Industrial Data Science: Developing a Qualification Concept for Machine Learning in Industrial Production

Nadja Bauer, Lukas Stankiewicz, Malte Jastrow, Daniel Horn, Jens Teubner, Kristian Kersting, Jochen Deuse and Claus Weihs

Abstract The advent of Industry 4.0 and the availability of large data storage systems lead to an increasing demand for specially educated data-oriented professionals in industrial production. The education of these specialists should combine elements from three fields: Industrial engineering, data analysis and data administration. However, a comprehensive education program incorporating all three elements has not yet been established in Germany.

The aim of the acquired research project, titled "Industrial Data Science" is to develop and implement a qualification concept for Machine Learning based on demands coming up in industrial environments. The concept is targeted at two groups: Advanced students from any of the three mentioned fields (Mechanical Engineering, Statistics, Computer Science) and industrial professionals.

Nadja Bauer, Malte Jastrow, Daniel Horn, Claus Weihs
TU Dortmund University, Faculty of Statistics, Chair of Computational Statistics
✉ bauer,jastrow,horn,weihs@statistik.tu-dortmund.de

Lukas Stankiewicz, Jochen Deuse
TU Dortmund University, Faculty of Mechanical Engineering, Institute of Production Systems
✉ lukas.stankiewicz, jochen.deuse@ips.tu-dortmund.de

Jens Teubner
TU Dortmund University, Faculty of Computer Science, Chair of Data Bases and Informational Systems
✉ jens.teubner@cs.tu-dortmund.de

Kristian Kersting
Technische Universität Darmstadt, Faculty of Computer Science, Chair of Machine Learning
✉ kersting@cs.tu-darmstadt.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/27
ISSN 2363-9881

In the qualification concept the needs of industrial companies are considered. Therefore, a survey was created to inquire the use and potentials of Machine Learning and the requirements for future Data Scientists in industrial production. The evaluation of the survey and the resulting conclusions affecting the qualification concept are presented in this paper.

1 Introduction

Many papers try to define the term Data Science. Demchenko et al (2016), Song and Zhu (2017), Horton and Hardin (2015) and Diggle (2015) e.g. provide here short overviews. In Diggle (2015) especially the relation between the terms Data Science and statistics is discussed. However, all definitions have in common that a Data Scientist should demonstrate interdisciplinary competences (analytical skills, engineering expertise, domain knowledge etc.), which poses a challenge for the education of such professionals.

The number of new Data Science courses has increased significantly over the last 10 years (Horton and Hardin, 2015). However, there is no consensus on how the curriculum of such courses should look like, so a lot of research is currently being done in this area. Demchenko et al (2016) present a framework, which can be used by universities to create data science curricula. In Song and Zhu (2017) three main focuses are identified as the basis of Data Science education: user-based (ability to solve the problems by knowing and applying Machine Learning (ML) techniques), tool-based (handling of high-level programming languages like R and Python as well as ML-tools like RapidMiner) and application-based (expertise in application domain). Horton and Hardin (2015) give a detailed overview of the structure and course of some Data Science courses at different universities. The authors summarize that such courses are very popular among students and lecturers, but bring with them the challenge of heterogeneity in students background.

TU Dortmund University was the first German university establishing a bachelor's program in "Data Analysis and Data Management" and a master's program in "Data Science" in 2002. Both programs are organized jointly by the faculties of Statistics, Computer Science and Mathematics. In addition, the Faculty of Mechanical Engineering also has extensive research work in the field of Data Science.

One of the strengths of the TU Dortmund University is its interdisciplinarity, which is expressed by the close cooperation between the faculties in teaching and research areas. For example, the faculty of Statistics offers bachelor's and master's courses for mechanical and industrial engineering students on statistical quality assurance and optimization in industry. Until now, only basic methods such as probability theory, correlation measures and linear regression could be treated. However, the experience shows that also advanced statistical approaches are often required for a complex analysis of industrial data.

Increasing digitalization in production areas is causing a rapid increase in complexity of business processes. As a result, more and more manufacturing processes are being monitored with sensors so that large data volumes are created that need to be stored and analyzed (often online). Machine Learning and knowledge discovery in data open up new opportunities for companies, e.g. in process analysis or development of new products and business ideas. Figure 1 shows how "Industrial Data Science" (abbreviated to InDaS below) can be located between these three domains. The lack of specialists with solid knowledge in all of these domains has long been complained about by employers.

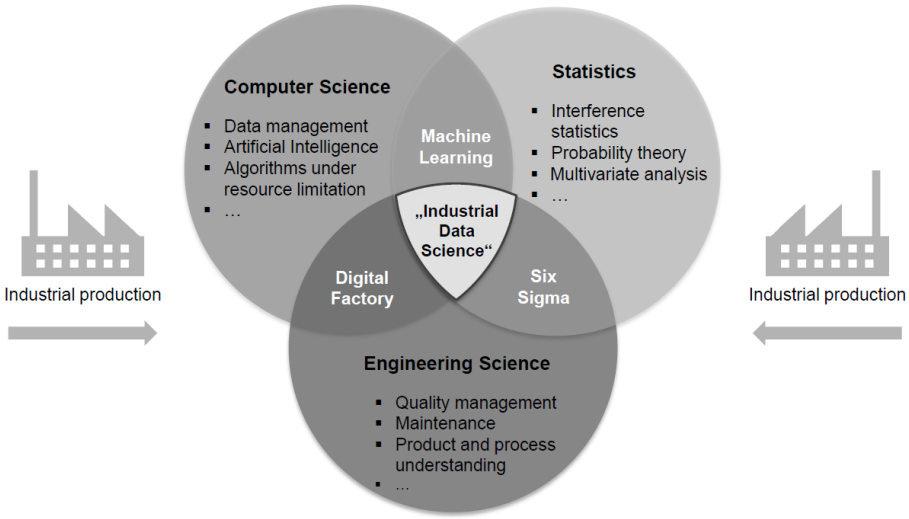

Figure 1: "Industrial Data Science" as the interface of the three domains.

The Federal Ministry of Education and Research (German: BMBF) responds to this development with a series of tenders and promotes the link of research and industry as well as a good education of ongoing experts. In this paper, we present our current results that we gained in the first phase of the two-year project "Industrial Data Science - Machine Learning Qualification Concept in Industrial Production" (November 2017 - October 2019). This project is one of 29 projects funded by the BMBF under the call for tenders for "Qualification Measures and Research Projects in Machine Learning" (February 2017). The main goals of the project can be defined as follows:

- Analysis of requirements of manufacturing companies.
- Development of an innovative InDaS qualification concept for master's students and specialists from industry.
- Involvement of real use cases in cooperation with several industrial partners.
- Project work in interdisciplinary groups.

This paper is structured as follows: Section 2 analyses the requirements of the manufacturing companies and consists of three points: literature review, a survey carried out as part of the research project and a requirement catalog. In Section 3 the structure and content of the InDaS education is discussed. Finally, Section 4 summarizes and concludes the paper.

2 Analysis of The Requirements of The Manufacturing Companies

In this section we first thoroughly analyze the existing corresponding company surveys. However, it has quickly become clear that they are too comprehensive to be able to explicitly derive the requirements of manufacturing companies. For this reason, a self-designed online survey was developed and a wide range of industrial companies from various sectors were asked to participate in the survey. The analysis of existing surveys as well as the self-designed survey are summarized in a requirement catalog for the education concept.

2.1 Summary of the literature review

In the context of this paper ML is seen as a generic term for artificial intelligence, data mining, data science, deep learning and big data. The term big data describes the case of the large data amounts that come about through digitization. The analyzed surveys focused on different target groups. Table 1 gives an overview of the number of participants (size) and composition of the surveys.

Table 1: Size and composition of the analyzed surveys (DACH: Germany, Austria and Switzerland).

Source	Size	Region	Target group
Böttcher, B. and Klemm, D. and Velten, C. (2017)	264	DACH countries	IT decision-makers, managing directors
Rexer et al (2016)	1220	Worldwide	Data science professionals
Lueth et al (2016)	151	Worldwide	Department heads, managers
Weber et al (2014)	507	Germany	IT director, managing director
Derwisch and Iffert (2017)	210	DACH countries	IT managers, managers

Added value through ML and challenges

In all studies, there is a high degree of approval and openness towards the use of ML methods as it creates added value for the company or is expected to do so in the future. The most important applications seem to be customer understanding, customer loyalty or increasing customer satisfaction. Also frequently mentioned are optimization of internal processes, predictive modeling (such as predictive maintenance and anomalies detection), improving product quality, development of new product ideas or analyses of competitors.

Regarding the current state of ML usage the following pattern can be seen: Only a few companies consider themselves to be excellent at practicing ML or are already actively implementing various projects (10% on average). The largest amount of companies are in pilot phases, planning or discussing possible ML implementations (50% - 80%). The remaining about 20% have not yet dealt with ML or are not planning to do so in the future (Böttcher, B. and Klemm, D. and Velten, C., 2017; Rexer et al, 2016; Lueth et al, 2016; Derwisch and Iffert, 2017).

From a technical point of view, many companies complain about problems with data administration. Göbel (2015) claim “handling of unstructured data

which until now has been brought into a suitable structure by manual recording" to be a problem. According to Derwisch and Iffert (2017), 55% of respondents consider analyzing large volumes of data as a challenge and 51% would like a faster deployment of data for analysis. Other identified issues are a lack of data cleansing, non-agile business intelligence infrastructure, lack of access to data sources, non-intuitive user interfaces or errors in data management. In Göbel (2015), therefore, the need for new database concepts which enable efficient parallel processing of data in cluster (such as NoSQL for modern non-relational databases or NewSQL for distributed workload) is emphasized.

From an analytical point of view, many respondents claim a challenge in choosing a suitable ML method. In Göbel (2015) the problem of recognizing statistical relationships between different parameters is mentioned, whereas in Rexer et al (2016) data visualization is mentioned as a major obstruction.

Another challange is the compliance with current data protection laws. According to Böttcher, B. and Klemm, D. and Velten, C. (2017), 31% of respondents see a potential privacy problem. Many companies forego data analysis because of fear of customer criticisms, image loss, and because of legal or ethical-moral reasons (Weber et al, 2014). Specifically, Göbel (2015) point out that data analysis can be used to create profiles of individual operators.

Data, ML algorithms and ML software

By Weber et al (2014), 36% of companies use "master" data for data analysis. Another 33% use transaction data, followed by log data (31%) and sensor data (25%). With a lower frequency, text (9%) and voice / video / audio data (6%) are used. Another list is given in Derwisch and Iffert (2017): 81% of respondents use "structured" data, followed by transaction (66%), log (33%) and clickstream data (32%). Again, the proportions of image / video (8%) and text data (9%) are relatively low. However, in Böttcher, B. and Klemm, D. and Velten, C. (2017) the following ML application fields are mentioned particularly frequently: Image recognition, pattern identification, speech recognition, text mining and video analysis.

The survey of Rexer et al (2016) figures out that most commonly used ML algorithms include regression and tree-based methods. Other frequently mentioned methods are clustering, ensemble methods, association analysis, time series analysis and neural networks. The software used covers a wide range of solutions. Rexer et al (2016) emphasize that R is being used more and more

frequently: The proportion of users among respondents has risen from 23% in 2007 to 76% in 2015. In this survey, R was the most commonly used software for the primary analysis. R also has a very high rate of satisfied users (95%). Other common tools include Java, Excel, Power BI, .Net, Python, SPSS, SAS, Matlab and Rapid Miner.

Continuing education / skills shortage / investment

The willingness to invest money in advanced analytics has increased over the last few years (Derwisch and Iffert, 2017). The investments go, in the first place, into IT solutions (software, hardware, storage space, cloud solutions) and, in the second place, into employment or continuing education of data scientists. However, 65% of companies see a lack of data scientists on the job market and 14% even refrain from certain data analyses due to this lack (Weber et al, 2014). Rexer et al (2016) point out that while universities are in the process of creating data science programs, it will take years to "produce" the highly skilled people.

2.2 Presentation and evaluation of our own online survey

The survey is divided into four sections: General company information, data management, machine learning and soft skills requirements. The results of the survey for the first three topics are analyzed and presented below.

Who participated in the survey?

A total of over 200 contacts (senior Data Scientist, senior engineers, CIOs) were contacted. These are mainly industry contacts of the Faculty of Mechanical Engineering (mostly manufacturing companies), but also of the Faculty of Statistics. The survey was answered anonymously by 57 companies, which corresponds to a non-response rate of about 72%. Figure 2 shows the industry affiliation of the respondents. The three most represented industry sectors cover automotive, electrical engineering and mechanical engineering.

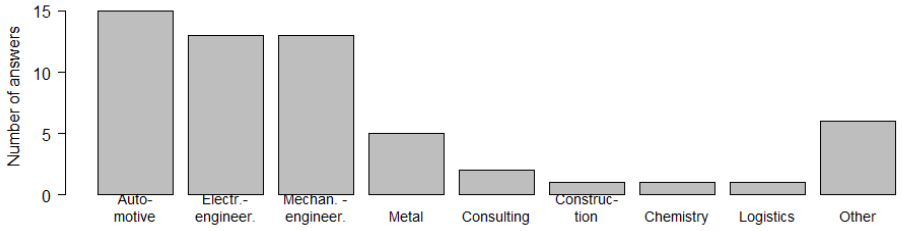

Figure 2: Sector affiliation of the survey participants.

More than 80% of the responses came from larger companies (with more than 1,000 employees). Furthermore, participants should assess the digitization level of their company. 14 respondents (25%) consider their company to be "digitized in a few areas", while the absolute majority (65%) classify their company as "digitized in many areas". Only in isolated cases the company is rated as completely (7%) or not at all (4%) digitized. This leads to the assumption that there is a greater ML affinity in lager companies.

How is ML used in the company?

The participants were asked in which form data is collected or evaluated. In the diagram on the left in Figure 3 it can be seen that in almost all companies numerical or character tables are collected, whereas audio data are only collected in 15 cases. All in all, there is a great potential for data that has not yet been evaluated. The chart on the right shows how many different types of data are collected in each company (typically more than two).

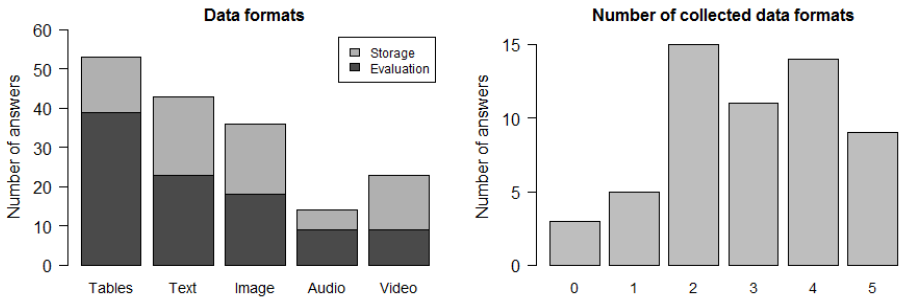

Figure 3: Collected and evaluated data formats in the companies.

A further question considers the level of ML usage in the companies. Only five companies (9%) use ML methods across the board, 27 others (47%) have already gained initial experience with the topic or are using the methods to a limited extent. The remaining 25 companies (44%) are not yet using ML but most plan to do so. These answers indicate that there is a high potential of data not being analyzed yet.

Which methods and software are used for ML?

For this question, only the 32 companies stating to use ML are considered. The first question in this section deals with different types of analysis using ML methods. Figure 4 (left) shows that most types of analysis are used by about half of the 32 considered companies. Only reliability and association analyzes are performed less frequently. The graph on the right shows that only three of the companies use exactly one type of analysis, while eight do not use any of the types of analysis requested, but nevertheless claim to use ML. The remaining companies use two or more types of analysis.

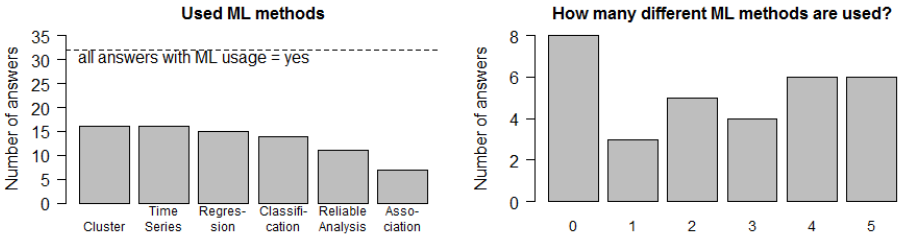

Figure 4: Analytical methods used in the 32 companies already using ML .

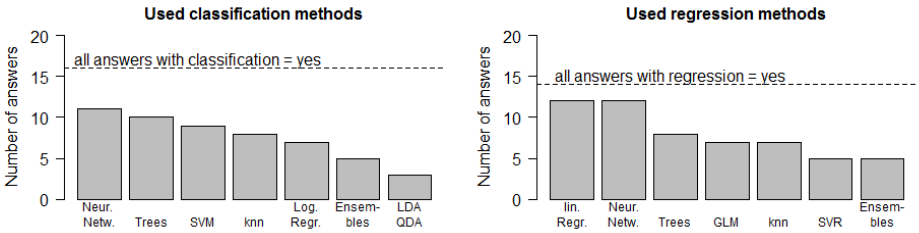

Figure 5: Classification and regression methods used (SVM: support vector machines, knn: k nearest neighbor, LDA/QDA: linear/quadratic discriminant analysis, GLM: generalized linear model, SVR: support vector regression, Neur. Netw. : Neural Network , lin. Regr. : linear Regression).

Figure 5 shows the answers to the conditional follow-up questions, here about classification and regression. Neural networks are most commonly used in classification procedures, followed by trees and support vector machines (SVM and SVR). Classical methods such as discriminant analysis (LDA and QDA) and logistic regression are less common, as well as ensemble methods, which include the popular Random Forests. In addition to neural networks, classical linear regression is most frequently used among regression methods. As a conclusion, a broad spectrum of ML methods is applied according to the needs, while the current trend to use neural networks can be observed.

The use of software to apply ML was analyzed using the answers shown in Figure 6. There are six programs or programming languages that are used with equal priority in companies. These are Excel, Matlab, Python, R, RapidMiner, and Minitab. Excel is indeed a commonly used tool for data processing which, however, does not provide profound ML methods. In general, the software solutions used are wide-ranging and there seems to be no clear favorite in the industry, however, script-oriented software (Matlab, Python and R) seems to be quite common and accepted.

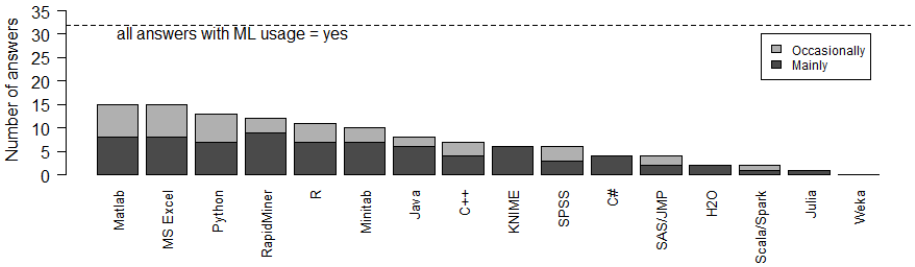

Figure 6: Software used for ML in the enterprises.

2.3 Derivation of a requirements catalog

Derived requirements for the design of the InDaS concept are presented in Figure 7. The foundation of knowledge transfer is a didactic-methodical concept comprising simple and comprehensible presentation of complicated methods, project-oriented work and soft skills training. On top of this, the content is subdivided into three superordinate thematic blocks: data management, ML and domain knowledge, which are shown in the figure as columns and represent the essential requirements for an Industrial Data Scientist.

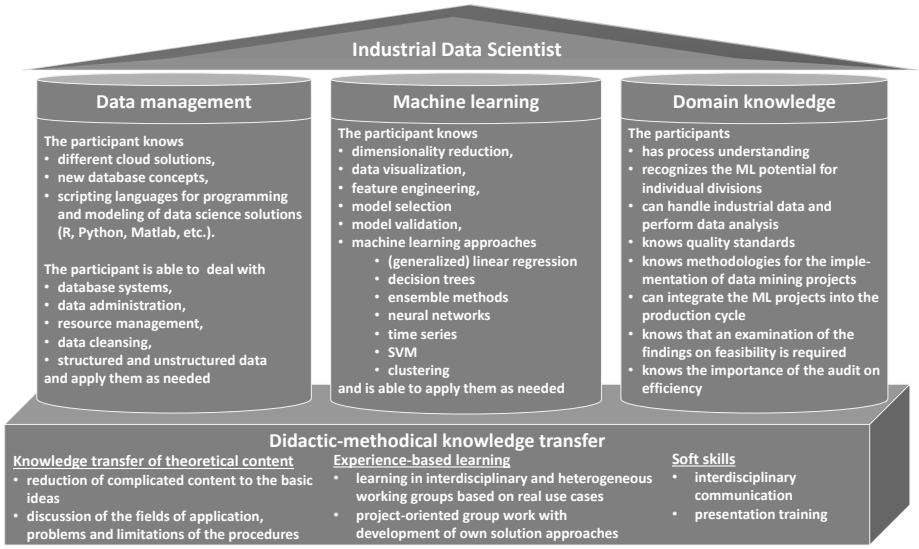

Figure 7: Requirements catalog.

3 Detailing the Qualification Concept

Our qualification concept - tailored to the needs of Germany's industry - consists of two parts: lecture (semester 1) and practical seminar (semester 2). Due to the heterogeneity of the participant groups and differences in their prior knowledge, it is neccessary to diversify the contents of the lecture. In this aspect, after a general introduction to InDaS, the fundamentals of data driven industrial engineering are provided including product life cycle, requirements and standards for data mining projects. Several real use cases will be presented and continued in the subsequent lecture units. Furthermore, students will be taught techniques for managing large amounts of structured and unstructured data. These include basic techniques of relational database systems (with SQL), basic strategies for efficient data processing in relational systems and the basis of "data cubes" modeling in relational systems. Building on this, the ML methods will be presented and explained in the following lecture units. Finally, guest lectures at the end of the semester show real industrial ML use-cases. The lecture plan is presented in Table 2. Each of the 14 learning units consists of a two-hour lecture and a two-hour practical exercise.

Table 2: Structure of the lecture content.

Unit	Lecture topic	Lecture content
1	Introduction to Industrial Data Science (InDaS)	• Definition: Machine Learning, InDaS • History, objectives and fields of application of InDaS • Presentation of the online survey and findings
2	Data in the industrial environment	• The product life cycle • Digital Factory and Industry 4.0 • Data availability / quality and infrastructure in companies • Requirements on / opportunities through InDaS
3	Data analysis in the industrial environment	• Requirements for data mining projects in the industry • Standards for the implementation of data mining projects • Real InDaS use-cases (e.g. using CRISP-DM standard)
4	Data administration of structured data	• Definition: structured and unstructured data • Basic techniques of relational databases • Cloud storage • Transparent data mapping
5	Data administration of unstructured data	• Handling unstructured data: MapReduce • Front-End languages: Apache Pig
6	Exploratory data analysis and data visualization	• Types of variables and measures for data description • Distribution of data and its visualization • Correlation, variance, multicollinearity
7	Data preprocessing	• Data cleansing, reduction and transformation • Feature engineering / extraction • Imputation of missing values • Handling of unbalanced data
8	Linear Models	• Definition: linear models for regression and classification • Generalized linear regression (logistic regression) • LDA, QDA, Penalization (Ridge, Lasso)
9	Tree-based methods	• Definition and characteristics of tree-based methods • CART approaches for classification and regression
10	Ensembles	• Definition and characteristics of ensembles methods • Bagging and Boosting (ADA boost, XG boost) • Random Forests
11	Further methods	• k nearest neighbors • SVM & One Class SVM • Neural networks • Naive Bayes
12	Model Selection Concepts	• Determination of the model quality • Resampling techniques • Hyperparameter tuning and variable selection
13	Clustering	• Partitioning procedures • Density-based methods • Evaluation of the performance
14	Guest lecture	• Presentation of real InDaS use cases

The practical seminar (two hours weekly) aims to simulate the work in real industrial projects. In small interdisciplinary teams of students from different departments several real use cases are dealt with. Those teams are supported and coached by lecturers and industrial professionals during the whole semester. In this way students do not only train professional skills, but also improve soft skills. Due to existing comprehensive expertise in the R language and its popularity in the industry it will be used as the analysis tool.

4 Conclusion

Education in Data Science is an important issue and is explicitly desired by industry and government. In this paper we present a novel qualification concept which is now being developed and practiced during a two years project supported by the Federal Ministry of Education and Research. TU Dortmund University provides a perfect basis for Industrial Data Science education due to its long history and broad expertise in these areas. Not only the internal cooperation between Statistics, Computer Science and Mechanical Engineering faculties are well-established at TU Dortmund but also close contacts to local and global industrial companies are maintained.

In this paper, a qualification concept based on a detailed study of demands of manufacturing industry for specialist staff is introduced. For this reason few related state-of-the-art publications as well as a conducted and evaluated own online survey were analyzed. The main principles of the qualification concept are interdisciplinarity, didactic mediation of complex Machine Learning concepts concentrating on its main ideas, application fields, limits and examples as well as complementing the theoretical knowledge by practical use cases.

After one year of successive education we expect the students to be able to identify the promising data mining projects in industrial production, communicate with team colleagues from other disciplines, choose and apply an appropriate Machine Learning strategy for a given task, be familiar with data administration and storage tools, communicate the results and recognize its value-added as well as implementing it into the industrial process.

Acknowledgements This work is kindly supported by the Federal Ministry of Education and Research in the context of “Qualification programs and research initiatives in the field of Machine Learning” program.

References

Böttcher, B. and Klemm, D. and Velten, C. (2017) Machine Learning im Unternehmenseinsatz. Künstliche Intelligenz als Grundlage digitaler Transformationsprozesse. Tech. Rep., CRISP Research.

Demchenko Y, Belloum A, Los W, Wiktorski T, Manieri A, Brocks H, Becker J, Heutelbeck D, Hemmje M, Brewer S (2016) EDISON Data Science Framework: A Foundation for Building Data Science Profession for Research and Industry. In: 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), pp. 620–626. DOI: 10.1109/CloudCom.2016.0107.

Derwisch S, Iffert L (2017) Advanced & Predictive Analytics: Data Science im Fachbereich. Tech. Rep., BARC. URL: https://barc.de/docs/advanced-und-predictive-analytics-data-science-im-fachbereich.

Diggle PJ (2015) Statistics: A Data Science for the 21st Century. Journal of the Royal Statistical Society. Series A (Statistics in Society) 178(4):792–813. DOI: 10.1111/rssa.12132.

Göbel R (2015) Leitfaden: Industrie 4.0 und Smart Data. Die Welt der großen Datenmengen in Unternehmen: Neue Möglichkeiten zur Erfassung und Auswertung großer Datenmengen am Beispiel der Qualitätssicherung von Konsumgütern . Tech. Rep., eBusiness-Lotse Oberfranken. URL: http://docplayer.org/1749165-Industrie-4-0-und-smart-data.html.

Horton NJ, Hardin JS (2015) Teaching the Next Generation of Statistics Students to "Think With Data": Special Issue on Statistics and the Undergraduate Curriculum. The American Statistician 69(4):259–265. DOI: 10.1080/00031305.2015.1094283.

Lueth KL, Patsioura C, Diaz Williams Z, Zahedi Kermani Z (2016) Industrial Analytics 2016 / 2017: The Current State of Data Analytics Usage in Industrial Companies. Tech. Rep., IoT Analytics. URL: https://digital-analytics-association.de/wp-content/uploads/2016/03/Industrial-Analytics-Report-2016-2017-vp-singlepage.pdf.

Rexer K, Gearan P, Allen H (2016) 2015 Data Science Survey. Tech. Rep., Rexer Analytics, Winchester, MA. URL: https://www.rexeranalytics.com/data-science-survey.html.

Song IY, Zhu Y (2017) Big Data and Data Science: Opportunities and Challenges of iSchools. Journal of Data and Information Science 2(3):1–18, Berlin. DOI: https://doi.org/10.1515/jdis-2017-0011.

Weber M, Dehmel S, Hampe K., Shahd M (2014) Potenziale und Einsatz von Big Data: Ergebnisse einer repräsentativen Befragung von Unternehmen in Deutschland. Tech. Rep., Bitkom, Berlin. URL: https://www.bitkom.org/sites/default/files/file/import/Studienbericht-Big-Data-in-deutschen-Unternehmen.pdf.