[1.]Robert BERNERSTÄTTER

DATA MATURITY FOR SMART FACTORY APPLICATIONS – AN ASSESSMENT MODEL

[1.]Montanuniversität Leoben, Chair of Economic and Business Management, Leoben, AUSTRIA

Abstract: Due to the Industry 4.0 initiative, data analytics gained importance even in the industrial environment in the past years. New and affordable technologies like sensors, RFID and wireless connectivity enable companies to collect huge amounts of data. In combination with already existing data a wide range of data is available for finding improvement potential with data analytics applications. Nevertheless the experience gained through different projects in the past years shows a lack of maturity in data and data systems suitable for automated analytical approaches. Because of that a model to assess the fitness of systems for data analytics is in development. A key factor, the assessment categories, will be introduced in this paper.
Keywords: data quality, smart factory, big data, big data analytics, industry 4.0

INTRODUCTION

Data is considered to be one of the most important resources of the 21st century. It is a resource which is created through internal processes as well as the utilization of products and services by the customer. The Smart Factory is a focal point of Industry 4.0. It generates and collects data on a large scale. [1] Figure 1 gives a visual representation which illustrates the importance of data, as universal data utilization is the foundation of the Smart Factory. This data can be further processed to knowledge which makes the Smart Factory smart at last. Sources for the data collection can be products, customers, services provided in addition to the products and the production process itself. The production process generates data via sensors and machines as well as administrative actions.

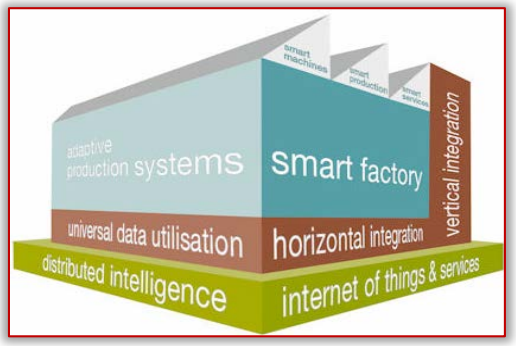

Figure 1 – Smart Factory [2]

As the end-products get ever more refined the resources used during the manufacturing process are subjected to a more and more rigorous quality process. Since data is a key resource in the Smart Factory and the knowledge generated is only as good as the data used, the same quality standards must apply there. Nevertheless, previous experiences show that most of the time data quality is not suitable for knowledge generation through new data analysis techniques. [1] This statement can be confirmed due to the work in at least twelve different projects. Quite often data, the data system and the processes concerning data generation were not mature enough.

This article introduces the first steps of an assessment model for data maturity explicitly aimed on Smart Factory applications like data analysis for the purpose of prognosis, suggestion systems or weak point analysis for sustainable process improvement.

BIG DATA AND DATA QUALITY

One of the most common used terms in combination with Smart Factory is Big Data. It was first mentioned in 1998 by John Mashey and the first scientific paper was published by Francis X. Diebolt in 2000. In these two publications, 'big' was only a lot of data - sheer volume. But volume was not the only key figure in Big Data, since it was always an issue. Gartner, former META Group, defined 2001 three dimensions which define Big Data. These 3 Vs, as the dimensions are known ever since, are still used to define the core essentials of Big Data. They stand for volume, velocity and variety. [3]

Volume

As mentioned above, this dimension represents the amount of data as a whole. Considering the fact that the information content in the world increases nearly in an exponential rate one can hardly try to grasp the full extent. A simple machine with a standard complimentary of sensors produces, depending on the sample rate, several gigabytes in a short period of time.

Velocity

Arguably the dimension that distinguishes the sheer volume from real Big Data. The speed by which data is produced is also a cause for the volume. It is a dimension imposed by the information age of the 21st century. Therefore, it is fitting, that the term Big Data was created at the beginning of the century. The sensors mentioned above might produce data every millisecond. Smartphones, GPS-trackers in vehicles and so forth send data in real time and nearly every second. Velocity is certainly a dimension that challenges every classical data management system.

Variety

As a dimension, which is mentioned above indirectly variety represents the fact that data is coming from a great range of different sources. This is also caused by the modern

information age, which allows for the instantaneous connection of different data sources. These sources can be machines, ERP Systems, products or service platforms. The variety of the sources also contribute to the volume of the generated data.

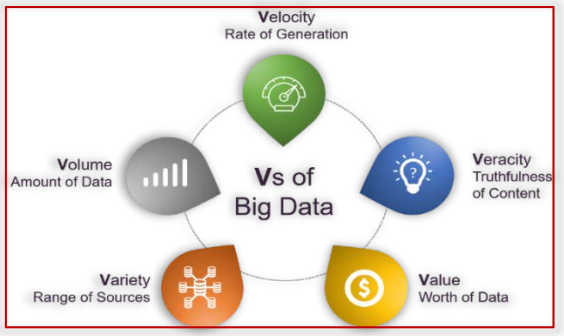

Figure 2 – Dimensions of Big Data

Although these three dimensions define Big Data and are sufficient for business intelligence applications, recent experiences made during projects in the manufacturing industry show that additional dimensions need to be considered when working with Big Data in a Smart Factory context. [4] This can be confirmed by on own experience gathered in various projects. Figure 2 gives an overview of the classical dimensions which define Big Data and the dimensions which need further consideration in combination with Smart Factory applications.

One of the most important dimensions concerns the quality of data, expressed in the dimension 'veracity'. And lately the contribution of data stated by the dimension 'value'.

Veracity

It describes the fact that data is in doubt. The fact might be triggered by inconsistencies, incompleteness or general ambiguity. The root cause might lie in the variety – one of the originally three Vs – of the data sources. If data can't be trusted, the value it contributes might be in doubt.

Value

Generally, the value is defined monetary for example what business models can be associated to the data. The different services or the internet of service (figure 1) come to mind, which are built on the compilation of knowledge out of data to provide a virtual product for which the customer is willing to pay. In the scope of this work, value is defined differently. It is considered, what value it might bring to classical data analytics applications. The better the quality and therefore the dimension veracity, the better is the result of a classification, a prognosis or a root cause analysis. That might bring a contribution to the continuous process improvement by eliminating weak points in the process.

ASSESSMENT MODEL

The new technologies developed in the past ten years contribute to a near exponential growth of the data stream and amount. This trend will intensify itself in the years to follow.

Yet a lot of companies, especially SMEs are overwhelmed by the possibilities and feel pressured to start Industry 4.0 initiatives on their own.[4] Data analytics projects are often the first step to gain experience with Smart Factory applications. Since data is produced in every company it is deemed to be the easiest way. Unfortunately, a lot of data found when starting with a data analytics project is not suitable for analytics projects right away. This problem will intensify in the future, because companies plan to invest in data recording. To get a quick overview of the data situation in a company prior starting the data analytics project, an assessment has been developed. In the following pages the categories of the assessment will be introduced.

Future Developments

The variety dimension and therefore the veracity dimension are getting more important in the future. A BARC study which was conducted in 2014, shows the intention of companies to further invest in data acquisition and gathering data from a growing number of sources. As can be seen in Figure 3, a significant increase will be in sources concerning Smart Factory applications, like machine data, sensor data or event streams from processes. [4]

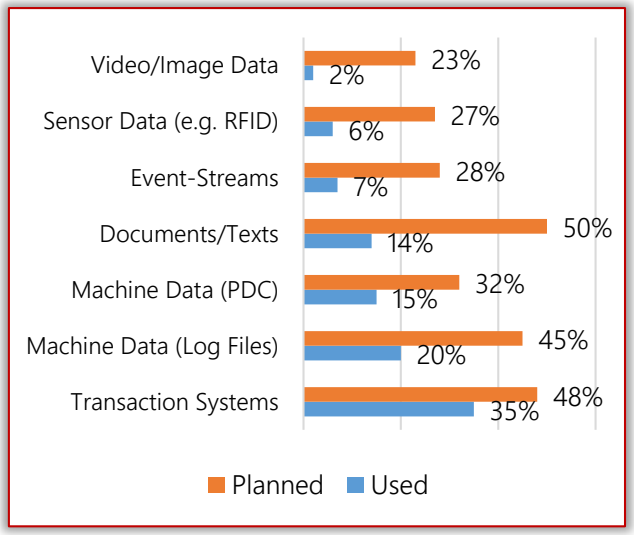

Figure 3 - Which data is used or is planned to be used to do big data analytics [4]

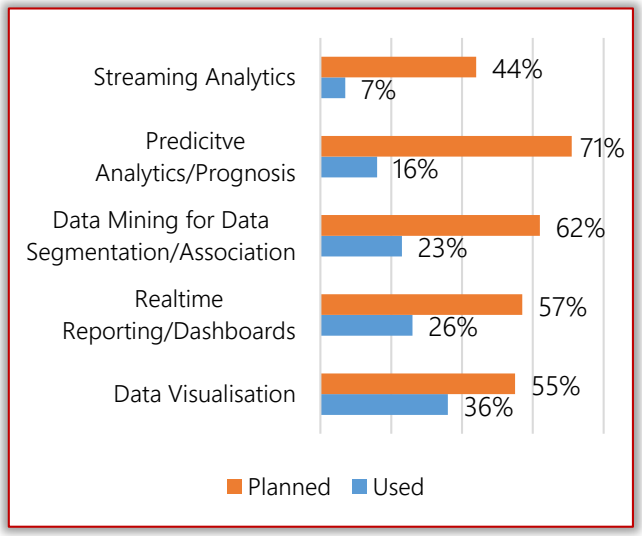

Figure 4 - Which big data analysis is currently performed and which is planned in the near future [4]

Figure 4 provides a summary of planned analytics applications. When looking at future applications of big data analytics, predictive analytics and prognosis, data segmentation and association or reporting and visualization of data are among the most mentioned applications. Especially the first two are prime applications of data analytics concerning Industry 4.0 and Smart Factory applications. The study also mentions the importance of data quality to harness the potential of the data and the possibilities of analysis applications. [4]

Assessment Model

Numerous projects offered the opportunity to verify the numbers above. The project portfolio includes, but is not limited to, production management. Here cluster algorithms where used to automatically form product families to use them in the improvement of planning robustness. Another project was conducted in the field of asset management with the goal to implement predictive maintenance in a steel mill. Upcoming projects, where the model will be further developed, will deal with an automated damage or root cause analysis to find problems concerning a welding robot.

Companies want to use the potentials mentioned in Figure 4 and they are using more and more data from different sources. The great variety brought problems with the data quality which delayed a project or changed the focus and scope of a project. Since many assessment models and methodologies have a different scope than Smart Factory applications, the development of an own model has been started. It specifically aims to assess categories, relevant to applications mentioned in figure 4 and it takes the fact of different sources and formats from figure 3 in account. [4]

The proposed model is still in development and being evaluated but has been tested as well in projects. When applied, it should help to estimate the effort in data acquisition and selection phase of the data mining process. At the current state the models allows a user, who has already gained some experience in data science, to estimate if a data mining project will lead to success. If the success might be in question the assessment categories, which a depicted in figure 5, help in improving the data quality by showing the current state of the data at hand. The operationalization is another point to consider. Often data analytics projects prepare data with considerable effort for the prototype phase. But the model developed can't be applied, because the data at hand can't be used without the recurring preparation phase.

The maturity or the necessary data quality depends, like the quality of each productions factor, on the application. Data analytics methods can be divided in various ways. In the case at hand we will use clustering, classification, prognosis and association or root cause analysis. Clustering is used to find prior unknown similarities in data while classification is used to learn given structures in data. Prognostic methods try to predict future outcomes based on past data and association analysis aims to find common occurrences of data values. [5]

The assessment model itself is based on different consecutive categories. The first three evaluate the data management and have relevance for the company in general. [6] The latter three rate the measures more specifically for data analytics applications in the Smart Factory.

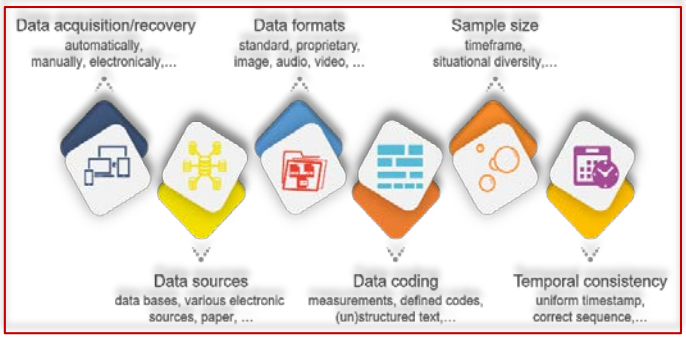

Figure 5 - Assessment categories [7]

Data acquisition/recovery

This category assesses the general organization of the data recovery process. Data can be recorded automatically or manually. An example for an automated recording of data would be sensors which send the measurements directly in a data management system or the automated recording of incidents or other process parameters.

The manual recovery of data needs a far more differentiated point of view. Generally speaking it means that data is recorded or entered by a user. A type of recording could be tablets, smartphones or other mobile solutions. The recording is done on site and the data is fed to the data warehouse in real time or at least on the first opportunity for synchronization. [7]

Data sources

Data sources can be differentiated in electronically and non-electronically sources like paper notes. Electronically sources are always preferred over paper sources. The most preferable source are databases. They allow for a systematic input and recovery of data. A lot of them apply routines to ensure redundancy free entries and therefore constitute a single point of truth. As mentioned in section 2 the variety of data sources is one defining aspect of Big Data. Data base systems increase the possibility to unambiguously merge data from different tables and sources by clear defined primary-foreign key relations. Closely related to the category data sources is the category data formats. [7]

Data formats

Data formats or more specific file formats are manifold. Well known standard formats are markup languages like HTML or XML. One big advantage is the meta information, that can be encoded and is machine readable as well. [8] Others are different spread sheet formats like Microsoft Excel or HDF (Hierarchical Data Format). HDF is very well suited to store huge data volumes in one- or multidimensional tables. The list could be continued with CSV files, an often used format to exchange structured text, or image, audio or video files.

A key fact about data formats is the compatibility. Different data formats should be convertible to a common one in order

to form a consistent data base to work with. Therefore proprietary formats are the least favorable, because they often aren't compatible and convertible to others. [7]
These three categories are important not just for data analytics applications, but for all data based applications in a company. The following three categories are more specific for big data analytics applications. They should give a better understanding which analysis methods are possible and what the results may be. [6]

Data coding

This category takes the semantics into account. Major fields are unstructured text, defined measurements and KPIs or standardized codes and messages. Despite modern text mining routines, unstructured text the way it was found in different projects, is still a challenge. Problems occurred through spelling errors, colloquial language or dialectic entries or simple phony entries with no sense at all. So it is important to take the entries to the next level and establish a hierarchy of codes, which represent the most common occurrences and therefore entries. Defined and standardized parameters are the most preferable for automated analysis. In the end they only differ in the scale of measurement and application. Measurements and KPIs are mostly metric scaled and are used as attributes. Standardized codes and messages are often nominal scaled and are used as attributes as well or for supervised learning methods as labels.
If the data fulfills most of the positive criteria - like automated recording, standard formats like CSV or HDF or a structured coding scheme - of the categories so far, a successful application of some data analytics methods is very likely. That might be a clustering algorithm for the segmentation of the product portfolio into product families. In lots of cases a classification algorithm might also be suitable, for example to automatically define quality classes but to be sure the next category might be necessary to have a look at. [5,7]

Sample size

For supervised learning algorithms, this is one of the most important categories to be assessed. The success of a reliable accuracy depends of course on one hand to use the right context that should be learned and on the other hand on the sample size of things and occurrences that should be learned. It is imperative that the necessary data is recorded reliable, which should be ensured by a high maturity in the prior categories. To name a definite number of samples is difficult. But of course every combination that should be learned and of course each label value should occur several times at least. If there are more label values, each should represent at least 10% of the entire sample size. The time horizon should encompass at least 2-3 years to account for possible seasonal effects. [6, 10]

Temporal consistency

It is a category important for supervised learning with time critical dependencies as well as for prognosis purposes or association and root cause analysis. When comparing time series from different sources it is of the utmost importance to have a common timestamp. So it is possible to combine them in a common data structure and find correlations and causations. In many cases when making a subsequent measurement, for example in product quality assurance, the time of the measurement is recorded and not the time of the production. So it is not possible to make a connection between a quality feature of a product and for example a machine failure. The two features, machine failure and product quality feature, are temporally not consistent in a way it might be necessary for a data analytics application. [7]
Table 1 sums up the categories to be assessed and gives an overview about the two possible extremes which might be found in each category. The graduations between them a numerous and will be developed to a capability maturity model in future research. The common case is a based on an evaluation of the data provided in the different projects of the Chair of Economic- and Business Management.

Table 1 – Assessment categories and possible characteristics

Category	Worst Case	Best Case	Common Case
Data acquisition/ recovery	Manually on paper with time delay	Automated with sensors with real time transfer	Manually but in electronical form
Data sources	Paper	Cloud based data base	Spreadsheets and ERP systems
Data format	(Several) proprietary formats	(one) Standard e.g. markup languages	A mixture of basically everything
Data coding	Unstructured text	Standardized codes or metric scaled values	Use case and data source dependent
Sample size	(Fragmented) test month with incomplete labels	Several years and entries to each label	Features enough, labels too little
Temporal consistency	No timestamps at all	Common timestamp, aliened to the cause of the analysis	Timestamps are available but often unreliable

One essential project where the categories were applied is called Maintenance 4.0 (Instandhaltung 4.0). It is publicly funded by the Austria Research Promotion Agency (FFG). The results of the application of the assessment categories helped to identify potentials in the sample size of the maintenance data and the acquisition of machine related data.
The value of a data assessment is manifold. A quick assessment gives an overview of the situation of the data quality. It helps therefore to improve the time and cost estimations for analytics projects. Knowing the maturity in the six categories makes it easier to find potentials for future improvements. The same way it helps to lay down tasks to reach higher levels in the assessment categories toward a state where automated real time analytics is possible. That might very well be the most important benefit, because if one does not improve the maturity long-term it is never

compatible with an automated analysis process. Every manual data preparation and cleansing step of a prototypical analytics project needs to be automated or better removed in order to operationalize the analytics process later on. This will also save a lot of time and therefore money. As can be seen in figure 6, up to 85% of the time in an analytics project is spend during the first two phases. The understanding encompasses the business or domain understanding, and the data understanding. The modelling, hence application of algorithms, and evaluation of the results take only 15%. The numbers are rounded averages of projects at the Chair of Economic- and Business Management. They are quite consistent to those found in different literature [11].

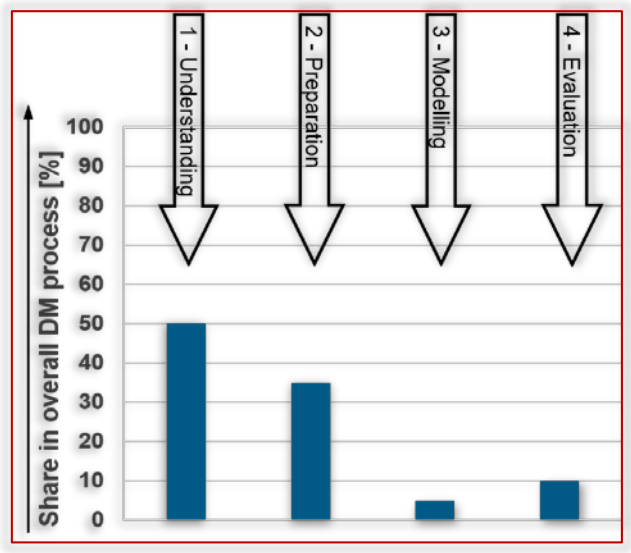

Figure 6 – Average time distribution in analytics projects

CONCLUSIONS

Data quality is an important perspective in every business situation and has been studied in BI applications during the last decade. Due to the chances initiated by new information technologies the factories change to Smart Factories where data becomes an ever more important resource. Therefore data quality is getting an integral part in the manufacturing industry as well. To fully embrace the possibilities given by Smart Factory and Big Data analytics it is important that data reaches a certain quality level or gain a certain maturity not just by the property of the data itself but also by the properties of the hard- and software systems.

This article discusses the nucleus of each assessment model, which are the assessment categories. They are not defined in general, but specific for Smart Factory and big data analytics applications in particular. The references and experiences were gathered in several projects with companies with the goal to use Big Data analytics on existing data to find valuable information for Smart Factory applications. In these projects the assessment categories where developed. They are the core of the further development of the model which should provide an easy and quick estimate about the situation of the data in a company concerning data analytics projects even before the projects gets fully started.

Note

This paper is based on the paper presented at 9th International Conference "Management of Technology – Step to Sustainable Production" – MOTSP 2017, organized by Faculty of Mechanical Engineering and Naval Architecture of the University of Zagreb, CROATIA and University North, Varaždin, CROATIA, in Dubrovnik, CROATIA, 5 – 7 April 2017.

References

[1] Brühl, V.: Wirtschaft des 21. Jahrhunderts – Herausforderungen in der Hightech-Ökonomie, Springer Gabler, Wiesbaden, 2015.
[2] ENGEL Austria Gmbh: inject 4.0 – die Antwort auf smart factory, Pressemitteilung, Schwertberg, October 2015, 5.
[3] Dean, J.: Big Data, Data Mining, and Machine Learning, John Wiley and Sons Inc., New Jersey, 2014.
[4] Bange, C., Janoschek, K.: Big Data Analytics – Auf dem Weg zur datengetriebenen Wirtschaft, BARC GmbH, Würzburg, 2014.
[5] Cleve, J.: Data mining, De Gryter Oldenbourg, München, 2014.
[6] Dippold, R.: Unternehmensweites Datenmanagement - von der Datenbankadministration bis zum Informationsmanagement, Vieweg, Braunschweig, 2005.
[7] Kinz, A., Bernerstätter, R.: Instandhaltungsoptimierung mittels Lean Smart Maintenance – Einführung des Lean Smart Maintenance Ansatzes, In: Lean Smart Maintenance, TÜV Media, 2016, 61-100.
[8] Born, G.: Dateiformate – die Referenz, Galileo Press, Bonn, 2001.
[9] Matzer, M., Lohse, H.: Dateiformate: ODF, DOCX, PSD, SMIL, WAV & Co., entwickler.press, Frankfurt am Main, 2007.
[10] Kuhn, M., Johnson, K.: Applied predictive modeling, Springer, New York, 2013.
[11] Sharafi, A.: Knowledge Discovery in Databases – Eine Analyse des Änderungsmanagements in der Produktentwicklung, Springer Gabler, Wiesbaden, 2013.

ISSN: 2067-3809