be practical for the Kaizen projects, because the number of projects and therefore opportunities to make savings is quite limited.

In the case study the engineers had a rich, but by no means enormous data set, but still had hoped to gain insights without having clear hypotheses. They provided the data scientist will all the data they could access and were very outstandingly friendly and encouraging colleagues. Like magical creatures in stories, they looked very well after the data scientist and in the hope that she would do her "magic" and provide insights that the engineers could not obtain from eye balling the data. Having produced an Excel spreadsheet, the data scientist was able to draw inferences from the data. These insights surprised the team, and they kept asking her for what she thought was going on within the plants and why this might be the case. Which were not questions that she could answer.

All this gives data an almost magical property in the eye of many potential users of data analyses and make the statistician or data scientist an almost magical creature, who can reveal connections and find causes when the domains experts themselves could not. Like with magical creatures in fairy stories, it had never been clear what exactly the data should reveal and what the consequences of any insights might be. The difficulty to predict what data analysis can bring and the heightened expectation have a profound effect on the sociotechnical process when applying data science.

5.1 The magic of technology

Viewing technology, in particular new technology, as magical has a long tradition. The science fiction author Arthur C. Clarke is often quotes for putting this succinctly as "any sufficiently advanced technology is indistinguishable from magic" ([67]p. 21). The phenomenon of technology as magic has been studied in different fields and comes to the forefront around applications of powerful AI. It gives us an additional lens with which to look at the current fascination with data science as a form of 'enchantment'.

In anthropology Gell [68] has argued that magic is a "commentary" on 'technical strategies in production, reproduction, and psychological manipulation' (p. 8). He also connects magic with the absence of apparent cost and effort: magic does not have the costs and disadvantageous of real production such as struggle and effort (p. 9). Technology is also magic in the sense that it has the power to enchant, in that it can make us do things that are against our personal interest and see things in an enchanted way [68]. This use of the term is thus linked to the issue of cost and struggle: the idea is that the data itself (and the magic work of the data scientist) will easily deliver results. It also refers to the power that people hope to achieve by using data science. The result is indeed a kind of 'enchantment' in the sense that individuals, professional groups (engineers), companies, and nations embrace AI and data science without really knowing how it works.

There has also been some work on how contemporary information and communication technologies enchant the world, and how this can be understood by means of comparisons with romantic science and stage magic. Technology becomes a device of wonder and deception. Devices such as robots, for example, seem to answer to our romantic longing for mystery and wonder[69]-[71].

AI seems to evoke similar responses. Our expectations about the technology and our experiences are influenced by desire for romantic wonder. We expect AI to perform like humans or better than humans, for example. The hype and mystique around AI systems and their ability to better human performance is exemplified by IBM Watson and its participation in the US TV show Jeopardy [9]. This is not only a romantic performance itself, but also creates high expectations in people that the technology will be human-like in the future soon. And for instance, an AI-powered device with a voice interface raises the question if users like young children may be deceived into thinking that it is a person and whether this is ethically acceptable.

Magic is a way to explain the gap between the expectations about what data can do especially from itself without much human intervention and what it actually can deliver in the absence of the necessary expertise to work with data or to interpret what data science delivers. Magical expectations are personal and rarely articulated as clear goals and imply an openness to be delighted as well as the risk of being shocked or disappointed. A magical understanding of data science can be a source of this misalignment of expectation between engineers and data scientists.

This almost magical view of data and the expectation gap it creates is reflected in practice in the expectations non-data scientists (non-statisticians) and non-computer scientists have concerning what (big) data can do. In the context of engineering design and production, this means that engineers may perceive data, data analysis, and more generally data science as magic.

To some extent this also happens within data science: in the case of black box machine learning algorithms, the different steps in the layers of the neural net are uninformative because the emergent processes that create the results are unintelligible. Nevertheless, they can perform them and make recommendation results from it, in which they trust whether or not they should. In this sense what happens can be perceived as “magic”.

The person who creates the data analysis becomes a type of magical creature, a benign fairy or a more morally ambiguous sorcerer, who can affect the fate of the company or of individuals who are affected by the decision resulting from the analysis of the data in ways that neither the data scientists or the engineers can predict. In the case study, the young statistician was seen as somebody who could do something useful but was essentially harmless – like a fairy, even though the results of the analysis could have a profound effect for the manufacturing plants that were analysed.

Conversely the engineers also don’t know what their sorcerer or fairy requires to perform their magic. The data alone is clearly not enough, rather they need to be provided with contextual information. However even though the engineers might be aware that how the data is presented affects the results, they are not do understand in which way this affects the results. They are not aware of the socio-technical nature of data analysis.

Engineers usually pride themselves in their rational understanding of their products and the performance. In particular in manufacturing, which unlike engineering design have largely uncertainty, the manufacturing engineers have usually a good causal understanding of the relationship between input data and output results. This a priori understanding of the relationship between input and output is precisely what is missing in data analysis that is carried out for them.

Both lay persons/ user and data scientists, each in their own way, have a magical understanding of what the technology can do (lay people) and of how the technology comes to a decision or recommendation (experts). Despite ignorance, the “magic” means that they trust in the technology and its magicians. Given these magical properties, both groups may also develop the will to subject everything to data analysis, and therefore to gather, analyse, and sell data. Conversely, they feel empowered by data in situations where they do not understand the causal relationships and are therefore hoping that data will give them insight they don’t have as experts in the field. Both leads to what we call “data grabbing”.

5.2 Data grabbing

Currently engineering companies collect data when they can and where they can, because they see the potential benefit. Few engineers are aware of the bias that they are introducing through the nature of the data, the way it is represented, and how it has been captured. Unless engineers have a clear picture of how the data is later analysed, they might capture the data they *can* capture in a straightforward way which might not be what is later required. Even if suitable data is captured, analysing data can be time-consuming and therefore needs to be planned into the process. If the people who analyse the data and benefit from it are different from those who capture it, valuable context information is lost. More complex analyses are sometimes postponed in the hope that the time and money to carry them out becomes available, so that the later analysis is biased by simpler earlier analysis and the process become very error prone. This also means that the data is sometimes passed over to young people, or interns as in the case study, because they are less involved in the day to day running of the company.

The ability to generate data is seen as something positive, however in the long run it is not clear how the data will be curated in a way that people can use it at a later date. For example, at the moment much valuable data is kept in a variety of systems, forms and formats, so that handing over data to those who could analyse it can become a practical challenge. In the case study some data arrived very late, because it had to manually be moved out of a legacy system. The data was also provided without the contextual information necessary to interpret the data. The engineers in the team could explain some of the idiosyncrasies of the different plants, however these caveats might be lost as the data spreadsheet was passed through the organisation.

The provenance of the data is often not clear, so that the assumptions behind the data gathering are not known. The most "dangerous" situation is where data can be accessed for free, i.e. the data generated elsewhere, sometimes for another purpose. The concern with using data that was generated in a different process is that the models built on it won't transfer as well as hoped when applied to different situations, leading to disappointment. However, provided the need to do this is recognised, so long as one has some data from the relevant system it can be used to recalibrate the model generated on the big set of similar data. This approach works quite well and is known as transfer learning. Engineers see this data as a potential gold-mine, but need "someone" to extract these valuable insights. Such magicians may be statisticians or data scientists.

Where the engineers create and/or own the data sets themselves, the relation is somewhat different. Design engineers create models and digital experiments, where a massive amount of data is generated. In such situations, the source of data generation is the engineers themselves. It is then more straightforward to have a healthy expectation of what one can expect to find. However, the risk remains that this data is passed on to others, who don't understand its context of creation or the limits of the data collection.

6 Discussion: Lack of understanding on multiple levels

Overcoming this almost magical belief in the power of data and data analysis is largely a sociotechnical issue need to be addresses as a social rather than a technical issue. It won't be overcome by better computers or cleverer algorithms, but a greater understanding of what data analysis can provide and what support data scientists need to provide useful results. This understanding is required both in the context of the specific problem, but in understand the capability of data science in general. This mutual lack of knowledge leads to problems on multiple levels.

To summarise, at present we are confronted with multiple levels of misunderstanding which can affect the ability to act on the results of the data analysis.

- Data origin: Lack of understanding about where the data comes from
- Data model: Lack of understanding about what data model is used and what assumptions went into the modelling

- Data analysis: Lack of understanding about how the analysis is done
- Machine learning: Lack of understanding about the machine learning algorithm and, more generally, about what machine learning is

AI, big data and other forms of data analysis have become fashionable. People want to argue and persuade with facts and equate data with facts. They are lured by the measurable and numerical, because it gives the impression of objectivity. This is part of the "magic" of the data and the related technologies and science. The idea that computer could analyse situations quickly and thereby reduce human expert time, attracts companies to data analysis in addition to the kudos they receive from being able to say that they use these state-of-the-art techniques. The potential of AI is currently still limited by three interrelated factors: the methodological and epistemological misconceptions about the capabilities of AI, restriction of the social context in which AI and machine learning are embedded, and the technical limitations in the development and use of AI [72].

However the present attitude shows a serious lack of critical thinking. The multiple levels of misunderstanding are not considered, because of the belief that data equals fact. There is a lure of numbers, which appear more "objective" and seem easier and clearer compared to doing the bigger, messy work of finding causal relations. The "magic" aspect returns here: humans reason about causal relations and interpret, but if we can take humans out of the process, we have a more objective view or representation of the world. The idea is to let the data speak for themselves, rather than have the human intervene. It is the magic promise of an unmediated world, a Big Data Eden of transparency compared to the dirty engineering world. This also explains the current distrust for human experts who bring understanding and insights into the context to the analysis of problems. In a sense, the experts are "in the way"; they block the direct representation of the world that the data analysis seems to promise [72].

Practically these issues can be overcome in a specific situation, if is it is treated consciously as a sociotechnical problem. Engineers, as the agents who commission data analysis, need to pick data scientists with suitable skills that meet their needs. If they need a causal analysis, they probably want to pick a team which includes someone with a strong understanding of statistics; if they want patterns or predictions, they need an expert on machine learning models. In either case they need to work together and explain the problem context. Both groups need to be explicit about their assumptions, the engineers about the sources of the data and the data scientists about the assumptions that affect their models. Both groups need to critically question the analysis. Is the method of analysis appropriate? Is the data set sufficient to draw particular inferences? Does the data really support the conclusion? This can only happen in an environment where both groups are able to challenge each other and are equal partners. As data scientists are often hired to analyse a particular situation or to "prove" a particular hypothesis, the power lies often with the engineers as customers. An analysis of the power relationships in the data analysis projects would be a fruitful extension of this paper.

7 Conclusion

This paper mainly reports on a short placement; therefore, it could be argued that the engineers did not put a lot of effort into this placement as they had low expectations that something useful would come out. However, this particular group of engineers was very happy to have a statistician looking at their data and tried very hard to be supportive and organise the information that was required and explain the data as required. However, the statistician did not know what questions to ask and the engineers did not quite know what to explain. Similar tasks are often carried out by expensive management consultants, who typically are not trained as statisticians, when their findings could be used as the basis for fundamental decisions in the organisation.

The case study also illustrates the sociotechnical nature of data analysis and points to the need for engineers and data scientists to communicate effectively when setting up the data models. The

creation of the data models needs to be seen as a design process in its own right, where an information artefact is being designed, which can have a long-term effect on the system it is modelling. If engineers want to benefit from data analysis, they also need to take ownership of the generation of the data models.

The engineers were largely unaware that the modelling decisions had a profound effect on the outcome and therefore might bias the decisions based on the data. This points to the ethical implications of making decisions based on data, in particular repurposed data, and to the vulnerability of those who become the subjects of data analysis.

This paper concludes, firstly, that engineers and data scientists have fundamentally different views of what constitutes a model; and the creation of the data model required to carry out these analyses lies outside the core expertise of either group. However, the engineers can do three basic steps either in preparation for meeting the data scientists or in collaboration with them: decide on the purpose of the model and the basic relationships they are interested in, and if they have data from multiple sources, decide on the basic vocabulary and logic of structuring the data, as this will have a fundamental effect on the actions taken based on the data analysis. Design ontologies address similar concerns; however in many situations the rigour and effort of an ontology might not be required, instead teams can take pragmatic decisions what works for them.

Secondly, the paper also points to the limitations of data science and what engineers and data scientists do and do not know, as well as the procedures they use to collaborate with each other. It is important that the magic of big data, machine learning, AI, etc. does not blind the parties involved – engineers, data scientists, management, politicians, general public – to these limitations and challenges, and that all actors involved are aware of the possibility that they might have misleading expectations about what data science can do. Further analysing the epistemic, political, and technological sources of ethical concern and focusing on the issues of bias, responsibility, and implications for people providing data, is work we are planning to do in the future. It is necessary to create more awareness of the potential problems and – going beyond awareness – to integrate ethics in the processes and organizations involved.

Acknowledgements

- *Intentionally left out in submission phase but will be entered.*

8 References

[1] Yeats, W.B. 1903, The Land of Heart's Desire, The Project Gutenberg EBook, http://www.gutenberg.org/files/15153/15153-h/15153-h.htm
[2] Digital Transformation Scoreboards, Available at https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/Digital%20Transformation%20Scoreboard%202018_0.pdf
[3] De Weck, O.L., Roos, D. and Magee, C.L., 2011. "Engineering systems: Meeting human needs in a complex technological world", Mit Press.
[4] Womack, J. P., Jones, D. T., & Roos, D. , 2007. "The machine that changed the world: The story of lean production--Toyota's secret weapon in the global car wars that is now revolutionizing world industry". Simon and Schuster.
[5] Schroeder, R. G., Linderman, K., Liedtke, C., & Choo, A. S., 2008, "Six Sigma: Definition and underlying theory". Journal of operations Management, 26(4), 536-554.
[6] Ustundag, A. and Cevikcan, E., 2018, Industry 4.0: Managing The Digital Transformation, Springer Series in Advanced Manufacturing, https://doi.org/10.1007/978-3-319-57870-5_9
[7] Varshney LR , 2016. Fundamental Limits of Data Analytics in Sociotechnical Systems. Front. ICT 3:2. doi: 10.3389/fict.2016.00002 18
[8] EU, 2020, "Shaping Europe's Digital Future", Luxembourg: Publications Office of the European Union, doi:10.2759/091014

[9] Elish, M. C. and Boyd, D. 2018. “Situating methods in the magic of Big Data and AI,” Communication monographs, 85(1), 57-80.
[10] Goode, L., 2018, “Life, but not as we know it: AI and the popular imagination” Culture Unbound: Journal of Current Cultural Research, 10(2), 185-207
[11] Gartner Inc., 2009, “Gartner at a glance”.
[12] Fenn, J. and Raskino, M., 2008, “Mastering the hype cycle: how to choose the right innovation at the right time,” Harvard Business Press
[13] Dosi, G., 1982, “Technological paradigms and technological trajectories,” Research policy, 1, pp. 147-162
[14] Steinert, M., and Leifer, L., 2010, “Scrutinizing Gartner's hype cycle approach”, Proceedings of Picmet 2010 Technology Management for Global Economic Growth, pp. 1-13. IEEE
[15] Domingos, P., 2012, “A Few Useful Things to Know about Machine Learning,” Communications of the ACM, 55 (10), pp.78-87.
[16] Eckert, C., Isaksson, O., Hallstedt, S., Malmqvist, J., Rönnbäck, A.Ö. and Panarotto, M., 2019, “Industry Trends to 2040”, Proceedings of the Design Society: International Conference on Engineering Design , 1(1), pp. 2121-2128. Cambridge University Press.
[17] Gartner, 2019, https://www.gartner.com/smarterwithgartner/gartner-top-10-data-analytics-trends/
[18] Bucciarelli, L. L., 1994, “Designing engineers”, MIT press, Cambridge.
[19] Star, S. L. (2010), “This is not a boundary object: Reflections on the origin of a concept”. Science, Technology & Human Values, 35(5), 601-617.
[20] Stacey, M. and Eckert, C., 2003. “Against ambiguity”. Computer Supported Cooperative Work (CSCW), 12(2), pp.153-183.
[21] Tushman, M., Tushman, M.L. and O'Reilly, C.A., 2002. Winning through innovation: A practical guide to leading organizational change and renewal. Harvard Business Press.
[22] Liker, J. K., 1997, “Becoming lean: Inside stories of US manufacturers”. CRC Press.
[23] Pande, P.S. and Holpp, L., 2001.” What is six sigma?”. McGraw-Hill Professional.
[24] Söderberg, R., Wärmefjord, K., Carlson, J.S. and Lindkvist, L., 2017. “Toward a Digital Twin for real-time geometry assurance in individualized production”. CIRP Annals, 66(1), pp.137-140
[25] Davis, R., 2015, “Industry 4.0. Digitalisation for productivity and growth”, European Parliament Briefing September 2015, EPRS | European Parliamentary Research Service
[26] European Union, 2018, “Digital Transformation Scoreboard 2018”, Luxembourg: Publications Office of the European Union,2018, DOI10.2826/821639.
[27] Soban, D., Thornhill, D., Salunkhe, S., & Long, A. (2016), “Visual Analytics as an enabler for manufacturing process decision-making”. Procedia Cirp, 56, 209-214
[28] Benson, P.R., 2019, “ISO 8000 Quality Data Principles”, An ECCMA White Paper, ECCMA
[29] Moldoveanu, M., 2015. “Unpacking the ‘Big Data’skill set”, European Business Review.
[30] https://medium.com/@ODSC/6-reasons-why-data-science-projects-fail-6240bf9326f6
[31] https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/
[32] Rose, D. (2016). Data science: Create teams that ask the right questions and deliver real value. Apress.
[33] https://medium.com/@ODSC/6-reasons-why-data-science-projects-fail-6240bf9326f6
[34] https://www.forbes.com/sites/bernardmarr/2018/03/12/forget-data-scientists-and-hire-a-data-translator-instead/#7e1acae2848a
[35] https://towardsdatascience.com/building-and-managing-data-science-teams-77ba4f43bc58
[36] https://www.forbes.com/sites/cognitiveworld/2019/09/11/the-full-stack-data-scientist-myth-unicorn-or-new-normal/#6fb90b132c60
[37] https://www.infoworld.com/article/3429185/stop-searching-for-that-data-science-unicorn.html
[38] Baškarada, S., & Koronios, A. (2017). Unicorn data scientist: the rarest of breeds. Program.
[39] Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. Briefings in Bioinformatics, 20(2), 398-404.
[40] Saltz, J. S., & Grady, N. W. (2017, December). The ambiguity of data science team roles and the need for a data science workforce framework. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 2355-2361). IEEE.
[41] https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them
[42] https://www.cio.com/article/2436638/project-management-what-do-business-analysts-actually-do-for-software-implementation-projects.html
[43] https://becominghuman.ai/what-is-the-difference-between-data-scientist-and-ml-engineer-703728fa21a0
[44] Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2019). Understanding data search as a socio-technical practice. Journal of Information Science, 0165551519837182.
[45] Gal, I. (2002), “Adults' statistical literacy: Meanings, components, responsibilities”, International statistical review, 70(1), 1-25.

[46] Wallman, K. K., 1993, “Enhancing statistical literacy: Enriching our society”, J. of the American Statistical Association, 88(421), 1-8.
[47] Watson, J. M., 2006, “Issues for statistical literacy in the middle school”, In ICOTS-7 Conference Proceedings. IASE, Salvador (CD-Rom).
[48] Kaplan, J., & Rogness, N. (2018). Increasing Statistical Literacy by Exploiting Lexical Ambiguity of Technical Terms. Numeracy: Advancing Education in Quantitative Literacy, 11(1).
[49] Watson, J. M., & Kelly, B. A. (2008). Sample, random and variation: The vocabulary of statistical literacy. International Journal of Science and Mathematics Education, 6(4), 741-767.
[50] https://ubc-mds.github.io/resources_pages/terminology/
[51] https://analyse-it.com/docs/user-guide/fit-model/fit-model
[52] https://developers.google.com/machine-learning/crash-course/framing/ml-terminology
[53] Malinowski, Bronisław. 1929. The Sexual Life of Savages in North-Western Melanesia. New York: Halcyon House.
[54] Mead, Margaret. 1928. Coming of age in Samoa: A Psychological Study of Primitive Youth for Western Civilisation. New York: William Morrow & Co.
[55] Hickey, S. and Mohan, G. (2004, eds.): Participation: From Tyranny to Transformation? Exploring New Approaches to Participation in Development (London) Zed Books
[56] Clark, Andrew; Holland, Caroline; Katz, Jeanne and Peace, Sheila (2009). Learning to see: lessons from a participatory observation research project in public spaces. International Journal of Social Research Methodology, 12(4) pp. 345–30.
[57] Agar, M. (1980) The Professional Stranger: An Informal Introduction to Ethnography. Academic Press, London.
[58] Atkinson, P. and Hammersley, M. (1994). Ethnography and Participant Observation. In Handbook of Qualitative Research eds N.K. Denzin and Y.S. Lincoln. Sage, Thousand Oaks, CA.
[59] O'Neil, C., 2016, “Weapons of math destruction: How big data increases inequality and threatens democracy”, Broadway Books.
[60] Bailer-Jones, D.M., 2009, “Scientific Models in Philosophy of Science”. Pittsburgh, PA: University of Pittsburgh Press.
[61] Giere, R.N., 1988, “Explaining Science: A Cognitive Approach”. Chicago: University of Chicago Press.
[62] Fahrmeir, L., Kneib, T., Lang, S and Marx, B., 2007, “Regression”Springer-Verlag Berlin Heidelberg.
[63] Eckert, C., & Hillerbrand, R. (2018). Models in engineering design: generative and epistemic function of product models. In Advancements in the Philosophy of Design (pp. 219-242). Springer, Cham.
[64] Vajna, S., Weber, C., Zeman, K., Hehenberger, P., Gerhard, D., and Wartzack, S., 2018,. “CAx für Ingenieure” (pp. 515-547). Springer Vieweg, Berlin, Heidelberg.
[65] Bowker, G., and Star, L., 2000, “Sorting things out: Classification and its consequences.” Cambridge, MA: MIT Press.
[66] Rahm, E. and Do, H. H, 2000, “Data cleaning: Problems and current approaches” IEEE Data Eng. Bull., 23(4), 3-13
[67] Clarke, A. C.,1973, “Profiles of the future: An enquiry into the limits of the possible”. New York, NY: Harper & Row
[68] Gell, Al., 1994, “The technology of enchantment and the enchantment of technology”. In J. Coote (Ed.), Anthropology, Art, and Aesthetics. Oxford: Clarendon Press.
[69] Coeckelbergh, M., 2017, “New romantic cyborgs: Romanticism, information technology, and the end of the machine.” Cambridge, MA/London: The MIT Press.
[70] Coeckelbergh, M., 2018, ‘How to describe and evaluate “deception” phenomena: Recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn‘, in: Ethics and Information Technology 20(2): 71–85, Springer
[71] Coeckelbergh, M., 2019, “Moved by Machines”. New York: Routledge.
[72] Hagendorff, T., Wezel, K., 2019,”15 challenges for AI: or what AI (currently) can’t do”. AI & Soc , https://doi.org/10.1007/s00146-019-00886-y