Big data analytics for proactive industrial decision support

Approaches and first experiences in the FEE Project

Big data technologies offer new opportunities for analyzing historical data generated by process plants. The development of new types of operator support systems (OSS) which help the plant operators during operations and in dealing with critical situations is one of these possibilities. The project FEE has the objective to develop such support functions based on big data analytics of historical plant data. In this contribution, we share our first insights and lessons learned in the development of big data applications and outline the approaches and tools that we developed in the course of the project.

KEYWORDS big data / data analytics / decision support

Big Data Analytics zur proaktiven industriellen Entscheidungsunterstützung: Lösungsansätze und erste Erfahrungen des Projekts FEE

Big-Data-Technologien eröffnen neue Optionen zur Analyse historischer Anlagendaten in der Prozessindustrie. Eine Möglichkeit ist die Entwicklung neuer Operator-Unterstützungssysteme (OSS), die dem Anlagenfahrer im Betrieb und bei der Behandlung kritischer Situationen assistieren. Das Projekt FEE hat das Ziel, derartige Unterstützungsfunktionen basierend auf Big Data Analytics unter Nutzung historischer Anlagendaten zu entwickeln. In diesem Beitrag teilen wir erste Erfahrungen und Lessons Learned hinsichtlich der Entwicklung von Big-Data-Applikationen. Weiterhin stellen wir im Projektentwickelte Lösungsansätze und Werkzeuge dar.

SCHLAGWÖRTER Big Data / Data Analytics / Entscheidungsunterstützung

MARTIN ATZMUELLER, UNIVERSITY OF KASSEL
BENJAMIN KLÖPPER, ABB CORPORATE RESEARCH CENTER GERMANY
HASSAN AL MAWLA, ABB CORPORATE RESEARCH CENTER GERMANY
BENJAMIN JÄSCHKE, UNIVERSITY OF KASSEL
MARTIN HOLLENDER, ABB CORPORATE RESEARCH CENTER GERMANY
MARKUS GRAUBE, TECHNISCHE UNIVERSITÄT DRESDEN
DAVID ARNU, RAPIDMINER
ANDREAS SCHMIDT, UNIVERSITY OF KASSEL
SEBASTIAN HEINZE, TECHNISCHE UNIVERSITÄT DRESDEN
LUKAS SCHORER, ABB CORPORATE RESEARCH CENTER GERMANY
ANDREAS KROLL, UNIVERSITY OF KASSEL
GERD STUMME, UNIVERSITY OF KASSEL
LEON URBAS, TECHNISCHE UNIVERSITÄT DRESDEN

The high degree of automation in the processing industries allows economical operations even in countries with high labour costs, such as Germany. However, it reduces the experience of the operators regarding the process dynamics. But know-how about the production process is crucial, especially if the production facility reaches an unexpected operation mode such as a critical situation. For example, such critical situations can lead to information overload (due to „alarm flood“), which can be overwhelming for the plant operator [1,2]. If control is lost it can result in serious damage to assets and costly downtime in the production process. Furthermore, this is not only expensive for the operating company but can also be a threat for humans and the environment. Therefore, it is important to support the plant operator in a critical situation with an assistant system using real-time analytics and ad-hoc decision support. The analysis of the historical data collected in process plants is an opportunity to develop such operator support systems (OSS).

A typical process plant like a paper mill, a hot-rolling mill or a petro-chemical plant, for example, generates a large amount of documentation and data throughout its entire life-cycle: I/O and tag lists, piping and instrumentation diagrams (P&ID), control logic, alarm configurations (during planning and commissioning), measurement values, alarm and event logs, shift books, laboratory results (during operation), maintenance notification, repair and inspection reports (during maintenance).

Analytics are already performed today [1,2,3] to analyze and improve the operation of plants. However, this is usually limited to data from single data sources and does not consider tight semantic integration. In contrast, an integration of all different types of data within process plants leads directly into the area of the so-called big data [4]. For instance, a refinery produces more than 300 GB measured values per year, from more than 60,000 sensors with sampling intervals between 1 and 60 seconds. Data may be structured (sensor readings, database tables), semi-structured (alarm and event logs) or unstructured (shift books, operation manuals). Data is often stored for ten years or more. The availability of such historical plant data makes big data analytics and machine learning interesting, also in the process industries. Overall, this application domain features the notorious big data criteria: high volume, high velocity and high variety. The development of such operator support functionality is the aim of the FEE project (http://fee-projekt.de), which is described in the following section.

1. THE FEE PROJECT

The objective of the BMBF-funded cooperative research project “Early detection and decision support for critical situations in production environments”(FEE) is the analysis of large and heterogeneous data volumes stored in petro-chemical production plants in a big data analytics platform with the aim of supporting plant operators. Big data technologies will enable data-driven OSS to warn the operator at an early stage about unexpected and uncommon situations and support an ad-hoc analysis, as well as the development of intervention strategies. The main goal is to enable the operators to act proactively and to overcome today’s reactive fashion of operating chemical plants. An early detection of critical situations will permit plant operators to analyse the situation and carry out corrective actions before a problem or deviation results in a major problem. Online what-if simulation will give plant operators the opportunity to simulate consequences of intervention strategies. Online process analysis will provide plant operators with information about major process couplings, dominant time constants and process gains. Together, these assistance functions will allow plant operators to develop a more appropriate response to process upsets.

The consortium of the FEE project includes application partners from the chemical industry. They provide use cases for the project and background

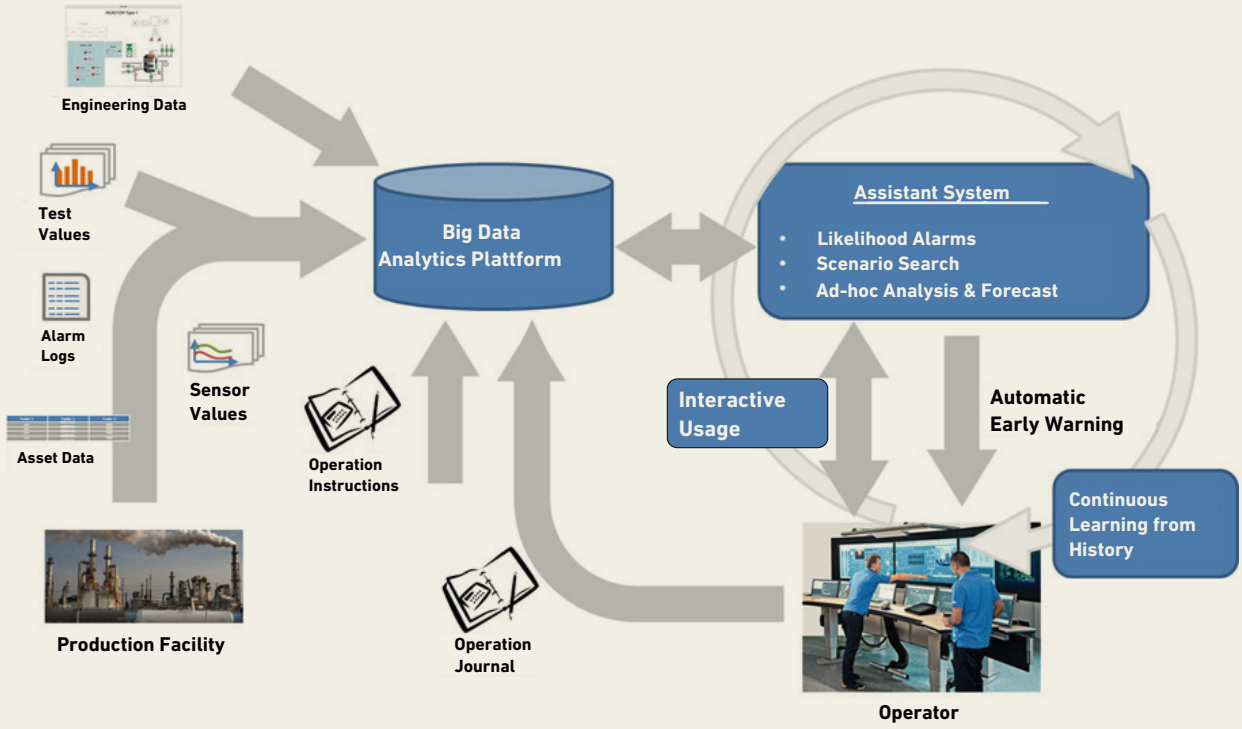

FIGURE 1: The big data analytics platform consolidates and integrates heterogeneous mass data collected over many years. The assistant system is built on top of the platform and utilizes analytical methods which automatically generate an early warning or even recommend how to handle a specific situation.

knowledge about the production process which is important for designing analytical methods. The data enabling the use cases was collected in petrochemical plants over many years from a variety of sources. The heterogeneous data is consolidated and integrated by the big data analytics platform (see Figure 1).

Petro-chemical production plants (see Figure 2 for an illustration) have some specific requirements that are not common for most of today's big data systems. First of all, such plants are safety-critical systems and consequently there are very important requirements regarding the reliability and clarity of operator's aid or guidance. Furthermore the plants and the corresponding process control systems are real-time systems with deterministic deadlines. Even if no hard deadlines will be applied for data analytics, the applicability of the results will depend on their timely availability. Overall, the development of data-driven OSS has to account for both the specific requirements of the future users as well as the requirements arising from the technical environment. Existing reference models of data analytics or software development do not completely cover these tasks.

As a consequence, even during the early phases of the project the project team identified a need for specific tools and methods tailored to the need of the application domain.

2. EXPERIENCE AND FIRST LESSONS LEARNED

In this section, we report about the first lessons learned during the execution of the FEE project regarding the organization of the development process, and we highlight the importance of data preprocessing and data exploration for successful project execution.

Multidisciplinary Development Process

The starting point for the development process applied in the FEE project was the well known CRISP-DM [5] process, a comprehensible reference process for data mining projects [6]. However, it turned out that such a data mining focussed process model is not sufficient and a multi-disciplinary approach will become necessary.

FIGURE 2: A Claus unit of a major German oil refinery, which is used for the recovery of sulfur from hydrogen sulfide, is used as a case study in the FEE project.

The development team needs to have expertise in the areas of data mining, user-centric design, software technologies and architectures, production management, as well as automation technology. In particular, we see the need for three areas of development or activities: user analysis, data analysis and big data architecture and infrastructure planning. The three areas can certainly be mapped back to the CRISP-DM phases, but require different levels of detailed involvement. User analysis can be mapped on business understanding but requires expertise in user-centred development and control system design. Data analysis can be mapped on data understanding, data exploration, data modelling and evaluation and is the actual domain of data scientists and data mining experts. Architecture and infrastructure planning addresses the deployment, but goes far beyond the submission of a final report; it deals with building and information systems that are able to utilize data mining models under the (soft) real-time requirements of process control systems. The different activity areas are strongly interdependent (e.g. a chosen visualisation is not feasible if the corresponding modelling fails, or architectural requirements like response time define constraints regarding possible modelling types); there is also an experimental character of the data analysis to be noted. The feasibility for data mining or other analytics approaches is limited by the historical data collected. In many cases, the data does not contain the necessary information for producing the desired results. This is also a reason why an agile development approach is better suited than a conventional one. In coordination with end users, the development team has to be able to adjust the direction of a project when environmental conditions change. Especially the creation of early proofs-of-concept is useful in order to identify potential failures as soon as possible.

Figure 3 shows the process that has been developed during the project to identify potential application scenarios, capture functional and non-functional requirements and to develop both suitable data analytics and user interfaces. The process is carried out in 6 steps, where steps 4 und 5 are executed in parallel and with close feedback loops between the corresponding teams.

1 | Scenario Identification: Here a specifically developed ‘Scenario Canvas’, (cf. [7] for a detailed description), was used in end-customer workshops to capture the specific plant situation along with consequences, possible intervention or prevention strategies and the available data – both historical and online. This information helped to rule out irrelevant scenarios (e.g. insufficient data available, consequences not critical, or no prevention/intervention possible from operator side). Furthermore, the current processes in the operator rooms have been described with the help of business process models and the basis of understanding the current situation. Based on

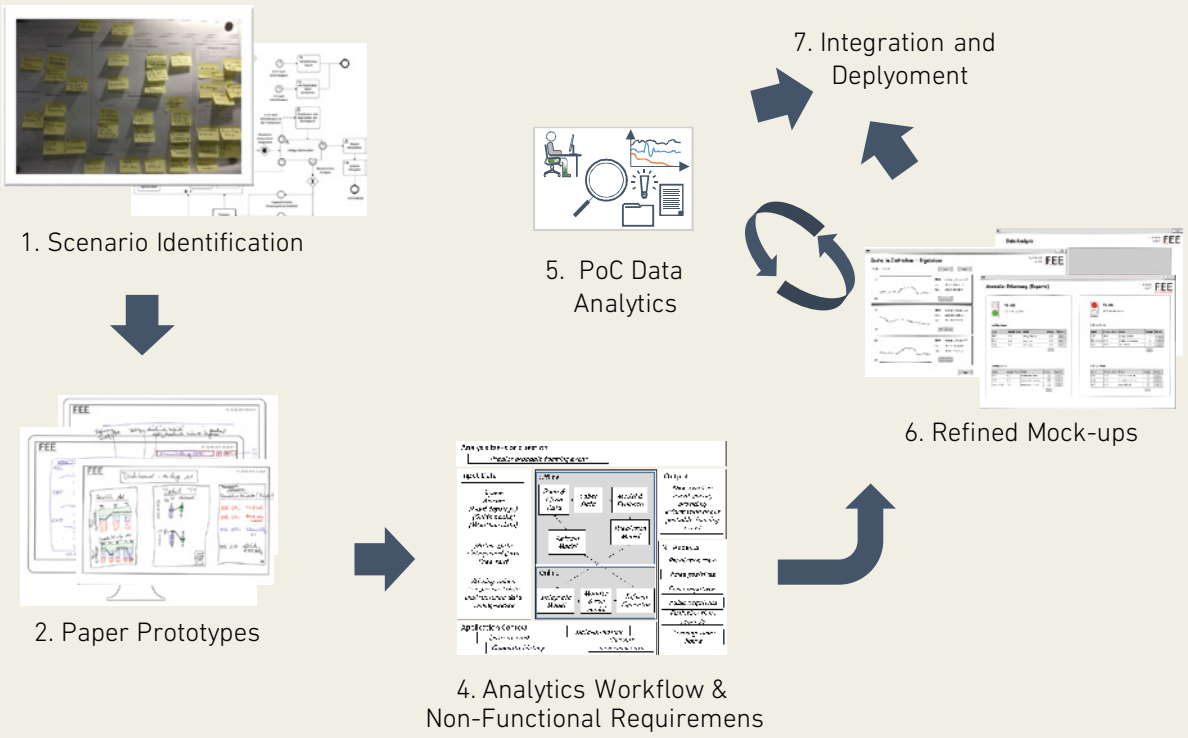

FIGURE 3: Activities and artifacts in the development of the application scenarios of the FEE project.

the learning collected in the 'Scenario Canvas' a short description is created of the scenario with the relevant actors (operators, shift-leader, process engineers) and the understanding of the desired situation.

2 | User Stories and Paper Prototype: Based on the desired situation an interdisciplinary team (HMI experts, data analysts, software architects) in a focus group develop both user stories and first paper prototypes of future user interfaces. The interdisciplinary setup of the team helps when considering the potentials and constraints of the different areas.

3 | Analytics Workflow and Non-Functional requirements: Based on the previous results (user stories and paper prototypes), the same interdisciplinary team starts to define a high level description of the required data analytics and the non-functional requirements arising from the scenario. In this step, big data forces like data volume, variety, velocity, required availability and consistency are discussed extensively.

4 | Proof of Concept (PoC) Data Analytics: Develop first models based on the historical data in order to demonstrate the feasibility of the intended support functionality

5 | Refined Mock-Ups: Refine the paper-prototypes into mock-ups appropriate for collecting early feedback from end-users

6 | Integration and Implementation: In this step, the PoC models have to be refined and improved to meet operational requirements, and integration with the automation system and the user interfaces has to be implemented. This phase has not yet been addressed in the project and remains future work.

Data Integration and Preprocessing

The importance of data integration and appropriate preprocessing for a successful implementation of data driven OSS can hardly be overestimated. In this section, the integration requirements for handling the plant data are described. It covers the necessary steps for preparing the data and implications for the chosen system architecture. Overall, techniques for structured data have been widely applied in the data mining community. Data preparation is a phase in the CRISP-DM

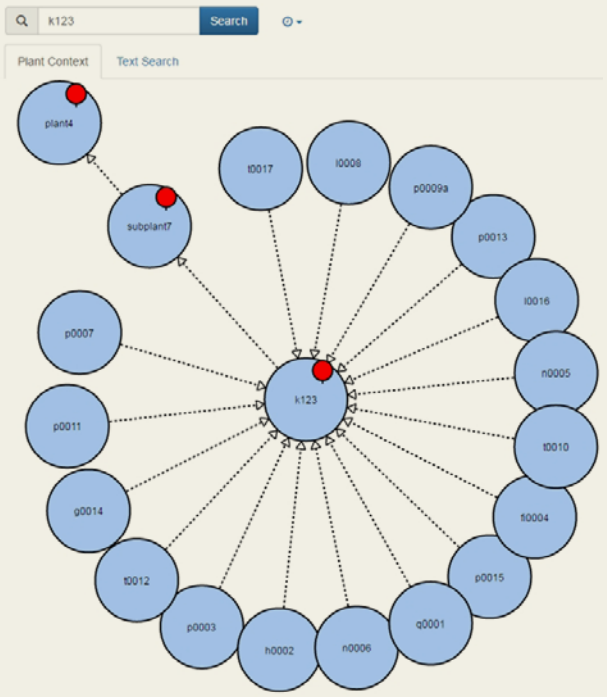

FIGURE 4: Example of a plant context graph, e.g. for browsing & search.

standard data mining process model [5] that is regarded as one of the key factors for good model quality.

The most important long term data store of a plant is called process historian or plant information management system (PIMS). Modern PIMS are able to store ten thousands of signals captured over many years [8]. It is almost certain that in a large plant with many thousands of sensors, some of the sensors will deliver wrong signals. As a consequence, some signals might not be suitable for the intended analysis [10]. Modern smart instruments implement sophisticated self-diagnosis mechanisms telling about the quality and reliability of the measured signal. Typical preprocessing problems that a data analyst faces include outliers [11], frozen signals (signal stops moving, typically the current value will stay at the last known good value in case of an error), noise (e.g. electromagnetic interference), or an unsuitable sampling rate (too high or too low).

Many analytical algorithms are based on very strong assumptions regarding the cleanness and validity of the data. A single outlier can lead to useless and misleading results of complex calculations. It is therefore very important to have an adequate preprocessing in place to either remove or at least identify intervals with measurement problems. Data reconciliation [9] uses mathematical models of the process to discover and remove errors in the measured signals.

In addition to the metric data outlined above, there are usually also semi-structured and unstructured data available, e.g. alarm logs and shift books. Typically, unstructured data is organized into multiple documents containing free text, for which relevant information needs to be extracted [12]. The preprocessing of unstructured data starts with a tokenization step. For each document, the text is cleaned by removing non-word characters, e.g. punctuation and special characters, and then split at each whitespace to create a set of words for the document. The union of all words in the document collection yields the dictionary. One of the most commonly used representations for a document is the bag-of-words model. A document is represented by a multiset of words, i.e. a set of words with corresponding frequencies. Since the order of the words is ignored, this type of model is not able to capture the relation between words, e.g. the co-occurrence of words in a sentence. Here, alternative representations that consider, for example, subsequent word pairs (bigrams) or longer word sequences can be used for capturing interdependencies. For further details and a detailed description of data integration and preprocessing methods in the context of big data, we refer to [13].

Data Exploration

The generation and validation of models to implement the desired support functions is a complicated task for a data analyst because usually different approaches need to be explored and validated. Gaining a strong grasp of the available data is crucial for an analyst to focus on the data that are of highest relevance to the situation at hand. Because of the amount of available data it is important to focus on the relevant parts and to give the operator a compact overview, cf. [14]. Supporting this task promises a faster development of data-driven models and thus increases quality and availability aspects of a plant.

A first step in the data understanding phase of the CRISP-DM model [5] is to calculate basic statistical key figures. This includes distribution information such as extreme values, mean and variance, but also information about the data types (metric, categorical, text) and quality (number of missing values). Those figures should be calculated once for the complete time frame but also for windows of finer granularity, like days or hours.

A further method for data exploration is the calculation of cross-correlations between the signals of sensors. These can give an understanding of how different parts of a plant interact and how situational

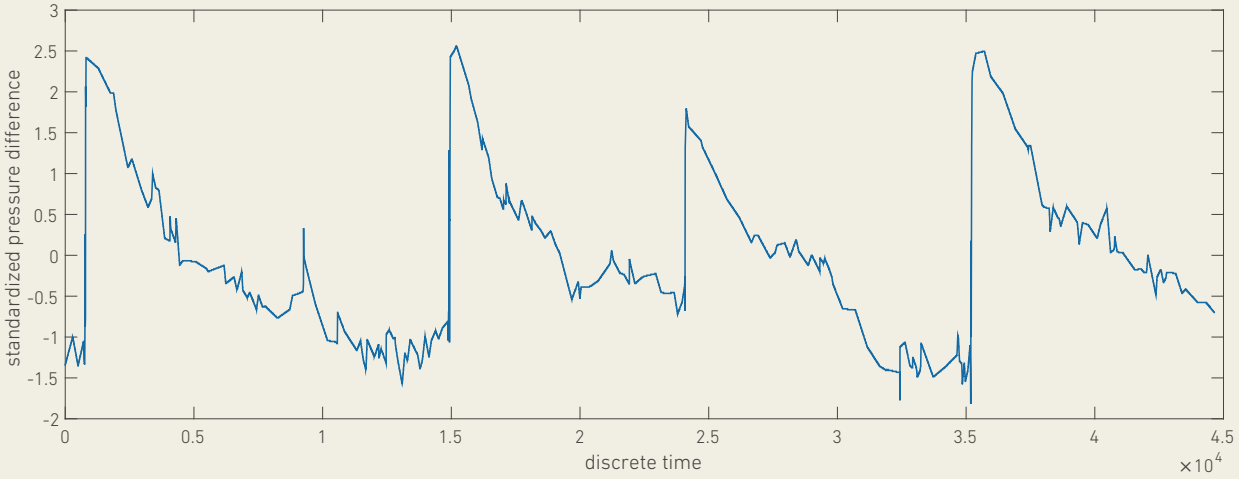

FIGURE 5: Standardized pressure difference measurement over a period of one month

changes propagate through the plant. A heat map can visualize the correlation values between all sensor signals (i.e. all tags of a pair of assets), whereas line plots are useful to show the effect a time shift has on the correlation between two signals. It is important to remember that the correlation indicates similar behaviour for the observed data but does not prove causality.

Another data exploration tool is provided by a hierarchical visualization model of the assets. The required information comes from the piping and information diagrams (P&ID). Figure 4 shows an abstracted P&ID graph; in this example, we can observe the hierarchical relations between the individual entities (modelled as nodes in the graph), and thus obtain an indication about the structural relations between the assets, which can be exploited for browsing, filtering, and providing details on demand according to the "Visual Information Seeking Mantra" [15].

To allow for a fast exploration of the data, one has to focus on the relevant aspects. To this end, additional filter options are necessary. Those filters have to be easy to use but at the same time powerful enough to give the user enough leeway to explore the data unrestricted. Therefore, semi-automatic methods that combine automatic network analysis methods for detecting exceptional patterns [16] with interactive exploratory approaches provide powerful tools for inspecting the network-based exploration tool. Using such methods we can, for example, detect strongly related (linked) groups that exhibit exceptional characteristics, e.g. relating to a temporal burst of alarms in a certain timeframe. In addition, customized views based on a network-based clustering of the assets can be provided.

3. BIG DATA ARCHITECTURE

Handling the large amounts of historical plant data requires a special IT infrastructure, e.g. based on Map/Reduce [17]. From a technical point of view, in most application cases of the FEE project, the data analysis can be divided into two parts: 1) training models with suitable algorithms (data modelling in CRISP-DM [5]), and 2) applying the models during operation of the plants or during deployment in order to fulfil the actual assistant function, (e.g. 'make predictions'). As part of the model training, the usually extensive historical data have to be analyzed as they are stored in the IT-system of the process plant (data-at-rest). There will be no direct interaction with the automation system during this phase and response time plays only a minor role. However, when the model is applied during operation only a small subset of data is required to calculate the input feature of the model and this will be available as a stream of data from the production process. Here the analytics process is typically subject to soft real-time requirements, i.e. "the shorter the response times the better". The Lambda Architecture [18] offers a layout for a robust system that is designed to prepare batch views of stored data (batch layer) and handling requests on those views as well as on rapidly arriving new data (speed layer). The separation of these layers matches the described phases of model training and model application very well. The analysis of historical data-at-rest takes place mainly in the batch layer suited for the execution of long-running processes. The models trained in the batch layer, such as classifiers, are applied to the speed layer with help of streaming technologies. Other outcomes of the batch layer, such

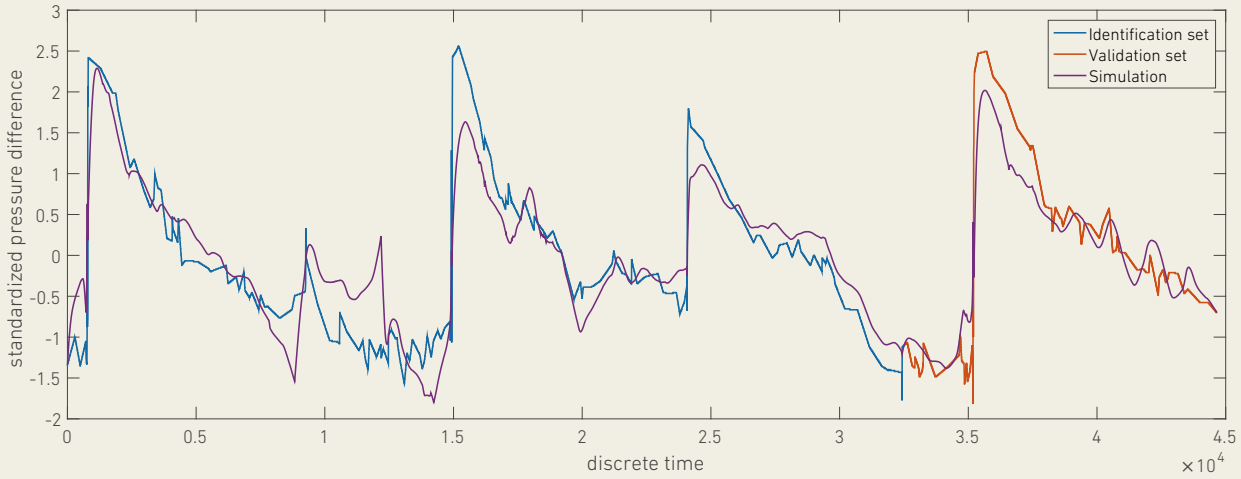

FIGURE 6: Simulation of foaming indicator

as search indexes, make up the basis of the serving layer and all interactive functions.

While the basic structure of the adjusted Lambda Architecture stays the same in different realizations of assistant systems, the requirements for each system must be considered. This way the design for each concrete case can be defined, e.g. by making the specific technology choices or by sizing the execution platform. Design tools include the so-called Quality Attribute Scenarios [19] which enable the communication of characteristics for quality attributes (such as performance, availability, security, adaptability) especially to stakeholders without IT or software background.

4. USE CASES FOR BIG DATA ANALYTICS

There are many actions in production that would benefit from better analysis. Concerning the most promising ones in the process industries we present two approaches for big data analytics in this area and possible outcomes. The single use cases can be implemented separately, but can provide a larger benefit if they are coupled.

4.1 Dynamic model-based predictive alarming

A unit of a major German oil refinery is considered as a case study. The unit under study consists of a Claus process, which is a gas desulfurizing process, and a Shell Claus off-gas treating (SCOT) process, which is connected downstream of the Claus process to remove sulfur compounds from Claus tail gas. An example of an abnormal operating situation is foaming in the SCOT process. The foaming occurs in the last stage of the SCOT process in which a solvent loaded with H_2S is regenerated. The cause of foaming is considered to be the accumulation of impurities and rust particles in the pipes. The formation of foam in a column causes it to overflow and eventually results in plant downtime. To avoid this undesirable consequence of foaming, an anti-foaming agent is introduced manually into the unit when peaks are observed in the pressure difference measurement in the column in which foaming takes place. Figure 5 depicts the standardized pressure difference measurement over a period of one month.

In the current situation, operators have to observe the relevant process values and react under time-pressure by informing their colleagues in the field to take appropriate countermeasures. An early warning giving the operators time to react without time-pressure is highly desired. However, a conventional alarm definition based on thresholds of single process values fails to provide this early warning in a reliable and useful manner.

Chosen methods for analytics PoC

The prediction of such re-occurring critical situations may be formulated as a system identification task which can be performed by a variety of methods, see [20] for an extended discussion.

Continuous processes in production plants are typically operated at fixed operating points, which allows the application of linear dynamic methods for modelling and control, as demonstrated by the numerous successful applications of linear model predictive control in the process industry [21].

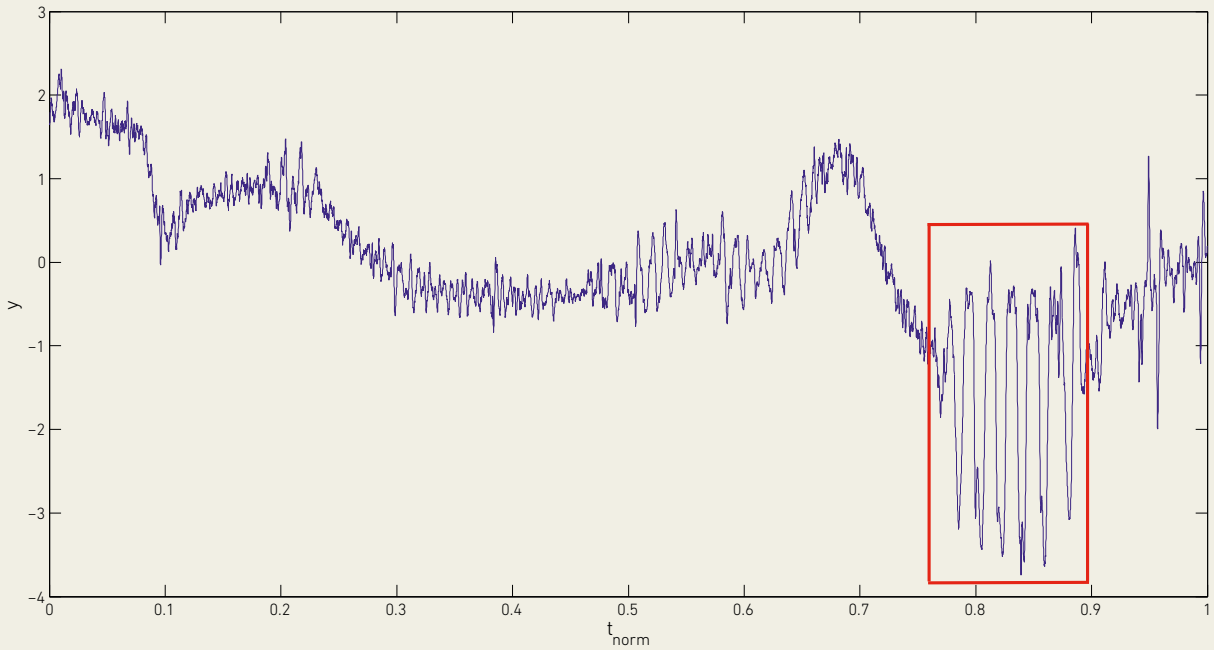

FIGURE 7: Example time series with anomaly (red box) and load change (e.g. at t = 0.1)

A possible way to detect foaming in good time is by predicting indicative process variables over a chosen prediction horizon and assessing these. The task of prediction may require making certain assumptions about process inputs over the prediction horizon. Furthermore, online what-if simulation could be beneficial and will require the modelling of multivariate input-output relationships of process systems. For this purpose, multi-input multi-output (MIMO) linear dynamic models can be identified locally around different operating points. When statistical assumptions are made, model estimation goes along with confidence information for the prediction to assess its trustworthiness. The identification of MIMO models from plant operational data will require the development of multivariate techniques for searching data segments with high information content with respect to system identification.

First results

Several linear system identification methods such as prediction error methods (PEM) and subspace identification methods (SIM) have been applied and assessed for modelling the pressure difference in the foaming column of the SCOT unit. Initial simulation results are promising. A total of 29 process variables were chosen as model inputs by inspecting the P&ID of the unit. The data set was preprocessed by reconstructing missing data points and standardizing it. Elastic net regularization was used to select 7 inputs and a total of 26 terms out of a chosen candidate set of 150 terms. The resulting sparse ARX model is 5th order. Figure 6 shows the model simulation on the identification data set (samples 1 to 3.25×10^4) and on the validation data set (remaining time series).

Future work will focus on making multi-step ahead predictions for giving early warning and developing additional methods for online process analysis for operator support.

4.2 Anomaly detection

Problem statement

Monitoring of the production process is one of the main tasks of operators. However, the large number of sensor values in chemical plants leads to a high workload if operators are supposed to monitor all sensor values. As an example, an industrial petrochemical plant is considered with about 1000 analog sensors providing numerical data, semi-structured categorical data such as alarms, and textual data such as operator logs. The plant operates continuously 24/7 and changes the load about twice a week.

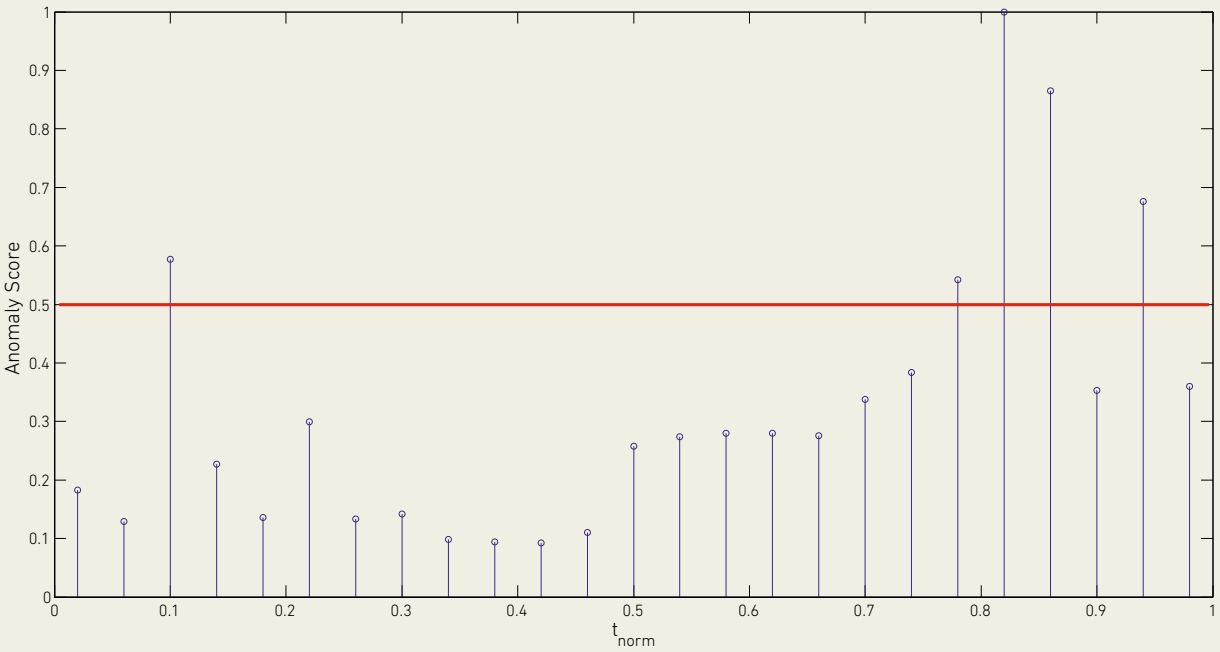

FIGURE 8: Anomaly scores and possible threshold

Often, the retrospective analysis of an event with major economic impact reveals that operators or process engineers would have been able to detect the problem earlier and to develop appropriate intervention strategies, if they had known which data and especially signals to focus on in monitoring and diagnostics. Consequently, an anomaly detection system is desirable. It should provide the users with early information if a process shows unusual behaviour and points to the specific signals which manifest the unusual process behaviour.

Anomalies are usually referred to as data points or sequences that differ significantly from data acquired under normal operating conditions [22]. For that purpose, a nominal operating condition model is extracted from historical data. An alarm will be set off if the current process deviates from normal behaviour. Figure 7 shows an example for anomalous behaviour of the butadiene plant (t = 0.75… 0.9). Only one time series is shown for reasons of clarity, but this anomaly can actually be observed in various signals in varying significance. Furthermore, Figure 7 demonstrates the behavior during a load change (e.g. t = 0.1), as well as the development of anomalous behaviour (t = 0.5…0.7). Ideally the detection method would already recognize the growing anomaly in order to provide the operator with time to think, plan and react before the anomaly occurs.

Chosen methods for analytics PoC

Anomaly detection has been of scientific interest throughout the past decade [23]. It can be described as a two-class-classification problem with data for only one class being available. Moreover particularly in complex systems not all possible anomalies might be known, or an anomaly will occur just once as plant authorities will take actions to prevent its repeated occurrence. Therefore an anomaly detection algorithm looks for data points or sequences which differ significantly from the normal behaviour of the system. Anomaly detection typically just uses numerical data but can also be enhanced by using non-metric data and asset information.

A broad variety of algorithms have been proposed to detect anomalous behaviour. For time series anomaly detection it is common practice to analyze sub-sequences, and where necessary time series transformations are used. The features represent the context rather than just a single observation and are a more sophisticated way of analyzing the behaviour of the system. Most of these methods use density-based descriptions of normal behaviour and analyze the query data by calculating the distance to the k-th nearest neighbour. Therefore it is assumed that under normal operating conditions the data points will form dense clusters, whereas in the case of an anomaly the data points will be located in regions of smaller

density. In most applications, Euclidian distances are used as a distance measure. Since a complex industrial system usually has a large number of sensors, these calculations become computationally expensive while the accuracy decreases with an increasing number of variables [24]. As a consequence, the data is often projected to a subspace assuming that the anomaly can still be detected therein. The most commonly used method for this purpose is principal component analysis (PCA). Other approaches for anomaly detection are also available but are not as flexible as the nearest neighbour method, which works

REFERENCES

[1] Folmer J., Vogel-Heuser, B.: Computing dependent industrial alarms for alarm flood reduction. In: Proc. 9th International Multi-Conference on Systems, Signals and Devices (SSD), pp.1-6, IEEE 2012

[2] Varga, T., Szeifert, F., Abonyi, J.: Detection of Safe Operating Regions: A Novel Dynamic Process Simulator Based Predictive Alarm Management Approach. Industrial & Engineering Chemistry Research 49(2), 658–668, 2010

[3] Windmann, S., Maier, A., Niggemann, O., Frey, C., Bernardi, A., Gu, Y., Pfrommer, H., Steckel, T., Krüger, M., Kraus, R.: Big Data Analysis of Manufacturing Processes. In: Proc. European Workshop on Advanced Control and Diagnosis (ACD2015), IOP 2015

[4] McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. Barton, D.: Big Data. The management Revolution. Harvard Bus Review 90(10), S. 61-67, 2012

[5] Shearer, C.: The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), S. 13-22, 2000

[6] Azevedo, A, Santos, M. F.: KDD, SEMMA and CRISP-DM: a parallel overview. In: Proceedings of the IADIS European Conference on Data Mining. IADIS 2008

[7] Klöpper, B., Dix, M., Schorer, L., Ampofo, A. Atzmueller, M., Arnu, D., Klinkenberg, R: Defining Software Architectures for Big Data Enabled Operator Support Systems. In: Proc. IEEE International Conference on Industrial Informatics (INDIN). IEEE Press, 2016

[8] Hollender, M.: Collaborative Process Automation Systems, ISA 2010

[9] Crowe, C. M.: Data reconciliation—progress and challenges. Journal of Process Control 6(2), 89-98, 1996

[10] Thornhill, N. F., Choudhury, M. S., Shah, S. L.: The impact of compression on data-driven process analyses. Journal of Process Control 14(4), 389-398, 2004

[11] Liu, H., Shah, S., Jiang, W.: On-line outlier detection and data cleaning. Computers & chemical engineering 28(9), 1635-1647, 2004

[12] Atzmueller, M., Kluegl, P., Puppe, F.: Rule-Based Information Extraction for Structured Data Acquisition using TextMarker. Proc. LWA, 2008

[13] Atzmueller, M., Schmidt, A., Hollender, M.: Data Preparation for Big Data Analytics: Methods & Experiences. In: Atzmueller, M., Oussena, S., Roth-Berghofer, T. (eds): Enterprise Big Data Engineering, Analytics, and Management, 157-170. Hershey,PA, IGI Global 2016

[14] Huss, J.: Prototypische Entwicklung einer Trend basierten Datenbank-Suchmaschine, Dissertation TU Berlin, 2008

[15] Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proc. IEEE Symposium on Visual Languages, pp. 336–343, IEEE 2016

[16] Atzmueller, M., Doerfel, S., Mitzlaff, F.: Description-Oriented Community Detection using Exhaustive Subgroup Discovery. Information Sciences 329, pp. 965-984, 2016

[17] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM 51(1), S. 107-113, 2008

[18] Marz, N., Warren, J.: Big Data: Principles and Best Practices of Scalable Realtime Data Systems, Manning 2013

[19] Barbaci, M., Clements, P., Lattanze, A., Northrop, L., Wood, W.: Using the Architecture Tradeoff Analysis Method (ATAM) to Evaluate the Software Architecture for a Product Line of Avionics Systems: A Case Study. In: Tech. Report CMU/SEI-2003-TN-012, Software Engineering Institute, Carnegie Mellon University, 2013

[20] Ljung, L.: Perspectives on system identification, Annual Reviews in Control 34(1), pp.1-12, 2010

[21] Qin, S. J., Badgwell, T. A.: A survey of industrial model predictive control technology, Control Engineering Practice 11(7), 733-764, 2003

[22] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection for discrete sequences: A survey. IEEE Transactions on Knowledge and Data Engineering 24(5), pp.823-839, 2012

[23] Pimentel, M. A., Clifton, D. A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Processing 99, pp.215-249, 2014

[24] Zhang, L., Lina, J., Karim, R.: An angle-based subspace anomaly detection approach to high-dimensional data: With an application to industrial fault detection. Reliability Engineering & System Safety 142, pp. 482-497, 2015

[25] Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information System 3(3), pp. 263-286, 2001

[26] Kluegl, P., Atzmueller, M., Puppe, F.: Meta-Level Information Extraction. Proc. 32nd Annual German Conference on AI, pp.233-240. Springer 2009

on original time series as well as feature vectors or subspace projections. For that reason, this approach was chosen for first tests that are supposed to provide further information on both the influence of different transformations or features and the feasibility of the anomaly detection for the available data.

First results

In order to gain experience, the above mentioned nearest neighbour method was implemented and tested on the data of the butadiene plant. At first, training data containing nominal operation behaviour was segmented to provide nominal operation patterns. Secondly, query data segments were compared to these patterns by computing the Euclidian distance. The distances can be regarded as anomaly scores.

Due to the flexibility of this approach the influence of time series representations and preprocessing could be investigated. Anomalous behaviour could be identified using either the original time series, piecewise aggregate approximation [25], or PCA scores as input to the nearest neighbour algorithm. Figure 8 illustrates anomaly scores for the same time window as used in Figure 7. At first, 13 variables were used that had been chosen by an asset expert. Since anomaly detection is supposed to supervise the process without a priori knowledge, in a second step 227 variables were used for the whole column . The results could be reproduced with dimension reduction by PCA but led to less difference between anomaly scores for the anomaly and for normal operation.

However there are two main drawbacks that need to be dealt with in the future: The method behind the results in Figure 8 is computationally expensive because of the distances that have to be computed in high-dimensional space for finding the k-th nearest neighbour. The computation time has to be improved by magnitudes for online analysis, thus appropriate big data techniques should be used. Furthermore the method leads to several false positives caused by load changes in the plant so that better and more meaningful features are required.

5. CONCLUSIONS & OUTLOOK

In the first phase of the FEE project, the team has spent a significant amount of time analyzing in detail the problems of operators in industrial control rooms. This has led to several use cases. For both ‚predictive alarming' and ‚anomaly detection' suitable methods have been identified and first results presented.

A future use case that poses different challenges and architectural requirements is the search for similar situations in the historical data. Such a functionality can support operators or process engineers with the retrospective or online diagnosis of the process. A possible approach is to look for signals that behave like a suspicious one in order to localize the problem. A major problem in diagnostics tasks is the large numbers of signals and alarms. A further interesting question is whether a situation similar to the current plant situation has occurred in the past. Then, the process manager or the operator can investigate differences in the further development of the process and differences to the current situation. Here we want to investigate the potential of search on time-series data. The search for similarities in time series is a known problem. In the setup phase, the feature space and similarity measures have to be defined. Afterwards the data has to be indexed accordingly and updated continuously with new data. Although the similarity of situations can be quite obvious for humans, who can easily focus on the relevant parts of the signals, the behaviour of similar signals on the time axis can differ a lot (e.g. a ramp after 2 hours instead of 3 hours). In these cases, advanced techniques like dynamic time warping (DTW) can be used to make the signals comparable again.

The search can also be enhanced by using non-metric data and asset information. This could include the occurrence and patterns of alarms as well as leveraging an ontology for selecting appropriate signals (e.g. all temperature signals in unit x and its adjacent units), also relying on hierarchical concepts and abstractions. Here, methods relying on network construction and analysis described in the data exploration use case provide promising results. Specifically, the network representation enables effective data integration and annotation, abstracting metric and unstructured data into a common representation. The setting of this potential use case again matches the Lambda Architecture [18] (cf. Section 2) with the batch, and speed layer, and serving layer. Compared with other FEE use cases, the integration of speed and serving layer is of particular importance. Also, interesting areas of future research consider the integration of further information about processes, for example concerning interactions by email or behavioural (sensor) data, or additional information from unstructured and semi-structured data, by adapting and integrating further advanced approaches for information extraction, e.g. [26].

ACKNOWLEDGEMENTS

The FEE project was sponsored by the German Federal Ministry of Education and Research (BMBF), reference number 01IS14006. The authors are responsible for the contents of this contribution.

The very close interaction with industrial end users enabled by the FEE project ensures that the most relevant problems are tackled and that all required industrial process data and information are available. In future work packages, the developed solutions will be demonstrated together with the industrial partners.

MANUSKRIPTEINGANG
12.02.2016
Im Peer-Review-Verfahren begutachtet

AUTHORS

PD Dr. **MARTIN ATZMUELLER** (born 1976) is adjunct professor (Privatdozent) at the University of Kassel and heads the Ubiquitous Data Mining Research Group at the Research Center for Information System Design (ITeG), Research Unit Knowledge and Data Engineering. His research areas include data mining, network analysis, ubiquitous social media, and big data.

FG Wissensverarbeitung, Wissenschaftliches Zentrum für Informationstechnik-Gestaltung & FB Elektrotechnik/Informatik, Universität Kassel, 34121 Kassel,
E-Mail: atzmueller@cs.uni-kassel.de

Dr. **BENJAMIN KLÖPPER** (born 1981) works for the Analytics and Software Applications group at ABB Corporate Research. His research areas are Big Data Architecture and Fleet Analytics.

HASSAN ENAM AL MAWLA (born 1986) is a scientific assistant at the Measurement and Control Department of the University of Kassel. His research interests include linear and nonlinear system identification methods, and data-driven process analysis.

BENJAMIN JÄSCHKE (born 1988) is a scientific assistant at the Measurement and Control Department of the University of Kassel. His research topics are data mining, time series analysis, and anomaly detection.

Dr.-Ing. **MARTIN HOLLENDER** (born 1963) is working for the Operations Management group at ABB Corporate Research. His research areas are alarm management and industrial analytics.

MARKUS GRAUBE (born 1985) is working for the Chair of Distributed Control Systems of the Technische Universität Dresden. His research areas are semantic information modeling and human-machine-interaction.

DAVID ARNU (born 1983) works as a data scientist at RapidMiner GmbH at the Funded R&D group in the R&D department. His research areas are predictive analytics and Big Data.

ANDREAS SCHMIDT (born 1982) is a scientific assistant at the Knowledge & Data Engineering Group and Interdisciplinary Research Center for Information System Design (ITeG) at the University of Kassel. His research interests include data mining, information retrieval and machine learning in the context of big data.

SEBASTIAN HEINZE (born 1990) is working for the Chair of Distributed Control Systems of the Technische Universität Dresden. His research area is human-machine-interaction in the process industry.

LUKAS SCHORER (born 1987) studied Media Technology at the Technical University of Ilmenau. His research interests are human-machine-interaction, user-centred design and requirements engineering. In 2015 he worked as research intern at the ABB Corporate Research Center Germany.

Univ.-Prof. Dr.-Ing. **ANDREAS KROLL** (born 1967) is head of the Measurement and Control Department at the University of Kassel. His research areas are nonlinear identification and control methods, computational intelligence, and complex systems.

Univ.-Prof. Dr. **GERD STUMME** (born 1967) is head of the Knowledge & Data Engineering Group and Executive Director of the Interdisciplinary Research Center for Information System Design (ITeG) at the University of Kassel. His research interests include data mining, information retrieval, social media, social network analysis.

Prof. Dr.-Ing. **LEON URBAS** (born 1965) is head of the Chair of Distributed Control Systems and the System Process Engineering Group of the Technische Universität Dresden. His main interest is the digital transformation in the process industry.