Melina Panzner

Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung

Geleitwort

Die Entwicklung neuer Methoden und Werkzeuge für das Engineering von Morgen steht im Mittelpunkt unserer Forschungsarbeiten. Am Heinz Nixdorf Institut der Universität Paderborn sowie am Fraunhofer-Institut für Entwurfstechnik Mechatronik IEM widmen wir uns intensiv dieser Thematik. Unser übergeordnetes Ziel besteht darin, die Innovationsfähigkeit von Industrieunternehmen nachhaltig zu steigern.

Die fortschreitende Digitalisierung und die neuesten technologischen Entwicklungen ermöglichen es, dass cyber-physische Systeme während ihres Betriebs umfangreiche Datenmengen generieren. Diese Daten eröffnen Herstellern mittels Data Analytics und Künstlicher Intelligenz völlig neue Möglichkeiten, tiefergehende Einblicke in die Nutzung ihrer Produkte zu gewinnen und diese Erkenntnisse für eine gezielte Produktplanung zu nutzen. Die betriebsdatengestützte Produktplanung stellt jedoch Unternehmen vor große Herausforderungen, da oft das notwendige interdisziplinäre Expertenwissen fehlt, um die verfügbaren Daten adäquat zu analysieren und in konkrete Maßnahmen zu überführen. Es bedarf daher neuer Methoden und Werkzeuge, um Domänenexperten in Unternehmen zu befähigen, das Potenzial der Datenanalyse bestmöglich auszuschöpfen.

Vor diesem Hintergrund hat Frau Panzner eine Systematik zur Datenanalyse für die betriebsdatengestützte Produktplanung entwickelt. Diese Systematik ermöglicht es Domänenexperten ohne große Analytics Kenntnisse, den Analytics-Prozess von der Planung der Anwendung bis zum Umsetzungsstart zu gestalten. Im Kern der Arbeit steht ein methodisches Vorgehensmodell, das Domänenexperten strukturiert an die Herausforderungen der Datenanalyse heranführt und ihnen geeignete Werkzeuge zur Verfügung stellt. Die Systematik berücksichtigt die spezifischen Anforderungen der Produktplanung und integriert sowohl geschäftliche als auch technische Aspekte dabei, um eine erfolgreiche Datenanalyse zu ermöglichen.

Die wissenschaftliche Auseinandersetzung legt eine wichtige Grundlage für zukünftige Forschungsarbeiten und bietet zugleich einen hohen praktischen Nutzen für die Industrie. Besonders die interdisziplinäre Verknüpfung von Datenanalyse und Produktplanung ist ein wertvoller Beitrag zur Weiterentwicklung datengetriebener Entwicklungsprozesse. Die Validierung der Systematik anhand eines prototypischen Assistenztools und industrieller Anwendungsfälle unterstreicht ihre Relevanz und Praxistauglichkeit.

Frau Panzner hat mit großem Engagement und fundiertem wissenschaftlichen Vorgehen einen bedeutenden Beitrag zur Befähigung von Unternehmen im Umgang mit Betriebsdaten und ihrer sinnvollen Nutzung geleistet. Sie hat wesentliche Pionierarbeit für datengetriebene Produktplanung geleistet.

Paderborn, im Februar 2025 Prof. Dr.-Ing. R. Dumitrescu

Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung

zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN (Dr. rer. nat.)
der Fakultät Elektrotechnik, Informatik und Mathematik
der Universität Paderborn

genehmigte
DISSERTATION

von
M.Sc. Melina Panzner (geb. Massmann)
aus *Bielefeld*

Tag des Kolloquiums:	18. Februar 2025
Referent:	Prof. Dr.-Ing. Roman Dumitrescu
Korreferent:	Prof. Dr.-Ing. Sebastian von Enzberg

Vorwort

Die vorliegende Dissertation entstand im Rahmen meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Fraunhofer-Institut für Entwurfstechnik Mechatronik IEM sowie Heinz Nixdorf Institut. Sie ist das Ergebnis meiner Arbeit an verschiedenen Forschungs- und Industrieprojekten, die mir immer wieder neue Einblicke und Herausforderungen geboten haben.

Mein besonderer Dank gilt **Prof. Dr.-Ing. Roman Dumitrescu**, der mich von Anfang an mit viel Vertrauen, wertvollem Feedback und der richtigen Mischung aus Förderung und Herausforderung begleitet hat. Deine Unterstützung und das inspirierende Umfeld, das du geschaffen hast, haben meine fachliche wie auch persönliche Entwicklung nachhaltig geprägt.

Ebenfalls bedanke ich mich ganz herzlich bei **Prof. Dr.-Ing. Sebastian von Enzberg** vom Fachbereich Ingenieurwesen und Industriedesign der Hochschule Magdeburg für die Übernahme des zweiten Gutachtens. Und nicht nur dafür – auch für deine fachliche und persönliche Begleitung während meiner gesamten Zeit am IEM gebührt dir mein herzlicher Dank. Deine wertvollen Impulse, dein Engagement und deine stetige Unterstützung haben mich immer wieder vorangebracht.

Mein Dank gilt zudem allen aktuellen und ehemaligen Kolleginnen und Kollegen am IEM. Die Zusammenarbeit mit euch, die vielen gemeinsamen Diskussionen und euer Teamgeist haben maßgeblich zu dieser Arbeit beigetragen. Ebenso möchte ich mich bei allen Studierenden bedanken, die mich mit ihren Abschlussarbeiten und ihrer Unterstützung als studentische Hilfskräfte begleitet haben.

Mein größter Dank gilt meiner Familie. Danke an meine Eltern und meine Schwester, die mich auf meinem gesamten Weg unterstützt und dafür so vieles hintenangestellt haben. Und natürlich danke ich von Herzen meinem Mann, der mich zur Promotion ermutigt und mir jederzeit den Rücken freigehalten hat. Danke für deine bestärkenden Worte.

Paderborn, im Februar 2025 Melina Panzner

Liste der veröffentlichten Teilergebnisse

[PED24] Panzner, M.; von Enzberg, S.; Dumitrescu, R.: Developing a data analytics toolbox for data-driven product planning: a review and survey methodology. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *38*, Article e18, 2024.

[MPW+23] Meyer, M.; Panzner, M.; Wiederkehr, I.; Fichtler, T.: Referenzprozess für die Betriebsdaten-gestützte Produktplanung. In Dumitrescu, R., & Koldewey, C. (Hrsg.), Datengestützte Produktplanung: Mit Betriebsdaten und Data Analytics zur faktenbasierten Planung zukünftiger Produktgenerationen im produzierenden Gewerbe (S. 29 – 77). HNI-Verlagsschriftenreihe 408, 2023.

[PEM+22] Panzner, M.; von Enzberg, S.; Meyer, M.; Dumitrescu, R.: Characterization of Usage Data with the Help of Data Classifications. Journal of the Knowledge Economy, 2022.

[MPK+22] Meyer, M.; Panzner, M.; Koldewey, C.; Dumitrescu, R.: 17 Use Cases for Analyzing Use Phase Data in Product Planning of Manufacturing Companies. Procedia CIRP, (107), 2022, S. 1053–1058.

[MWP+22] Meyer, M.; Wiederkehr, I.; Panzner, M.; Koldewey, C.; Dumitrescu, R.: A Reference Process Model for Usage Data-Driven Product Planning, 2022.

[PME+22] Panzner, M.; Meyer, M.; Enzberg, S. von; Dumitrescu, R.: Business-to-Analytics Canvas - Translation of Product Planning-Related Business Use Cases into Concrete Data Analytics Tasks. Procedia CIRP, (109), 2022, S. 580–585.

[MPK+21] Meyer, M.; Panzner, M.; Koldewey, C.; Dumitrescu, R.: Towards identifying data analytics use cases in product planning. Procedia CIRP, (104), 2021, S. 1179–1184.

[MMD+19] Massmann, M.; Meyer, M.; Dumitrescu, R.; Enzberg, S. von; Frank, M.; Koldewey, C.; Kühn, A.; Reinhold, J.: Significance and Challenges of Data-driven Product Generation and Retrofit Planning. Procedia CIRP, (84), 2019, S. 992–997.

Zusammenfassung

Die neuesten technischen Entwicklungen ermöglichen es, während des Betriebs große Datenmengen von cyber-physischen Systemen (CPS) zu erfassen und auszuwerten. Diese Analysen geben Herstellern tiefere Einblicke in die Produktnutzung und Funktionsweise, was zur Ableitung neuer Anforderungen oder Produktideen genutzt werden kann. Allerdings stellt der Prozess der Datenanalyse, insbesondere für kleine und mittlere Unternehmen, eine Herausforderung dar, da oft das nötige gebündelte Expertenwissen in den Bereichen Produkt, Produktplanung und Datenanalyse fehlt. Ein geeigneter Ansatz, um Domänenexperten und Citizen Data Scientists für diese Datenanalysen zu befähigen, ist bisher nicht vorhanden.

Zur *Datenanalyse in der betriebsdatengestützten Produktplanung* wird daher eine *Systematik* erarbeitet. Das Fundament der Systematik bildet ein Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung. Ein Analytics-Baukasten stellt entlang des Prozesses relevante Lösungskomponenten zur Verfügung. Das Vorgehen zur betriebsdatengestützten Produktplanung befähigt seine Anwender mittels verschiedener Werkzeuge, die passenden Lösungskomponenten des Analytics-Baukastens zu bestimmen. Die genannten Elemente der Systematik werden in einem digitalen Lernassistenten prototypisch umgesetzt. Die Systematik wird anhand zweier Fallbeispiele demonstriert und ihr Nutzen in ersten Ansätzen evaluiert.

Summary

The latest technical developments enable large amounts of data from cyber-physical systems (CPS) to be recorded and analysed during usage. These analyses give manufacturers deeper insights into product use and functionality, which can be used to derive new requirements or product ideas. However, the process of analysing data poses a challenge, especially for small and medium-sized companies, as they often lack the necessary bundled expert knowledge in the areas of product, product planning and data analysis. A suitable approach to empower domain experts and citizen data scientists for these data analyses is not yet available.

A *systematic* approach is therefore being developed to *analyse data in use phase data-driven product planning*. The foundation of the system is a reference process for data analysis in data-driven product planning. An analytics toolkit provides relevant solution components along the process. The procedure for data-driven product planning enables its users to determine the appropriate solution components of the analytics toolbox using various tools. The aforementioned elements of the system are prototypically implemented in a digital learning assistant. The systematic approach is demonstrated using two case studies and its benefits are evaluated in initial approaches.

Inhaltsverzeichnis

Seite

Anhang

1 Einleitung

"*Die wahre Entdeckungsreise besteht nicht darin, neue Landschaften zu suchen, sondern darin, neue Augen zu haben.*"
- MARCEL PROUST

Moderne Produkte wie intelligente technische Systeme sammeln während ihres Betriebs wertvolle Daten. Durch den Einsatz von Data Analytics zur Verarbeitung und nutzbaren Interpretation dieser Betriebsdaten eröffnen sich für die Hersteller vielversprechende Möglichkeiten zur Optimierung von Entscheidungsprozessen in der Produktplanung. So erhalten jene Hersteller beispielsweise Erkenntnisse darüber, wie ihre Produkte genutzt werden und wie sie in der Praxis funktionieren. Sie können daraus neue Produktanforderungen sowie erste Produktideen ableiten. Eine datengetriebene Produktplanung erfordert jedoch sehr seltene hochspezialisierte Arbeitskräfte mit gebündeltem Fachwissen über das technische Produkt, Produktplanungsprozesse und Data Analytics. Ein geeigneter Lösungsansatz zur Befähigung von Produktexperten[1] zur Planung und Umsetzung der meist explorativen Datenanalysen fehlt bisher. Daher befasst sich die vorliegende Arbeit mit der Datenanalyse in der betriebsdatengestützten Produktplanung. In Abschnitt 1.1 wird die zugrunde liegende Problematik erläutert. Abschnitt 1.2 legt die Zielsetzung der vorliegenden Arbeit fest. In Abschnitt 1.3 wird das Forschungsdesign und die Vorgehensweise vorgestellt. Dabei legt Abschnitt 1.3.1 den Schwerpunkt auf den Design-Science-Research-Ansatz und 1.3.2 auf das Vorgehen der Arbeit.

1.1 Problematik

Die strategische Produktplanung ist die erste Stufe des Produktentwicklungsprozesses. In dieser Phase werden zukünftige Erfolgspotenziale identifiziert, vielversprechende Produktideen aufgedeckt und das Geschäft geplant [GDE+19]. Das Ergebnis dieser Phase ist ein Portfolio von Produktentwicklungsprojekten [UE16] einschließlich der Anforderungsliste für neue Produkte [BGP+21]. Bestehende Methoden zur Potential- und Ideenfindung, darunter Marktforschungsstudien, werden von Praktikern als zeitaufwändig und teuer bewertet [TH02, OP06]. Als ein vielversprechender Ansatz erscheint hier die Rückführung der Daten und Informationen aus den Produktlebenszyklusphasen, insbesondere der Betriebsphase, in die Produktplanung, um die tatsächlichen Produkteigenschaften im Betrieb zu beurteilen und Verbesserungspotentiale aufzudecken [HSW+18, EM13]. Dies wird insbesondere durch zwei Faktoren ermöglicht: (1) Die Entwicklung von mechatronischen Produkten zu cyber-physischen Systemen (CPS), welche mit Sensorik und Aktorik ausgestattet sowie mit dem Internet verbunden sind [GDE+19]. (2) Fortschritte im

[1] Die Inhalte dieser Arbeit beziehen sich in gleichem Maße auf jedes Geschlecht. Aus Gründen der besseren Lesbarkeit wird die männliche Form verwendet.

Data-Analytics-Bereich, welche die effiziente Auswertung auch von großen Datenmengen über eine Vielzahl an Methoden u.a. aus der Statistik, dem maschinellen Lernen und der künstlichen Intelligenz zulassen [Run12].

CPS, Produktplanung und Data-Analytics bilden das Forschungsgebiet der betriebsdatengestützten Produktplanung [MWK+21]. Die betriebsdatengestützte Produktplanung bietet viele Potentiale, darunter ein besseres Produkt-, Kunden- und Nutzungsverständnis [HJ20, OB13], sowie validere Entscheidungen in der Planung neuer Produktgenerationen [HNU+17].

Die Implementierung und Integration von Data Analytics in die Entscheidungsprozesse der Produktplanung stellt Unternehmen jedoch vor große Herausforderungen [HJ20, WTH+17]. Auf organisatorischer Ebene bestehen die Herausforderungen in einem Mangel an Know-how und qualifizierten Mitarbeitern, der Dominanz von Domänenspezialisten und einem geringeren Bewusstsein für den Nutzen von Data Analytics und Künstliche Intelligenz (KI). Dies ist vor allem für kleine und mittlere Unternehmen (KMU) zutreffend [CGM+16, HB21], denn die erfolgreiche Implementierung einer Datenanalyse erfordert umfassendes Wissen und Verständnis in jedem Schritt des Datenanalyseprozesses (z. B. CRISP-DM).

Entlang des Data-Analytics-Prozesses mit seinen Kernphasen Anwendungsdefinition, Aufbau von Datenverständnis, Modellauswahl und Umsetzung [She00] ergeben sich jeweils einzelne technische Herausforderungen, die es zu berücksichtigen gilt. Im Rahmen der **Anwendungsdefinition** ist es die Aufgabe der Data Scientists zunächst das Data-Analytics-Ziel zu bestimmen und dieses in ein adäquates Data-Analytics-Problem zu übersetzen [CCK+00]. Dieser Übersetzungsprozess erfordert in der Regel ein Überblick über mögliche Ziele der Produktplanung und Kenntnisse darüber, wie diese mit den relevanten Problemen zusammenpassen. Daher müssen Nicht-Experten mit relevanten Faktoren für einen erfolgreichen Übersetzungsprozess unterstützt werden.

Der Schritt **Aufbau von Datenverständnis** umfasst die Identifikation und Sammlung von Betriebsdatenquellen sowie die Beschreibung der Daten, um sie besser zu verstehen und wichtige Erkenntnisse für die anschließende Methodenauswahl zu gewinnen. Problematisch für die Datensammlung ist, dass oftmals über die vorliegenden Daten an den vielen verschiedenen Stellen im Unternehmen keine übersichtlichen Informationen bestehen [OB13, WTH+17]. Eine systematische Identifikation muss an dieser Stelle möglich gemacht werden. Auf der anderen Seite gibt es Herausforderungen in der Heterogenität der Datenmerkmale (z. B. [NAL21, RLS05, MBS+14]). Verschiedene Typologien beschreiben die Struktur, den Inhalt und die Verarbeitung der Daten. Die Bestimmung der Datenqualität erfolgt anhand verschiedener Kriterien [Eng99], was die Aufgabe der Datenbeschreibung zusätzlich komplex macht. Daher besteht der Bedarf für eine einheitliche und vereinfachte Bestimmung von Merkmalen.

Ein entscheidender Aspekt im Analyseprozess ist weiterhin die **Methodenauswahl** und damit die Gestaltung der Kern-Pipeline, d. h. die Zusammenstellung geeigneter

Analysekomponenten, die beispielsweise die Datensätze bereinigen, Daten weiter vorverarbeiten, domänenspezifische Merkmale aus den Daten extrahieren, sie angemessen modellieren und die Modellausgabe nachverarbeiten [RKD17, SHB19]. Es gibt derzeit keine Allzweckmethode, bzw. einen Algorithmus, der alle anderen in einer Vielzahl von Problemklassen und Domänen übertrifft [HKN+09, SHB19]. Die Auswahl dieser Komponenten hängt im Wesentlichen vom Anwendungsziel und den verfügbaren Daten und deren Eigenschaften ab [BS95, NYO+19]. Es ist wichtig, dass der gesamte Prozess von der Anwendungsfalldefinition über das Datenverständnis bis hin zur Modellevaluation ganzheitlich betrachtet und bestimmt wird.

Zuletzt ist die **Umsetzung** zu nennen, wo insbesondere die Toolauswahl als eine relevante Herausforderung hervorsticht, da die enorme Vielzahl von Tools und Bibliotheken die Bestimmung der passenden Werkzeuge für definierte Pipelines erschwert [NS23]. Eine systematische Auswahlunterstützung ist daher notwendig.

Um diese organisatorischen und auch technischen Herausforderungen zu adressieren, besteht einer der Hebel in der Demokratisierung von Data Science durch automatisiertes maschinelles Lernen, Meta-Lernen und No-Code-Tools [Gug23, TT23, VRW+23] . Aber auch Schulungen und Lernkonzepte gehören zu den Maßnahmen, um einen breiteren Zugang zu den Praktiken und Werkzeugen der Datenwissenschaft zu ermöglichen und auch Nicht-Experten oder Domänenexperten in die Lage zu versetzen, die mit Analytics verbundenen Aufgaben zu bewältigen und zu Citizen Data Scientists zu machen. Insbesondere das Thema Training und Lernen, also die fundierte Befähigung ist von großer Bedeutung und unerlässlich, um das Risiko von Fehlschlägen zu vermeiden [BS22-ol, MEK+, MEK+22]. Richtlinien, Best Practices und Vorlagen, die Fragen beantworten und es den Data-Analytics-Anfängern ermöglichen, kontinuierlich zu lernen, können besonders hilfreich sein [BS22-ol]. Vor diesem Hintergrund bedarf es einer Systematik, welche die skizzierten Herausforderungen aufgreift und Produktexperten zur Datenanalyse in der betriebsdatengestützten Produktplanung qualifiziert.

1.2 Zielsetzung

Ziel der Arbeit ist eine *Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung*. Sie soll produzierende Unternehmen befähigen, den Data-Analytics-Prozess für die Produktplanung von der Anwendungsdefinition, über den Aufbau von Datenverständnis hin zur Methoden- und Toolauswahl erfolgreich, bewusst und lernförderlich zu durchlaufen. Das daraus resultierende Konzept soll die Unternehmen schließlich in die konkrete Umsetzung begleiten und beim Aufdecken von neuen Trends und Verbesserungspotenzialen unterstützen. Diese können anschließend interpretiert und für die strategische Produktplanung verwertet werden (nicht Teil dieser Arbeit). Ein Beispiel für das Nutzen von Data Analytics in der strategischen Produktplanung kann die Optimierung einer aktuellen Generation an Haushaltskühlschränken sein, welche, durch die Anbindung ans Internet, Daten im Betrieb Daten sammeln. Mit Hilfe von Data-Analytics-

Verfahren kann beispielsweise das Verhalten der Nutzer untersucht werden. Ein mögliches Ergebnis einer solchen Analyse ist ein allgemeines niedriges Interesse an der Funktion „personalisierte Einstellungen für Lebensmittel“ oder das verstärkte Aufkommen von Serviceanfragen, nachdem die Kühlschränke mit Smart-Home-Systemen verbunden wurden. Im Rahmen einer Verwertung für die Produktplanung können diese Erkenntnisse in Verbesserungsideen wie das Verzichten auf die Personalisierungsfunktion in der nächsten Generation oder einem speziellen Update der Smart-Home-Kopplung resultieren. Die Systematik beabsichtigt in diesem Beispiel folgende Aufgaben zu unterstützen:

- Definition und Tieferlegung der Anwendung (Nutzerverhaltensanalyse)
- Aufbau von Verständnis über die Betriebsdaten der Haushaltskühlschränke (z. B. System- und Servicedaten)
- Auswahl geeigneter Analytics-Methoden für die Nutzerverhaltensanalyse und die System- und Servicedaten (z. B. statistische Kennzahlen)
- Auswahl geeigneter Tools für die Umsetzung der definierten Methoden (z. B. Python-Bibliothek *pandas*)

Die Systematik richtet sich in erster Linie an Analytics-affine Produktexperten, z. B. Produktmanager, welche datengetriebene Entscheidungen in Bezug auf die Weiterentwicklung ihrer Produkte wünschen und in ihrem Team selbst die Datenanalyse vornehmen wollen, bzw. müssen. Aber auch Data-Science-Anfänger sollen angesprochen werden, die einige Vorkenntnisse im Bereich Data-Analytics mitbringen, jedoch noch bei konkret zu beantwortenden Fragestellungen aus einer ggf. für sie fremden Domäne einer Begleitung bedürfen.

1.3 Forschungsdesign und Vorgehensweise

Die vorliegende Arbeit entstand im Rahmen der beiden Forschungsprojekte DizRuPt – Datengestützte Produktplanung sowie KI-Marktplatz. Im vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Konsortialforschungsprojekt DizRuPt wurde ein Instrumentarium für die datengestützte Produktgenerationen- und Retrofitplanung entwickelt. Das vom Bundesministerium für Wirtschaft und Energie (BMWi) geförderte Projekt KI-Marktplatz entwickelte ein Ökosystem für Künstliche Intelligenz in der Produktentstehung. Die in der vorliegenden Arbeit vorgestellten Ergebnisse stellen wesentliche Teilergebnisse der Forschungsprojekte dar. Abschnitt 1.3.1 erläutert die Forschungsmethodik, die dieser Arbeit zugrunde liegt. In Abschnitt 1.3.2 wird die Vorgehensweise in Form der Gliederung der Arbeit vorgestellt.

1.3.1 Der Design-Science-Research-Ansatz

Forschungsmethodisch orientiert sich die vorliegende Arbeit an der Design Science Research Methodology (DSRM) nach PEFFERS ET AL. [PTR+07]. Die DSRM ist eine

Methode der Design Science Research (DSR), welche versucht, neue Mittel für das Handeln in der Welt zu erfinden (Design), um die Realität zu verändern und zu verbessern [VPB17]. Die Design Science (DS), wie sie von SIMON konzeptualisiert wurde, unterstützt ein pragmatisches Forschungsparadigma, das die Schaffung innovativer Artefakte zur Lösung von Problemen der realen Welt fordert [Sim96]. Somit verbindet die Design-Science-Forschung einen Fokus auf das IT-Artefakt mit einer hohen Priorität für die Relevanz im Anwendungsbereich. Die DSRM stellt ein wissenschaftliches Rahmenwerk dar, das den Anspruch erhebt, den Designprozess in der Informationssystemforschung effektiver und effizienter zu gestalten. Ein zentrales Leitprinzip dieser Methode ist die integrative Entwicklung eines besseren Verständnisses einer Aufgabe oder eines Problems sowie einer geeigneten Unterstützungsmethode. Um diesem Anspruch gerecht zu werden, schlagen PEFFERS ET AL. ein sequenzielles und iteratives Vorgehen mit sechs Hauptphasen vor (s. Bild 1-1): 1) Problemidentifikation, 2) Zielformulierung, 3) Design und Entwicklung, 4) Demonstration, 5) Evaluation sowie 6) Kommunikation und Weiterentwicklung. Diese Phasen bilden den Kern des Designprozesses und bieten einen strukturierten Rahmen für die Lösungsentwicklung.

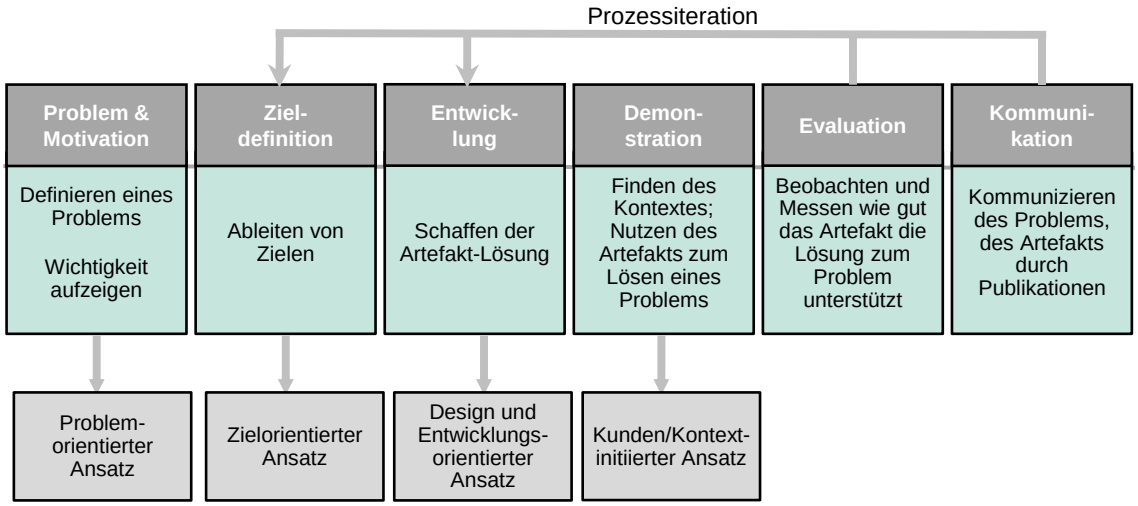

Bild 1-1: DSRM-Prozess [PTR+07]

1) Problemidentifikation und Motivation: In dieser Phase wird ein spezifisches Forschungsproblem definiert und der Wert einer Lösung begründet. PEFFERS ET AL. schlagen vor, das Problem konzeptionell zu zerlegen, um die Komplexität zu erfassen. Die Motivation erfüllt zwei Zwecke: sie begeistert den Forscher und verleiht ihm Entschlossenheit, die Forschung voranzutreiben, und sensibilisiert das Publikum für die Bedeutung der Lösung. Ressourcen wie Kenntnis des Problemzustands und seiner Lösung sind erforderlich [PTR+07].

2) Definition der Ziele für eine Lösung: Basierend auf der Problemidentifikation werden klare, spezifische und erreichbare Ziele für die Lösungsentwicklung festgelegt. Diese dienen als Leitlinien für den Designprozess und legen fest, welche Ergebnisse angestrebt werden. Die Ziele können quantitativ sein, wie die

Verbesserung gegenüber der aktuellen Lösung, oder qualitativ, indem sie beschreiben, wie ein neues Artefakt bisher ungelöste Probleme unterstützen soll. Sie sollten rational aus der Problemspezifikation abgeleitet werden und erfordern Kenntnisse über den aktuellen Stand der Probleme und möglicher Lösungen sowie deren Wirksamkeit [PTR+07].

3) Design und Entwicklung: In dieser Phase werden konzeptionelle Modelle, Konstrukte, Methoden oder andere Artefakte entwickelt, die zur Lösung des identifizierten Problems beitragen. Diese Tätigkeit umfasst die Bestimmung der gewünschten Funktionalität und der Architektur des Artefakts sowie die anschließende Erstellung des eigentlichen Artefakts. Der Designprozess basiert auf vorhandenem Wissen und bewährten Methoden, aber auch auf kreativem Denken und innovativen Ansätzen [PTR+07].

4) Demonstration: Die entwickelte Lösung wird in dieser Phase in der realen Welt demonstriert. Dies kann durch Fallstudien, Simulationen, Experimente oder andere Methoden erfolgen, um die Funktionalität und den Nutzen der Lösung zu validieren. Die Demonstration dient auch dazu, potenzielle Anwender oder Stakeholder von der Wirksamkeit des Artefakts zu überzeugen. Zu den für die Demonstration erforderlichen Ressourcen gehören effektive Kenntnisse über die Verwendung des Artefakts zur Lösung des Problems [PTR+07].

5) Evaluation: Die Evaluation beinhaltet die Bewertung der entwickelten Lösung hinsichtlich ihrer Effektivität, Effizienz und Nützlichkeit. Hier werden verschiedene Methoden und Metriken verwendet, um die Leistung des Artefakts zu messen und zu vergleichen. PEFFERS ET AL. identifizierten acht relevante Typen von Evaluierungsmethoden: Logische Argumente, Expertenevaluation, technische Experimente, Subjekt-basierte Experimente, Action Research, Prototyp, Fallstudie und illustrative Szenarien [PRT+12]. Die Ergebnisse der Evaluation dienen als Grundlage für Verbesserungen oder Weiterentwicklungen des Artefakts (ggf. Rückkehr zu Phase 3) [PTR+07].

6) Kommunikation: Die letzte Phase der DSRM besteht darin, das Problem und seine Bedeutung, die entwickelte Lösung und ihren Nutzen und ihre Effektivität zu kommunizieren und zu verbreiten. Dies kann durch wissenschaftliche Publikationen, Präsentationen auf Konferenzen oder andere Kanäle erfolgen [PTR+07].

Der DSRM-Prozess sieht vier unterschiedliche Forschungseinstiege vor: eine problemzentrierte Initiierung, eine zielorientierte Lösung, eine design- und entwicklungsorientierte Initiierung und eine Kunden/Kontext initiierte Lösung [PTR+07]. Der Problemorientierte Ansatz beginnt mit der ersten Phase. Ein Vorgehen in dieser Reihenfolge ist sinnvoll, wenn die Idee für die Forschung aus der Beobachtung des Problems oder aus einem Vorschlag für zukünftige Forschung aus einem früheren Projekt resultiert. Eine zielorientierte Lösung, die mit Phase 2 beginnt, könnte durch einen Industrie- oder Forschungsbedarf ausgelöst werden, der erfüllt werden kann, indem ein Artefakt entwickelt wird.

Ein design- und entwicklungsorientierter Ansatz würde mit Phase 3 beginnen. Er könnte sich aus der Existenz eines Artefakts ergeben, das noch nicht formal als Lösung für den expliziten Problembereich, in dem es eingesetzt werden soll, konzipiert wurde. Ein solches Artefakt kann z. B. aus einem anderen Forschungsbereich stammen, bereits zur Lösung eines anderen Problems verwendet worden sein, oder als analoge Idee entstanden sein. Zuletzt kann eine vom Kunden/Kontext initiierte Lösung auf der Beobachtung einer praktischen Lösung basieren, die funktioniert hat; sie beginnt mit Phase 4 und führt zu einer Design-Science-Lösung, wenn die Forscher rückwärts arbeiten, um rückwirkend Strenge auf den Prozess anzuwenden. Dies könnte ein Nebenprodukt einer Beratungserfahrung sein.

Die DSRM bietet einen strukturierten Ansatz für die Durchführung von Design Science Research im Feld der Informationssysteme (IS) und Informatik. Es umfasst Grundsätze, Praktiken und Verfahren, die für die Durchführung einer solchen Forschung erforderlich sind. Es bietet ein nominales Prozessmodell für die Durchführung von DS-Forschung und ein mentales Modell für die Präsentation und Evaluierung von DS-Forschung in IS. Außerdem wird die DSRM derzeit als der dominierende Ansatz angesehen [FVP22].

1.3.2 Einordnung und Vorgehen in der Arbeit

Die DSRM wurde im Rahmen der vorliegenden Arbeit als problemorientierter Ansatz verfolgt. Bild 1-2 stellt die Zusammenhänge der DSRM-Phasen und der Gliederung der Arbeit dar.

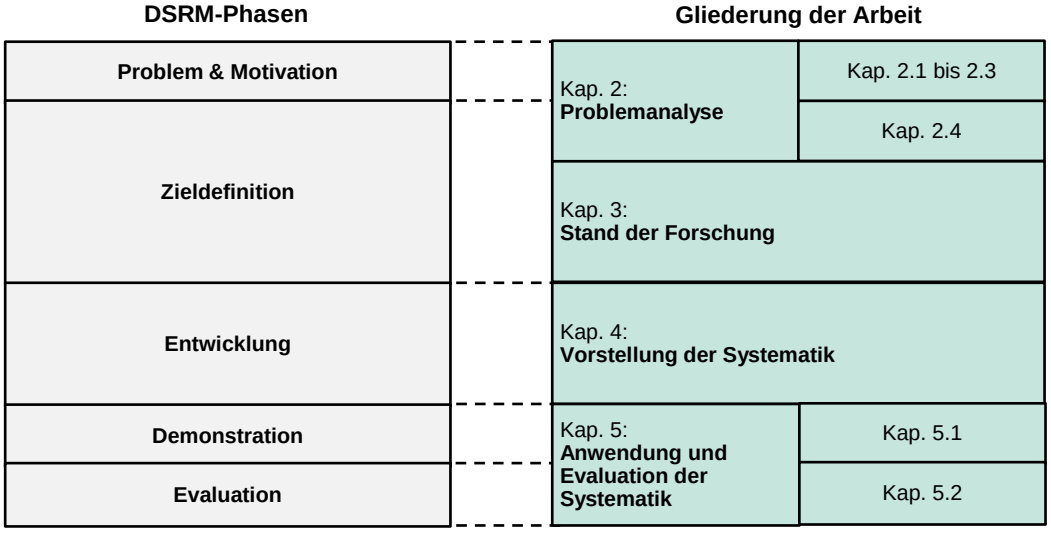

Bild 1-2: DSRM-Phasen und korrespondierende Gliederungsstruktur

1) **Problemidentifikation und Motivation**: Mithilfe einer umfassenden Literaturrecherche wurden die Probleme und Herausforderungen im Datenanalyseprozess innerhalb der betriebsdatengestützten Produktplanung identifiziert und strukturiert. Nach der Einführung in die Problematik in **Kapitel 1** wird diese Phase in

Kapitel 2 thematisiert. Dafür werden zunächst die Begriffe definiert, die für das Verständnis der Arbeit zentral sind. Anschließend wird das Forschungsfeld der betriebsdatengestützten Produktplanung anhand seiner einzelnen Disziplinen aufgespannt und dessen Potenziale und Herausforderungen vorgestellt. Auf dieser Grundlage werden die domänenspezifischen Probleme entlang des allgemeinen Data-Analytics-Prozesses analysiert. Es wird auch ein Blick auf verschiedene automatisierte Ansätze zur Demokratisierung von Data Analytics und ihre Grenzen geworfen.

2) **Definition der Ziele**: Aus diesen Untersuchungsergebnissen resultieren die Ziele an die Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung am Ende von **Kapitel 2**. Die Ziele werden erneut entlang des Data-Analytics-Prozesses strukturiert. **Kapitel 3** beschreibt im Anschluss den gegenwärtigen Stand der Forschung. Dabei werden zu Beginn allgemeine konkurrierende Ansätze vorgestellt und bewertet. Aufbauend werden Ansätze zu jeder betrachteten Data-Analytics-Prozess-Phase beschrieben. Die Bewertung der diskutierten Ansätze anhand der definierten Ziele führt zum Handlungsbedarf für die zu entwickelnde Systematik.

3) **Design und Entwicklung**: In **Kapitel 4** wird die Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung vorgestellt. Zunächst wird ein Überblick über die Systematik und ihre Bestandteile gegeben. Danach werden die einzelnen Bestandteile und ihre Entwicklung detailliert erklärt. Zur Entwicklung der Systematik wird ein Mixed-Methods-Ansatz genutzt, d.h. eine Kombination von qualitativen und quantitativen Forschungsmethoden, um verschiedene Perspektiven abzudecken. Tabelle 1-1 gibt einen Überblick über die entwickelten Artefakte der Systematik und die jeweils verwendeten Forschungsmethoden.

4) **Demonstration**: Die Demonstration der entwickelten Systematik erfolgt im ersten Teil von **Kapitel 5** in Form von zwei Fallstudien.

5) **Evaluation**: Im zweiten Teil von **Kapitel 5** folgt die Evaluation, welche auch anhand eines Prototypens in Gestalt eines digitalen Assistenten für die Datenanalyse in der betriebsdatengestützten Produktplanung und Nutzerinterviews umgesetzt wird.

6) **Kommunikation**: Die Kommunikation der Probleme und der Systematik erfolgte während der Erstellung der gesamten Arbeit im Rahmen von verschiedenen Konferenz- und Journalbeiträgen.

Kapitel 6 schließt die Arbeit mit einer Zusammenfassung, den Limitationen und einem Ausblick auf zukünftige Forschungsarbeiten ab.

Tabelle 1-1: Übersicht der neu entwickelten Artefakte und der eingesetzten Forschungsmethoden

(neue) Artefakte	Forschungsmethode	Abschnitt
Referenzprozess für die betriebsdatengestützte Produktplanung	Eigenes Vorgehen auf Basis von verschiedenen Ansätzen zur Entwicklung von Referenzprozessen sowie Literaturrecherche	4.2
Data-Analytics-Baukasten für die betriebsdatengestützte Produktplanung	Systematische Literaturrecherche + Methode zur Taxonomieentwicklung + Umfrage	4.3
Business-to-Analytics-Canvas	Action-Design-Research-Ansatz	4.4.1
Python-Tool-Übersicht	Bewertung nach dem 4-Augen-Prinzip	4.4.4
Baukasten-Wissensbasis	Literatur + Expertenbewertung	4.4.3 bzw. 4.5.1
Prototyp für ein digitales Assistenz- und Lerntool	Eigenes Vorgehen	4.5.1

2 Problemanalyse

Das Ziel der vorliegenden Arbeit ist eine Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung. Für deren Entwicklung ist ein umfassendes Problemverständnis erforderlich. Abschnitt 2.1 dient zunächst der Begriffsabgrenzung. In Abschnitt 2.2 werden die drei Forschungsfelder vorgestellt, in deren Schnittpunkt die betriebsdatengestützte Produktplanung entsteht. In Abschnitt 2.3 wird darauf aufbauend der Datenanalyseprozess für die betriebsdatengestützte Produktplanung im Detail betrachtet sowie Herausforderungen und wichtige Ausgestaltungsfaktoren beschrieben. Abschließend werden in Abschnitt 2.4 die daraus abgeleiteten Ziele an eine Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung präsentiert.

2.1 Begriffsdefinitionen und -abgrenzung

Voraussetzung für eine effektive wissenschaftliche Auseinandersetzung mit der Betriebsdatenanalyse in der strategischen Produktplanung ist ein einheitliches Begriffsverständnis. Daher werden in den nachfolgenden Abschnitten 2.1.1 bis 2.1.3 die relevantesten Begriffe der vorliegenden Arbeit definiert und voneinander abgegrenzt.

2.1.1 Datenanalyse

Der Begriff „Datenanalyse“, bzw. „Data Analytics“, wurde erstmals Anfang der 2000er populär. Er definiert zusammenfassend den Prozess des Zugriffs auf große Datenmengen aus verschiedenen Quellen sowie deren Zusammenführung und Analyse mit dem Ziel Wissen aus Daten zu extrahieren, um historische Ereignisse zu verstehen und zukünftige Ereignisse vorherzusagen [Tya03]. RUNKLER beschreibt Data Analytics als ein sehr interdisziplinäres Feld, das Aspekte aus vielen anderen wissenschaftlichen Disziplinen wie Statistik, maschinelles Lernen, Mustererkennung, Systemtheorie, Operations Research oder künstliche Intelligenz übernommen hat [Run20].

Von diesen Disziplinen sind vor allem zwei hervorzuheben: Künstliche Intelligenz (KI) und Maschinelles Lernen (ML). Für KI existieren zahlreiche Definitionen und eine breit akzeptierte Definition scheint nicht zu existieren [CMX+21, Pei19]. Eingeführt wurde der Begriff bereits 1955 durch JOHN MCCARTHY [MMR+06]. Eine Definition nach WINSTON lautet: „Künstliche Intelligenz ist die Untersuchung von Berechnungsverfahren, die es ermöglichen, wahrzunehmen, zu schlussfolgern und zu handeln“ [Win92]. Die europäische Kommission definiert KI wie folgt: „KI bezieht sich auf Systeme, die intelligentes Verhalten zeigen, indem sie ihre Umgebung analysieren und - mit einem gewissen Maß an Autonomie - Maßnahmen ergreifen, um bestimmte Ziele zu erreichen.“ [Phi20]. Diese Definition schließt die Technologien aus, die zur Erreichung von Intelligenz verwendet werden. Ein Ansatz zur Implementierung von Intelligenz in Maschinen ist ML – „Maschinen entwickeln, die lernen können Aufgaben zu bewältigen“ [RRC19]. Genauer formuliert bezieht sich der Begriff auf ein Computerprogramm, das lernen kann, ein

Verhalten zu zeigen, das nicht ausdrücklich vom Autor des Programms programmiert wurde. Vielmehr ist es in der Lage, ein Verhalten zu zeigen, dessen sich der Autor möglicherweise gar nicht bewusst ist. Dieses Verhalten wird erlernt basierend auf folgenden drei Faktoren: (1) Daten, die von dem Programm verbraucht werden, (2) eine Metrik, die den Fehler oder eine Form des Abstands zwischen dem aktuellen Verhalten und dem idealen Verhalten quantifiziert, (3) ein Feedback-Mechanismus, der den quantifizierten Fehler nutzt, um das Programm so zu steuern, dass es bei den nachfolgenden Ereignissen ein besseres Verhalten zeigt [Jos20]. Etwas knapper beschreiben WROBEL ET AL. ML als „[…] ein Forschungsgebiet, das sich mit der computergestützten Modellierung und Realisierung von Lernphänomenen beschäftigt [Gör00].

Nach MOHAGHEGH liegt der Unterschied zwischen statistischen Ansätzen und Ansätzen der KI und des ML darin, dass erstere von einer Reihe vorgegebener Gleichungen (Satz an Hypothesen) ausgehen, z. B. einfache oder multivariable lineare Regression oder nichtlineare Regression, die genau definiert sind (z. B. logarithmisch, exponentiell) [Sha19-ol]. Aufbauend wird versucht, die am besten geeignete Gleichung zu finden, die zu den gesammelten Daten passt. KI und ML beginnen nicht mit vorgegebenen Modellen oder Gleichungen. Ihr Merkmal ist die Entdeckung von Mustern aus vorhandenen Daten und in großen Mengen von Variablen. Das Endergebnis lässt sich meist nicht in einer oder wenigen Gleichungen zusammenfassen. Eine weitere Abgrenzung zwischen Statistik und ML wagen BZDOK ET AL. und fassen zusammen, dass die Statistik aus einer Stichprobe Rückschlüsse auf die Grundgesamtheit zieht und ML verallgemeinerbare Vorhersagemuster findet [BAK18].

Häufig wird in der Literatur im Zusammenhang mit Data Analytics eine Ergänzung vorgenommen und von Big Data Analytics geschrieben (z. B. in [NT17, Ro11, ZSB15]). Das soll betonen, dass Data Analytics auf „Big Data“ angewendet wird – Daten, die in erster Linie durch (1) großes Datenvolumen, (2) hohe Geschwindigkeit der Datenentstehung und (3) hohe Heterogenität charakterisiert sind [Lan01]. Durch diese Eigenschaften werden teilweise neue Ansätze zur Wissensgenerierung erforderlich, die unter dem Begriff Big Data Analytics gebündelt werden.

Stark verwandt mit Data Analytics ist der Begriff Data Mining. Dieser etablierte sich schon deutlich früher in den 1980ern [Lov83]. Wörtlich übersetzt bedeutet Data Mining „Schürfen oder Graben in Daten“. BISSANTZ und HAGEDORN verstehen darunter die Extraktion implizit vorhandenen, nicht trivialen und nützlichen Wissens aus großen, dynamischen, relativ komplex strukturierten Datenbeständen [BH09]. Die Ergebnisse ermöglichen die Identifikation von Mustern in den Daten, daher wird Data Mining auch als Datenmustererkennung übersetzt [AN00]. FAYYAD ET AL. definieren Data Mining als die Anwendung spezifischer Algorithmen zur Extraktion von Mustern aus Daten [FPS96b]. Hier steht also die Erkennung von neuen Mustern oder Hypothesen im Vordergrund. Gemäß der Definition von MOHAGHEDH (s. o.) wurde dies erst durch die Ansätze der Künstlichen Intelligenz und des Maschinellen Lernens ermöglicht. Hierbei ist es nicht erforderlich, wie es in der Statistik üblich ist, zunächst Hypothesen über Datenzusammenhänge

aufzustellen. TAN ET. AL. verdeutlichen hingegen, dass das Feld Data Mining zunächst auf den Methoden und Algorithmen aufbaute, die die Forscher zuvor eingesetzt hatten [TSK16]. Dies waren Methoden wie Sampling, Schätzung, Hypothesentesten, Suchalgorithmen, Lerntheorien aus der KI, Mustererkennung und ML. Data Mining integrierte schnell auch Ideen aus anderen Bereichen, wie Optimierung, Signalverarbeitung, Visualisierung und Informationstheorie und erweiterte diese.

Ein weiterer Begriff, der häufig im Zusammenhang mit den aufgeführten Begriffen fällt, ist Data Science. Data Science bedeutet wörtlich übersetzt „Datenwissenschaft". STADELMANN ET AL. beschreiben diese Wissenschaft weiterhin als einzigartige Verbindung von Prinzipien und Methoden aus den Bereichen Analytics, Technik, Unternehmertum und Kommunikation, die darauf abzielt, aus den Daten selbst einen Wert zu schaffen [SSB+13]. Die DATA SCIENCE ASSOCIATION definiert diese Wissenschaft wie folgt: „Data Science means the scientific study of the creation, validation and transformation of data to create meaning. […] Data science uses scientific principles to get meaning from data and uses machine learning and algorithms to manage and extract actionable, valuable intelligence from large data sets [Dat23-ol].

Ein Versuch zur bildlichen Abgrenzung der verschiedenen Begrifflichkeiten ist in Bild 2-1 abgebildet.

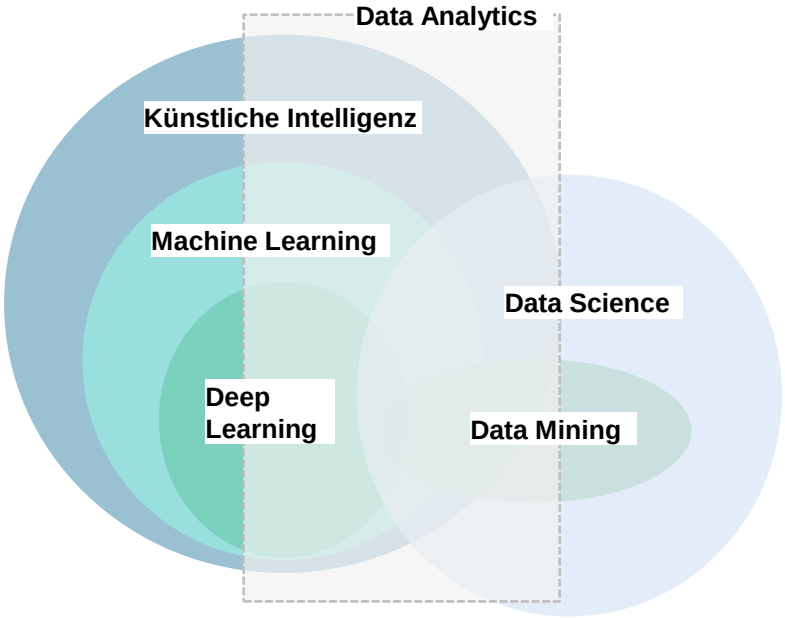

Bild 2-1: Abgrenzung der Begriffe Data Science, Data Mining, KI, ML und Deep Learning in Anlehnung an KULIN ET AL. [KKM+20]

Die Ausführungen zeigen deutlich, dass der Data-Analytics-Begriff viele verschiedene Disziplinen wie Maschinelles Lernen, Statistik und Data Mining vereint, welche wiederrum Schnittpunkte mit den genannten Disziplinen aufweisen. Im Vordergrund steht die Aufdeckung von Wissen. Dies kann durch viele unterschiedliche Techniken ermöglicht werden. Im Rahmen der vorliegenden Arbeit soll Data Analytics daher als der gesamte

Prozess der Wissensaufdeckung aus großen Datenmengen durch Nutzung von vielseitigen Techniken aus Statistik, ML, Data Mining und KI verstanden werden. Der Prozess wird innerhalb der Data Science erforscht.

2.1.2 Data-Analytics-Pipeline

Data Science basiert auf Prinzipien und Techniken, die für die Durchführung von Aufdeckungsaktivitäten erforderlich sind. In der Regel erfolgt die Definition jener in Form einer Abfolge von Schritten, einem so genannten Workflow oder einer Pipeline [Bra19].

Der Begriff "Pipeline" wurde von GARLAN über seine Boxen-und-Linien Diagramme und erklärende Prosa geprägt, die Softwareentwicklern beim Design und der verständlichen Beschreibung komplexer Systeme unterstützen [Gar00].

Unter einer Data-Science-Pipeline (DS-Pipeline) verstehen BISWAS ET AL. eine Reihe von sequentiellen Verarbeitungsstufen, die mit Daten interagieren und integraler Bestandteil vieler Softwaresysteme sind [BWR22]. Die sequenziellen Data-Science-Stufen sind in einer DS-Pipeline organisiert, wobei Daten von einer Stufe der Pipeline zur nächsten fließen können. Eine DS-Pipeline kann aus mehreren Stufen und Verbindungen zwischen ihnen bestehen. Die Stufen sind definiert, um bestimmte Aufgaben wie z. B. Datenvorverarbeitung, Feature Engineering und Modellierung auszuführen, und sind über Input-Output-Beziehungen mit anderen Stufen verbunden [ABB+19]. Die Definition der Stufen ist allerdings in der Literatur nicht einheitlich gewählt. Die Terminologie variiert je nach Anwendungskontext und Schwerpunkt.

Oft werden Data-Science-Pipelines nur im Hinblick auf die Analyse betrachtet, z. B. die Verfahren des maschinellen Lernens, die zur Ableitung der Ergebnisse im Rahmen der Datenanalyse (Schritt 3) verwendet werden. So sprechen z. B. HAPKE und NELSON von einer ML-Pipeline, die mit der Aufnahme neuer Trainingsdaten beginnt und mit dem Erhalt einer Art von Feedback über die Leistung des neu trainierten Modells endet [HN20]. Sie enthält die Schritte Datenvorverarbeitung, Modell-Training, Modell-Analyse und Deployment[2] des Modells. Wieder andere verstehen unter Pipelines den gesamten Prozess von Data-Science-Projekten, einschließlich der Datenerfassung, Verwaltung, Analyse und dem Ziehen von Schlussfolgerungen [NDB+19, OBU+16].

Während der Begriff „Pipeline“ häufig also eher als automatisierter oder zu automatisierender Fluss von Daten zu verstehen ist, unterstützt der „Workflow“ eher die menschlichen Arbeitsschritte in einem Data-Science-Projekt. So spricht die NATIONAL SCIENCE FOUNDATION (NSF) von dem „Workflow“ als zentrales Organisationsprinzip einer Data-Science-Aktivität und Vorgehen [BRH+18]. Demnach ist ein Data-Science-Workflow

[2] Deployment: Im Kontext von Machine Learning meint Modell-Deployment die Integration des trainierten Modells in die Software-Infrastruktur, die für seine Ausführung erforderlich ist. In dieser Phase geht es auch um Fragen der Modellpflege und -aktualisierung [ACP22].

eine durchgängige Abfolge von Arbeitsschritten von der Datenerfassung bis zur Veröffentlichung der Ergebnisinterpretation in Form eines Datenproduktes. Welche Schritte umfasst werden, wird in der Literatur nicht eindeutig beschrieben. Häufig unterscheiden sich Workflows in Details und werden in unterschiedlicher Granularität repräsentiert. So führt die NSF die folgenden Schritte auf:

1. Entdeckung, Erfassung, Aufbereitung und Speicherung von Rohdaten
2. Auswahl und Erwerb von kuratierten Daten aus Datenbeständen für die Datenanalyse
3. Analyse der Daten
4. Interpretation der Ergebnisse
5. Veröffentlichung der Ergebnisse und ggf. Operationalisierung der Pipeline für kontinuierliche Analysen

Arbeitsschritte im Sinne eines Workflows definiert auch das wohl bekannteste Prozessmodell für Data Mining, CRISP-DM [She00]. Dieses umfasst sechs Schritte: unternehmerisches Verständnis, Datenverständnis, Datenaufbereitung, Modellierung, Auswertung und Bereitstellung. Diese Stufen führen im Allgemeinen verschiedene Aufgaben durch. Die üblichen Projektaufgaben und -schritte kombinieren verschiedene Methoden, um eine Pipeline zu formen [vV15, Juo17].

Aus diesen Definitionen ergibt sich für die vorliegende Arbeit folgendes Verständnis von einer Data-Analytics-Pipeline: Sie beschreiben das Resultat eines ganzheitlichen Data-Analytics-Prozesses in Form einer sinnvollen Abfolge von (automatisierbaren) Aufgaben für die Datenanalyse, welche durch den Einsatz verschiedener Methoden, Techniken und Algorithmen (zusammengefasst unter dem generischen Begriff „Bausteine“ oder „Komponenten“) ausgestaltet werden.

2.1.3 Daten, Datenquellen und Betriebsdaten

Daten ist im Wörtlichen die Pluralbildung von dem Begriff *Datum*, das als Lehnwort aus dem Lateinischen zurückgeht auf *datum* ‚gegeben‘ (Partizip Perfekt Passiv zu lat. *dare* ‚geben‘) bzw. substantiviert, das Gegebene. Der Begriff bezeichnet laut dem Duden (Zahlen)**werte**, die z. B. durch Beobachtungen, Messungen und statistische Erhebungen gewonnen werden [Dud-ol]. Oft werden Daten spezifischer definiert als unverarbeitete oder unorganisierte Fakten, als "rohe Fakten" oder als "Rohmaterial von Informationen" [AG99, EH04]. Es gibt weitere, leicht abweichende Definitionen. Die deutsche Norm DIN 44300 definiert Daten als aus Zeichen bestehende Einheiten, die Informationen darstellen und im Wesentlichen zur Verarbeitung bestimmt sind. Die internationale Norm (ISO/IEC 2382-1) fasst den Begriff Daten weiter und definiert ihn als: "eine reinterpretierbare Darstellung von Informationen in einer formalisierten Weise, die für die Kommunikation, Interpretation oder Verarbeitung geeignet ist". In [Int17] heißt es, dass ein

Kontext benötigt wird, damit Daten sinnvoll sind. Der Begriff Kontext kann als das Repräsentationssystem von Daten betrachtet werden. Diese Definitionen zeigen die starke Verflechtung mit dem Begriff *Information*, was eine Abgrenzung dieser beiden Begriffe erfordert. Um die Zusammenhänge darzustellen, eignen sich das Modell der Daten – Information – Wissen – Weisheit (DIKW) Hierarchie oder Pyramide [Row07, Bat05] und die Wissenstreppe nach NORTH ET AL. [NM18]. Demnach können Daten als eine Sammlung von objektiven, meist messbaren oder beobachtbaren Fakten in einer Rohform wie Zahlen oder Symbolen verstanden werden. Die Verarbeitung von Daten und ihre Einordnung in einen relevanten Kontext ergeben Informationen. Informationen zeichnen sich dadurch aus, dass die zugrundeliegenden Daten einen bestimmten Bezug erhalten. Informationen können in Wissen umgewandelt werden, indem verstanden wird, wie Daten zur Erreichung relevanter Ziele beitragen. Alle Informationen sind so organisiert, dass sie nützlich sind und Erkenntnisse daraus gezogen werden können. Im letzten Schritt kann Weisheit erreicht werden, wenn das "Warum" hinter all den Mustern verstanden wird.

Im Kontext mit dem Begriff „Daten“ stehen weitere Begriffe wie *Datensatz*, *Datenobjekt* und *Datenquelle*, die teilweise auch synonym verstanden werden. Eine eindeutige Definition für den Begriff *Datensatz* gibt es nicht. RENEAR ET AL. vergleichen verschiedene in der Literatur zu findende Definitionen und kommen zu dem Schluss, dass die meisten Definitionen vier Merkmale aufweisen: Gruppierung von Daten, Inhalt, Zusammenhang und Zweck [RSW10]. Bild 2-2 zeigt diese Eigenschaften in der Übersicht.

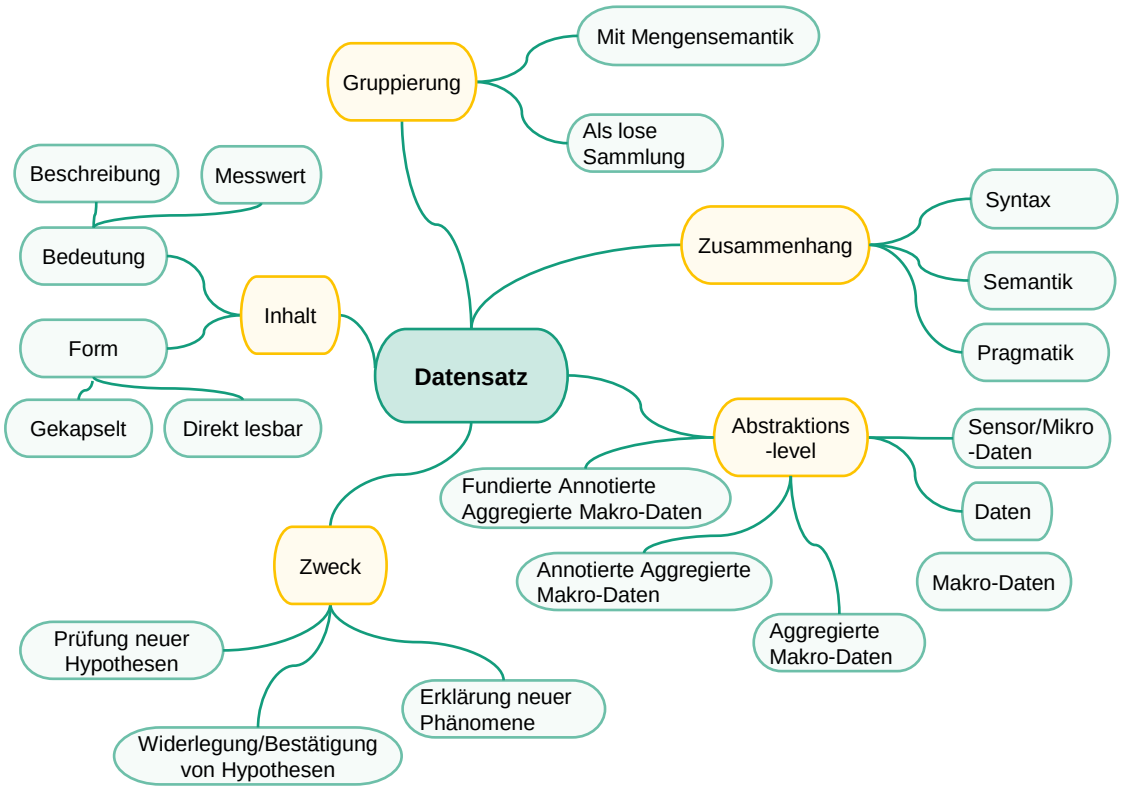

Bild 2-2: Eigenschaften der Definitionen von Datensatz [RSW10]

TAN ET AL. definieren Datensätze als Sammlungen von Datenobjekten [TSK16]. Andere Begriffe für Datenobjekte sind z. B. Einträge, Punkte, Vektoren, Muster, Events, Fälle und Beobachtung. Datenobjekte werden durch eine Anzahl an Variablen beschrieben [Sau19]. Variable beschreibt dem Duden zufolge eine veränderliche Größe. Sie ist demnach eine Eigenschaft eines Objekts, die variieren kann, entweder von einem Objekt zum anderen oder von einem Zeitpunkt zu einem anderen. Beispielsweise variiert die Augenfarbe von Person zu Person, während die Temperatur eines Objekts über die Zeit hinweg variiert. Synonyme hierfür sind die Begriffe Attribute, Features oder Merkmale [TSK16]. Die Variablen sind in einem strukturierten Tabellenformat üblicherweise die Spalten. In den Zeilen stehen die Datenobjekte. Es gibt auch Daten, die sich nicht (so einfach) in das Format der Tabelle pressen lassen, wie z. B. Textdokumente oder Bilder. Diese nicht tabellarisierten Daten werden auch als unstrukturierte Daten bezeichnet, Daten in Tabellenform als strukturiert [Sau19].

In dem Kontext ist ebenfalls der Begriff Datenquelle zu nennen. Die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) definiert diesen als einen spezifischen Datensatz, Metadatensatz, eine Datenbank oder Metadaten-Speicherort, von dem aus Daten oder Metadaten[3] verfügbar sind. Gemäß sind Datenquellen der Ort, an denen die Daten erzeugt werden. Damit können Datenquellen sehr divers sein, wie eine Exceldatei, eine Datenbank oder ein Sensor. Gleichzeitig ist eine Datenquelle der physische oder digitale Ort, an dem die betrachteten Daten in Form einer Datentabelle, eines Datenobjekts oder eines anderen Speicherformats gespeichert werden [Gom-ol]. Die Datenquellen können je nach Anwendung oder Bereich unterschiedlich sein. Hardware wie Eingabegeräte und Sensoren nutzen die Umgebung als primäre Datenquelle. Daten wie Temperatur und Druck der Flüssigkeit werden regelmäßig von Sensoren erfasst und in einer Datenbank gespeichert, die dann zur primären Datenquelle für eine andere Computeranwendung wird, die diese Daten verarbeitet und darstellt.

Im Folgenden werden Daten als Synonym zu Datensatz, also als eine Sammlung von aufgezeichneten Datenobjekten verstanden, welche potenziell Informationen in einem bestimmten Kontext oder für einen bestimmten Zweck liefern. Sie entstammen verschiedenen Datenquellen.

Betriebsdaten:

Beim Begriff Betriebsdaten tritt zusätzlich der Begriff *Betrieb* in Erscheinung. Dieser steht im Zusammenhang mit dem intrinsischen Produktlebenszyklus, welcher die produkteigenen Lebenszyklus-Phasen der ersten Idee bis zum Recycling des Produkts beleuchtet [GK12, Gau00]. Für diesen Produktlebenszyklus existieren verschiedene Modelle, die sich in der Benennung und der Anzahl der Lebenszyklusphasen leicht unterscheiden. Zum Beispiel nennen GAUSEMEIER und KOKOSCHKA die Phasen Strategische

[3] Metadaten ("Daten über Daten") beziehen sich auf strukturierte Daten, die zur Beschreibung und Spezifizierung von Fakten über ein Informationsobjekt verwendet werden können [DMS+05].

Produktplanung, Produktentwicklung, Produktionssystementwicklung, Fertigung, Distribution, Nutzung und Rücknahme/Entsorgung [GK12]. EIGNER und STELZER führen die Phasen Anforderungen, Produktplanung, Entwicklung, Prozessplanung, Produktion, Betrieb und Recycling auf [ES09]. Der Betrieb ist somit eine Phase im intrinsischen Produktlebenszyklus. In dieser Phase wird das Produkt betrieben bzw. genutzt.

So, wie es verschiedene Bezeichnungen für die Phase gibt, existieren auch unterschiedliche Begriffe für Betriebsdaten, die mehr oder weniger synonym verwendet werden. So sprechen KAMMERL ET AL. von Nutzungsdaten (engl. product usage data), welche Daten umfassen, die in der Nutzungsphase des Produkts (Daten in Bezug auf Performance und Wartung sowie aktueller Standort des Produkts) entstehen sowie Daten, die sich auf den Benutzer beziehen, wie z. B. persönliche Daten und Browser-Historie [KNH+16]. Um stärker den zeitlichen Aspekt der Nutzungsphase hervorzuheben, nutzen WILBERG ET AL. den Begriff Nutzungsphasen-Daten (engl. use phase data) [WFH+18]. Genauer sprechen sie dabei von Daten, die während der Nutzungsphase von dem Produkt selbst (z. B. durch Sensoren oder Mikroprozessoren) oder von zugehörigen Produktservices (z. B. Wartung, Kundenservice oder mobile Anwendungen) generiert werden [WTH+17]. HORVATH UND EDDAHAB nutzen den Begriff Daten der Lebensmitte (engl. „Middle-of-life Data“), welche durch Vor-Ort-Beobachtungen und Befragungen der Nutzer, durch das Studium von Fehlerprotokollen und Wartungsberichten oder aus einschlägigen Webressourcen wie sozialen Medien und Nutzerforen zusammengetragen werden können [HA19]. EDLER hingegen verwendet den Begriff Felddaten und versteht darunter Daten, die im Zusammenhang mit der Nutzung eines Produktes im Feld oder der Inanspruchnahme einer Dienstleistung durch den Kunden anfallen [Edl01]. Darunter fasst er nicht nur Fehler- und Störmeldungen, sondern auch Nutzungsdaten wie Betriebsstoffverbräuche, sowie Rückmeldungen von Nutzern oder Bedienern, beispielsweise in Form von Verbesserungsvorschlägen, die als Anforderungen für die nächste Produktgeneration dienen können. KREUTZER baut darauf auf und bezeichnet die Daten, die nach dem Verkaufszeitpunkt des Produkts generiert werden, als Felddaten [Kre19].

Es wird deutlich, dass in der Literatur unterschiedlich enge Definitionen des Betriebsdatenbegriffs vorgenommen werden. Während einige Autoren eher die produkteigenen Daten fokussieren, fassen andere auch vom Nutzer generierte Daten wie Social-Media-Daten unter den Begriff. Der Betriebsdatenbegriff soll folglich Daten bezeichnen, die in der Betriebsphase des Produkts von dem Produkt selbst, einem zugehörigen Service oder seinen Nutzern generiert und gesammelt werden.

2.2 Von der strategischen zur betriebsdatengestützten Produktplanung

Die vorliegende Arbeit basiert auf der Zusammenführung dreier eigenständiger Forschungsbereiche: die strategische Produktplanung, die Digitalisierung von Produkten und Data Analytics. Um das Thema der Arbeit besser zu verstehen, werden die verschiedenen

Forschungsfelder näher beleuchtet. Dazu wird in Abschnitt 2.2.1 die Bedeutung der strategischen Produktplanung erläutert sowie deren Weiterentwicklung beschrieben. Abschnitt 2.2.2 stellt die betriebsdatengestützte Produktplanung als Schnittpunkt zwischen den drei genannten Disziplinen vor. Abschließend widmet sich Abschnitt 2.2.3 den übergreifenden Potentialen und Herausforderungen des neuen Forschungsgebiets.

2.2.1 Die strategische Produktplanung

Neue komplexe Produkte entstehen in einem Prozess, der sich von einer Produkt- und Geschäftsidee bis zum Serienanlauf erstreckt [GDE+19]. Die Aktivitäten der Produktentstehung lassen sich über das Referenzmodell der strategischen Produktplanung und integrativen Entwicklung von Marktleistungen, auch das 4-Zyklen Modell der Produktentstehung genannt, in vier Hauptaufgabenbereiche strukturieren: Strategische Produktplanung, Produktentwicklung, Dienstleistungsentwicklung und Produktionssystementwicklung (s. Bild 2-3).

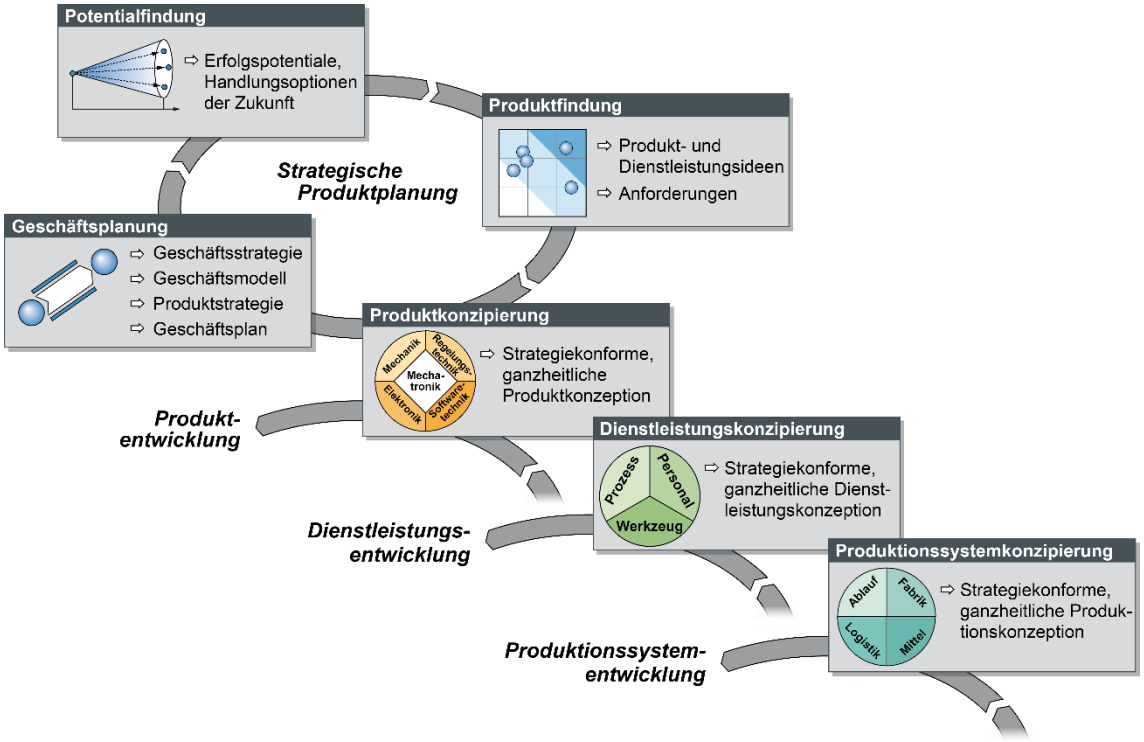

Bild 2-3: Referenzmodell der Strategischen Planung und integrativen Entwicklung von Marktleistungen mit nur angedeuteten Entwicklungszyklen [GDE+19]

Erster Zyklus – Strategische Produktplanung: Die strategische Produktplanung umfasst den Prozess von der Ermittlung der Erfolgspotentiale der Zukunft bis zur Erstellung von Entwicklungsaufträgen. Sie adressiert die Aufgabenbereiche: Potentialfindung, Produktfindung und Geschäftsplanung. Das Ziel der Potentialfindung ist die Identifizierung von Erfolgspotenzialen für die Zukunft sowie die relevanten Geschäftsoptionen. Kern der Produktfindung ist es, neue Produktideen zu finden, die die erkannten Erfolgspotenziale

ausschöpfen. Die Geschäftsplanung beschäftigt sich zunächst mit der Unternehmensstrategie, d.h. sie beschäftigt sich mit der Frage, welche Marktsegmente abgedeckt werden sollen. Darauf aufbauend wird die Produktstrategie und der Geschäftsplan ausgearbeitet.

Zweiter Zyklus – Produktentwicklung: Die Produktentwicklung umfasst das konzeptionelle Produktdesign, das domänenspezifische Design und die entsprechende Aufbereitung sowie die Integration der Ergebnisse aus den einzelnen Domänen zu einer Gesamtlösung.

Dritter Zyklus – Dienstleistungsentwicklung: Das Ziel der Dienstleistungsentwicklung ist die Umsetzung einer Dienstleistungsidee in eine Marktleistung. Zyklisch zu durchlaufen sind dabei die Aufgaben Dienstleistungskonzipierung, Dienstleistungsplanung und Dienstleistungsintegration.

Vierter Zyklus – Produktionssystementwicklung: Gegenstand der Produktionssystementwicklung sind die Bereiche Produktionssystemkonzipierung, Arbeitsplanung und Produktionssystemintegration. Bei der Arbeitsplanung sind die Arbeitsablaufplanung, Arbeitsstättenplanung, Materialflussplanung und Arbeitsmittelplanung zu betrachten.

Dieses Modell verdeutlicht die starke Vernetzung der verschiedenen Aufgabenbereiche. Die Produktplanung steht in enger Beziehung mit der Produktentwicklung und der Produktionssystementwicklung; sie sind aufeinander abgestimmt durchzuführen.

Der Produktplanung wird eine große Bedeutung beigemessen, da die Ergebnisse der strategischen Produktplanung der Produktentwicklung vorgeben, wie die Produkte auszugestalten und effizient zu entwickeln sind [BGP+21, UE16]. Dadurch entscheiden die in dieser Phase verorteten Aktivitäten über den Erfolg eines neuen Produktes [Rob19]. Neben ihrem Einfluss auf den Produkterfolg hat die strategische Produktplanung auch einen erheblichen Effekt auf die anderen Phasen des Produktlebenszyklus [IAG+15, VAE+19]. HERSTATT und VERWORN sowie MACHAC und STEINER betonen, dass die Produktplanung zwar die größte Gestaltungsfreiheit aufweist, gleichzeitig jedoch dort die größten Unsicherheiten auftreten. In der Produktplanung werden viele Entscheidungen getroffen, z. B. hinsichtlich der im Produkt einzusetzenden Technologien. Damit bestimmt sie auch maßgeblich die Kosten für die weiteren Phasen des Produktlebenszyklus [HV14, JF14]. Änderungen in späteren Lebenszyklusphasen führen zu höheren Kosten. Daher ist es wichtig, dass frühzeitig kundenseitiger Input in die Produktentwicklung einfließt, damit die Produkte den Kunden einen optimalen Mehrwert liefern [CL04].

Wie zuvor erläutert, liegt der Schwerpunkt der Produktplanung auf der Identifikation, Generierung und Bewertung von Potenzialen und Ideen für Produkte. Dabei stehen in der Praxis aber weniger Ideen für ganz neue Produkte im Vordergrund, sondern mehr bestehende Produkte und ihre **Verbesserungspotentiale** [ABU+14, CEK04]. Ursächlich ist hier das hohe Entwicklungsrisiko von Neuentwicklungen [ARB+17]. Um Verbesserungspotenziale, die auch auf den Markt und die Kunden ausgerichtet sind, zu identifizieren, stehen den Unternehmen z. B. Methoden der Marktforschung zur Verfügung. Diese

werden jedoch häufig als sehr aufwändig, teuer und lückenhaft empfunden [TH02]. Alternativ werden Entscheidungen erfahrungsbasiert getroffen [Die14, QLZ+23]. Dabei stellen gerade faktenbasierte Produktdefinitionen vor der eigentlichen (Weiter-)Entwicklung einen wichtigen Erfolgstreiber dar [Rob19]. Demzufolge wünschen sich Unternehmen und Führungskräfte häufig, Entscheidungen in der Produktplanung stärker auf Daten zu basieren und die Produktnutzung im Feld zu analysieren [LPW+16, Har18-ol].

2.2.2 Die betriebsdatengestützte Produktplanung

Um Entscheidungen in der Produktplanung datenbasiert zu unterstützen und Verbesserungspotenziale zu identifizieren, bietet sich die Rückkopplung von Daten und Informationen aus den Produktlebenszyklusphasen, insbesondere der Betriebsphase, in die **Produktplanung** an [HSW+18]. Gerade die Daten aus der Betriebsphase bieten einen Mehrwert, da sie hier meist in großer Menge vorliegen [BB15]. Sie zeigen z. B. in Form von Nutzer- und Nutzungsdaten auf, wie das Produkt genutzt wird und wie es sich im Betrieb verhält. Sie ermöglichen so eine Beurteilung der Produkte und ihrer Eigenschaften [Edl01, Kre19].

Ermöglicht werden die Bereitstellung und Nutzung der Betriebsdaten durch die **Digitalisierung** und Entwicklung der Produkte von mechatronischen zu **cyber-physischen Systemen**. Mechatronische Systeme bestehen aus mechanischen, elektronischen und informationstechnischen Subsystemen; ihre Grundstruktur wird aus dem Grundsystem, Sensorik, der Informationsverarbeitung und Aktorik gebildet [Ise08]. Während sich mechatronische Systeme durch eine nicht-kognitive Regulierung auszeichnen, sind intelligente technische Systeme durch eine Erweiterung um eine assoziative und kognitive Regulierung lernfähig [Ise08, GTD13]. Wenn solche intelligenten technischen Systeme zusätzlich über das Internet miteinander kommunizieren und kooperieren können, handelt es sich um cyber-physische Systeme (CPS) [Bro10, BG19]. Über die Verbindung zu Netzinfrastrukturen können die physischen Vorgänge überwacht und gesteuert werden. Außerdem sind sie dazu fähig, über das Internet bereitgestellte Daten und Services zu nutzen. Sensorik und Aktorik kann so Daten aus der physischen Welt über Netzwerke an eine verarbeitende Software melden. Mithilfe von Technologien wie dem Cloud Computing und darauf aufsetzenden IT-Plattformen können die bereitgestellten, häufig sehr großen Datenmengen gespeichert und für die weitere Verarbeitung bereitgestellt werden [CEK+16, ALW+18]. Sobald die Daten gesammelt wurden, wird die sinnvolle Nutzung der Daten zu einem der wichtigsten Aspekte. Hier kommt **Data Analytics** ins Spiel. Data Analytics verspricht verbesserte Entscheidungsfindung, erhöhte Effizienz, Kosteneinsparungen und gesteigerte Wettbewerbsfähigkeit [PH15, Tya03]. Die Datenanalyse kann mit einer Vielzahl von Techniken durchgeführt werden: (1) die deskriptive Analyse, bei der Daten zusammengefasst und beschrieben werden, (2) die diagnostische Analyse, bei der Muster und Beziehungen in Daten identifiziert werden, (3) die prädiktive Analyse, bei der Daten verwendet werden, um Vorhersagen über zukünftige Ereignisse zu treffen, und

(4) die präskriptive Analyse, bei der Daten verwendet werden, um Maßnahmen oder Entscheidungen vorzuschlagen [SSE+14].

Strategische Produktplanung, Digitalisierung und Data Analytics, die für sich alle eigene Forschungsfelder darstellen, bilden die Grundidee der betriebsdatengestützten Produktplanung. MEYER ET AL. nennen drei Grundprinzipien, die sich durch die Schnittmengen der Felder ergeben (s. Bild 2-4), und fassen damit die wichtigsten Faktoren der betriebsdatengestützten Produktplanung zusammen [DK23].

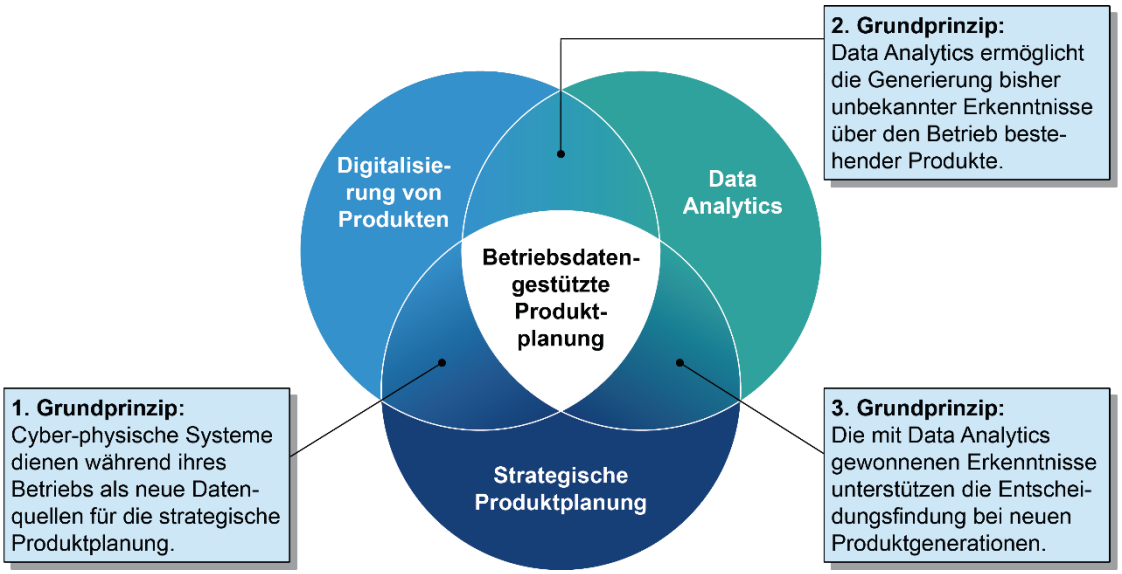

Bild 2-4: Grundprinzipien der betriebsdatengestützten Produktplanung [DK23]

1. Grundprinzip: *Cyber-physische Systeme dienen während ihres Betriebs als neue Datenquellen für die strategische Produktplanung.*

Kundenbefragungen und -gespräche sowie Marktforschungen sind traditionelle Informationsquellen in der strategischen Produktplanung, um Verbesserungspotenziale für Produkte zu identifizieren (vgl. 2.2.1). Der Wandel von mechatronischen Systemen hin zu CPS sowie Fortschritte bei digitalen Technologien ermöglichen die Ergänzung der traditionellen Datenquellen um Betriebsdaten. Diese liefern zusätzliche Informationen über die tatsächliche Nutzung des Produkts und das reale Produktverhalten.

2. Grundprinzip: *Data Analytics ermöglicht die Generierung bisher unbekannter Erkenntnisse über den Betrieb bestehender Produkte.*

All diese zusätzlichen Informationen bergen wertvolles Wissen. Dieses wird jedoch erst durch die geeignete Datenanalyse aufgedeckt. Data Analytics bietet verschiedenen Ansätze, um die unterschiedlichen Quellen und ggf. großen Datenmengen zu analysieren sowie Zusammenhänge, Klassifikationen und Gruppierungen zu erhalten, die den Ausgangspunkt für neue Potentiale darstellen.

3. Grundprinzip: *Die mit Data Analytics gewonnenen Erkenntnisse unterstützen die Entscheidungsfindung bei neuen Produktgenerationen.*

Die Ergebnisse der Datenanalyse können genutzt werden, um die Unsicherheiten der strategischen Produktplanung zu minimieren als auch um neue Funktionen und verbesserte Eigenschaften abzuleiten.

2.2.3 Potentiale und Herausforderungen

Gegenüber der traditionellen strategischen Produktplanung und üblichen Methoden zur Identifizierung von Verbesserungspotentialen bietet die betriebsdatengestützte Produktplanung eine Vielzahl an Potentialen. Gleichzeitig entstehen in den Schnittmengen der Forschungsfelder neue Herausforderungen.

Potentiale

Im Gegensatz zu anderen auf Experimenten beruhenden Techniken können die notwendigen Daten während des normalen Betriebs gesammelt werden [CHT09]. Die Bereitstellung wertvoller Informationen über Aktivitäten in der Nutzungsphase des Produkts, über die Anwender (z. B. Maschinenbediener) und deren Nutzung von Produkten ermöglicht es, wertvolle und objektive Informationen über den aktuellen Zustand, Störungen und Fehlbedienungen zu erhalten [Die14]. Mit diesen Informationen ist es möglich, die nächste Produktgeneration zu entwickeln, die besser an die tatsächlichen Bedürfnisse angepasst und weniger fehleranfällig ist. Gleichzeitig besteht auch die Möglichkeit, die im Einsatz befindenden Maschinen nachzurüsten [PJ20]. Die Objektivität des datengetriebenen Ansatzes bewirkt eine größere Aussagekraft von Untersuchungsergebnissen und bietet so ein besseres Verständnis über die Nutzung der eigenen Produkte. Darüber hinaus erlaubt die Datenintegration Kundenbedürfnisse miteinzubeziehen, welche für die strategische Planung von großer Bedeutung sind [BK11]. Durch automatisierte Analysen ist ein schnelles Feedback möglich. Informationen können zielgerichteter zwischen Benutzern und Entwicklern ausgetauscht werden [Moz17]. Dies ermöglicht nicht zuletzt die Verkürzung von Entwicklungszeiten und damit eine verkürzte Time-to-market; es führt ebenso zu Wettbewerbsvorteilen und eine erhöhte Wettbewerbsfähigkeit [HC05, HNU+17].

Neben diesen genannten Aspekten, identifizierten MEYER ET AL. in einer Literaturstudie Vorteile auf drei Ebenen: Analyse, Prozess und Geschäft [MWK+21].

Analyse: (1.1) Der Einsatz von Analyseverfahren wie Data Mining ermöglicht es, in den Daten verborgene Informationen zu finden, die manuell nicht ermittelt werden könnten. Dies gilt insbesondere dann, wenn viele Produkte und Tausende von Sensormessungen im Laufe der Zeit verglichen werden. (1.2) Die Daten führen zu einem besseren Verständnis des Produkts im Betrieb. Die betriebsdatengestützte Produktplanung nutzt diese Erkenntnisse, um das Produktdesign zu verbessern und so Ausfälle zu vermeiden, anstatt Daten zu nutzen, um sie vorherzusagen. (1.3) Die Analyse von Nutzungsdaten führt zu einem besseren Verständnis der Kunden- und Nutzerbedürfnisse und liefert mit höherer Wahrscheinlichkeit bessere Ergebnisse als traditionelle Ansätze. Die Daten können

genutzt werden, um Kundenverhalten und -präferenzen zu analysieren und Kundensegmente abzuleiten. Für diese können individuelle, zukünftige (Holmström Olsson und Bosch, 2013), unausgesprochene und latente Bedürfnisse identifiziert werden. (1.4) Darüber hinaus helfen quantitative Produktnutzungsdaten bei der Kontextualisierung und Bewertung qualitativer und subjektiver Daten.

Prozess: (2.1) Die Integration von Nutzungsdaten in den Entwicklungsprozess trägt erheblich zur Verbesserung der Kunden- und Nutzereinbindung bei, was zu einer besseren Zusammenarbeit führt. (2.2) Feedback-Daten ermöglichen eine kontinuierliche Anforderungsanalyse (z. B. durch die kontinuierliche Live-Erhebung von Kundenbedürfnissen) und eliminieren momentane Verzerrungen. (2.3) Da die Rückmeldung von Nutzungsdaten einen nutzungszentrierten und faktenbasierten Entscheidungsprozess fördert, kann der Bedarf an Hardware-Prototyping und Feldtests deutlich reduziert werden. (2.4) Folglich ermöglicht eine nutzungsdatenbasierte Produktplanung einen schnelleren Produktentwicklungsprozess.

Geschäft (3.1) Die Integration von Data-Analytics-Methoden in die Produktplanung ermöglicht daten- und faktenbasierte Entscheidungen bei gleichzeitiger Reduzierung von annahme- und erfahrungsbasierten Entscheidungen und verbessert damit die Entscheidungsprozesse. (3.2) Die Analyse von Nutzungsdaten fördert ein nutzungszentriertes Produktportfolio. (3.3) Die kontinuierliche Anforderungsanalyse und der schnellere Entwicklungsprozess helfen den Unternehmen, die Häufigkeit der Bereitstellung von Funktionen zu erhöhen und damit auf Erkenntnisse zu reagieren, die im bisherigen Design nicht vorgesehen waren. (3.4) Die Analyse von Nutzungsdaten ermöglicht Unternehmen letztendlich die Schaffung von qualitativ hochwertigen Innovationen. Dies geschieht, indem die Daten neue Ideen mit höherer Geschwindigkeit fördern und die Akzeptanz erleichtern, da sie die Unternehmen offen für Veränderungen hält.

Herausforderungen

Dass die Umsetzung einer betriebsdatengestützten Produktplanung auch viele Herausforderungen für die Unternehmen mit sich bringt, liegt aufgrund der Interdisziplinarität des Forschungsfeldes fast auf der Hand. MEYER ET AL. arbeiteten im Rahmen einer Interviewstudie zentrale Herausforderungen heraus [MFK+22b]. Die Produkte sind beispielsweise meist so komplex, dass einfache Datenanalysen und Schlussfolgerungen nicht ausreichen. Es gibt zum Beispiel eine große Anzahl an möglichen verschiedenen Fehlerfällen, was schnelle Datenanalysen und richtige Schlussfolgerungen erschwert. Dies macht viele Iterationen, Expertenfeedback und das Kennen passender Analytics-Verfahren, wie z. B. Ursache-Wirkungs-Analysen erforderlich. Darüber hinaus haben die Unternehmen sehr häufig zu wenig Erfahrung auf dem Gebiet der Datenanalyse. Einige Unternehmen greifen daher auf externe Experten zurück. Das ist aber aufgrund des notwendigen technischen Produkt- und Gesamtverständnisses nicht immer ausreichend. Daher ist die durchgängige Integration von Produkt- und Domänenverständnis in den Data-Analytics-Prozess notwendig.

Eine weitere Herausforderung liegt darin, dass die Produkte häufig nur geringe Datenmengen zur Verfügung stellen. Dies hat unterschiedliche Gründe: Zum einen haben viele Produkte im Feld zu wenige Sensoren oder es fehlt an einer Netzwerkverbindung, um ausreichend Daten zu sammeln. Zum anderen wollen oder können Kunden dem Hersteller die Daten aus der Betriebsphase nicht bereitstellen. So haben Kunden entweder Angst vor einem Abfluss von Know-How oder der Hersteller ist nicht in der Lage eine sichere Datenverbindung zum Kunden aufzubauen. Im B2B-Sektor kommt erschwerend hinzu, dass meist nur wenige Instanzen eines Produkts verkauft werden oder Produkte sogar ganz individuell für einen Kunden hergestellt werden, sodass Rückschlüsse auf eine ganze Produktgeneration nicht immer möglich sind. Für die Umsetzer der Data-Analytics-Projekte bedeutet der Mangel an Daten, dass sie die verfügbaren und die fehlenden Daten identifizieren müssen, sodass ggf. notwendige Daten akquiriert werden können. Außerdem müssen sie die Analytics-Verfahren auswählen, die die besten Ergebnisse aus den (wenigen) Daten herausziehen und sowohl aggregierte als auch einzelne Datensätze verarbeiten können. Letztendlich bedeuten diese Herausforderungen auch, dass in der Praxis erst wenige Erfahrungen mit der betriebsdatengestützten Produktplanung gemacht wurden, sodass Best-Practices als wichtiger Orientierungspunkt fehlen.

Auffällig ist folglich die Komplexität des Prozesses der Datenanalyse im Rahmen der betriebsdatengestützten Produktplanung, die sich nach MEYER ET AL. insbesondere durch unbekannte Use Cases und Zielstellungen für komplexe Produkte, den „neuen“ und vielseitigen Betriebsdaten sowie den daraus erforderlichen Data-Analytics-Verfahren kennzeichnet. Dafür fehlt es den Unternehmen häufig an Kompetenz und Know-How. Das ist auch das Fazit anderer Studien. COLEMAN ET AL. heben die Schwierigkeiten von KMU bei der Implementierung von Data Analytics hervor [CGM+16]. Die liegen insbesondere in fehlendem Know-How, der Dominanz von Domänenexperten, einem Mangel an eigenen Data-Analytics-Experten, einem generellen Fachkräftemangel und gleichzeitig dem Mangel an nützlichen und erschwinglichen Beratungs- und Analytics-Services und Produkten. Im Industriekontext bzw. Industrial Data Science ist laut BAUER ET AL. der Mangel an Data Scientists noch stärker, da dieser Bereich eine Kombination an Fähigkeiten aus der Informatik, Statistik und Ingenieurswissenschaft erfordert [BSJ+18]. Um diese Lücke zu schließen, setzen Unternehmen verstärkt nutzerfreundliche Analytics-Tools zur Demokratisierung von Data Science ein [Mat19-ol] und lassen ihre Fachexperten auf dem Gebiet der Datenanalyse weiterbilden. GARTNER definiert solche Experten ohne formelle Ausbildung im Bereich Informatik und Statistik als „Citizen Data Scientists“ [Moo17].

HOPKINS und BOOTH fanden heraus, dass Unternehmen jenseits der riesigen Technologieunternehmen neben der Schwierigkeit ML-Talente zu rekrutieren auch fehlende Erklärbarkeit und herausfordernde Stakeholder-Kommunikation beklagen [HB21].

Fazit: In Kapitel 2.2 wurde die hohe Bedeutung der strategischen Produktplanung innerhalb der Produktentstehung deutlich. Insbesondere die kontinuierliche Verbesserung bestehender Produkte stellt eine Voraussetzung für den nachhaltigen Erfolg von Produkten dar und erfordert die Generierung von Verbesserungspotentialen. Die Rückführung von

Betriebsdaten in die Produktplanung stellt eine wertvolle datengetriebene Ergänzung zu traditionellen Ansätzen zum Finden von Produktpotenzialen dar. Das Konzept dahinter ist die betriebsdatengestützte Produktplanung, welche sich aus dem Schnittpunkt der Disziplinen Digitalisierung, strategische Produktplanung und Data Analytics ergibt. Neben vielen Vorteilen bietet dieser Ansatz allerdings auch einige Herausforderungen. Dabei sticht vor allem die hohe Komplexität des Datenanalyseprozesses hervor. Diese fordert viel Know-How und Expertenwissen. Da Analytics-Experten den Unternehmen jedoch nur sehr begrenzt zur Verfügung stehen, bedarf es Ansätzen zur Demokratisierung von Data Analytics und der Ansprache der sogenannten „Citizen Data Scientists“.

2.3 Data Analytics in der betriebsdatengestützten Produktplanung

Im letzten Abschnitt wurde deutlich, dass die Datenanalyse und ihr Prozess Unternehmen vor große Herausforderungen stellt. Um diese im Detail besser zu verstehen, wird in diesem Abschnitt der Data-Analytics-Prozess für die betriebsdatengestützte Produktplanung genau beleuchtet. Dazu werden in Abschnitt 2.3.1 zunächst bekannte Prozessmodelle analysiert. In den Abschnitten 2.3.2 bis 2.3.5 werden die in dieser Arbeit betrachteten Prozessschritte detailliert untersucht.

2.3.1 Aufgaben und Vorgehen im Data Analytics Prozess

BECKER ET AL. nennen als eine Herausforderung, dass viele der Unternehmen die Prozessschritte der Datenanalyse (Sammlung, Aufbereitung, Analyse und Bewertung) nicht umsetzen oder nicht kennen [BUB16]. Dabei existieren eine Vielzahl an Prozessmodellen, welche die verschiedenen Aufgaben in der Datenanalyse in einem sequenziellen, teils iterativen Prozess betrachten (vgl. Abschnitt 2.1).

Das wohl bekannteste Prozessmodell stellt das **CRISP-DM** (Cross-industry standard process for data mining) Modell dar [She00]. Es soll helfen die Interaktionen entlang des komplexen Data-Mining-Prozesses zu verstehen und zu bewältigen [RJ00]. CRISP-DM besteht im Wesentlichen aus sechs iterativen Phasen (s. Bild 2-5): (1) die erste Phase, das Verstehen der Domäne, zielt darauf ab, die Geschäftsziele und -anforderungen zu verstehen und sie in eine konkrete Data-Mining-Problemdefinition umzuwandeln. (2) In der Phase Verstehen der Daten werden die Daten gesammelt und untersucht, um z. B. Datenqualitätsprobleme zu erkennen. (3) Die Datenverarbeitung umfasst Aufgaben wie die Transformation und Bereinigung der Daten, um den finalen Datensatz zu erstellen. (4) Bei der Modellierung werden Modelle erstellt und mit den Daten verfeinert. (5) In der Evaluation müssen die Modelle im Hinblick auf die Geschäftsziele bewertet werden. (6) Die Nutzung bringt ein Modell schließlich in den Betrieb; ein Bericht wird erstellt.

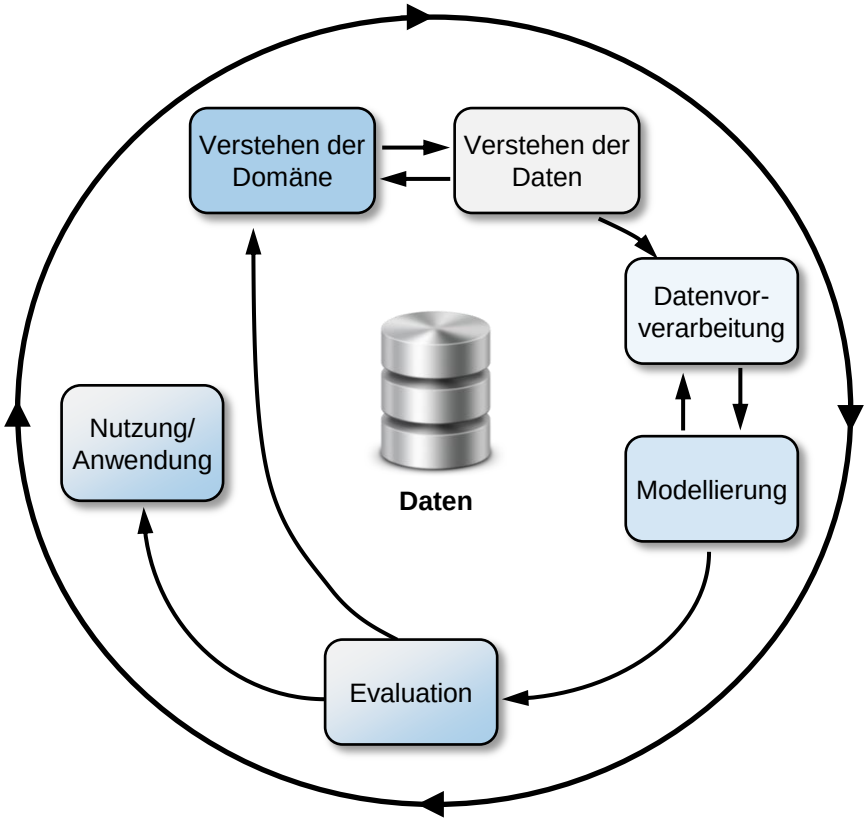

Bild 2-5: CRISP-DM Modell nach [She00]

Das **CRISP-DMME** (Data Mining Methodology for Engineering Applications) Modell stellt eine Erweiterung zum CRISP-DM Standard dar und betrachtet die Datenanalyse innerhalb der Produktionsdomäne [HWS+19]. Es adressiert in erster Linie Ingenieure, um sie bei der Optimierung von Produktions- und Wartungsprozessen zu unterstützen. Der CRISP-DMME-Prozess berücksichtigt nach dem Verstehen der Domäne zusätzlich das technische Verständnis mit dem Ziel die Geschäftsziele in messbare technische Ziele zu transformieren und einen Experimentierplan für die Messungen zur Datenerhebung zu entwickeln. Danach folgt als zusätzliche Erweiterung die technische Realisierung. Diese Aufgabe umfasst das Testen und Auswählen des Messkonzepts und die Durchführung der Experimente. Nach der Evaluation wird die technische Implementierung eingeschoben, welche die Datenerhebungsmethode aus der Phase „technische Realisierung“ in eine lauffähige Infrastruktur überführt.

Ein weiteres Prozessmodell auf Basis von CRISP-DM schlagen MERKELBACH ET AL. vor [MEK+22]. Dieses Modell integriert das Training von Domänenexperten, um sie zu Citizen Data Scientists zu befähigen, um unabhängig Data-Analytics-Anwendungen zu entwickeln und implementieren. Die erste Phase des Prozessmodells ist die Vorbereitungsphase. In dieser Phase wird der Anwendungsfall spezifiziert, die beteiligten Rollen werden definiert und die Planung wird vorgenommen. In der anschließenden Domänen- und Datenverständnis Phase untersuchen die Data Scientists die Daten und gewinnen im Austausch mit den Domänenexperten weiteres Wissen über die Domäne. Die Phasen

Domänenverständnis und Datenverständnis werden folglich in einer Phase zusammengefasst. Die nächste Phase besteht 1) aus dem Design der Evaluation und der Analytics Pipeline und 2) aus einem Data Science Training für die Domänenexperten, die später die Implementierung vornehmen werden. Anschließend findet die Datenaufbereitung statt. In der darauffolgenden Phase wird die zuvor entworfene Auswertungs- und Analysepipeline implementiert. Die letzte Phase ist das Deployment.

Als weiteres bekanntes Prozessmodell ist das **KDD** (Knowledge Discovery in Databases) – Modell zu nennen. FAYYAD et al. definieren den KDD-Prozess als "den nicht trivialen Prozess der Identifizierung gültiger, neuartiger, potenziell nützlicher und letztlich verständlicher Muster in Daten" [FPS96a]. Der Prozess besteht aus neun iterativen Schritten: (1) Beim *Erlernen der Anwendungsdomäne* werden das relevante Domänenwissen und die gewünschten Ziele erfasst. (2) Die *Erstellung eines Zieldatensatzes* zielt auf die Auswahl des zu analysierenden Datensatzes. (3) Bei der *Datenbereinigung und -vorverarbeitung* finden grundlegende Operationen wie Rauschunterdrückung und Ausreißerentfernung statt. (4) Die *Datenreduktion und -projektion* konzentrieren sich auf Aufgaben wie Feature Engineering und Dimensionsreduktion. (5) Bei der Wahl der Data-Mining-Funktion wird der Zweck des gewünschten Modells festgelegt (z. B. Klassifizierung, Regression). (6) Die *Wahl der Data-Mining-Algorithmen* befasst sich mit der Auswahl geeigneter Methoden zur Suche von Mustern in den Daten, z. B. durch den Vergleich verschiedener Modelle und Parameter. (7) *Data Mining* beschreibt die Suche nach Mustern in den Daten. (8) Bei der *Interpretation der Ergebnisse* werden die entdeckten Muster interpretiert und in die Domänensprache übersetzt. (9) Der letzte Schritt, die *Nutzung des entdeckten Wissens*, umfasst die Dokumentation des neuen Wissens, die Berichterstattung und das Ergreifen von Maßnahmen.

Die **VDI/VDE-Richtlinie 3714** stellt einen Standard für die Implementierung und den Betrieb von Big-Data-Anwendungen in der Fertigungsindustrie dar. Sie zielt darauf ab, die zahlreichen Beiträge zu Big Data Analytics in der Fertigungsindustrie zu bündeln und in einem Modell zu vereinheitlichen [VDI22]. Die Richtlinie beschreibt sieben iterative Phasen: (1) In der *Definition*sphase werden die zu beantwortenden oder zu erreichenden Fragestellungen und Ziele festgelegt. (2) Die anschließende *Exploration der Datenlage* zielt darauf ab, die vorhandenen Daten zu beschreiben und zu strukturieren sowie zusätzlich benötigte Daten zu definieren. (3) Beim *Datenmanagement* werden Daten aus verschiedenen Quellen zusammengeführt. (4) Bei der *Modellierung* geht es darum, aus den Daten ein evaluierbares Modell zu erstellen. (5) Anschließend ist eine erste *Bewertung der Ergebnisse* der Datenanalyse im Hinblick auf die Projektziele notwendig. (6) *Implementierung und Rollout* zielen darauf ab, die Big-Data-Anwendung in den Dauerbetrieb zu überführen. (7) Die abschließende Phase *Nachhaltigkeit* befasst sich mit der Projektdokumentation sowie einer Bewertung der wirtschaftlichen, technischen und sozialen Aspekte, um eine nachhaltige Wirkung des Big-Data-Projekts zu gewährleisten.

Ein anderes Modell, welches die Aufgaben nicht in einem Prozess darstellt, sondern in mehreren Ebenen, ist das **4-Ebenen Modell** zur Beschreibung von Analytics Use Cases

(s. Bild 2-6) [RKD17, KJR+18]. Es bündelt verschiedene Komponenten zur Erfüllung der Aufgaben und Vorgehensmodelle. Das Modell unterscheidet zwischen den Ebenen „Analytics Use Case“, „Datenanalyse“, „Daten-Infrastruktur“ und „Datenquellen“ [KJR+18]. Die erste Ebene sieht vor das Problem zu verstehen und die Domäne zu analysieren. Das Ziel und das Ergebnis dieser Problemanalyse sind die Analytics Use Cases, wie bspw. die Durchführung eines Prozess-Monitorings. In der zweiten Ebene steht die Definition der Datenquellen im Vordergrund. Hier wird definiert, welche Sensoren oder Dateisysteme die Daten in welcher Form zur Verfügung stellen. Die Experten aus den jeweiligen Bereichen werden aktiv miteinbezogen. Im Anschluss werden die Daten erhoben, in Daten-Pools gespeichert und vorverarbeitet. Die Ebene Data Analytics umfasst die eigentliche Verarbeitung der Daten, wozu unterschiedliche Vorverarbeitungs- und Modellierungskomponenten wie z. B. die Trendanalyse zum Einsatz kommen.

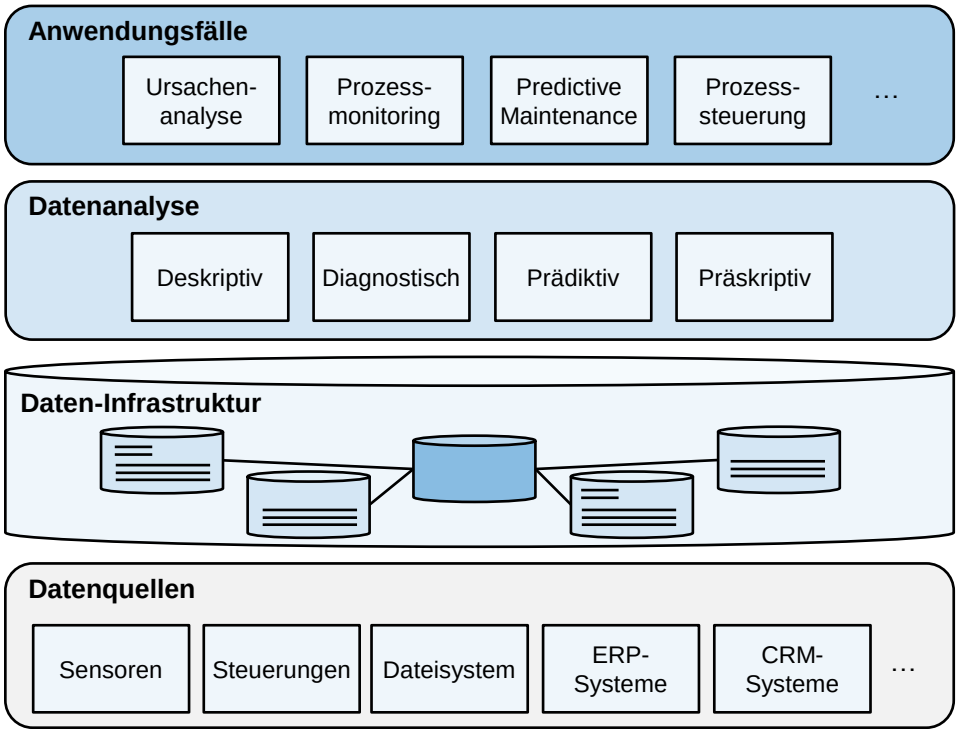

Bild 2-6: 4-Ebenen Modell zur Beschreibung von Analytics Use Cases [KJR+18]

KURGAN und MUSILEK stellen in einer Studie verschiedene Prozessmodelle vergleichend gegenüber [KM06]. Neben dem KDD und dem CRISP-DM-Modell betrachten sie noch vier weitere Modelle im Detail. Dabei stellen sie fest, dass es einige Merkmale gibt, die allen Prozessen gemeinsam sind. Die meisten Prozessmodelle folgen der gleichen Abfolge von Schritten mit ähnlichen Inhalten. Bei der Gegenüberstellung fällt auf, dass viele Ansätze eine Auswahl- bzw. Identifikationsphase für Modelle und Algorithmen vorsehen. Das KDD-Modell und der achtstufige Prozess nach ANAND und BUCHNER [SA98] berücksichtigen dafür einen eigenen Schritt, andere Modelle, wie CRISP-DM enthalten dies als Aufgabe in einer der Phasen.

Fazit: Gegenübergestellt folgen die meisten Modelle einer ähnlichen Abfolge von Schritten. Alle Modelle haben Aktivitäten wie Domänenverständnis, Datenvorverarbeitung, Analyse/Modellierung und Evaluierung gemeinsam. Der Großteil berücksichtigt zusätzlich einen eigenständigen oder zumindest untergliederten Schritt zur Datensammlung und zur Beschreibung bzw. zum Verstehen der Daten. Außerdem erachtet es sich als sinnvoll die Aufgabe der Methodenauswahl und Pipeline-Konzipierung vor der eigentlichen Umsetzung als separaten Schritt zu integrieren, insbesondere im Kontext einer Befähigung von Citizen Data Scientists. Damit ergeben sich folgende domänenübergreifende Kernaufgaben und Arbeitsschritte (vgl. 2.1.2) der Datenanalyse, die zur Strukturierung der folgenden Abschnitte genutzt werden sollen:

1) **Definition der Anwendung**
2) **Aufbau von Datenverständnis** (Datensammlung und -beschreibung)
3) **Methodenauswahl** (Vorverarbeitung, Modellierung und Evaluierung)
4) **Umsetzung**

2.3.2 Definition der Data-Analytics-Anwendung

Bei dem Einsatz von Data Analytics in der Produktplanung ist die Definition von Anwendungsfällen (Use Cases) ein wichtiger Schritt, allerdings ist dieser auch herausfordernd und oft sehr schwierig, weil die potentiellen Anwendungen so divers sind und Interdisziplinarität im Kontext der Produktentstehung erforderlich ist [WTH+17]. Idealerweise soll zudem der Use Case auch einen greifbaren Return-on-Investment und operativen Mehrwert bieten. In diesem Sinne erfolgreiche Data-Analytics-Anwendungen zu definieren, fällt dem Großteil der Unternehmen schwer [FMS19].

Nach CHAPMAN und CLINTON umfasst die erste Phase, das Domänenverständnis, das geschäftliche Ziel (z. B. "Steigerung der Katalogverkäufe an bestehende Kunden") in ein Data-Analytics-Ziel oder eine Data-Analytics-Aufgabe umzuwandeln ("Vorhersage der Anzahl von gekauften Artikeln auf Basis der Kundenverkäufe der letzten drei Jahre, demographischer Informationen und dem Preis des Artikels") [CCK+00]. Das geschäftliche Ziel wird häufig in der Produktplanung durch den Produktmanager definiert. Diesem Beispiel nach besteht das Analytics-Ziel und die Aufgabe in der „Vorhersage“ eines bestimmten Wertes bzw. Objektes unter Nutzung verschiedener Variablen. Außerdem erfolgt eine Konkretisierung der Aufgabenstellung durch Aufführen der notwendigen Variablen. Ziel dieser Phase sind ausspezifizierte Analytics-Use-Cases, die alle relevanten Informationen für den Data Scientist bereithalten, um die weiteren Pipeline-Schritte zu planen und auszugestalten.

In der Literatur werden einige Anwendungen und ihre Analytics-Ziele mit dem übergreifenden Ziel der Produktverbesserung genannt. Beispielsweise untersucht DIENST das geschäftliche Ziel der Produktverbesserung durch Fehlerdiagnose und stellt als Analytics-

Ziel die Bestimmung von Abhängigkeiten für Defekte eines Lagers mit verschiedenen Parametern wie der Umgebungstemperatur auf [Die14]. BENTLAGE und ZENKER verfolgen das Ziel der Produktverbesserung durch das Aufdecken von Schwachstellen [KBZ14]. Das konkretisieren die Autoren weiter in der „Identifikation von relevanten Mustern in z. B. Servicedaten in Form von Regeln“. ZHANG ET AL. sprechen im Kontext der Produktverbesserung von Zielen wie Bewältigung von Marktanforderungen und Verständnis des Kundenverhaltens, welche sie durch „Online Review Data Mining“ erreichen möchten [ZRF18]. Dahinter steckt die Extraktion von wichtigen Produktattribut-Features zusammen mit der Kundenzufriedenheit. Diese Beispiele verdeutlichen, dass einer konkreten Data-Analytics-Anwendung neben einem Ziel, welches den Analysegegenstand konkretisiert, auch das dahinterstehende Data-Analytics-Problem angehört, wie z. B. Musteridentifikation oder Bestimmung von Abhängigkeiten, welches auf den Ansatz zur Lösung hindeutet.

Gemäß CHAPMAN ET AL. umfasst ein Data-Analytics-Ziel die Outputs, die das Erreichen der geschäftlichen Ziele ermöglichen [CCK+00]. Dieses Ziel nimmt also die Datenperspektive ein. Data-Analytics-Ziele werden in der Literatur als deskriptiv, diagnostisch, prädiktiv und präskriptiv kategorisiert [SSE+14]. Dahinter verbergen sich die vier Leistungsklassen von Data Analytics nach STEENSTRUP ET AL. (s. Bild 2-7). Deskriptive Analysen werden in der Regel in der Anfangsphase der Analyse durchgeführt, um ein gutes Verständnis der Daten und der darin enthaltenen Muster zu entwickeln und sich auf das "Was" zu konzentrieren - „Was sind die Verhaltensmuster unserer Kunden“? "Was sind typische Fehlerarten?“

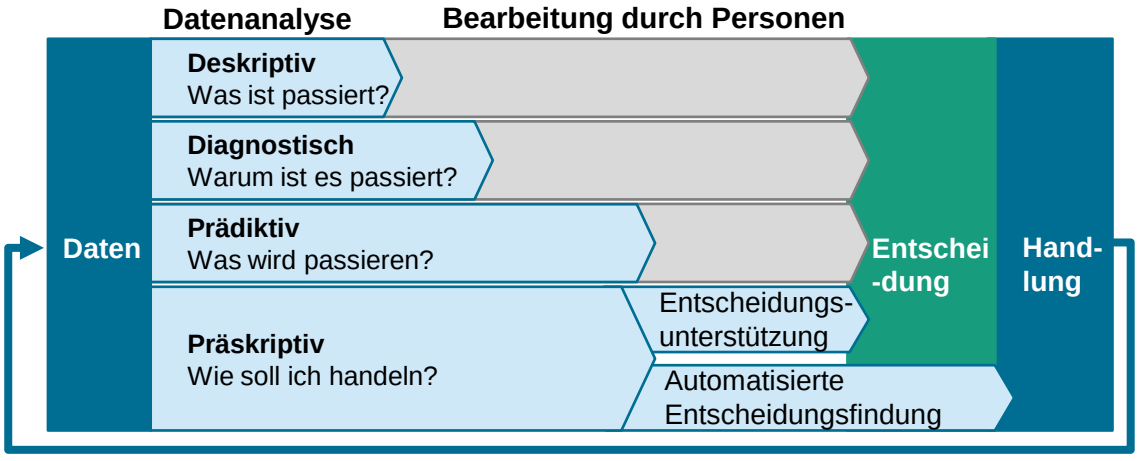

Bild 2-7: Leistungsklassen von Data Analytics in Anlehnung an STEENSTRUP [SSE+14]

Die nächste Stufe ist die diagnostische Analyse, die versucht, das "Warum" (z. B. „Warum kam es zu dem Fehler?“) zu beantworten und mehr Erkenntnisse zu gewinnen, indem sie Korrelationen und Kausalitäten aufdeckt. Bei der prädiktiven Analyse handelt es sich um die Nutzung von Informationen aus der Vergangenheit zur Vorhersage von Ereignissen in der Zukunft, z. B.: „Was wird wahrscheinlich als Nächstes passieren?“. Die präskriptive Analyse schließlich liefert Erkenntnisse darüber, was gemacht werden kann, um

die Wahrscheinlichkeit eines gewünschten Ergebnisses zu erhöhen – „was kann unternommen werden, damit das Produkt fehlerfrei betrieben wird?“. Bei deskriptiven, diagnostischen und prädiktiven Analysen werden Entscheidungen durch Menschen noch selbst herausgearbeitet. Präskriptive Analysen können hingegen entweder eine konkrete Alternative vorschlagen oder diese Entscheidung sogar automatisiert treffen und die sich anschließende Handlung selbstständig einleiten.

Um neue Potenziale und Anforderungen für die Produktplanung ableiten zu können, ist es zunächst notwendig, Wissen über den Ist-Zustand oder Zustände aus der nahen Vergangenheit zu gewinnen. Auch Ursachen für einen Zustand können helfen, konkrete Anforderungen zu definieren. Unter Umständen können auch prädiktive und präskriptive Fragestellungen von Interesse sein, je nachdem, welcher Automatisierungsgrad bei der Identifikation von Potenzialen gewünscht ist oder wie stark der Mensch in die Interpretation der Erkenntnisse einbezogen werden soll.

Weiterführend lassen sich betriebswirtschaftliche bzw. geschäftliche Fragestellungen mindestens einem der folgenden Problemtypen der Datenanalyse zuordnen: Datenbeschreibung, Klassifikation, Regression, Clustering und Abhängigkeitsanalyse (z. B. [CCK+00]).

- **Datenbeschreibung:** Die Datenbeschreibung zielt auf die prägnante Beschreibung von Merkmalen der Daten in elementarer und aggregierter Form ab. Dadurch erhält der Benutzer einen Überblick über die Struktur der Daten. Auch die Datenreduktion ist hier zu nennen. Sie dient dazu große Datensätze mit vielen Variablen anhand weniger prägnanter Merkmale zu beschreiben. Typischerweise ist die Datenbeschreibung ein Teilziel eines Data Analytics Use Cases, meist in den frühen Phasen.
- **Klassifikation:** Die Klassifikation ist die Suche nach Mustern anhand eines Klassifikationsmerkmals, z. B. der Bauteilgröße. Ziel ist das Erlernen der Zuordnung von Beobachtungen zu vorgegebenen Klassen, z. B. Bauteil defekt/ nicht defekt. Diese Klassenlabel sind kategorische Werte und für jede Beobachtung bekannt. Klassifikationsmodelle werden meistens für prädiktive Modellierungen verwendet. Viele Data Analytics-Probleme können in Klassifikationsprobleme transformiert werden, um Klassifikationsverfahren anwenden zu können.
- **Regression:** Die Regression ist der Klassifikation sehr ähnlich. Der einzige Unterscheidungspunkt besteht darin, dass ein Regressionsmodell versucht, eine kontinuierliche Größe vorherzusagen und nicht eine kategorische.
- **Clustering/Segmentierung:** Bei der Clusteranalyse werden Daten in sinnvolle oder nützliche Gruppen (Cluster) eingeteilt. Auf der einen Seite können Cluster für das Verständnis von Daten verwendet werden, indem automatisch potenzielle Klassen gebildet werden (z. B. Segmentierung in kleinere Gruppen mit ähnlichen Merkmalen) und somit Objekte in den Daten strukturieren. Auf der anderen Seite können Cluster für die weitere Verwendung, d. h. als Grundlage für zusätzliche Datenanalysen, genutzt werden. In diesem Fall findet die Clusteranalyse die

repräsentativsten Clusterprototypen (repräsentative Datenobjekte in den Clustern), z. B. durch Zusammenfassung von Daten oder Komprimierung. Diese Datenrepräsentationen können dann in ein weiteres Modell als Input gegeben werden.

- **Abhängigkeitsanalyse:** Bei der Abhängigkeitsanalyse wird ein Modell gesucht, das signifikante Abhängigkeiten (oder Assoziationen) zwischen Datenelementen oder Ereignissen beschreibt. Meist werden Abhängigkeiten zum Verständnis von gegenwärtigen Zusammenhängen verwendet, können aber auch für die prädiktive Modellierung verwendet werden. Vorteil gegenüber der Klassifikation ist, dass man eine Erklärung erhält und nicht nur einen vorhergesagten Wert. Assoziationen sind ein Spezialfall von Abhängigkeiten. Assoziationen beschreiben Affinitäten von Datenelementen (d. h. Datenelemente oder Ereignisse, die häufig zusammen auftreten). Zu unterscheiden sind hier die Begriffe und Ansätze der Korrelation und Kausalität. Korrelation ist eine statistische Technik, die angibt, wie stark Paare von Variablen linear voneinander abhängen. Kausalität geht einen Schritt weiter und bedeutet, dass eine Veränderung in einer Variablen eine Veränderung in einer anderen Variablen verursacht; es besteht eine kausale Beziehung zwischen den Variablen. Im Allgemeinen reicht das Vorhandensein einer Korrelation nicht aus, um auf das Vorhandensein einer kausalen Beziehung zu schließen (d. h. eine Korrelation impliziert keine Kausalität).

Konkrete Beispiele von betriebswirtschaftlichen Zielen und zugehörigen Analytics-Zielen und Problemen, wie zuvor aufgezeigt, sind sehr verteilt und nicht immer klar der strategischen Produktplanung zuzuordnen. Einige Anwendungen werden auch in anderen Kontexten genannt, wie z. B. Fehlerdiagnosen zur Prozessverbesserung (u.a. [HDG+20]), sind aber auf die Ziele der betriebsdatengestützten Produktplanung übertragbar.

Außerdem bleibt unklar, wie die Übersetzung der betriebswirtschaftlichen und Data-Analytics-Ziele in die passenden Data-Analytics-Probleme vorgenommen wird. Dies ist eine wichtige Aufgabe von Data Scientists, allerdings auch eine schwierige [NY18]. An diesem Schnittpunkt müssen sowohl Business- und Produktmanager bzw. Domänenexperten als auch Data Scientists eng zusammenarbeiten und miteinander kommunizieren, um sicherzustellen, dass die Geschäftsziele klar definiert sind und zu den Analytics-Aktivitäten passen. Während Produktmanager oft ein klares Verständnis ihrer strategischen Ziele haben, sind sie sich nicht im Klaren darüber, wie datengetriebene Analysemethoden dabei helfen können, diese Ziele zu erreichen und geeignete Anwendungsfälle und Fragen oder Hypothesen zu formulieren. Auf der anderen Seite mangelt es Data Scientists häufig an einem tieferen Verständnis für die Problemstellung der Anwendungsdomäne und den Kompetenzen, diese als Analytics-Aufgabe zu formalisieren [MPK+21].

Nicht grundlos spricht McKinsey von der Bedeutung einer neuen Rolle, dem Analytics-Übersetzer, welche das technische Fachwissen von Data Engineers und Data Scientists mit dem operativen Domänenwissen verbindet [HLM18]. Analytics-Übersetzer

vermitteln den Datenexperten die geschäftlichen Ziele und stellen sicher, dass die Lösung Erkenntnisse liefert, die das Unternehmen interpretieren und umsetzen kann.

Für die Übersetzung des Ziels in das passende Analytics-Problem können bestimmte Faktoren und benötigte Informationen relevant sein. NALCHIGAR ET AL. z. B. nutzen zur Definition der Geschäftssicht in Analytics-Projekten spezifische Fragestellungen, den gewünschten Output eines Analytics-Ansatzes sowie Metriken zur Performance-Messung und „weiche Ziele“ (z. B. Skalierbarkeit) [NY18]. Es stellt sich folglich die Frage, welche Parameter bedacht werden müssen. Die Bestimmung der geeigneten Analyseziele oder die Übersetzung von Geschäftszielen in Analyseziele, die durch Datenanalyseansätze umgesetzt werden können, bleibt folglich eine wichtige und nicht triviale Aufgabe.

Fazit: Die Definition der Data-Analytics-Anwendung als optimaler Ausgangspunkt des weiteren Prozesses besteht für den Data Scientist im Wesentlichen aus der Bestimmung des Data-Analytics-Ziels, welches das Analyseobjekt konkretisiert (z. B. Fehlerdiagnose, Abhängigkeitsanalyse, Deskription), und der Übersetzung in ein adäquates Data-Analytics-Problem (z. B. Klassifikation) als grober Lösungsansatz. Eine umfassende Sammlung an realisierbaren Analytics-Zielen für Betriebsdaten-Analysen in der strategischen Produktplanung fehlt allerdings, was eine schnelle Ausgestaltung von konkreten Data-Analytics-Anwendungen erschwert. Außerdem ist der Prozess der Übersetzung der Ziele in die Probleme herausfordernd und erfordert in der Regel sowohl Domänenwissen als auch Analytics-Know-How und Erfahrung. Zusätzlich läuft die notwendige Kommunikation zwischen den Beteiligten oftmals unstrukturiert und mit vielen Iterationen ab [Zah20-ol]. Teil des Übersetzungsprozesses sind auch verschiedene Faktoren, die es im Rahmen dieser Arbeit zu untersuchen und in den Übersetzungsprozess zu integrieren gilt.

2.3.3 Aufbau von Datenverständnis (Datensammlung und -beschreibung)

Ziel der **Datensammlung** ist die Zusammenstellung der **relevanten Datenquellen** und Daten(sätze) für die definierte Anwendung (vgl. 2.3.1). Dazu ist im ersten Schritt die Identifizierung der relevanten Quellen und sich darin befindlichen Datensätze notwendig. Wie in 2.1.3 angedeutet sind die Betriebsdaten sehr divers, da sie innerhalb der Betriebsphase eines Produktes an unterschiedlichen Stellen und auf unterschiedliche Weise generiert werden. In der Literatur werden verschiedene Beispiele für Betriebsdaten genannt. KASSNER ET AL. nennen hier CRM (Customer Relationship Management)-, Nutzungs- und Social-Media-Daten [KGM+15]. LI ET AL. erwähnen in dem Kontext Benutzerhandbücher, Produktinformationen, Informationen zum Produktstatus und Informationen zur Nutzungsumgebung [LTC+15]. Weitere Beispiele für Betriebsdaten sind Fehler, Störungen und Defekte sowie Nutzungsinformationen wie Maschinenlaufzeiten und Betriebsmittelverbrauch, Sensor- und Aktordaten, Benutzer- und Systemdaten [Edl01, Kre19]. Außer solchen meist vereinzelten Beispielen existiert keine umfassende strukturierte

Übersicht über Betriebsdaten und potenziell nutzbare Datensätze, welche wertvolle Informationsquellen darstellen.

Um die Diversität und die möglichen Ursprünge der Betriebsdaten im Kontext der Produktplanung besser zu verstehen, wird im Nachfolgenden genauer analysiert, wie Betriebsdaten (und industrielle Daten allgemein) weiterhin klassifiziert werden können.

KURBEL unterteilt Daten in einem produktionsorientierten Unternehmen in organisatorische und technische Betriebsdaten (s. Bild 2-8) [Kur05]. Zu den organisatorischen Betriebsdaten gehören Auftragsdaten und Personaldaten. Technische Betriebsdaten sind Maschinendaten, Werkzeugdaten und Materialdaten. Maschinendaten werden in Produkt- und Prozessdaten unterschieden. Letztere umfassen alle Daten, die während des Betriebs einer Maschine anfallen. Produktdaten beschreiben den Zustand des gefertigten Teils. In Kombination mit den Prozessdaten umfassen sie Informationen über den Produktionsprozess als Ganzes.

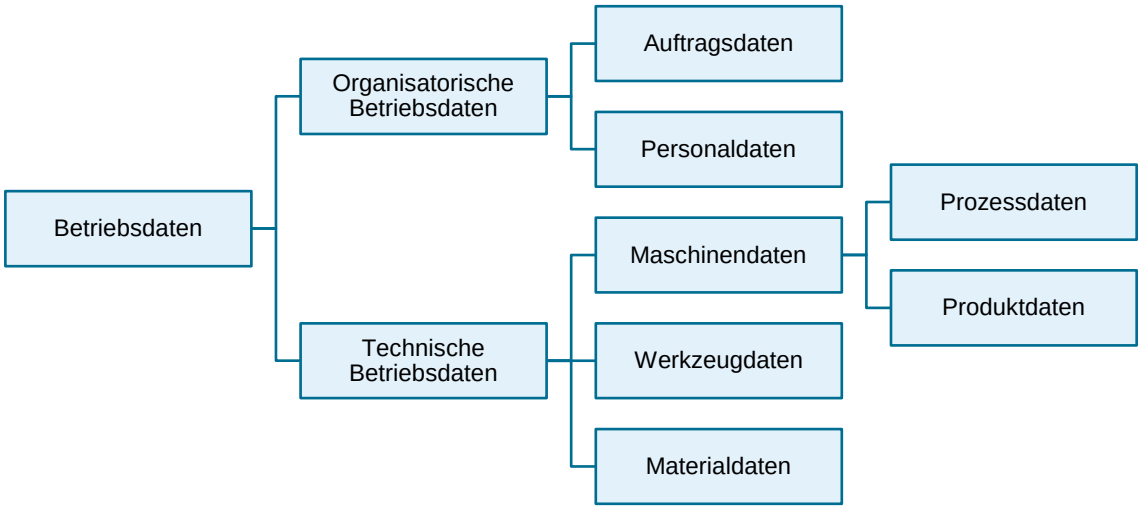

Bild 2-8: Betriebsdaten in produktionsorientierten Unternehmen [Kur05]

Nach SCHÄFER ET AL. lassen sich die Datenquellen je nach Herkunft bzw. Erzeugungsquelle der Daten grob in drei Gruppen einteilen: maschinengenerierte Inhalte, menschengenerierte Inhalte und Geschäftsdaten [SKM+12]. Sensordaten und Logdaten, die gebündelt unter dem Schlagwort Internet of Things (IoT)[4] Einzug gehalten haben und Datendienste sowie Sprach-, Audio- und Videoinformationen zählen dabei zu den von Maschinen generierten Daten. Daten einer Transaktion aus Bereichen wie Supply Chain, Logistik und Produktion oder aus (Online-)Geschäftsvorgängen sowie Stammdaten fallen im Rahmen eines Geschäftsprozesses an und zählen damit zu den Geschäftsdaten. Von

[4] Unter IoT wird grundsätzlich die fortschreitende Verbreitung vernetzter Klein- und Kleinstcomputer in verschiedenen Arbeits- und Lebensbereichen zusammengefasst [Ao09].

Menschen generierte Daten werden durch eine aktive Beteiligung von Menschen erzeugt. Die Generierung dieser Daten wird durch die starke Verbreitung von mobilen Endgeräten (Smartphones, Tablet, Wearables) unterstützt [Sut21]. Bei diesen Daten handelt es sich zu einem bedeutenden Anteil um personenbezogene Daten, zu deren Auslösung die Handlung einer Person benötigt wird. Innerhalb dieser Kategorie kann gemäß dem World Economic Forum eine weitere Unterscheidung in freiwillige Daten, beobachtete Daten und abgeleitete Daten vorgenommen werden [Wor11-ol]. Unter die erste Kategorie fallen z. B. Inhalte in sozialen Netzwerken und Sprach- und Audiodaten sowie E-Mails etc. – Daten, die von Personen selbst erstellt und geteilt wurden. Beobachtete Daten entstehen im Kontext von Handlungen, wie beispielsweise Verhaltensdaten. Diese sind für Unternehmen besonders wertvoll, da sie als Nebenprodukt einer Handlung entstehen und somit eine hohe Authentizität ausweisen. Abgeleitete Daten stammen aus einer Analyse der beiden ersten Kategorien.

RAFFEINER schlägt eine Klassifizierung vor, die zwischen erstellten, erhaltenen, bezahlten und öffentlichen Daten unterscheidet. Diese Einteilung bezieht Drittanbieter ein, über die auch externe Daten erworben werden können [Raf19-ol].

Eine weitere Unterteilung von Daten, die sowohl in der Informatik als auch in der Betriebswirtschaftslehre vorgenommen wird, ist eine Unterscheidung hinsichtlich des Zeitbezugs. Im Hinblick auf diese Datenkonstanz kann zwischen "Stammdaten" und "Transaktionsdaten" unterschieden werden. Unter Stammdaten sind Daten zu verstehen, die über einen längeren Zeitraum konstant bleiben. Dazu gehören z. B. Unternehmensdaten wie Gebäude oder Anlagen. Im Gegensatz zu Stammdaten sind Transaktionsdaten zeitbezogen und ändern sich nach bekannten oder unbekannten Vorgängen. Transaktions- und Bewegungsdaten verweisen in der Regel auf Stammdaten [SB08].

Eine weitere Klassifizierung bietet die Automatisierungspyramide, die die Informationsverarbeitung in einem automatisierten Produktionsunternehmen in Ebenen hierarchisch strukturiert (s. Bild 2-9). In die Ebenen können die verschiedenen IT-Systeme, wie ERP[5], MES[6], SCADA[7] oder SPS[8] Systeme eingeordnet werden [DGK+15]. Über diese Systeme können verschiedene Daten gebündelt zur Verfügung gestellt werden.

---

[5] ERP-Systeme (Enterprise Ressource Planning) sind Softwarelösungen, die die betriebswirtschaftlichen Prozesse, z. B. in Vertrieb, Produktion, Logistik und Personal steuern und auswerten

[6] MES (Manufacturing Execution Systems) sind Produktionsmanagementsysteme, welche zur Produktionsplanung und -steuerung eingesetzt werden.

[7] SCADA (Supervisory Control and Data Acquistion) sind Netzleitsysteme für die Überwachung, Steuerung und Optimierung von Industrieanlagen.

[8] SPS (Speicherprogrammierbare Steuerungen) verarbeiten Eingangssignale von Sensoren und steuern damit Aktoren.

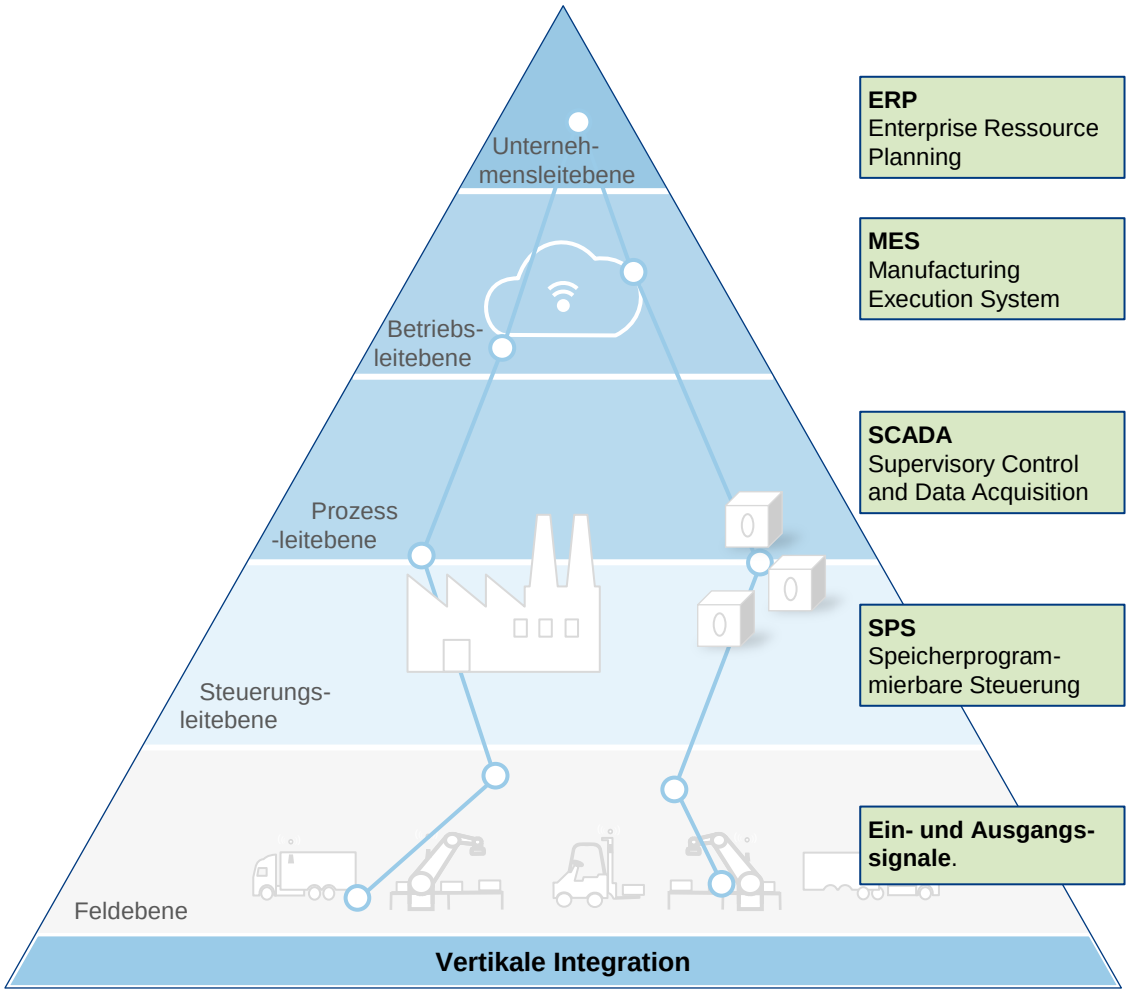

Bild 2-9: Ebenen der Automatisierungspyramide [DGK+15]

Darüber hinaus können Industriedaten und ihre Generierung nach den Funktionsbereichen Service, Marketing, Arbeitsvorbereitung, Entwicklung, Einkauf, Produktion, Qualitätssicherung und IT klassifiziert werden [GPW+14]. Servicedaten sind beispielsweise Kundenanfragen und Reklamationen [Aue04]. Produktionsdaten umfassen z. B. Produktionsabläufe, Maschinenbelegungen und Fertigungsdaten wie Ausschuss und Produktivität der Anlagen. Diese Sicht überschneidet sich teilweise mit der Strukturierung anhand des Produktlebenszyklus. TAO ET AL. schlagen eine weitere Klassifizierung nach Inhalt der Daten vor und unterscheiden zwischen Managementdaten, Gerätedaten, Benutzerdaten, Produktdaten und öffentlichen Daten [TQL18]. Bild 2-10 fasst die vorgestellten Ansätze zur Datenklassifizierung zusammen.

Funktion	Service	Marketing	Arbeits-vorbereitung	Entwicklung	Einkauf	Produktion	Qualitäts-sicherung	IT
Ressource	Sensor/Aktor (Feldebene)		SPS (Kontrollebene)		SCADA/HMI (Prozess-kontrollebene)	MES (Werksleitungsebene)		ERP/CRM (Managementebene)
Konstanz	Stammdaten					Transaktionsdaten		
Zugang	geschaffen			erhalten		öffentlich		bezahlt
Ursprung	Maschinengeneriert			Menschengeneriert			Geschäftsdaten	
PLM	Beginning of Live (BOL)			Middle of Life (MOL)			End of Life (EOL)	
Produkt-lebenszyklus	Produktplanung	Design & Entwicklung		Produktions-planung	Produktion		Nutzung und Support	Wiederverwen-dung & Recycling
Inhalt	Management Daten		Ausrüstungsdaten		Nutzerdaten	Produktdaten		Öffentliche Daten
Sonstiges	organisational					technisch		

Bild 2-10: Ansätze zur Datenklassifizierung

Die Existenz all dieser Klassifizierungen zeigt, dass eine einheitliche Klassifizierung nicht existiert, anhand dessen Betriebsdaten strukturiert und aus Domänensicht mit Bezug zur Geschäftssicht beschrieben werden. Deutlich wurde jedoch, dass die Daten potenziell aus vielen diversen und vielseitigen Quellen stammen können, wie Sensorik, Aktorik, den Systemen selbst, Benutzerschnittstellen, Soziale Medien und IT-Systemen, welche an verschiedenen Stellen im Unternehmen angesiedelt sind. Darüber hinaus resultieren die Datensätze aus verschiedenen Prozessen, sind teilweise von der Maschine, teilweise vom Menschen generiert und enthalten verschiedene Inhalte zum Produkt, zum Nutzer oder zum Prozess. Diese diversen Orte ermöglichen das Entstehen und die Generierung von zahlreichen unterschiedlichen und potenziell nutzbaren Datensätzen und erschweren so ihre Identifikation und das Zusammentragen. Die Vielseitigkeit der intrinsischen Eigenschaften kann dadurch bereits erahnt werden.

Das Hauptziel des Datenverständnisses ist es, allgemeine Erkenntnisse über die Daten zu gewinnen, die für die weiteren Schritte im Datenanalyseprozess hilfreich sein können [BBH+10]. Die **Datenbeschreibung** bzw. **Datencharakterisierung** ermöglicht diese Erkenntnisse. "Die Datencharakterisierung beschreibt die Daten in einer Weise, die für den Benutzer nützlich ist, und beginnt mit dem Prozess des Verstehens, was in den Daten enthalten ist – das heißt, sind sie zuverlässig und für den Zweck geeignet?" [Pyl99]. Um die Art der Daten zu beschreiben, werden Merkmale bzw. Eigenschaften benötigt [KM16]. In diesem Zusammenhang wird oft auch von Metadaten gesprochen. Metadaten ("Daten über Daten") beziehen sich auf strukturierte Daten, die zur Beschreibung und Spezifizierung von Fakten über ein Informationsobjekt verwendet werden können [DMS+05]. Metadaten werden verwendet, um Datenmerkmale zu definieren. Diese Idee ist im Bereich des Meta-Lernens verbreitet, wo die für das Problem relevanten Attribute von besonderem Interesse sind.

Mit der Vielzahl der unterschiedlichen Datenquellen von Betriebsdaten geht auch die Heterogenität ihrer Eigenschaften einher. NIETO ET AL. erörtern **Sensordaten** und ihre Eigenschaften im Detail [NAL21]. Sie schlagen vor, bei der Auswahl und Anwendung von

Analysetechniken die Kategorien 1) Sensortypen, 2) Maßeinheiten und Bereiche, 3) Datenverteilung, 4) Ausreißer und Homogenität sowie 5) Korrelation zu berücksichtigen.

1) Die Annahme, dass Messungen kontinuierlich und ohne große Lücken sind, trifft nicht immer zu, insbesondere bei verschiedenen Sensortypen. Temperatursensoren zeigen **saisonale Schwankungen** und klare **Trends**. Helligkeitssensoren haben Unterschiede zwischen Innen- und Außenbereichen. Die Berücksichtigung solcher Aspekte ist entscheidend, da bestimmte Analysemodelle Ausreißer entfernen können, was die Genauigkeit beeinträchtigt.

2) Unterschiedliche Sensoren und **Maßeinheiten** werden je nach zu messendem Aspekt verwendet. Selbst bei der Messung desselben Aspekts können verschiedene Einheiten verwendet werden, z. B. für die Temperatur in Celsius, Fahrenheit oder Kelvin. Unterschiedliche Skalen können ebenfalls auftreten, z. B. für den Druck in Millibar oder Pascal.

3) Die **statistische Verteilung** der Daten beeinflusst die Ergebnisse von statistischen Tests, insbesondere bei der Bereinigung von Daten und der Suche nach Ausreißern. Reduzierte Zeitfenster können die Datenverteilung beeinflussen, was die Wahl der Methoden zur Berechnung von Mittelwert und Varianz beeinflussen kann.

4) **Ausreißer** spielen eine Rolle bei anderen Merkmalen der Daten, können die Datenverteilung beeinflussen und entstehen auch durch Rauschen in den Daten.

5) Die **Korrelation** zwischen Sensoren muss von Fall zu Fall analysiert werden und kann nicht verallgemeinert werden. Einige Sensoren können sogar über große Entfernungen korreliert sein, während andere keine Korrelation aufweisen, selbst wenn sie nahe beieinander liegen.

Für **unstrukturierte Daten** innerhalb der Betriebsphase von Produkten, wie z. B. Serviceberichte und Social-Media-Daten, sind wiederum andere Eigenschaften und Herausforderungen entscheidend: z. B. semantische Inkonsistenzen, in unterschiedlichem Maße vorverarbeitete Daten und Rauschen [SRF+09, RLS05]. Unter Rauschen (engl. „noise“) werden meist zufälliger Fehler in gemessenen Variablen verstanden [GLH15]. Manche Forscher zählen auch fehlende Werte, Ausreißer und redundante Daten dazu [SA19, GLD00, LCL09]. MENON ET AL. untersuchten textuelle Service und Kundendaten, z. B. Reparaturmaßnahmen, Kundenbeschwerden und individuelle Produktdetails. Sie stellten fest, dass sich solche Daten dadurch auszeichnen, dass sie festdefinierte Felder (z. B. „ID_Nummer“ und „Geschlecht“) und unstrukturierte Freitextfelder enthalten. Die Qualität der Freitextfelder war meist nicht gut, da Felder nicht ausgefüllt sind oder keine hilfreichen Informationen enthalten. Die sozialen Medien zeichnen sich durch viele unterschiedliche Formate aus (z. B. Text, Bild, Video oder Audio). Auch hier mischen sich strukturierte Formate dazu, wie z. B. Bewertungswerte, Datenkategorien und geografische Koordinaten [PLA19]. Die Datenqualität scheint ebenfalls eine Herausforderung zu

sein. So identifizierten STIEGLITZ ET AL. geringe Qualität als ein Kernproblem von Daten aus den sozialen Medien, welches gerade bei der Datenvorverarbeitung zu Schwierigkeiten führen kann [SMB+18]. Am schwersten wiegen unvollständige und verrauschte Daten (z. B. Rechtschreibfehler, nicht standardisierte Wörter, Wiederholungen usw.). Auch fehlende Informationen sind ein Thema, weil Nutzer beschlossen haben diese nicht zur Verfügung zu stellen. Darüber hinaus kennzeichnen sich Daten aus sozialen Medien im Gegensatz zu anderen klassischen Medien durch Zeitsensibilität (Text verändert sich durch den Echtzeitgedanken von den Services), kurze Länge der Texte und unstrukturierte Phasen, welche durch viele Abkürzungen und Akronyme auffallen [VA16].

Reale **Event-Logs** von beispielsweise Zustandsdaten enthalten verrauschte oder beschädigte Datensätze, die durch verschiedene Faktoren verursacht werden können: Einige Spuren sind doppelt, unvollständig, inkonsistent oder spiegeln ein anderes falsches Verhalten wider [MT21].

Viele der genannten Daten teilen einige Charakteristika von **Big Data**, sehr große Datenmengen, die spezielle Verfahren zur Verarbeitung erfordern. Dass gerade Big Data neue Herausforderungen mit sich bringen, ist ein großes Thema in der Literatur (z. B. [ABB+11, Ale13, JGL+14]). Als erstes nannte DOUG LANEY drei Dimensionen von Big Data, auch bekannt als die „drei V's" [Lan01]. Sie können wie folgt zusammengefasst werden:

- Volume: Der enorme Datenumfang ist wohl das wichtigste und charakteristischste Merkmal von Big Data, das zusätzliche und spezifische Anforderungen an alle traditionellen Technologien und Tools stellt [DGL+13a]. Big Data Volume umfasst Merkmale wie Größe, Umfang, Menge und Dimension. Das Volumen kann sich auf Zettabytes, Yottabytes oder darüber hinaus erstrecken [PA16].
- Velocity: Die Geschwindigkeit der Datenerzeugung in Verbindung mit dem Vorteil der Echtzeitanalyse ermöglicht es, Informationen in Echtzeit zu gewinnen und zu nutzen.
- Variety: Die Vielfalt der Daten äußert sich in vielen verschiedenen Formaten. Sie sind oft unstrukturiert oder ihre Struktur ist spezifisch für die Datenquelle.

Weitere „V's" wurden ergänzend vorgeschlagen. Häufig ist von Glaubwürdigkeit und Wahrhaftigkeit („Veracity") die Rede. Einiger Forscher verwenden diese Eigenschaft nur, um sich auf Fragen der Informationssicherheit wie Datenintegrität und Authentizität zu beziehen [DGL+13b, KGM+14]. Andere verwenden eine breitere Definition, die Unsicherheiten in Bezug auf Datenqualität umfasst [SS14, AEF+12]. Dazu gehört, dass Daten ungeordnet und verrauscht sein können und Unsicherheit und Fehler enthalten. *Value* bezeichnet den Mehrwert der Daten für die Unternehmen, dadurch, dass viele Erkenntnisse aus den Daten gezogen werden können [Mar14]. Die 5 V's von Big Data sind in Bild 2-11 dargestellt. MURTHY ET AL. kategorisieren Big Data anhand einer sechsstufigen Taxonomie, die sich auf die Handhabung und Verarbeitung konzentriert: (1) Daten (a)

zeitliche Latenz für die Analyse: Echtzeit, nahezu Echtzeit, Batch; und (b) Struktur: strukturiert, halbstrukturiert, unstrukturiert); (2) Recheninfrastruktur (Batch oder Streaming); (3) Speicherinfrastruktur (SQL, NoSQL, NewSQL); (4) Analyse (überwachtes, halbüberwachtes, unüberwachtes oder verstärkendes maschinelles Lernen; Data Mining; statistische Verfahren); (5) Visualisierung (Karten, abstrakt, interaktiv, Echtzeit); und (6) Datenschutz und Sicherheit (Datenschutz, Verwaltung, Sicherheit) [MBS+14]. KITCHIN identifiziert sieben Eigenschaften: Volumen, Geschwindigkeit, Vielfalt, Vollständigkeit, Auflösung und Indexikalität, Relationalität, Erweiterbarkeit und Skalierbarkeit [Kit13, Kit17]. Durch eine Analyse, bei der diese Typologie auf 26 Datensätzen angewendet wurde, zeigte sich, dass Big Data nicht alle dieselben und nicht alle identifizierten Merkmale aufweisen [KM16] .

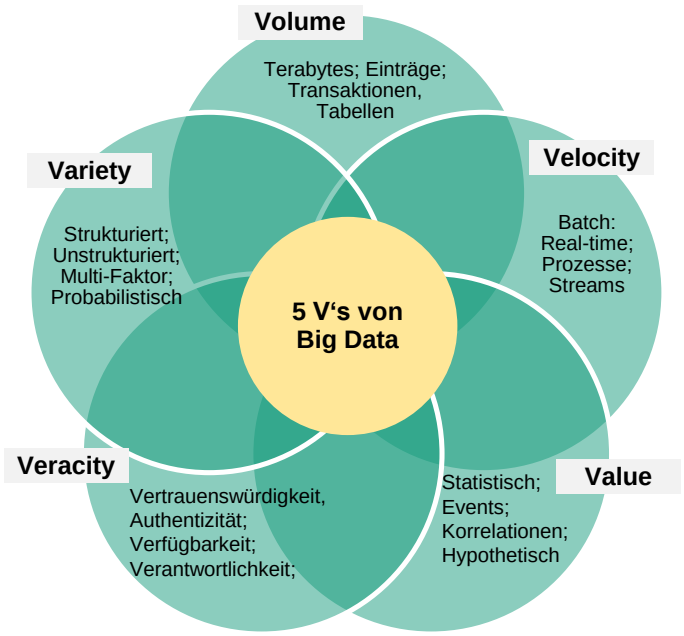

Bild 2-11: 5 V's von Big Data [DGL+13a]

Während in der Literatur im Kontext von Industrie, IoT und Cyber-Physical Systems oft von Big Data die Rede ist [LTC+15, YK15], welche teilweise neue und fortgeschrittene Verfahren[9] erfordern [KEA+15], klagen Unternehmen in der Praxis über einen Mangel an Daten [MFK+22]. Die Gründe dafür liegen darin, dass bei vielen Produkten im Feld noch zu wenig Sensoren installiert sind oder der Zugriff auf die Daten aufgrund fehlender

[9] Fortgeschrittene Analytik ist die Anwendung von mehreren Analysemethoden, die die Vielfalt von Big Data adressieren, um beschreibende Ergebnisse zu liefern und um umsetzbare und prädiktive Ergebnisse zu liefern, die die Entscheidungsfindung erleichtern. Fortgeschrittene Analytik geht über Data Mining und statistische Verarbeitungsmethoden hinaus und umfasst logikbasierte Methoden, qualitative Analysen und nicht statistische quantitative Methoden.

Netzwerkanbindung nicht möglich ist. Selbst wenn viele Daten vorhanden sind, fehlen sehr oft annotierte Daten oder es sind nur wenige davon verfügbar, da die Annotation sehr teuer und zeitaufwändig ist [CMX+21]. In diesen Fällen ist die Anwendung fortgeschrittener Klassifizierungsalgorithmen aufgrund von Überanpassungs- und Nichtkonvergenzproblemen möglicherweise nicht möglich [GDY20, GBC16].

Neben den Big-Data-Charakteristika gibt es weitere allgemeine Ansätze zur Datenbeschreibung; sie werden nachfolgend erläutert. Die Idee Datensätze zu charakterisieren, wird schon lange im Forschungsfeld Meta-Lernen[10] verfolgt [Van19]. BILALLI ET AL. unterscheiden zwei Gruppen von Merkmalen: (1) allgemeine Maße (allgemeine Informationen zum vorliegenden Datensatz, wie Anzahl der Instanzen und Dimensionalität) und (2) statistische und informationstheoretische Maße (Attributstatistiken und Klassenverteilungen wie Mittelwert und Standardabweichung) [BAT+16].

Nach HILDEBRAND ET AL. können Datenarten anhand von verschiedenen Beschreibungskriterien definiert werden [HGH+15]. Die Kriterien gehen auf inhärente Eigenschaften der Daten und den Kontext, in dem sich die Daten befinden, ein. Die Eigenschaften werden durch Format (z. B. numerisch, integer, etc.), Struktur (strukturiert, semistrukturiert, unstrukturiert), Inhalt (Sachverhalte beinhaltende Daten oder Sachverhalte beschreibende Daten), Stabilität (fixe und variable Daten), Verarbeitung (Eingabedaten, Speicherdaten und Ausgabedaten) und dem Business Object beschrieben. Die Kontextinformationen von Daten bilden die Angaben zu den Prozessen, in denen die Daten benötigt werden (In welchem Prozess werden die Daten generiert? In welchem Prozess wird die Information genutzt?) und zu verschiedenen Verwendungszwecken (z. B. Kundennummer für die Rechnungsabwicklung oder Auftragsübersicht).

ZIEGENBEIN ET AL. stellen eine Liste von Datensatzmerkmalen zur Verfügung, die mit maschinellen Lernverfahren in Verbindung stehen (s. Tabelle 2-1) [ZSM+19].

Tabelle 2-1: Datensatzmerkmale für maschinelle Lernverfahren [ZSM+19]

Datensatz-eigenschaften	**Indikator**	**Exemplarische Skalen**		
		niedrig	mittel	hoch
Datensatzgröße	# Datenpunkte	<500	500-5000	>5000
Vollständigkeit	% verfügbare Daten	<80	80-90	>90

[10] Meta-Lernen oder lernen, wie man lernt, ist die Wissenschaft des systematischen Beobachtens, wie verschiedene Ansätze des maschinellen Lernens bei einer Vielzahl von Lernaufgaben abschneiden, und des anschließenden Lernens aus diesen Erfahrungen oder Metadaten, um neue Aufgaben viel schneller zu lernen, als es sonst möglich wäre.

Verständlichkeit	% verständliche Daten	<40	40-90	>90
Rauschen	% Rauschen in den Daten	>70	40-70	<40
Ausreißer	% Ausreißer in den Daten	>20	5-20	<5
Monotonie	binär	nein		ja
Linearität	binär	nein		ja
Dimensionalität	# Variablen/Features	>50	2-50	<2
Volatilität	# Histogrammbalken	>10	5-10	<5
Stationarität	binär	nein		ja
Multikolinearität	binär	ja		nein
Heteroskedastizität	binär	ja		nein

Die bereits genannten Qualitätsprobleme und -unterschiede bilden eine weitere Dimension [CLC15]. Ein bekannter Ansatz zur Strukturierung der Datenqualitätsmerkmale ist von Larry English [Eng99]. Er unterscheidet drei Hauptmerkmale mit insgesamt zehn Untermerkmalen:

1) Qualität der Datendefinition (Definition der Daten über die Metadaten)
    a. Datenspezifikation (beschreibt die einzelnen Objekte wie z. B. Tabellen und Felder in ihrer absoluten fachlichen Bedeutung und Modellierung im System)
    b. Geschäftsregeln (beschreiben Abhängigkeiten und Beziehungen der modellierten Objekte untereinander)
    c. Integritätsbedingungen
2) Inhaltliche Datenqualität (Korrektheit der Datenwerte)
    a. Vollständigkeit
    b. Eindeutigkeit
    c. Einhaltung der Geschäftsregeln
    d. Genauigkeit und Fehlerfreiheit
3) Qualität der Datenpräsentation (ist durch Fragen der (zeitlichen) Verfügbarkeit, Angemessenheit des Formats und Verständlichkeit geprägt)
    a. Rechtzeitige Bereitstellung
    b. Angemessenheit des Formats
    c. Verständlichkeit des Formats

WANG definiert vier Datenqualitätskategorien Intrinsisch, Kontextbezug, Repräsentation und Verfügbarkeit, die jeweils weitere Datenqualitätsdimensionen aufweisen [WS96]. Da diese Merkmale im Kontext vom Informationsmanagement definiert wurden, bleibt unklar, welche davon relevant für die anderen Schritte der Data-Analytics-Pipeline sind. Nur teilweise überschneiden sich diese mit den weiter vorne genannten Aspekten im Zusammenhang der verschiedenen Betriebsdatenarten. Im Produktionskontext sind laut BATINI und SCANNAPIECO folgende sechs Qualitätskriterien die wichtigsten [BS06]:

1) Vollständigkeit: Ein Datensatz muss alle erforderlichen Attribute enthalten. Die Attribute müssen alle erforderlichen Daten enthalten.
2) Einzigartigkeit: Jeder Datensatz muss eindeutig interpretierbar sein.
3) Korrektheit: Die Daten müssen der Realität entsprechen.
4) Aktualität: Alle Datensätze müssen dem aktuellen Stand der abgebildeten Realität entsprechen.
5) Genauigkeit: Die Daten müssen mit der geforderten Genauigkeit verfügbar sein.
6) Konsistenz: Ein Datensatz darf keine Widersprüche in sich oder zu anderen Datensätzen aufweisen.

Weitere fünf können im Laufe der Zeit ergänzt werden, um die erreichte Grunddatenqualität zu sichern und weiter zu verbessern:

1) Nicht-Redundanz: Es darf keine Duplikate innerhalb der Datensätze geben.
2) Relevanz: Der Informationsgehalt der Datensätze muss den jeweiligen Informationsanforderungen entsprechen.
3) Einheitlichkeit: Die Informationen in einem Datensatz müssen einheitlich strukturiert sein.
4) Verlässlichkeit: Die Herkunft der Daten muss nachvollziehbar sein.
5) Verständlichkeit: Die Terminologie und Struktur der Datensätze muss mit den Vorstellungen der Informationsempfänger (z. B. Abteilungen) übereinstimmen.

Da es bei der datengestützten Produktplanung darum geht, möglichst eine ganze Produktgeneration und nicht nur eine Produktinstanz zu verbessern, ist die aggregierte oder globale Sicht über alle einzelnen Produktinstanzen und deren Daten im Feld wichtig [Die14]. Damit kommt eine weitere wichtige Eigenschaft ins Spiel – der Aggregationsgrad. Wenn Betriebsdaten direkt aus ihrer Ursprungsquelle kommen, wie z. B. den Sensoren, liegen sie im Rohformat vor. Sie bieten dann meist feinaufgelöste individuelle Produktinformationen. Werden diese Daten mit Daten aus anderen Quellen und Produktinstanzen integriert und Informationen zusammengefasst, handelt es sich um aggregierte Daten. In der Fahrzeugbranche werden diese Daten als „Flottendaten“ bezeichnet. Diese Daten haben in der Regel einen sequentiellen Charakter und listen wichtige Ereignisse und Kennzahlen, wie z. B. Geschwindigkeit, Reparaturzeitpunkt und diagnostizierter Fehler, sowie den Zeitpunkt für verschiedene Fahrzeuge über eine Fahrzeugnummer auf [CCM+18, AAA+14]. Daten im Rohformat können eine größere Vorverarbeitung in Form einer

Datenintegration oder initiale individuelle Auswertungen erforderlich machen [BPD+22]. Häufig fehlen auf der Ebene der Produktinstanzen wichtige Informationen, wie Fehlerursachen. Hier ist es meist sinnvoller, diese auf den individuellen Daten zu ermitteln, da eine Aggregation zu Datenreduzierungszwecken wichtige Informationen entfernen könnte. Eine anschließende Aggregation kann im zweiten Schritt Erkenntnisse über die gesamte Produktflotte liefern. Aber auch reine individuelle Auswertungen auf einzelnen Instanzen können im Vordergrund stehen, wenn maßgeschneiderte Anpassungen gewünscht sind [PGP+22].

Fazit: Aufgaben des Schrittes *Aufbau von Datenverständnis* sind zum einen die Identifikation und Sammlung der unterschiedlichen Betriebsdatenquellen und ihrer Datensätze, und zum anderen die Beschreibung dieser Daten, um die Daten so weit zu verstehen, dass die weiteren Schritte wie die Vorverarbeitung und Modellierung besser geplant werden können. Die Herausforderungen liegen auf der einen Seite in der Unübersichtlichkeit und Vielzahl der potenziellen Quellen, die an vielen verschiedenen Stellen im Unternehmen liegen und über ihre Inhalte, Ursprünge und Zugänge etc. verschiedene Zusammenhänge zu Domänenwissen aufdecken. Oftmals bestehen über sie keine Informationen in übersichtlicher oder aggregierter Form [OB13, WTH+17]. Um diese systematisch zu identifizieren, ist eine Strukturierung der Datenursprünge und Klassifizierung der für die Produktplanung relevanten Betriebsdatenquellen und ihrer Datensätze notwendig, die wichtige Sichten aus der Domäne der Produktplanung berücksichtigen. Gemäß gängigen Klassifikationen von Industriedaten können diese Sichten die Datenerzeugung durch menschliche oder maschinelle Quellen, den inhaltlichen Fokus (wie Produkt oder Nutzer) sowie die zugrundeliegenden IT-Systeme (wie CRM und MES) umfassen.

Auf der anderen Seite liegen Herausforderungen in der starken Heterogenität der intrinsischen Merkmale dieser Daten. Es wurde gezeigt, dass Betriebsdaten, darunter Sensordaten, textuelle Daten und Event-Logs, viele verschiedene Eigenschaften aufweisen. Um Verständnis über die Daten aufzubauen, existieren einige Typologien, Taxonomien und Zusammenstellungen von Dateneigenschaften und -kriterien. Diese beschreiben die Daten hinsichtlich ihrer Struktur, ihres Inhalts, ihrer Verarbeitung. Auch die Datenqualität kann anhand einer Vielzahl von Kriterien bestimmt werden. Einige lassen sich leicht feststellen, andere nur durch Berechnungen, wie es z. B. bei den Merkmalen für das Meta-Lernen der Fall ist. Nur wenige Ansätze stellen einen Bezug zu anderen Data-Analytics-Komponenten her, sodass unklar bleibt, welche Kriterien für den (weiteren) Datenanalyseprozess in erster Linie wichtig sind. Die Vielzahl und Vielseitigkeit der Merkmale erschweren die Aufgabe der Datenbeschreibung zusätzlich. Daher ist ein Beschreibungsrahmen zur einheitlichen Beschreibung von Betriebsdaten erforderlich, welcher auch eine vereinfachte und abstrahierte Bestimmung der Merkmale für vorliegende Betriebsdaten ermöglicht.

2.3.4 Methodenauswahl (Vorverarbeitung, Modellierung, Evaluierung)

Um eine Auswahl der passenden Methoden für die Aufgaben Vorverarbeitung und Modellierung zu treffen, ist ein umfassenden Verständnis dieser Aufgaben notwendig. Insbesondere die Vorverarbeitung und Modellierung stehen in enger Beziehung zueinander. Daher werden im Folgenden die einzelnen Schritte innerhalb der betriebsdatengestützten Produktplanung genauer beleuchtet.

2.3.4.1 Vorverarbeitung

Die Vorverarbeitung von Daten ist ein Prozess, der die Daten der realen Welt für den Datenanalyseprozess aufbereitet und in Qualitätsinput transformiert [SM17]. Gemäß CRISP-DM deckt die Datenvorverarbeitungsphase alle Aktivitäten ab, um den finalen Datensatz aus den initialen rohen Daten zu konstruieren. Laut CHAPMAN ET AL. werden Datenvorverarbeitungsaufgaben häufig mehrmals und nicht in einer vorgeschriebenen Reihenfolge durchgeführt [CCK+00].

Reale Daten, insbesondere Betriebsdaten, sind hinsichtlich ihrer Eigenschaften und Datenqualität sehr unterschiedlich aufgestellt (vgl. Abschnitt 2.3.3). Die genannten Faktoren beeinträchtigen die Qualität der Ergebnisse nach dem Mining oder der Modellierung. Daher ist es notwendig, vor dem Mining oder der Modellierung Verbesserungstechniken, die sogenannte Datenvorverarbeitung, anzuwenden. Neben der Ergebnisverbesserung hat Vorverarbeitung weitere Vorteile [GLH15]:

- Anpassung und Spezifizierung der Daten für jeden Data-Analytics-Algorithmus
- Verringerung der Datenmenge, die für eine geeignete Lernaufgabe erforderlich ist
- Verringerung der Zeitkomplexität
- „Ermöglichung des Unmöglichen“ mit Rohdaten, sodass Data-Analytics-Algorithmen auf großen Datenmengen angewendet werden können
- Unterstützung beim Verstehen der Daten

Es gibt innerhalb dieses Prozesses verschiedene Techniken, um die Daten für Analysezwecke zu optimieren. Beispiele sind Operationen wie **Datenbereinigung** (z. B. Behandlung fehlender Werte, Ausreißererkennung), **Datentransformation** (Numerisierung, Diskretisierung, Normalisierung, numerische Transformationen), **Dimensionalitätsreduktion** und **Merkmalsextraktion** [Li19].

Die Datenbereinigung befasst sich mit Fragen der Datenqualität. CORRALES ET AL. fassen verschiedene Datenbereinigungsansätze zusammen, um Qualitätsaspekte wie **fehlende Werte**, **Ausreißer**, Redundanz und **Rauschen** zu adressieren: z. B. durch Imputation, Ausreißererkennung sowie Entfernung von Duplikaten [CCL18].

Fehlende Werte, die durchaus typisch für Betriebsdaten sind (vgl. Abschnitt 2.3.3), erschweren die Durchführung von Datenanalysen. Ungeeignete Behandlung der fehlenden Werte kann zu Verzerrungen und irreführenden Schlussfolgerungen führen. Folgende Probleme werden mit fehlenden Werten in Verbindung gebracht: der Verlust an Effizienz, Komplikationen im Umgang und bei der Analyse von Daten und Verzerrungen aufgrund von Unterschieden zwischen fehlenden und vollständigen Daten [GLH15]. Üblicherweise werden fehlende Werte auf drei verschiedenen Arten behandelt [FKP07]:

1) Verwerfen der Datenpunkte mit fehlenden Werten in den Attributen
2) Nutzung von Maximum-Likelihood-Verfahren, bei denen die Parameter eines Modells für den vollständigen Teil der Daten geschätzt werden und später für die Imputation mittels Stichproben verwendet werden
3) Die Imputation beschreibt eine Klasse von Verfahren, die darauf abzielt, die fehlenden Werte mit geschätzten Werten aufzufüllen. In den meisten Fällen sind die Attribute eines Datensatzes nicht unabhängig voneinander, weshalb fehlende Werte durch die Identifizierung von Beziehungen zwischen Attributen bestimmt werden können.

Die Imputation bietet den Vorteil, dass dieser Ansatz unabhängig von dem Lernalgorithmus ist, sodass der Nutzer „nur“ aufgrund der Situation die am besten geeignete Methode auswählen kann. Es gibt eine breite Palette von Imputationsmethoden, von einfachen Imputationstechniken wie Mittelwertsubstitution, und K-Neirest-Neighbour. bis hin zu solchen, die die Beziehungen zwischen den Attributen analysieren, wie SVM-basierte, Clustering-basierte, logistische Regressionen, Maximum-Likelihood-Verfahren und Mehrfach-Imputation. Für verschiedene Domänen, wie z. B. die Medizin [UW01] und Produktion [SZY+16] werden unterschiedliche Imputationsmethoden vorgeschlagen, die an die gemeinsamen Merkmale der dort analysierten Daten angepasst sind.

Unerkannte **Ausreißer** haben einen negativen Effekt auf die Modellperformance [WE17]. Mit Hilfe von Techniken der Ausreißererkennung können solche unüblichen Werte erkannt und auch entfernt werden. Alternativ sind robuste Modellierungsmethoden zu verwenden, die unempfindlich gegenüber Ausreißern sind [Kan19]. Ein Beispiel ist der C4.5 Algorithmus [Sal94]. Für bestimmte Datentypen gibt es auch konkrete Empfehlungen, wie z. B. statistische Tests für die Ausreißererkennung bei Sensordaten zu nutzen [NAL21].

Zu erkennen, ob Daten verrauscht sind, ist keine triviale Aufgabe und eine falsche Detektion kann den Daten schaden. LIBRALON ET AL. nennen vier Ansätze, um Rauschen zu detektieren [LCL09]: (1) Statistische Ansätze, welche auf Datenverteilungsmodellen beruhen, (2) Clustering-Ansätze, die potenzielles Rauschen als kleine Gruppen von Daten identifizieren, die sich auf die vorhandenen Beispiele verteilen, (3) ML-Klassifikationsalgorithmen, die verrauschte Beispiele auf Basis von Trainingsbeispielen erkennen und (4) abstandbasierte Techniken, welche Ähnlichkeitsmaße verwenden, um den Abstand

zwischen Instanzen aus einem Datensatz zu berechnen und diese Informationen nutzen, um mögliche verrauschte Daten zu identifizieren. Bei Letzteren dreht sich eine Frage hauptsächlich um das Ähnlichkeitsmaß, welches durch die Daten beeinflusst wird [AHK01]. Zum Beispiel eignet sich die häufig genutzte euklidische Metrik nicht für hoch-dimensionale Daten. Nach der Identifikation von verrauschten Instanzen, können drei Techniken genutzt werden, um Rauschen zu behandeln. Der erste Ansatz besteht im Ignorieren des Rauschens, wobei die Analysetechniken robust genug sein müssen. Beim zweiten Ansatz werden die Daten gefiltert – Instanzen, die nach bestimmten Bewertungskriterien als verrauscht gelten, werden verworfen. Der dritte Ansatz („Polieren“) repariert verrauschte Instanzen, indem die beschädigten Werte durch geeignetere ersetzt werden. Für jeden dieser Ansätze gibt es Vor- und Nachteile [Ten01].

Betriebsdaten, insbesondere Sensordaten, sind häufig verrauscht (vgl. 2.3.3). Um die bedeutungsvollen Informationen aus diesen zu extrahieren, werden meist Filtertechniken angewendet. Auch hier sind diverse Techniken, insbesondere aus der Signalverarbeitung, aufzuführen, wie z. B. Kalman Filter, Wiener Filter oder Least Mean Squared Error (LMS) Filter [Vas96]. Ihre Vor- und Nachteile sind nur schwer für Nichtexperten abzuwägen. Adaptive Filter wie LMS sind beispielsweise nur für nicht-stationäre Signale geeignet. Bei Textdaten kommen weitere spezielle Bereinigungstechniken aus dem Natural Language Processing hinzu, wie Tokenisierung (Segmentierung eines Textes in seine Wörter) und Stemming (Zurückführung verschiedener Varianten eines Worts auf seine Stammform) [GRL+16, Moh15].

Die **Datentransformation** ist der Prozess zur Änderung des Formats, der Struktur oder der Werte von Daten in ein nutzbares Format, das von einem Modell analysiert werden kann. Die meisten Algorithmen für Data Analytics und maschinelles Lernen haben bestimmte Anforderungen an die Eingabedaten. So bestehen beispielsweise einige Betriebsdaten aus numerischen Sensordaten, wie z. B. Leistung, Temperatur und Druck. Herkömmliche Assoziationsregel-Algorithmen (z. B. A-priori) sowie viele andere Data-Analytics- Methoden können jedoch nur kategorische Daten wie hoch, mittel und niedrig verarbeiten [FXY15]. In diesem Fall sollte eine Datentransformation durchgeführt werden, um die Kompatibilität zwischen Daten und Algorithmen zu gewährleisten [CMX+21].

Die **Diskretisierung** ist eine essenzielle Transformation, die häufig als Vorverarbeitungsmethode in der Datenanalyse (für Betriebsdaten) eingesetzt wird. Ihr Hauptziel ist es einen Satz an kontinuierlichen Attributen in diskrete umzuwandeln, indem kategorische Werte mit Intervallen verknüpft werden und so quantitative Daten in qualitative Daten umgewandelt werden [GLH15]. Diskretisierung kann auch als Datenreduktionsmethode betrachtet werden, da sie Daten aus einem riesigen Spektrum numerischer Werte auf eine stark reduzierte Teilmenge diskreter Werte abbildet. Die Notwendigkeit Daten zu diskretisieren kann durch mehrere Faktoren bedingt sein [GLH15]. Viele Data Analytics Algorithmen sind in erster Linie auf die Verarbeitung nominaler Attribute ausgerichtet oder verarbeiten z. T. sogar nur diskrete Attribute. Zum Beispiel erfordern drei von zehn sehr beliebten Methoden [Wu09], eine eingebettete oder externe Diskretisierung: C4.5

[Sal94], Apriori [AS94] und Naive Bayes [FGM+11]. Darüber hinaus werden Daten durch die Diskretisierung reduziert und vereinfacht, wodurch das Lernen schneller wird und genauere sowie kompaktere Ergebnisse liefert; auch das Rauschen kann reduziert werden. Diskrete Attribute sind zudem leichter zu verstehen und zu erklären. Allerdings führt jede Diskretisierung im Allgemeinen zu einem Informationsverlust. In der Literatur findet sich eine Vielzahl von Diskretisierungstechniken. GARCIA identifizierte mehr als 80 Diskretisierungsmethoden, darunter immer noch 30, die sie als sehr relevant einstuften. Die Wahl eines Diskretisierers bestimmt den Erfolg der anschließenden Lernaufgabe. Doch die Identifikation des besten Diskretisierers für jede Situation ist eine sehr schwierige Aufgabe, die meist nur durch Durchführung umfassender Experimente bewältigt werden kann [GLH15].

Die **Dimensionalitätsreduktion** (DR) wird oft als ein separater wichtiger Aspekt der Datenvorverarbeitung angesehen, insbesondere wenn es sich um Big Data oder Textdaten handelt, die eine große Anzahl von Beobachtungen und/oder Variablen umfassen [DSV+21, RRL+20]. Sie reduziert hochdimensionale Daten auf eine geringere Dimensionalität, indem sie die wichtigsten Attribute extrahiert und so die zeitliche Komplexität der Trainingsphasen minimiert [vPH07]. Die Transformation oder Projektion der originalen Daten in einen kleineren Raum kann beispielsweise mit einer Hauptkomponentenanalyse (PCA für Principal Component Analysis), Multidimensionalen Skalierung (MDS) oder Faktorenanalyse erfolgen. Das Problem hochdimensionaler Daten ist als „Fluch der Dimensionalität“ bekannt [Bel16]. Der Fluch der Dimensionalität beeinflusst die Daten je nach der folgenden Data-Analytics-Aufgabe oder Algorithmus unterschiedlich [ZPŠ14]. Beispielsweise können Techniken wie Entscheidungsbäume keine sinnvollen und verständlichen Ergebnisse liefern, wenn die Anzahl der Dimensionen zunimmt, obwohl die Geschwindigkeit in der Lernphase kaum beeinträchtigt wird. Im Gegensatz dazu ist das instanzbasierte Lernen stark von der Dimensionalität abhängig [GLH15]. Daher wurden im Laufe der Jahre einige Methoden für die DR entwickelt. Einige DR-Methoden haben das Ziel, vor allem irrelevante und redundante Features zu entfernen und so die Anzahl an Variablen zu reduzieren. Diese gehören zur Familie der Merkmalsselektionsmethoden. Merkmalsselektionsmethoden haben positive Effekte, wie z. B. die Beschleunigung der Verarbeitung des Data-Analytics-Algorithmus, die Verbesserung der Datenqualität und die erhöhte Verständlichkeit der Ergebnisse [GLH15].

Die **Merkmalsextraktion** und **-konstruktion** sind wichtige Vorverarbeitungsschritte, die das ML-Modell vereinfachen und auch die Qualität der Ergebnisse eines Algorithmus für maschinelles Lernen verbessern. Während die Extraktion einen Satz an neuen Features aus den ursprünglichen Features durch eine funktionale Abbildung herauszieht und den Feature-Raum in der Regel reduziert, entdeckt die Konstruktion fehlende Informationen über die Beziehung zwischen Features und erweitert den Raum an Features über Ableitung oder Schaffung neuer Features [LM98]. Um Features zu konstruieren, kann eine Vielzahl an verschiedenen Ansätzen genutzt werden. Diese lassen sich in vier Gruppen einteilen: datengetrieben, hypothesen-getrieben, wissensbasiert und hybrid [ML02].

Beim datengetriebenen Ansatz werden neue Merkmale auf der Grundlage der Analyse der verfügbaren Daten durch Anwendung verschiedener Operatoren konstruiert. Der hypothesengetriebene Ansatz konstruiert neue Features auf Grundlage zuvor erstellter Hypothesen. Beim wissensbasierten Ansatz werden existierendes Wissen und Domänenwissen angewendet. Welche Merkmalsextraktions- und Konstruktionsansätze sich in verschiedenen Domänen, Analytics-Aufgaben und für welche Daten anbieten, wurde in verschiedenen Studien untersucht. So existieren einige Veröffentlichungen zu Feature-Extraktion für Textdaten und Anwendungen wie Textklassifikation und Sentiment Analysen (z. B [Lew92], [DS19], [ACK+19]). Auch Sensor- und Signaldaten wurden dahingehend genauer untersucht (z. B. [PGK+09]). Hier finden sich allerdings schon weniger Übersichtspapiere, eher spezialisierte Ansätze für einzelne Datensätze wie z. B. Beschleunigungssensoren am Smartphone [DZS20]. Allein die beiden verschiedenen Datenarten Text und Sensor erfordern sehr unterschiedliche Ansätze (bei Sensordaten kommen häufig Features aus der Frequenz- sowie Zeit-Frequenz-Domäne hinzu, während bei textuellen Daten oftmals Vektorraummodelle und Natural Language Processing (NLP) Modelle wie Latent Semantic Analysis (LSA) im Fokus stehen, was die Umsetzung von Betriebsdatenanalysen erschwert.

Fazit: Die Datenvorverarbeitung ist komplex und in der Regel auf ein bestimmtes Problem zugeschnitten. Dies ist auf die Heterogenität der Daten, Anwendungen und Kontexte zurückzuführen [MTG+19, CCL18]. Aufgrund der Vielfalt von Daten und Anwendungen in der datengestützten Produktplanung stellt die Datenvorverarbeitung eine erhebliche Herausforderung dar. Hierbei kommen zahlreiche Methoden wie Datenbereinigung, Datentransformation und Merkmalsextraktion zum Einsatz, von denen jede ihre spezifischen Vor- und Nachteile aufweist. Um jedoch zu verhindern, dass sich Fehler in der Pipeline ausbreiten und negative Auswirkungen und Verluste verursachen, ist eine Vorverarbeitung zur Verbesserung der Datenqualität unabdingbar [TKE20]. Es ist auch unbestritten, dass die Qualität der Endergebnisse durch die Auswahl der richtigen Vorverarbeitungsoperatoren erheblich verbessert werden kann [JVD+14]. In einigen Fällen können beispielsweise eine Min-Max-Skalierung und Merkmalskreuzungen hilfreich sein, während in anderen Fällen eine Standardskalierung und Merkmalsauswahl zur Vermeidung von Überanpassung die bessere Wahl ist. Welche genau für die Betriebsdaten relevant sind, wird nur ansatzweise anhand einzelner Studien analysiert. Häufig wird in der wissenschaftlichen Literatur nur sehr knapp auf Vorverarbeitungsmethoden eingegangen, was rein literaturbasierte Auswertungen dazu noch erschwert. Durch die aufgeführten Beispiele konnte gezeigt werden, dass durch Methoden der Datenbereinigung vor allem Datenqualitätsaspekte wie Rauschen und fehlende Werte adressiert werden. Die Datentransformation nimmt bestimmte Format- und Strukturanforderungen von bestimmten Modellen, die eingesetzt werden sollen, in Angriff. Damit besteht eine zweiseitige Abhängigkeit, zu den Daten und Modellen, die bei der Auswahl der passenden Vorverarbeitungsverfahren zum Tragen kommt.

Es bedarf daher sowohl einer Übersicht an vorausgewählten Verfahren für Betriebsdaten und einer Auswahlunterstützung unter Berücksichtigung aller Abhängigkeiten.

2.3.4.2 Modellierung

Im Abschnitt 2.3.2 wurden bereits unterschiedliche Ansätze zur Wissensgewinnung, wie beispielsweise Prädiktion und Beschreibung, sowie typische Data-Analytics-Probleme, darunter Klassifikation und Clustering, vorgestellt. Darunter lassen sich viele Ansätze und Algorithmen einsortieren. Es existieren eine Vielzahl an verschiedenen Taxonomien und Paradigmen. GARCIA ET AL. unterteilen Prädiktionsmodelle in statistische und symbolische Methoden, Beschreibungsmodelle in Clustering und Assoziationsregeln (s. Bild 2-12) [GLH15]. Statistische Methoden zeichnen sich in der Regel durch die Darstellung von Wissen durch mathematische Modelle mit Berechnungen aus. Im Gegensatz dazu bevorzugen symbolische Methoden die Darstellung des Wissens mittels Symbole und Verknüpfungen, welche für Menschen leichter zu interpretieren sind.

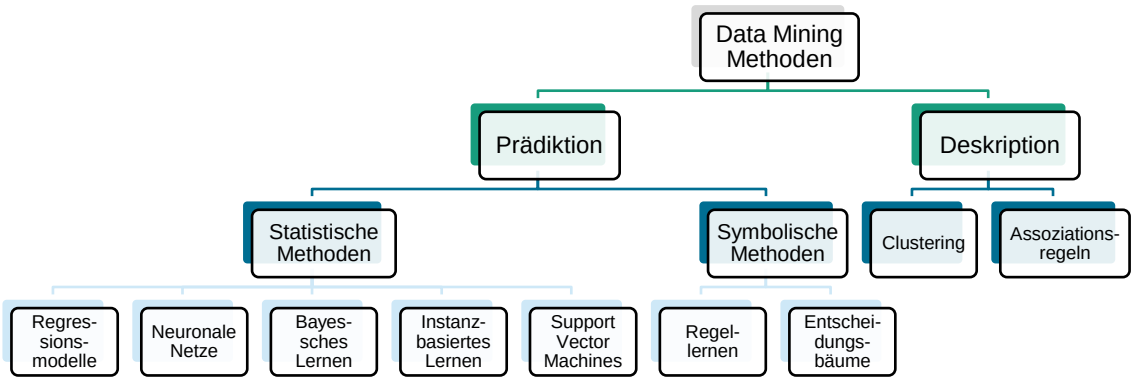

Bild 2-12: Data-Mining-Methoden [GLH15]

GAVRILOVSKI ET AL. entwickelten eine Taxonomie zur Organisation der wichtigsten Data-Mining-Techniken in der Literatur (s. Bild 2-13) [GJM+16]. Sie klassifizieren Algorithmen nach ihrem Verwendungszweck auf der obersten Ebene: Verifizierung und Entdeckung. Verifizierung bezieht sich dabei auf Techniken, die typischerweise mit der traditionellen Statistik in Verbindung gebracht werden, wie z. B. Hypothesentests. Techniken der Entdeckung entsprechen dem allgemeinen Verständnis von Data Mining, bei dem Muster entdeckt und Wissen aus den Daten extrahiert wird. Entdeckungstechniken können weiterhin in Vorhersage und Beschreibung oder überwachtes und unüberwachtes Lernen unterteilt werden. Auf der dritten Ebene nutzen sie die Lernaufgabe als weitere Unterteilung der Techniken und nennen hier Klassifikation, Clustering, Regression und Assoziationsanalyse. Innerhalb jeder dieser vier Gruppen ist es möglich, die Techniken nach der Modellrepräsentation, d. h. nach der Form und den Merkmalen der Data-Mining-Ergebnisse, weiter zu gliedern. So können beispielsweise Clustering-Algorithmen Datenstrukturen ausgeben, die entweder partitioniert, hierarchisch, dichtebasiert oder

gitterbasiert sind. In ähnlicher Weise können Algorithmen zur Klassifizierung in Techniken unterteilt werden, die die Klassen durch Regeln, Bäume, lineare und nichtlineare Modelle und Instanz basierte Repräsentation darstellen.

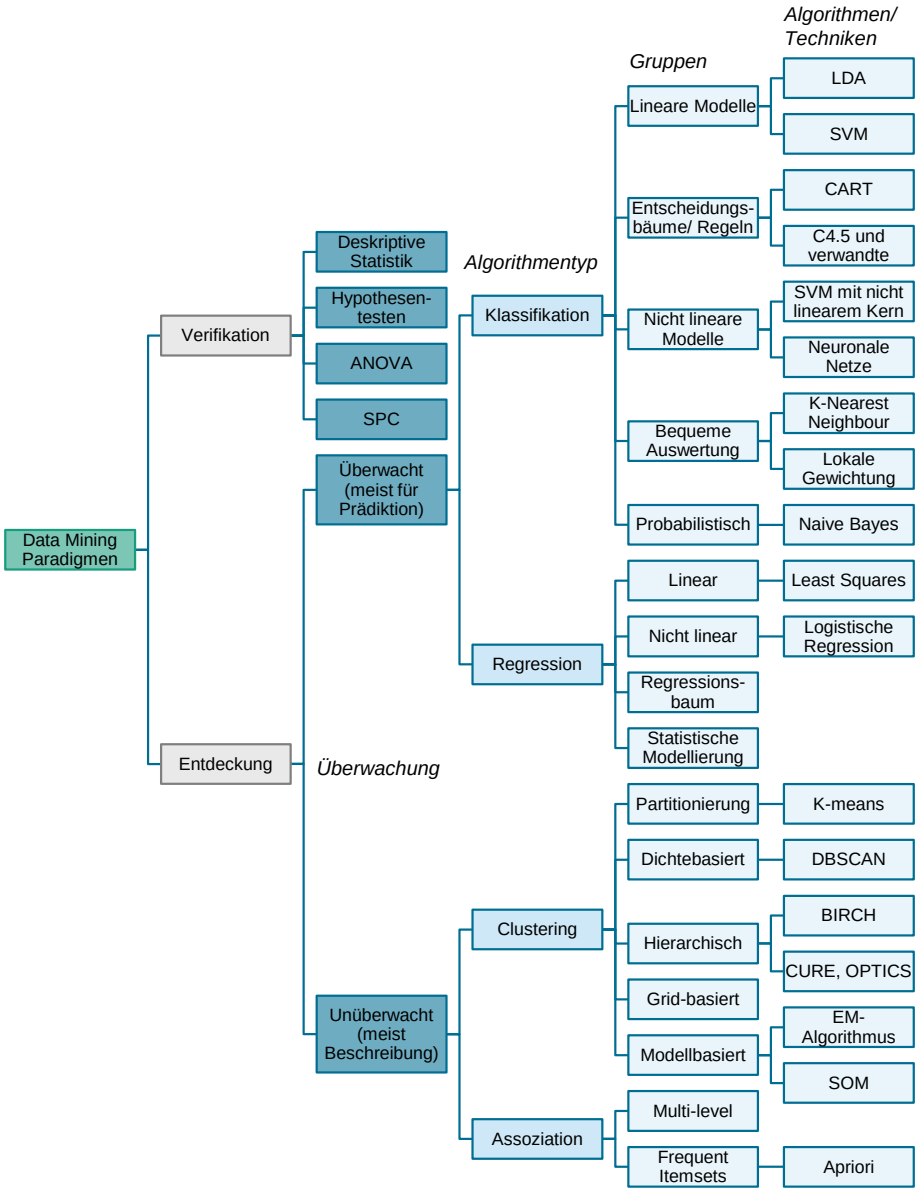

Bild 2-13: Taxonomie von Data-Mining-Algorithmen [GJM+16]

Für Machine-Learning-Verfahren schlagen SASSI ET AL. eine Taxonomie vor, welche zwei Möglichkeiten zur Klassifizierung von Algorithmen bietet: (1) Gruppierung von Algorithmen nach Lernstil (überwacht, unüberwacht, halbüberwacht und verstärkendes Lernen) und (2) Gruppierung nach der Ähnlichkeit oder der Funktion der Algorithmen wie z. B. bayesianische und instanzbasierte Algorithmen [SAB19]. Neben diesen allgemeinen Klassifikationen existieren auch spezialisierte Taxonomien zur Klassifikation bestimmter Lernaufgaben. Beispielsweise lassen sich Clustering-Methoden in hierarchische und partitionierende Verfahren einteilen [EIO+22]. Bei hierarchischen Verfahren werden die

Datenobjekte in einem hierarchischen Format in Ebenen unterteilt. Sie lassen sich weiterhin in agglomerative[11] und divisive[12] Verfahren aufteilen. Bei einem partitionalen Clustering-Algorithmus werden die Daten in einer verschachtelten Folge von Gruppen ohne hierarchische Struktur organisiert. Es folgt eine weitere Unterscheidung in hartes Clustering, Mischungsauflösung und Fuzzy Clustering.

Da stetig neue Algorithmen entwickelt werden, sind Gesamtübersichten schwer zu erhalten; einige Auflistungen existieren [Ham16-ol, Muk15, Sar21a, Sar21b]. Allein die dort genannten Algorithmen ergeben eine Gesamtanzahl von über 200, jeder mit seinen eigenen Modellannahmen, Stärken und Schwächen. GUGGENHEIM spricht davon, dass allein über 180 verschiedene Klassifikationsalgorithmen existieren [Gug23]. Folglich sieht sich ein Umsetzer von Data Analytics mit einer enormen Anzahl von potenziellen Verfahrensgruppen und speziellen Algorithmen konfrontiert. Nach der „No free lunch"-Theorie gibt es aktuell keine universell optimale Methode, die in allen Situationen anwendbar ist [Wol96]. Die Auswahl geeigneter Modelle ist daher ein bekanntes Problem im Data Analytics.

Erschwert wird die Auswahl durch verschiedene Einflussfaktoren, wie die Anwendung, die Charakteristika der verfügbaren Daten und Algorithmen sowie Randbedingungen. Die Einflussfaktoren auf Klassifikationsaufgaben wurden bereits genauer analysiert. REDA ET AL. unterscheiden hier (1) technische und (2) nicht-technische Faktoren (s. Bild 2-14) [MNS19]. Zu den technischen Faktoren zählen zum einen die Metadaten des Datensatzes, die einer der drei Gruppen einfache, statistische und informationstheoretische Merkmale zugeordnet werden können. Einfache Metadaten oder allgemeine Datenmerkmale sind Messungen, die leicht berechnet werden können, da sie direkt aus den Daten gewonnen werden. Statistische Metadaten sind hauptsächlich Diskriminanzanalysen und andere Messungen, die nur für kontinuierliche Attribute berechnet werden können. Statistische Metadaten stellen die statistischen Eigenschaften der Daten dar, z. B. die Abweichung des Verlaufs einer Verteilung vom Verlauf einer Normalsverteilung (Kurtosis – Wölbung). Informationstheoretische Metadaten sind Metadaten, die nur für kategoriale Attribute berechnet werden können. Zum anderen können den technischen Merkmalen *Landmarking*, *Modell-basierte Metadaten* und *Charakterisierung von Klassifikationsalgorithmen* zugeordnet werden. Landmarking versucht die Position eines bestimmten Lernproblems im Raum aller Lernprobleme zu bestimmen, indem die Leistung einiger einfacher und effizienter Lernalgorithmen direkt gemessen wird [Pfa01]. Bei modellbasierten Metadaten handelt es sich um ein Entscheidungsbaummodell mit verschiedenen

[11] Bei der agglomerativen Methode werden Cluster aus einzelnen Objekten gebildet, die iterativ zu größeren Clustern zusammengeführt werden, die die verschiedenen Ebenen der Hierarchie bilden, bis das gesamte Objekt ein einziges Cluster enthält oder bis das Abbruchkriterium erfüllt ist.

[12] Das divisive hierarchische Clustering ist eine Umkehrung des agglomerativen Clustering-Prozesses, bei dem jeder Cluster in kleinere Teile aufgeteilt wird, wobei jedes Objekt in einem einzigen Cluster beginnt, bis die erforderliche Anzahl von Clustern erreicht ist.

Eigenschaften, die aus dem Datensatz erstellt werden [RSG+14]. Beispiele für Eigenschaften von Entscheidungsbaummodellen sind die Anzahl der Blätter, die Anzahl der Knoten, die Knoten pro Attribut, die Knoten pro Probe und die Blattkorrelation. Die Annahme ist, dass das aus den Datensätzen abgeleitete Entscheidungsbaummodell die Eigenschaften besitzt, die in hohem Maße vom Datensatz abhängen. Die Charakterisierung von Klassifikationsalgorithmen nutzt die Eigenschaften der Modelle, wie z. B. ihre Performance: Genauigkeit, Komplexität und Trainingsdauer. Einige Studien vergleichen verschiedene Algorithmen anhand solcher Merkmale (z. B. [DT13], [LLS00]).

Zu den nicht-technischen Faktoren gehören die Fachkenntnisse des Data Scientists in der Geschäftsdomäne, die Vertrautheit mit einem Algorithmus sowie die Benutzerfreundlichkeit des Algorithmus und die Verständlichkeit der Ergebnisse.

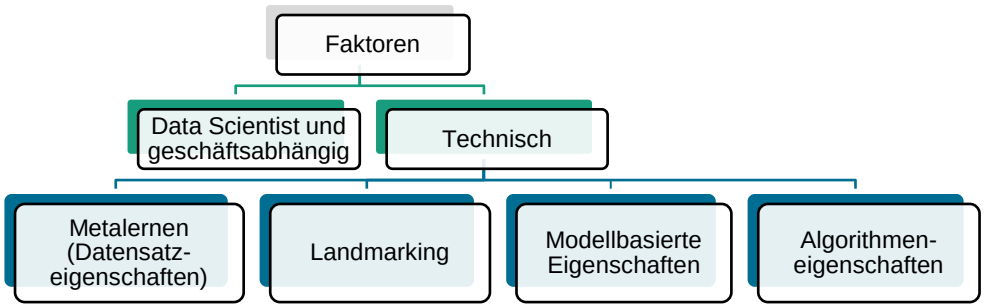

Bild 2-14: Einflussfaktoren auf Klassifikationsaufgaben [MNS19]

In der Literatur besteht Konsens darüber, dass insbesondere die Dateneigenschaften einen großen Einfluss auf die Algorithmenauswahl haben [BTG12, TP17, PC19, PC20, TZ21]. Andere betonen die Vielfalt von Einschränkungen und Kriterien. BRODLY und SMYTH empfehlen einen Abgleich der identifizierten problemspezifischen Faktoren mit allgemeinen Merkmalen der Modelle, um ein bestimmtes Modell bzw. einen bestimmten Algorithmus auszuwählen [BS95]. Sie nennen neben den Datenfaktoren zwei weitere problemspezifische Hauptfaktoren. Für jeden Hauptfaktor gibt es wieder mehrere relevante Faktoren und Kriterien. Diese Faktoren können entweder verändert werden (wie z. B. mehr Daten erheben) oder nicht (wie z. B. die Einschränkung, dass der Klassifikator in ein größeres System eingebettet sein muss und Schätzungen der Klassenwahrscheinlichkeiten anhand der Merkmalsdaten erstellen muss). Die Hauptfaktoren sind:

- **Anwendungsfaktoren**, darunter die Projektziele, das verfügbare Domänenwissen über das Problem und operationale Faktoren, die bestimmen, wie das Modell in der Praxis genutzt wird.
- **Menschliche Faktoren**, die das unterschiedliche Wissen der Data Scientists, der Kunden/Nutzer und der Domänenexperten berücksichtigen.

- **Datenfaktoren**, die Aspekte wie die Repräsentation der Daten (kontinuierlich oder kategorisch), die Dimensionalität, die Anzahl der Klassen und die Art der Datenerhebung (fehlende Daten, Rauschen usw.) berücksichtigen.

NALCHIGAR ET AL. sprechen von drei Kontexten, die die Anwendung eines Algorithmus einschränken und angeben, wann er zu verwenden ist: **Benutzer** (persönliche Anforderungen und Aufgaben, z. B. Einfachheit des Algorithmus), **Daten** (Soll- und Ist-Eingangsdatenmerkmale wie Typ) und **Modell** (z. B. Parameterkonfiguration) [SE18, NY18]. Beispielsweise kann der Entscheidungsbaum eine geeignete Wahl sein, wenn der Benutzer ein leicht interpretierbares Modell bevorzugt, während sich ein Naive-Bayes-Klassifikator gut für die Verarbeitung von Textdaten eignet. Zudem zeigen sich Unterschiede darin, wie verschiedene Algorithmen mit fehlenden Werten umgehen – einige sind weniger tolerant als andere.

Fazit: Es gibt zahlreiche Methoden und Algorithmen mit jeweiligen Vor- und Nachteilen, und kontinuierlich kommen neue hinzu. Dies erweitert den Lösungsraum erheblich, wenn es um potenzielle Algorithmen für die datengestützte Produktplanung und die verschiedenen relevanten Lernaufgaben (wie Clustering, Klassifikation und Abhängigkeitsanalyse, vgl. Abschnitt 2.3.2). Es existieren zwar erste anwendungsspezifische Klassifikationsparadigmen, wie z. B. für die Bioinformatik, die den Lösungsraum etwas einschränken [LCS+06]. Eine solche Auflistung und Struktur gibt es jedoch nicht für die datengestützte Produktplanung oder verwandte Domänen. Es bedarf daher einer Übersicht, welche Modelle und Algorithmen für die strategische Produktplanung relevant sind.

Darüber hinaus wird die Auswahl neben der Vielzahl durch diverse zu berücksichtigende komplexe Einflussfaktoren erschwert, wie Anwendung, Dateneigenschaften und Charakteristika der Algorithmen. Sie müssen zum Großteil im Prozess vor der Modellierung adressiert und mit den dazugehörigen Algorithmen zusammengebracht werden. Vorgeschlagene Faktoren sind zudem teilweise nur mit Hilfe von Vorkenntnissen zu bestimmen oder in einem automatisierten Umfeld, wie es im Kontext des Meta-Lernens der Fall ist. Insbesondere die menschlichen bzw. nicht technischen Faktoren spielen für die datengestützte Produktplanung eine Rolle, da besondere Anforderungen durch die Nutzung der Ergebnisse in der Produktentwicklung bestehen. Daher wird eine Unterstützung der Modell- und Algorithmenauswahl benötigt, welche die genannten Einflussfaktoren berücksichtigt und das Zusammenspiel nachvollziehbar und einfach zugänglich macht.

2.3.4.3 Evaluation

Das Thema Evaluation ist zentral für jeden Datenanalyseprozess. Die Evaluation dient den folgenden Zwecken [Ras18]:

- **Leistungsmessung**: die Überprüfung, wie gut das Modell generalisiert, also unbekannte Daten vorhersagt.

- **Modellauswahl**: der Prozess, das beste Modell aus einer Menge von Modellen zu finden, die mit unterschiedlichen Hyperparametereinstellungen erstellt wurden, zur Optimierung einer Leistungsmetrik.
- **Modell- und Algorithmenvergleich**: die Auswahl des Algorithmus, der für das jeweilige Problem am besten geeignet ist.

Die Modelle bzw. die gesamten Pipelines werden folglich in der letzten Phase des Hauptprozesses im Hinblick auf den technischen Erfolg bewertet. In der Evaluierung wird auch die Unterscheidung zwischen überwachten und unüberwachten Methoden relevant. Wenn Labels vorhanden sind, wie es beim überwachten Lernen erforderlich ist, ist es häufig einfacher, die Modellleistung zu bewerten, da die korrekten Ausgaben bekannt sind. Beim unüberwachten Lernen ist es oft der Fall, dass das dahinterstehende Modell nur der erste Schritt in einer längeren Analysepipeline war, und die Nützlichkeit des Ergebnisses nur durch den Erfolg der größeren Operation, wie z. B. eines Vorhersagemodells, getestet werden kann. Wenn das unüberwachte Modell an sich der Gewinnung von Erkenntnissen dient und der Output für den Menschen interpretierbar sein soll, können entweder externe Daten oder intrinsische Validierungsverfahren helfen, dies zu bewerten. Letztere stehen beispielsweise beim Clustering zur Verfügung und basieren auf den durch die Daten verfügbaren Informationen [Zim20].

LIU ET AL. fanden heraus, dass für die Auswahl der geeigneten Metriken eine Strategie hilfreich ist, da die Bewertung der Modellleistung eine entscheidende Rolle bei der Konstruktion und Auswahl von Modellen spielt [LZW+14]. Je nach Anwendungsszenario werden verschiedene Metriken vorgeschlagen. So ist beispielsweise die Genauigkeit die primäre Metrik zur Bewertung der Leistung von Klassifikatoren. Häufig schneiden Modelle in Bezug auf eine Metrik gut und in Bezug auf eine andere schlecht ab. Neuronale Netze optimieren in der Regel den quadratischen Fehler, so dass die Metrik des mittleren quadratischen Fehlers die tatsächliche Leistung eines Klassifizierers besser wiedergeben kann als andere Metriken. Eine Auswahlstrategie scheint darin zu bestehen, die Metrik in Abhängigkeit von den praktischen Anforderungen einer bestimmten Anwendung zu wählen. Solche spezifischen Kriterien sind jedoch oft nicht bekannt, so dass Praktiker dazu neigen, sehr allgemeine Metriken zu verwenden.

Da der Raum möglicher Modelle in der datengestützten Produktplanung derzeit sehr groß ist, sind auch die potenziellen Metriken zahlreich, was die Auswahl der Bewertungsmetriken ebenfalls erschwert.

Fazit: Da in der betriebsdatengestützten Produktplanung sowohl überwachte als auch unüberwachte Algorithmen zum Einsatz kommen und adäquate Bewertungen wichtig sind, ist die richtige Auswahl aus einer Vielzahl an potenziell einsetzbaren Evaluierungsmetriken, die teilweise auch eher unbekannt sind, entscheidend. Daher bedarf es einer Strukturierung der Metriken.

Bild 2-15 fasst die im Kapitel 2 identifizierten Schritte und -unterschritte des gesamten betrachteten Prozesses sowie die aufgedeckten Abhängigkeiten bzw. Auswahlfaktoren zwischen diesen zusammen.

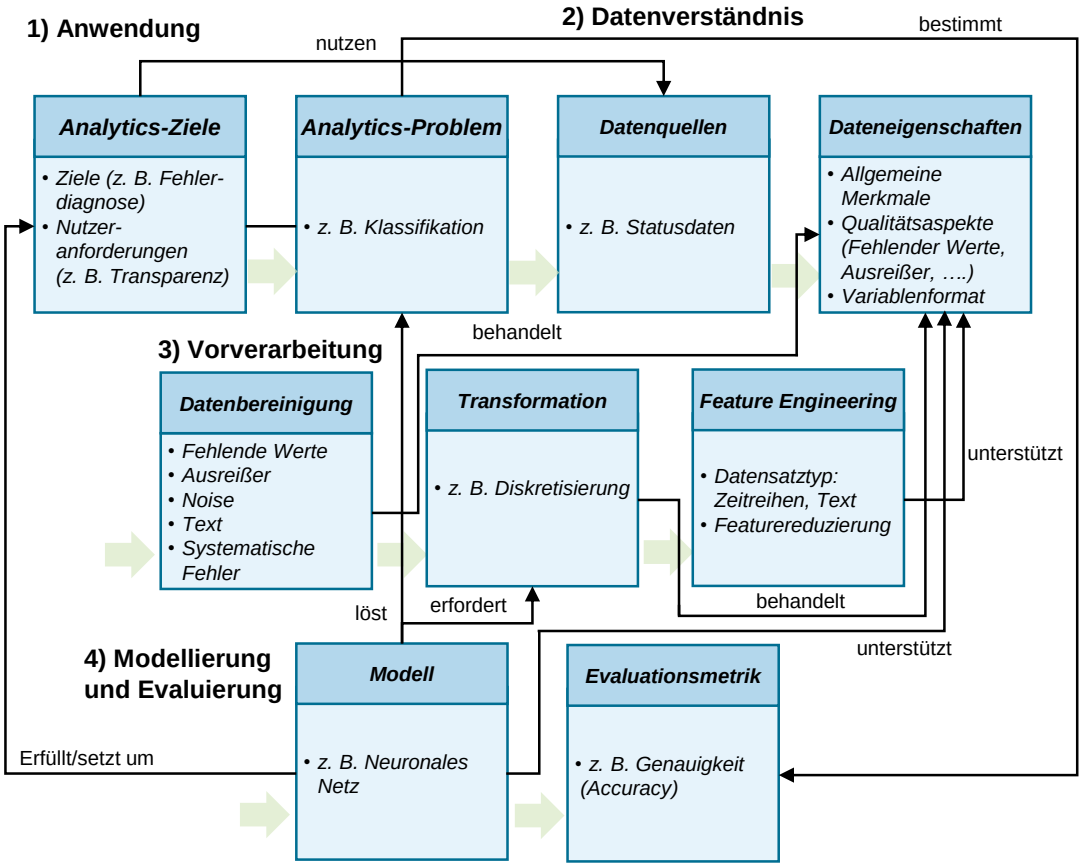

Bild 2-15: Abhängigkeiten in der Data Analytics Pipeline

2.3.5 Umsetzung von Data Analytics Pipelines

Im Rahmen der Umsetzung von Data Analytics und Machine Learning, insbesondere für Big Data, sind eine Vielzahl an Herausforderungen zu nennen, mit denen sich die Forschungs-Community beschäftigt. Beispielhafte Herausforderungen sind die Hyperparameteroptimierung oder der hohe Rechenbedarf im Kontext der Modellierung [BJS19]. Innerhalb dieser Arbeit werden Tools näher betrachtet, mit denen Data-Analytics-Verantwortliche die Data Analytics Pipeline, insbesondere die Vorverarbeitung und Modellierung, umsetzen können. Tools spielen eine wichtige Rolle, da diese eine Vielzahl an Herausforderungen adressieren und den Nutzer unterstützen. Jedoch wurden mit dem wachsenden Bedarf für Data Science und Analytics eine Vielzahl an Tools und Technologien für diverse Aufgaben im Analytics-Prozess und Funktionen entwickelt, sodass die Tool-Auswahl schon für Data-Analytics-Experten herausfordernd ist [NS23]. So kommen einige Forscher und Blog-Autoren zu dem Schluss, dass Data Scientists Unterstützung bei der Auswahl benötigen und vergleichen diverse Tools [NDB+19, DDS+17]. Um die Vielseitigkeit von Tools besser zu verstehen und einen Startpunkt für die Auswahl zu haben,

schlagen BARTSCHAT ET AL. eine Kategorisierung in neun verschiedene Typen vor [BRM19]: Data-Mining-Suites, Business-Intelligence-Pakete, Mathematische Pakete, Integrationspakete, Erweiterungen, Data-Mining-Bibliotheken, Forschungsprototypen, Spezialisierungen und Lösungen.

NADI und SAKR führten eine große Umfrage mit Data Scientists durch, um zu verstehen, welche Faktoren Data Scientists beeinflussen, wenn sie Softwarebibliotheken (Python und R) auswählen [NS23]. Dazu nutzten sie die 26 Auswahlfaktoren von LARIOS VARGAS ET AL., die das Ergebnis einer Untersuchung zu Beeinflussungsfaktoren von Softwareentwicklern bei der Auswahl von Drittanbieter-Bibliotheken waren [LAT+20]. Diese Faktoren lassen sich in technische, menschliche und ökonomische Faktoren gruppieren:

- **Technische Faktoren** umfassen Faktoren in Bezug auf Funktionalität, Qualität, Art des Projekts und Freigabeprozess. Beispiele sind Nutzbarkeit, Dokumentation, Sicherheit, Performance und Häufigkeit von Release-Zyklen.
- **Menschliche Faktoren** beziehen sich auf Stakeholder, die Organisation, das Individuum und die Community. Dazu gehören beispielsweise Kunden, Typ der Industrie, Erfahrung und Beliebtheit der Bibliothek.
- **Ökonomische Faktoren** betreffen die Gesamtbetriebskosten, die Lizenz und das Risiko eines Projekts.

Sieben Faktoren davon sind für Data Scientists wichtig: die fünf Faktoren Nutzbarkeit/Benutzerfreundlichkeit, Zweckmäßigkeit, Dokumentation, Reife und Stabilität sowie Performance sind technischer Natur. Zwei sind den menschlichen Faktoren zuzuordnen: Aktivitätsgrad und Erfahrung. Darüber hinaus berücksichtigen Data Scientists noch drei weitere Faktoren: statistische Stichhaltigkeit, Konsistenz und Skalierbarkeit sowie Individualisierung. NADI und SAKR betonen in dem Kontext den großen Mehrwert von Unterstützungstools zur Auswahl von Bibliotheken. Die Umfrage ergab schließlich auch, dass viele Data Scientists Probleme mit der Auswahl von Tools haben, die sich am besten für die jeweiligen Aufgaben und Anwendungen eignen oder ein System schätzen würden, welches Bibliotheken auf Grundlage von Faktoren, die für sie von Interesse sind, empfiehlt.

Fazit: Tools spielen bei der erfolgreichen Umsetzung von Data Analytics eine große Rolle, jedoch erschwert die enorme Vielzahl an Tools und Bibliotheken, welche sich in verschiedene Kategorien einteilen lassen, sowohl Data Scientists als auch Nicht-Experten die Auswahl der passenden Werkzeuge. Eine systematische Auswahlunterstützung verspricht hier die Identifikation der Tools, welche für die jeweiligen Anwendungen am besten geeignet sind. Außerdem sollten bei der Auswahl eine reduzierte Anzahl an technischen und menschlichen Faktoren berücksichtigt werden.

2.4 Demokratisierung von KI und Data Analytics

Ein allgemeiner Ansatz, um die Herausforderungen zu lösen, die sich daraus ergeben, dass es nur wenige Data-Science-Experten gibt und wichtige Aufgaben im Analyseprozess, wie z. B. die Auswahl von Algorithmen, erfolgreich durchgeführt werden müssen, ist die Demokratisierung von KI und maschinellem Lernen durch effiziente Tools, die Teile des Analyse-Workflows automatisieren [PSV+21, MS21]. Hier sind die beiden Forschungsfelder Automated Machine Learning (AutoML) und Meta-Lernen zu berücksichtigen:

AutoML: Beim automatisierten maschinellen Lernen (AutoML) werden Teile der maschinellen Lernpipeline wie Feature-Engineering, Architektursuche und kombinierte Modellauswahl als auch Hyperparameteroptimierung vollständig automatisiert, so dass die Leistung für einen bestimmten Anwendungsfall optimiert wird [TAR+19, BRH+22]. Die Automatisierung des Prozesses erspart Experten nicht nur die Zeit und den Aufwand umfangreicher, oft mühsamer Experimente, sondern ermöglicht es auch Nicht-Experten, eine wesentlich bessere Leistung zu erzielen, als dies sonst möglich wäre. AutoML-Systeme erreichen diese Vorteile oft mit recht hohen Rechenkosten. Im Zusammenhang mit der KI-Entwicklung wird auch von AutoAI [WRW+20] gesprochen.

Techniken zur Lösung von AutoML-Problemen können in grundlegende und erfahrene Techniken unterteilt werden [QMH+18]. Grundlegende Techniken reichen von einfachen Suchansätzen wie Grid Search (Brute Force), über Gradient Descent bis hin zu komplexeren Ansätzen wie Reinforcement Learning. Erfahrene Techniken lernen und akkumulieren Wissen aus vergangenen Suchen oder externen Daten. Zwei beliebte Methoden sind hier Meta-Lernen und Transfer-Lernen.

Meta-Lernen beschreibt die Anwendung von ML-Techniken auf frühere ML-Experimente und zielt darauf ab, bestimmte Aspekte des Lernprozesses zu verändern, um die Leistung der Ergebnisse zu verbessern [JDG11]. Beim traditionellen Meta-Lernen wird der Lernalgorithmus als Black Box behandelt, wobei die beobachtete Leistung des Ausgabemodells mit den Merkmalen der Eingabedaten korreliert wird.

Die meisten AutoML-Systeme enthalten jedoch nur wenige oder gar keine Informationen über den Prozess der Auswahl und Erzeugung von Modellausgaben. Daher verstehen die Nutzer weder den Prozess noch vertrauen sie den Ergebnissen [WWO+20]. Dies kann im Ingenieurwesen problematisch sein, wo Vertrauen und Transparenz entscheidend sind.

Darüber hinaus haben sich die meisten Arbeiten zu AutoML auf überwachtes Lernen konzentriert. Aufgaben wie unüberwachtes Lernen haben sich als wesentlich schwieriger zu automatisieren erwiesen, da die Optimierungsziele eher subjektiv und abhängig von der Domäne sind und viele Kompromisse erfordern [BRH+22].

Außerdem werden oft viele Daten benötigt, z. B. beim Meta-Lernen, das Attribute wie die Größe des Datensatzes sowie die Anzahl und Aspekte der Merkmale mit den

Leistungsdaten verwendet, um gute Lernmodelle aus früheren Aufgaben zu lernen. Fehlen diese, wird das "Lernen zu lernen" unmöglich.

LEE und MACKE behaupten, dass die vollautomatische Einstellung der heutigen AutoML-Systeme für viele Benutzer und Problemdomänen nicht die optimale Lösung ist. In einigen Szenarien, in denen Vertrauen und Transparenz von entscheidender Bedeutung sind und in explorativen Situationen, in denen das Problem nicht genau definiert ist, kann der Mangel an menschlicher Kontrolle und Interpretierbarkeit problematisch sein [LMX+19]. Zusätzlich besteht bei solchen automatisierten Tools die Gefahr, dass insbesondere "Data-Science-Neulinge" oder ungeschulte Citizen Data Scientists diese ohne angemessenes Kontextverständnis verwenden, was die Wahrscheinlichkeit von Fehlern erhöht. AutoML-Tools kompensieren dabei keine Defizite in Bezug auf Fachwissen, Ausbildung und Erfahrung [BS22-ol].

Aufgrund dieser Problematiken setzen andere (Sub-)Disziplinen auf den Einbezug von Fach- bzw. Domänenwissen oder menschlichen Nutzern. Das neuere Forschungsfeld namens Human-guided machine learning (HGML) konzentriert sich darauf, wie Nutzer dabei unterstützt werden können, Domänenwissen zu nutzen, um ein AutoML-System bei der Auswahl von maschinellen Lernalgorithmen zu leiten und mehrstufige Lösungen zu finden [GHG+19]. Die automatisierte Datenwissenschaft (Automated Data Science) umfasst Versuche, jeden Teil des Data-Science-Prozesses (wie CRISP-DM) zu automatisieren. Dazu gehört auch das maschinelle Lernen als eines der Werkzeuge des Datenwissenschaftlers. AutoML kann also als ein Werkzeug im Kontext der automatisierten Datenwissenschaft verwendet werden. Doch gerade die Vollautomatisierung birgt noch viele Herausforderungen. DE BIE ET AL. haben mehrere Herausforderungen bei der Automatisierung des gesamten Prozesses zusammengefasst, die verschiedene Schritte der Data-Science-Pipeline betreffen [BRH+22]:

- Verbesserung der Zusammenarbeit zwischen Mensch und KI durch Einbeziehung des Domänenkontextes zur interaktiven Definition und Verfeinerung des Ziels von Data-Science-Aktivitäten
- Ausweitung von AutoML auf Aufgaben, die über überwachtes Lernen hinausgehen
- Generierung aussagekräftiger Merkmale unter Berücksichtigung von Domänenkontext und Aufgabe
- Beschleunigung der Prozesse zur Datenbereinigung, Ausreißererkennung und Imputation von Daten
- Entwicklung von kollaborativen Systemen zwischen Mensch und KI für die Daten- und Musterexploration
- Erleichterung der Abfrage, Validierung und Erklärung von Modellen und Ergebnissen

Ein weiterer Ansatz ist Semantic Data Mining (SDM). SDM bezieht sich auf Data-Mining-Aufgaben, die systematisch Domänenwissen in den Prozess einbeziehen, insbesondere formale Semantik [DWL15]. Hier wird argumentiert, dass Domänenwissen eine große Rolle in allen Schritten des Data-Mining-Prozesses haben kann. Forschung auf dem Gebiet hat ergeben, dass z. B. während des Mustergenerierungsprozesses, Domänenwissen als Vorwissen über Einschränkungen dienen kann, um den Suchraum zu reduzieren und den Suchpfad zu lenken [BBM13, BFG+07]. Darüber hinaus können die entdeckten Muster, also die Ergebnisse, bereinigt oder durch Kodierung besser sichtbar gemacht werden [MOR11, MG10, WD10], [MG10]. Um solches Domänenwissen im Data-Mining-Prozess nutzbar zu machen, verwendet SDM-Techniken, mit denen das Wissen formal und maschinenverstehbar abgebildet werden kann, wie beispielsweise mit Hilfe von Ontologien. DOU ET AL. fanden hier drei Aspekte, zu deren Zweck Ontologien im SDM verwendet werden [DWL15]:

1) Überbrückung der semantischen Kluft zwischen den Daten, Anwendungen, Data-Mining-Algorithmen und Data-Mining-Ergebnissen

2) Versorgung von Data-Mining-Algorithmen mit A-Priori-Wissen, das entweder den Prozess leitet oder den Suchraum einschränkt

3) Bereitstellung einer formalen Methode zur Darstellung des Data-Mining-Flusses

HILARIO ET AL. sprechen in einem ganz ähnlichen Kontext von Semantic Meta Mining (SMM) und beschreiben dieses Feld als Data-Mining-Prozess- oder Workflow-Mining, was gleichzeitig durch Meta-Daten und durch das kollektive Fachwissen der Experten, verankert in einer Wissensbasis, getrieben wird [HKN+09]. Die SMM-Technik geht davon aus, dass Data Mining Workflows durch die Auswahl von Algorithmen/Modellen unter Verwendung eines Beschreibungsrahmens aufgebaut werden, der die komplexen Beziehungen zwischen Aufgaben, Daten und Algorithmen in verschiedenen Phasen des DM-Prozesses verdeutlicht. Auch hier wird die Ontologie als maschinenverständliche Beschreibungssprache für die Darstellung der Wissensbasis verwendet [JZL+22].

Nach TIANXING ET AL. sind die Rollen des klassischen Meta-Lernens und des semantischen Meta-Minings nicht gegensätzlich [TZ21]. Die Lernziele des Meta-Learnings sind detaillierter (z. B. die Parameter der Algorithmen). Das semantische Meta-Mining liefert die geeignete Algorithmusauswahl und formuliert den Ausführungsprozess. Diese Vorschläge sind allgemeiner gehalten. Ein solches semantisches Meta-Mining kann in der Regel auch das Kaltstartproblem des Meta-Learnings lösen, um sicherzustellen, dass der Lernprozess in die richtige Richtung läuft.

Neben solchen technischen Lösungen sind zusätzlich die bildungsorientierten Lösungen nicht außer Acht zu lassen, um das notwendige Hintergrundwissen aufzubauen und so das Risiko von Fehlern durch ahnungsloses Anwenden von genannten automatisierten Lösungen zu reduzieren [BS22-ol]. Der Sprung von einem datenaffinen Mitarbeiter hin zu einem Citizen Data Scientist erfordert eine neue Denkweise: sie müssen wissen, wie

sie Analysemöglichkeiten bewerten, über oberflächliche Dateneinblicke hinausgehen und die richtigen Fragen stellen können, um die effektivsten Ergebnisse zu erzielen [Zio19-ol]. Verschiedene Ansätze wie Online-Kurse, Software-Trainings, Konferenzen und Meetups versprechen das notwendige tiefe Verständnis zu vermitteln. Allein solche intensiven, meist teuren Einzelmaßnahmen können vielfach nicht die notwendigen, kontinuierlichen praktischen Erfahrungen abbilden; hier können zusätzlich Leitfäden, Best Practices und Templates für „On-the-Job-Trainings“ helfen [BS22-ol].

Fazit: Die Automatisierung des maschinellen Lernens und der Datenwissenschaft verspricht, die Aufgaben des Datenwissenschaftlers erheblich zu vereinfachen und Nicht-Experten dabei zu helfen, schnell Ergebnisse zu erzielen. Die Ansätze und Konzepte, die den Menschen und seine Anforderungen stärker in den Mittelpunkt stellen und den gesamten Prozess einschließlich der Bestimmung relevanter Anwendungsfälle begleiten, befinden sich jedoch weitgehend in einem sehr frühen Stadium [NP19]. Zusätzlich bieten Ansätze, die Fachwissen verarbeiten und somit nicht als Black-Box-Modelle fungieren, in Kombination mit bildungsorientierten Ansätzen eine weitere interessante Möglichkeit, den Nutzer beim Data-Analytics-Prozess bestmöglich zu unterstützen. Sie sollten daher für die Systematik eingesetzt werden.

2.5 Ziele an die Systematik

Aus der vorangegangenen Problemanalyse wird ersichtlich, dass der Datenanalyseprozess, im Wesentlichen bestehend aus der Definition der Anwendung, der Datensammlung und -beschreibung, der Methodenauswahl und Umsetzung, für die betriebsdatengestützte Produktplanung verschiedene Herausforderungen und zu berücksichtigende Faktoren zur Ausgestaltung des Gesamtprozesses sowie der einzelnen Prozessschritte mit sich bringt. Diese lassen sich, wie in Bild 2-16 dargestellt, zusammenfassen:

Data-Analytics-Prozess übergreifend

Aufgabe	Konzipierung und Umsetzung einer erfolgsversprechenden DA-Pipeline, von der Use-Case-Definition, über den Aufbau von Datenverständnis bis zur Modellierung und Evaluierung
Herausforderung	1) Komplexität und mangelnde Prozesskenntnisse (Kap. 2.2.3) 2) Fachkräftemangel, fehlendes Know-How, Mangel an erschwinglichen Beratungs- und Analytics-Services und Produkten (Kap. 2.2.3)
Ausgestaltungs-faktoren	1) Berücksichtigung der folgenden Aufgaben (Kap. 2.3.1): Definition der Anwendung, Datenverständnis, Methodenauswahl und Umsetzung 2) Ansprache und Know-How-Aufbau der Citizen Data Scientists; Transparenz (Kap. 2.3)
Ziele an die Systematik	1) Bereitstellung eines Strukturierungsrahmens für die Datenanalyse von Betriebsdaten in der strategischen Produktplanung 2) Befähigung von Nicht-Experten aus der Industrie

Anwendung

Aufgabe	Definition von Data-Analytics-Anwendungen für die Produktplanung (aus Sicht der Data Scientisten)
Herausforderung	Manager und Data Scientists sprechen nicht dieselbe „Sprache“ (Kap. 2.3.2)
Ausgestaltungs-faktoren	Berücksichtigung von Geschäftszielen, Data-Analytics-Zielen, weiteren Faktoren und Data-Analytics-Problemen (Kap. 2.3.2)
Ziele an die Systematik	1) Bereitstellung von relevanten Analytics-Zielen und -Problemen für die Produktplanung 2) Übersetzung der Business Use Cases in Analytics-Aufgabenstellungen

Datenverständnis

Aufgabe	1) Datenidentifikation und –sammlung 2) Aufbau von Datenverständnis
Herausforderung	1) Vielzahl potenzieller Betriebsdatenquellen mit diversen Datensätzen (konkrete Übersicht auf Ebene der Datensätze fehlt) (Kap. 2.3.3) 2) Heterogenität der intrinsischen Eigenschaften (Kap. 2.3.3)
Ausgestaltungs-faktoren	1) Betriebsdaten können anhand verschiedener extrinsischer Merkmale (z.B. Ursprung, Inhalt etc.) klassifiziert werden (Kap. 2.3.3) 2) Intrinsische Charakteristika wie Struktur, Format, Verarbeitung, Qualität zum Aufbau von Datenverständnis als Vorbereitung der nachfolgenden Analyse (Kap. 2.3.3)
Ziele an die Systematik	1a) Bereitstellung einer strukturierten und detaillierten Betriebsdatenübersicht 1b) Bestimmung der relevanten Betriebsdaten 2a) Bereitstellung eines Beschreibungsrah-mens für Betriebsdaten 2b) Bestimmung der Dateneigenschaften

Methodenauswahl

Aufgabe	Auswahl der richtigen Vorverarbeitungs-, Modellierungs- und Evaluierungstechniken
Herausforderung	1) Vielzahl an potenziellen Techniken (Kap. 2.3.4) 2) Abhängigkeiten durch den gesamten Prozess (Kap. 2.3.4)
Ausgestaltungs-faktoren	1) Berücksichtigung von Techniken der Datenbereinigung, Transformation und Merkmalskonstruktion sowie Algorithmen aus dem Clustering, der Klassifikation und der Abhängigkeitsanalyse (Kap. 2.3.4) 2) Anwendungsfaktoren, Menschliche Faktoren und Datenfaktoren (Kap. 2.3.4.2)
Ziele an die Systematik	1) Bereitstellung von relevanten Vorverarbeitungsmethoden, Algorithmen, und Evaluierungsmetriken 2) Auswahl geeigneter Methoden unter Berücksichtigung der Abhängigkeiten

Umsetzung

Aufgabe	Umsetzung der Vorverarbeitung, Modellierung und Evaluierung
Herausforderung	Vielzahl an potenziellen Tools (Kap. 2.3.5)
Ausgestaltungs-faktoren	Berücksichtigung wichtiger Entscheidungsfaktoren (technisch und menschlich) (Kap. 2.3.5)
Ziele an die Systematik	Auswahl geeigneter Tools anhand der Entscheidungsfaktoren

Bild 2-16: Zusammenfassung der Aufgaben, Herausforderungen, Ausgestaltungsfaktoren und Ziele entlang des DA-Prozesses

Nachfolgend werden die Ziele der Systematik zur Datenanalyse von Betriebsdaten in der strategischen Produktplanung im Detail vorgestellt. Sie werden anhand des Data-Analytics-Prozesses strukturiert und adressieren die Herausforderungen und Ausgestaltungsfaktoren.

Data-Analytics-Prozess übergreifende Ziele

Z1) Befähigung von Nicht-Experten/Citizen Data Scientists in der Industrie

Der Fachkräftemangel erschwert vielen Unternehmen den (erfolgreichen) Einsatz von Data Analytics. Insbesondere KMUs haben mit den Herausforderungen der Datenanalyse zu kämpfen, da oftmals Datenanalyseexperten und Know-How fehlen. Hinzu kommt, dass die betriebsdatengestützte Produktplanung im Kontext von cyber-physischen Produkten neben den erforderlichen Fähigkeiten aus den Disziplinen Informatik und Statistik zusätzliche Kompetenzen im Bereich der Ingenieurwissenschaft erfordert, was den Mangel an geeigneten Experten zusätzlich verstärkt. Daher soll die Systematik insbesondere Nicht-Experten, Einsteigern und sog. „Citizen Data Scientists“ zur Aufstellung und Umsetzung einer erfolgsversprechenden Data-Analytics-Pipeline befähigen, welche auch das erforderliche Wissen im Kontext der Anwendung aus der Produktplanung berücksichtigt. Um das Bedürfnis in der Industrie nach Transparenz zum Aufbau von Vertrauen zu erfüllen, soll die Systematik zusätzlich Verständnis über die Methoden und Zusammenhänge fördern. Die Nachvollziehbarkeit der getroffenen Auswahl und der resultierenden Ergebnisse soll gewährleistet werden, indem die Nutzer und ihre Anforderungen berücksichtigt werden.

Z2) Bereitstellung eines Strukturierungsrahmens für die Datenanalyse von Betriebsdaten in der strategischen Produktplanung

Die Problemanalyse hat gezeigt, dass ein Prozess aus Anwendungsdefinition, Datenverständnis, Methodenauswahl und Umsetzung für ein Data-Analytics-Projekt notwendig ist. Gleichzeitig schreckt die Komplexität von Datenanalyseprozessen im Zusammenspiel mit den Besonderheiten der strategischen Produktplanung viele Unternehmen ab, wie z. B. die teils unbekannten Anwendungsmöglichkeiten, die Vielseitigkeit der Betriebsdaten und die stetige Integration von Domänenwissen (in Form dieser Faktoren) zur Bestimmung passender Analytics-Verfahren. Daher soll die Systematik einen Strukturierungsrahmen für die Analyse von Betriebsdaten in der strategischen Produktplanung bereitstellen. Dieser soll als Grundlage für die betriebsdatengestützte Produktplanung dienen und eine einfache, lernförderliche Ausgestaltung ermöglichen. Um eine einfache Bedienbarkeit und Umsetzbarkeit zu gewährleisten, soll der Rahmen mit seinen Ausgestaltungsbausteinen als ein digitaler Assistent umgesetzt werden.

Ziele an die Definition von Analytics-Anwendungen für die strategische Produktplanung

Z3) Bereitstellung von relevanten Data-Analytics-Zielen und -Problemen

Um realisierbare Data-Analytics-Anwendungen zu definieren, können die Analytics-Ziele und -Probleme eine Brücke zu den Geschäftszielen schlagen. Daher soll die Systematik relevante Ziele und Probleme für die Produktplanung bereitstellen. Diese sollen sowohl die Sprache der Produktmanager als auch die der Data Scientists berücksichtigen.

Z4) Übersetzung der Business-Use-Cases in Analytics-Aufgabenstellungen

Die Übersetzung der geschäftlichen Ziele in konkrete Analytics-Anwendungen, welche als Startpunkt für die weiteren Aufgaben dienen können, ist nicht trivial und erfordert die zu vorige Zusammenstellung verschiedener Informationen. Die Systematik soll diesen Prozess und die Bestimmung der relevanten Faktoren dazu unterstützen.

Ziele an das Datenverständnis

Z5) Bereitstellung einer strukturierten und detaillierten Betriebsdatenübersicht

Betriebsdaten müssen anhand domänenrelevanter Aspekte wie Inhalt, Ursprung und Zugang strukturiert werden, um die erforderliche Übersicht zur Identifikation der relevanten Daten zu erhalten und die notwendigen Daten domänengetrieben identifizieren zu können. Als Teil der Systematik soll daher eine Übersicht über Betriebsdaten für Analytics-Use-Cases in der strategischen Produktplanung bereitgestellt werden, welche die Quellen entlang wichtiger Dimensionen der Produktplanung strukturiert und Datensätze einsortiert.

Z6) Bestimmung der relevanten Betriebsdaten

Aus der Problemanalyse ist hervorgegangen, dass der Schritt der Datenidentifikation und -sammlung mit einem nicht unerheblichen Aufwand verbunden ist. Die Daten sind oftmals über die ganze Unternehmenslandschaft verteilt und es fehlen Übersichten, was den Zugang deutlich erschwert. Daher soll die Dateninventur bei der Identifikation und Sammlung der interessanten Daten systematisch unterstützen.

Z7) Bereitstellung eines Beschreibungsrahmens für Betriebsdaten

Umfassendes Datenverständnis ist für die Bestimmung der nächsten Schritte essenziell. Zum Aufbau ist die Betrachtung intrinsischer Merkmale, wie z. B. Struktur, Format und Qualität, der Betriebsdaten sinnvoll, welche im Zusammenhang mit der Vorverarbeitung und Modellierung stehen. Daher soll die Systematik mittels eines Beschreibungsrahmens den Aufbau eines umfassenden Datenverständnisses für die anschließende Datenanalyse fördern.

Z8) Bestimmung der Dateneigenschaften

Die vielen möglichen analyserelevanten Eigenschaften haben wiederrum eine Vielzahl an Ausprägungen, die oftmals nicht schnell und einfach für die eigenen Datensätze zu bestimmen sind. Daher soll die Systematik bei der Bestimmung der relevanten Eigenschaften unterstützen und dabei den Beschreibungsrahmen sinnvoll einsetzen.

Ziele an die Methodenauswahl

Z9) Bereitstellung von für die betriebsdatengestützten Produktplanung relevanten Vorverarbeitungsmethoden sowie Algorithmen und Evaluierungsmetriken

Die großen Lösungsräume der möglichen Analytics-Methoden und Algorithmen müssen für die betriebsdatengestützte Produktplanung eingeschränkt und gefiltert werden, um die relevanten Verfahren schneller zu erkennen und die notwendigen Experimente zu reduzieren. Mit der Systematik sollen daher relevante Vorverarbeitungsmethoden, Data-Analytics-Algorithmen und Evaluierungsmetriken bereitgestellt werden.

Z10) Auswahl geeigneter Methoden unter Berücksichtigung der Abhängigkeiten

Um die richtigen Methoden der Vorverarbeitung und Modellierung auszuwählen, müssen ihre Abhängigkeiten untereinander als auch ihre Abhängigkeiten zu vorher adressierten produktplanungsspezifischen Einflussfaktoren wie der Anwendung und den Dateneigenschaften berücksichtigt werden. Die Systematik soll die Auswahl von wenigen erfolgsversprechenden Techniken für die Schritte Vorverarbeitung, Modellierung und Evaluierung unterstützen und die jeweiligen Abhängigkeiten zwischen diesen und auch zuvor festgelegten Komponenten einbeziehen, um ganzheitliche Data-Analytics-Pipelines für die betriebsdatengestützte Produktplanung aufzustellen.

Ziele an die Umsetzung

Z11) Auswahl geeigneter Tools

Um die passenden Tools für die Vorverarbeitung und Modellierung zu finden, sind wichtige Entscheidungsfaktoren zu berücksichtigen. Die Systematik soll diese Faktoren bündeln und eine systematische Auswahl ermöglichen.

3 Stand der Forschung

Unter Berücksichtigung der Ziele aus Abschnitt 2.5 widmet sich dieses Kapitel der Analyse des aktuellen Forschungsstandes. Hierbei erfolgen eine detaillierte Beschreibung und Bewertung ausgewählter, weit verbreiteter, häufig zitierter und inhaltlich relevanter Ansätze. Das Hauptziel dieser Untersuchung besteht darin, den Handlungsbedarf für die Entwicklung einer Systematik zur Datenanalyse in der betriebsdatengestützten strategischen Produktplanung zu identifizieren. Zunächst werden in Abschnitt 3.1 spezifische Ansätze zur Betriebsdatenanalyse in der Produktplanung und -entwicklung untersucht. Anschließend werden in Abschnitt 3.2 ausgewählte Ansätze zur Definition von Data-Analytics-Anwendungen vorgestellt. In Abschnitt 3.3 folgen ausgewählte Ansätze zum Datenverständnis, in Abschnitt 3.4 Ansätze zur Methodenauswahl und in Abschnitt 3.5 ausgewählte Ansätze zur Toolauswahl. Durch diese Vorgehensweise wird geprüft, ob eine einfache Kombination eines speziellen Ansatzes aus dem Data-Analytics-Prozess und eines allgemeines zur datengestützten Produktplanung oder -entwicklung ausreicht, um sämtliche Ziele zu erfüllen. Im abschließenden Abschnitt 3.6 wird der resultierende Handlungsbedarf im Einklang mit den gestellten Zielen abgeleitet.

3.1 Spezifische Ansätze zur Betriebsdatenanalyse in der Produktplanung und -entwicklung

Die betriebsdatengestützte Produktplanung stellt ein vergleichsweise neues Forschungsgebiet dar. Dennoch existieren bereits einige Arbeiten, die sich mit der Integration von Daten, im Speziellen von Betriebsdaten, in die Prozesse der Produktplanung und -entwicklung beschäftigen. In diesem Abschnitt werden fünf ausgewählte Ansätze näher betrachtet, in denen Betriebsdaten und deren Analysen bereits mit den Aktivitäten der Produktplanung und -entwicklung verknüpft wurden.

3.1.1 Feedback Assistenz System zur Entscheidungsunterstützung bei der Produktverbesserung nach DIENST

DIENST stellt einen der ersten Ansätze zur Rückführung von Produktnutzungsdaten in die Produktentwicklung in Form eines Feedback Assistenz Systems (FAS) vor [Die14]. Dieses System verwaltet und verarbeitet Produktnutzungsinformationen von Industriegütern mit dem Ziel Verbesserungspotenziale durch Aufdeckung von Schwachstellen der bisherigen Produktgeneration *n* zu identifizieren. Der Fokus liegt dabei auf strukturierten Nutzungsdaten wie Sensor- und Instandhaltungsdaten. Um nutzbares Verbesserungswissen zu generieren, wendet DIENST wissens- und entscheidungsbasierte Methoden und Strategien (eine wissensbasierte Methodenkombination im System) zur Strukturierung und Verdichtung der Daten ein. Dabei unterscheidet sie zwei Schwerpunkte: (1) die Analyse der aktuellen Situation ($Produkt_n$) und (2) die Entscheidungsunterstützung bei der Verbesserung der nächsten Generation ($Produkt_{n+1}$). Für den ersten Schwerpunkt werden zwei Methoden, bzw. Kernfunktionen, angewandt:

1) Über eine Evaluierung der Anforderungen auf Basis von Kennzahlen wird eine Produktverbesserung ausgelöst. Statistische Analysen, welche Kennzahlen (z. B. Bereitschaftsdauer einer Maschine, Kosten für Ersatzteile) aus den Nutzungsdaten berechnen und mit den definierten Anforderungen vergleichen, sollen feststellen, ob die aktuelle Generation die Anforderungen nicht mehr erfüllt und somit eine Produktverbesserung erforderlich ist.

2) Um Schwachstellen und Fehlerursachen aufzudecken, wird ein Data-Mining-Diagnosemodell eingesetzt. Im Speziellen werden dazu Bayes'sche Netze (BN) genutzt, welche bedingte Wahrscheinlichkeiten verwenden, um über Rückwärtsinduktion von Ereignissen auf Ursachen zu schließen. Ziel sind BNs, welche für den Produktentwickler intuitiv verständlich und interpretierbar sind, d.h. sie sollen möglichst übersichtlich sein.

Die Ergebnisse werden dann im zweiten Schwerpunkt wie folgt zur Entscheidungsunterstützung genutzt:

3) Es wird eine automatische Bewertung und Darstellung der möglichen Verbesserungsalternativen auf Basis der Zielsetzungen der Produktentwicklung durchgeführt. Die Kriterien dazu setzen sich aus Kosten, Zeit und Qualität zusammen. Da die Kriterien anhand der Feedbackdaten bestimmt werden sollen, müssen sich die verschiedenen Alternativen bereits im Betrieb befinden.

4) Die verschiedenen Charakteristika jeder Alternative (z. B. Material) werden in Abhängigkeit der im Betrieb gegebenen Umgebungstemperaturen untersucht. Auf Basis einer Erweiterung der Diagnosemodelle aus 2) werden ihre jeweils erwarteten Erfolge prognostiziert.

Die Funktionen sind als eigenständige Softwaremodule im Feedback Assistenzsystem implementiert.

Bewertung: DIENST setzt in ihrem Ansatz zur Nutzung von Betriebsdaten in der Produktentwicklung den Schwerpunkt auf die Verwaltung und Analyse strukturierter Betriebsdaten, welche sie softwareseitig umsetzt. Das resultierende Assistenzsystem unterstützt Produktentwickler dabei, den Informationsrückfluss in die Produktverbesserung zu strukturieren und teilweise zu automatisieren. Analyseseitig konzentriert sich DIENST im Wesentlichen auf einen Use Case, die Fehlerdiagnose. Zur Umsetzung der Diagnose zeigt sie eine spezifische Pipeline auf, welche aus einer Datenselektion, Vorverarbeitung, Modellierung durch Kombination verschiedener Bayes'schen Netze und der Visualisierung des Graphen besteht. Für die vorliegende Arbeit zeigt dieser Ansatz demnach mehr ein gutes Beispiel für eine erfolgsversprechende Pipeline für einen potenziellen Use Case der Produktplanung auf. Ein eigenständiges und individuelles Pipeline-Design wird nicht unterstützt.

3.1.2 Smarte Data Analytics Toolbox für Produktdesigner nach ABOU EDDAHAB

ABOU EDDAHAB stellt im Rahmen ihrer Dissertation demonstrative, funktionale Elemente einer smarten Data Analytics Toolbox vor. Diese sollen Produktdesigner bei der Produktverbesserung basierend auf Daten der Lebensmitte („MoL-Daten“) unterstützten [Abo20]. Im Rahmen ihrer Arbeit stand eine wesentliche Forschungsfrage im Zentrum: *Welche Funktionen sollten in einer intelligenten Datenanalyse-Toolbox der nächsten Generation enthalten sein, um Produktdesigner bei der Verbesserung von Produkten und Dienstleistungen auf der Grundlage von MoL-Daten zu unterstützen?* Aus einer Sammlung an identifizierten Anforderungen, wählte sie drei Funktionen zur weiteren Implementierung aus: 1) die Zusammenführung verschiedener MoL-Datenströme aus mehreren Sensorquellen und Empfehlungen für den Designer in Form eines Aktionsplans, was mit dem Produkt zu tun ist. 2) Empfehlung von aufgabenrelevanten Datenanalysetools, wie z. B. Support Vector Machines, und 3) eine intelligente Benutzeridentifizierung, um eine sichere Analyseumgebung zu schaffen. Zur Umsetzung dieser Funktionen wurden verschiedene eigene und existierende Algorithmen eingesetzt. Ein Algorithmus für die zweite Funktion analysiert beispielsweise verschiedene Designeraufgaben (z. B. Verbesserung des Produktdesigns durch Analysieren der meistgenutzten Features), die in einer Datenbank abgelegt und beschrieben sind. Er ordnet ebenfalls gespeicherte Tools (z. B. Support Vector Machines, K-means) gemäß ihrer Übereinstimmung hinsichtlich der Datenquellen, Datenkategorien und möglichen Outputs. Für die Zusammenführung der MoL-Daten und der Empfehlung von passenden Aktionen werden „Wenn - dann“- Regeln zur Zuordnung von verschiedenen detektierten Anomalien (z. B. die Heizung ist oft eingeschaltet) und möglichen Empfehlungen in Form von Aktionen (z. B. Wasserheizelement sollte gesäubert oder ersetzt werden) eingesetzt. Eine Ansicht aus dem Tool ist in Bild 3-1 dargestellt.

Bewertung: Die Data-Analytics-Toolbox nach ABOU EDDAHAB liefert wertvolle Impulse zur technischen Umsetzung von wissensbasierten Empfehlungssystemen mit regel- und Matching-basierten Algorithmen. Aus der Analyse wird jedoch deutlich, dass die Toolbox einen Fokus auf der Empfehlung von ganz konkreten Aktionen auf Basis von Anomalien in Sensordaten legt. Die Aufdeckung von Trends und ersten Tendenzen in beliebigen Betriebsdaten wird nicht betrachtet. Darüber hinaus hegt die Toolbox nicht den Anspruch, ihre Nutzer für den Data-Analytics-Prozess zu befähigen; die Entscheidungsfaktoren werden beispielsweise für die Anwender gar nicht ersichtlich; lediglich das Ergebnis ist relevant. Darüber hinaus ist das Tool durch seine fixierte Anzahl an Designeraufgaben eingeschränkt bezüglich der Empfehlungen.

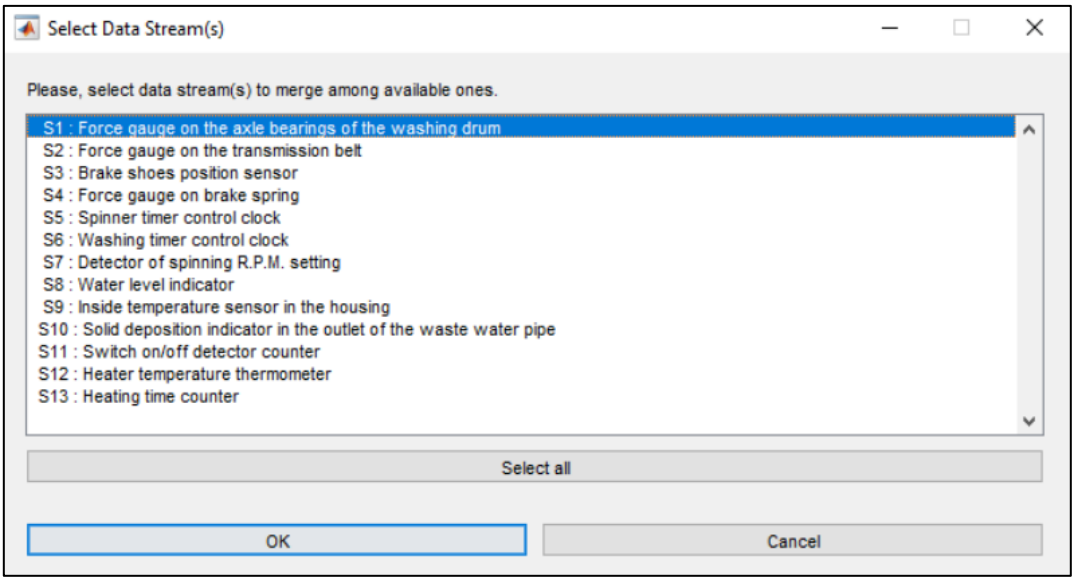

Bild 3-1: „Auswahl von Datenströmen“ – Screen des Moduls Zusammenführung von MoLD [Abo20]

3.1.3 Taxonomie für Feedback-getriebene Produktentwicklungsszenarien nach HOLLER ET AL.

HOLLER ET AL. präsentieren ein Klassifikationsmodell in Form einer Taxonomie für Feedbackgetriebene Produktentwicklungsszenarien in der verarbeitenden Industrie [HNU+17]. Mit diesem Modell soll die Landschaft der Feedback-getriebenen Produktentwicklung mit ihren möglichen Szenarien und Möglichkeiten umfassend beschrieben werden und damit Produktentwickler und Entscheider unterstützen. Die Taxonomie unterscheidet neun Dimensionen (s. Bild 3-2): (1) Ansatz zur Datensammlung, (2) Produktdatenquelle (Abstraktionslevel), (3) Produktdatenquelle (Format der Erscheinung), (4) Komplexität der Feedbackverarbeitung, (5) Grad der Feedbackverarbeitungsautonomie, (6) Grad der Produktneuheit, (7) Adressierte Produktentwicklungsphase, (8) Ermöglichter geschäftlicher Vorteil, (9) Ermöglichte Wertsteigerung.

Dimension	Charakteristika				
D1 – Ansatz zur Datensammlung	C1.1 – Reaktiver Ansatz (ex post)		C1.2 – Proaktiver Ansatz (ex ante)		
D2 – Produktdatenquelle (Abstraktionslevel)	C2.1 - Produktinstanz		C2.2 - Produktklasse		
D3 – Produktdatenquelle (Erscheinungsformat)	C3.1 – Strukturierte Daten	C3.2 – Semi-strukturierte Daten	C3.3 – Unstrukturierte Daten		
D4 – Komplexität der Feedback-Verarbeitung	C4.1 – Geringe Komplexität	C4.2 – Mittlere Komplexität	C4.3 – Hohe Komplexität		
D5 – Autonomiegrad der Feedback-Verarbeitung	C5.1 – Manuelle Feedback-Verarbeitung	C5.2 – Teilweise automatisierte Feedback-Verarbeitung	C5.3 – Automatisierte Feedback-Verarbeitung		
D6 – Grad der Produktneuheit	C6.1 Neue Produktentwicklung		C6.2 – Produktverbesserung		
D7 – Adressierte Produktentwicklungs-ebene	C7.1 – Produktkonzept-ionalisierung	C7.2 - Produktdefinition	C7.3 - Produktrealisierung		
D8 – Ermöglichter Geschäftsnutzen	C8.1 – Anforderungs-spezifikation	C8.2 – Kunden-zentr. Produkt-portfolioplanung	C8.3 – Design für den Gebrauch	C8.4 – Ver-kürzung d. phys. Prototypings	
D9 – Ermöglichter Wertzuwachs	C9.1 - Technisch	C9.2 - Ökonomisch	C9.3 - Umwelt	C9.4 - Sozial	C9.5 - Kombination

Bild 3-2: Taxonomie für Feedback-getriebene Produktentwicklungsszenarien in Anlehnung an [HNU+17]

Bewertung: Die Taxonomie bietet einen schnellen Überblick darüber, wie Produktnutzungsdaten für die Produktentwicklung genutzt werden können und hilft die unterschiedlichen Typen von Feedback-getriebener Produktentwicklung zu verstehen und verschiedene Szenarien zu entwickeln. Diese Szenarien bleiben allerdings recht abstrakt. Die Datenanalyse wird lediglich auf einer sehr hohen Ebene anhand der Komplexität betrachtet. Als methodischer Ansatz ist die Taxonomie nach HOLLER ET AL. aber vor allem für die Bereitstellung von Übersichten nützlich.

3.1.4 Ansatz zur Anforderungserhebung durch explorative Analyse von Nutzungsdaten nach RIESENER ET AL.

RIESENER ET AL. schlagen einen Ansatz zur Analyse von Nutzungsdaten von Produktionssystemen vor, um Anforderungen durch Anwendung von explorativen Datenanalysen zu identifizieren und priorisieren [RDL+21]. Die Besonderheit dieses Ansatzes zur Verbesserung des Anforderungsmanagements besteht darin, dass es nicht existierende

Annahmen und Hypothesen verifiziert oder falsifiziert, sondern neue, bisher unbekannte Beziehungen aufdeckt. Der Ansatz besteht im Wesentlichen aus vier Schritten:

1) **Definition des Verbesserungsziels und Identifikation von Zieldatensätzen:** Um nicht nur triviale Korrelationen aus der Analyse zu erhalten, muss am Anfang ein Ziel definiert werden, wie z. B. die Gesamteffizienz der Anlage (OEE). Für die Analyse bedeutet das die Suche nach Faktoren, die beispielsweise zu Ineffizienzen beitragen. Auf Basis des Ziels sind die Zieldatensätze zu identifizieren, d.h. eine Teilmenge aller verfügbaren Datensätze, welche das Verbesserungsziel (OEE) beschreiben. Über Gewichtungen sollen nur die Datensätze bestimmt werden, die einen Beitrag leisten. Der Ansatz kann angewendet werden, solange das Ziel durch KPIs gemessen werden kann, die wiederum durch verfügbare Daten, die Zieldatensätze, ausgedrückt werden können.

2) **Definition des Datenumfangs**: Im zweiten Schritt werden die Input-Datensätze definiert, welche zur Analyse herangezogen werden sollen. Dies ist notwendig, weil eine Analyse mit allen verfügbaren Datensätzen technisch nicht machbar ist und zu viele nicht-einflussreiche Daten das Risiko von Korrelationen ohne Kausalitäten erhöhen. Zur Auswahl von Input-Datensätzen (Nutzungsdaten, periphere und wirtschaftliche Daten) wird ein Punktesystem vorgeschlagen, welches die Daten anhand verschiedener Kriterien, wie die physische Nähe zum Produktionsprozess oder die funktionale Verknüpfung mit dem Zieldatensatz, bewertet und anhand eines gesetzten Schwellenwerts selektiert.

3) Identifizierung von Hypothesen und Auswahl von Potenzialen für technische Verbesserungen: Um Potenziale für technische Verbesserungen zu erschließen, werden in diesem Schritt Korrelationen zwischen Nutzungsdaten mit einem globalen Zeitstempel identifiziert. Dabei werden nur die gefundenen Korrelationen zurückgegeben, die sich als relevant für die Verbesserungsziele erweisen. Anschließend werden die sich daraus ergebenden Potenziale verifiziert. Dies geschieht zum einen über eine Granger-Analyse zur Aufdeckung von Kausalität. Zum anderen werden die Potenziale über eine Bewertung der fachlichen Relevanz durch Kriterien wie funktionale Nähe zum Herstellungsprozess oder Reproduzierbarkeit überprüft.

4) Ableitung von Produktanforderungen und Analyse von Effekten: Um über Kontext Informationen in Wissen umzuwandeln, wird im vierten Schritt die Ursache-Wirkungs-Kette der Hypothese modelliert und auf die funktionale Architektur des Produkts projiziert. Die Zuordnung zwischen funktionaler und physischer Architektur gibt einen Hinweis, wo eine technische Verbesserung in der Produktarchitektur angesiedelt sein könnte. So werden die technischen Anforderungen abgeleitet. Die Bewertung des Nutzens des Ansatzes kann durch eine Analyse der Zieldatensätze nach Umsetzung der neuen Anforderungen erfolgen, beispielsweise durch eine Kosten-Nutzen-Analyse

Bewertung: RIESENER ET AL. stellen mit ihrer Methode ein vollständiges und konsistentes Vorgehen zur betriebsdatengestützten Anforderungserhebung innerhalb der Produktentstehung vor, welche nicht auf vordefinierten Hypothesen basiert. Damit werden alle Schritte der Datenanalyse-Pipeline adressiert. Es liegt hier allerdings ein klarer Schwerpunkt vor – auf Abhängigkeitsanalysen durch Korrelations- und Kausalitätsanalysen. Wie solche Analysen erfolgreich auf den Eingangsdaten angewendet werden können, wird nicht thematisiert. Eine Konzipierung, bzw. das Durchlaufen eigener anwendungsspezifischer Pipelines wird nicht unterstützt.

3.1.5 Konzept der technischen Vererbung nach LACHMAYER ET AL.

Das Konzept der Technischen Vererbung (TI), das im Sonderforschungsbereich 653 "Gentelligente Komponenten in ihrem Lebenszyklus" [DM17] entwickelt wurde, basiert auf einer algorithmisierten Rückführung von Informationen aus den Lebenszyklusphasen eines Produkts in die nächste Produktgeneration. Der Grundgedanke ist die Entwicklung oder Modifikation einer neuen Generation von Produkten oder Dienstleistungen unter Berücksichtigung der gesammelten Informationen aus den Lebenszyklen der vorherigen Produktgenerationen. Hierfür wurden Materialien, Sensoren, Technologien und Methoden entwickelt [LM22]. Ein herausragendes Merkmal ist, dass die gesammelten Daten von intelligenten Produkten selbstständig erfasst und durch einen genetischen Code auf ihnen gespeichert und verarbeitet werden [DMQ+16, Moz17]. Für die Entwicklung des Informationstransferprozesses realisierten die Autoren eine algorithmische Datenrückführung, die sowohl statistische Methoden und Operationen als auch eine Design-Evolution für Produktanpassungen im Entwicklungsprozess umfasst. Der Prozess beinhaltet die folgenden Schritte (s. Bild 3-3) [Moz17]:

1) **Datenaufbereitung:** Wichtig ist im ersten Schritt die Identifikation von relevanten Informationen aus den großen Datenmengen, um den Umfang des Datensatzes zu reduzieren. Sinnvolle Aktivitäten können in dem Kontext eine intelligente Aggregation der Daten oder die Aufteilung der Daten in Segmente sein.

2) **Vorverarbeitung:** Zur Vorverarbeitung der Daten können Methoden der Transformation, Aggregation und Standardisierung verwendet werden.

3) **Analyse:** Der Analyseprozess gliedert sich in zwei Teile: die Konstruktion des Modells und die Anwendung des Modells auf den neuen Daten. Umfasst werden insbesondere Methoden der Clusteranalyse und der Mustererkennung.

4) **Nachverarbeitung:** Die gewonnenen Informationen können im Wissensspeicher abgelegt werden und sind somit für die Entwicklung einer neuen Produktgeneration nutzbar.

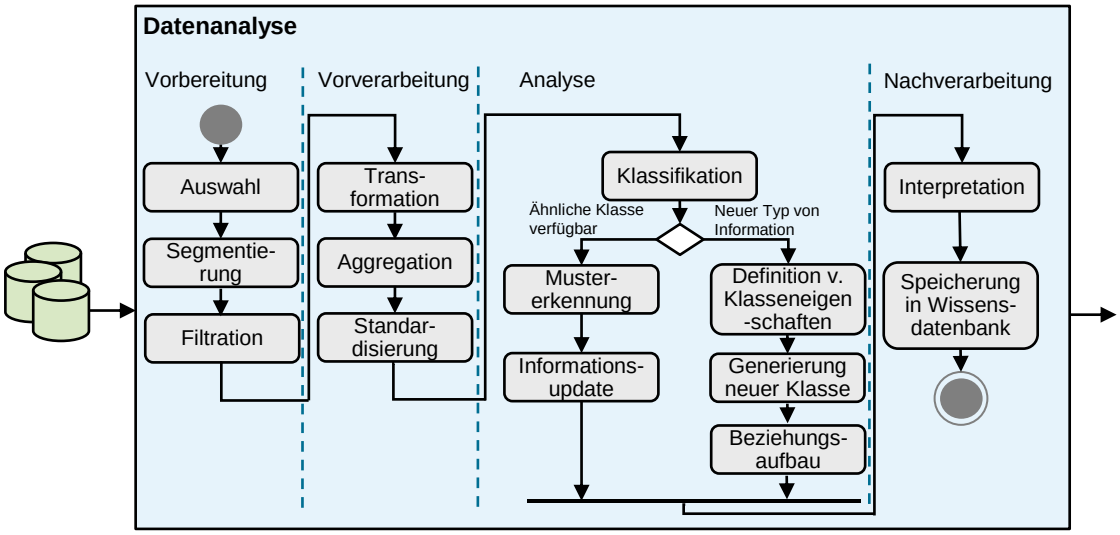

Bild 3-3: Datenanalyse im Konzept der technischen Vererbung in Anlehnung an [Moz17]

Bewertung: Das beschriebene Konzept der Technischen Vererbung (TI) ist ein generationenorientierter Ansatz für die Entwicklung von technischen Systemen und Produkten. Es basiert auf der Anwendung der Gesetze der technischen Evolution und evolutionärer Mechanismen. Auch wenn eine Vielzahl an Methoden, teilweise auch Data-Analytics-Methoden, Bestandteil des Konzeptes sind, liegt der Schwerpunkt eher auf dem geschlossenen Kreislauf und dem standardisierten Datenaustausch. Das Vorgehen zur Datenanalyse ist sehr abstrakt gehalten und enthält keine weiteren Artefakte zur Auswahl und Umsetzung.

3.2 Ansätze zur Definition von Data-Analytics-Anwendungen

Die Analyse der allgemeinen Ansätze in Abschnitt 3.1 zeigt, dass kein Ansatz alle Ziele aus Abschnitt 2.6 erfüllt. Nach diesen werden ausgewählte Ansätze für die einzelnen Phasen des betrachteten Data-Analytics-Prozesses betrachtet. Dieser Abschnitt fokussiert dabei zunächst drei Ansätze zur Definition von Data-Analytics-Anwendungen.

3.2.1 Data Analytics Canvas nach KÜHN ET AL.

Das Data Analytics Canvas ist eine semiformale Spezifikationstechnik zur Beschreibung eines Analytics-Anwendungsfalls, der notwendigen unternehmensweiten Dateninfrastruktur, sowie Anforderungen für interdisziplinäre Domänen [KJR+18]. Das Canvas unterstützt in der Phase der Identifikation und Planung von Analytics-Anwendungen und ermöglicht eine zielgerichtete Kommunikation zwischen den Projektbeteiligten. Da das Canvas auf dem 4-Ebenen-Modell basiert (vgl. Abschnitt 2.3.1), orientiert sich der Aufbau an diesem Modell. Lediglich in der Ebene der Daten-Infrastruktur und der Datenquellen ergeben sich Veränderungen: Die Daten-Infrastruktur wird im Canvas in die zwei Schichten Daten-Pools und Datenbeschreibung aufgeteilt. Die Datenquellen-Schicht

ermöglicht im Canvas durch eine Dreiteilung in *Ressourcen, manuelle und automatisierte Datensammlung* eine Konkretisierung der Datenursprünge. Darüber hinaus bietet das Canvas neun Konstrukte, mit denen es befüllt werden kann.

Startpunkt der Analytics Canvas bildet die Schicht "Datenanalyse". Sie fasst die Typen der Datenanalyse zusammen. Die vier Stufen deskriptiv, diagnostisch, prädiktiv und präskriptiv dienen als Leitfaden für die Bestimmung von Anwendungsfällen. Im ersten Schritt füllt der Nutzer die Ebene Analytics-Lösung aus und nutzt das gleichnamige Konstrukt zur Definition einer Anwendung, wie z. B. Condition Monitoring[13] - basierend auf identifizierten Verbesserungspotenzialen. Anschließend wird die Ressourcen-Ebene innerhalb der Datenquellen-Schicht durch Aufführen der relevanten Produktionsressourcen spezifiziert. Außerdem wird definiert, ob die Daten manuell oder automatisiert von der Ressource gesammelt werden. In der Datenbeschreibung wird im Detail spezifiziert, welche Daten benötigt werden. In der Folge wird in der Ebene Datenpools durch ein Datenpool-Konstrukt definiert, wo die Daten gespeichert sind oder gespeichert werden sollen. Bild 3-4 demonstriert die Anwendung des Data Analytics Canvas anhand eines Anwendungsfalls aus dem Condition Monitoring.

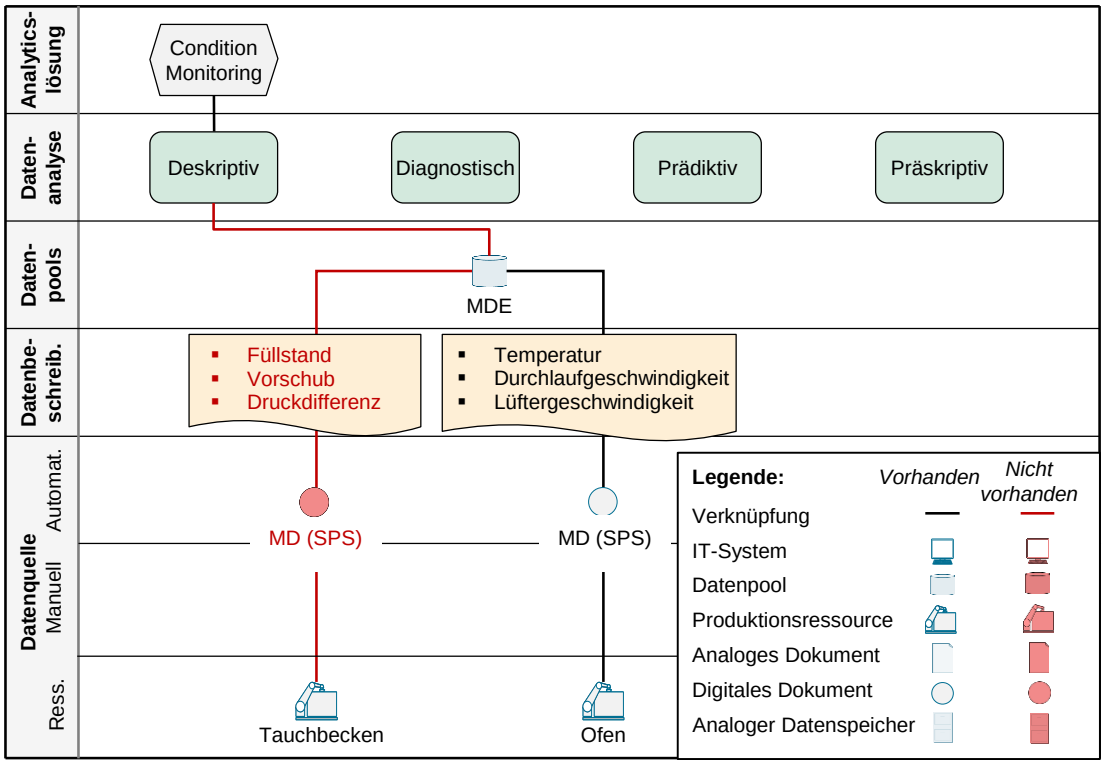

Bild 3-4: Ausgefülltes Data Analytics Canvas für eine Condition-Monitoring-Anwendung in Anlehnung an KÜHN ET AL. [KJR+18]

[13] Das Prinzip des Condition Monitoring, auch als Zustandsüberwachung bezeichnet, beruht auf der kontinuierlichen oder dauerhaften Erfassung des Zustands einer Maschine durch die Messung und Analyse von physikalischen Größen wie Schwingungen, Temperaturen und Lage/Näherung [KW17].

Bewertung: Das Data Analytics Canvas von KÜHN ET AL. ist ein hilfreiches Werkzeug zur Spezifizierung von Data-Analytics-Anwendungen und -Projekten. Auch im Bereich der Betriebsdaten-Analyse in der Produktplanung bietet es sowohl für die Umsetzer der Datenanalyse als auch für die Produktexperten den ganzheitlichen Blick auf einen Use Case. Somit wird eine gute Kommunikationsbasis und Dokumentation der Analyse ermöglicht. Das Canvas stellt folglich ein nützliches Hilfsmittel an verschiedenen Stellen des Analyseprozesses dar, insbesondere bei der Bestimmung der relevanten Daten.

3.2.2 Use Case Modellierung in Data Mining Projekten auf Basis von BDM nach MARBAN und SEGOVIA

MARBAN und SEGOVIA stellen eine Erweiterung der UML-Modellierungssprache für Data-Mining-Projekte (DM-UML) vor, die den Dokumentationsbedarf für ein Projekt gemäß CRISP-DM abdeckt [SM13]. Diese kann als Werkzeug für die Modellierung und als Verbindung des Geschäftsverständnisses mit der Modellierungsphase und anderen Phasen eines Analytics-Projektes genutzt werden. DM-UML unterscheidet zwischen geschäftsbezogenen und Data-Mining-bezogenen Modellen bzw. Diagrammen (s. Bild 3-5).

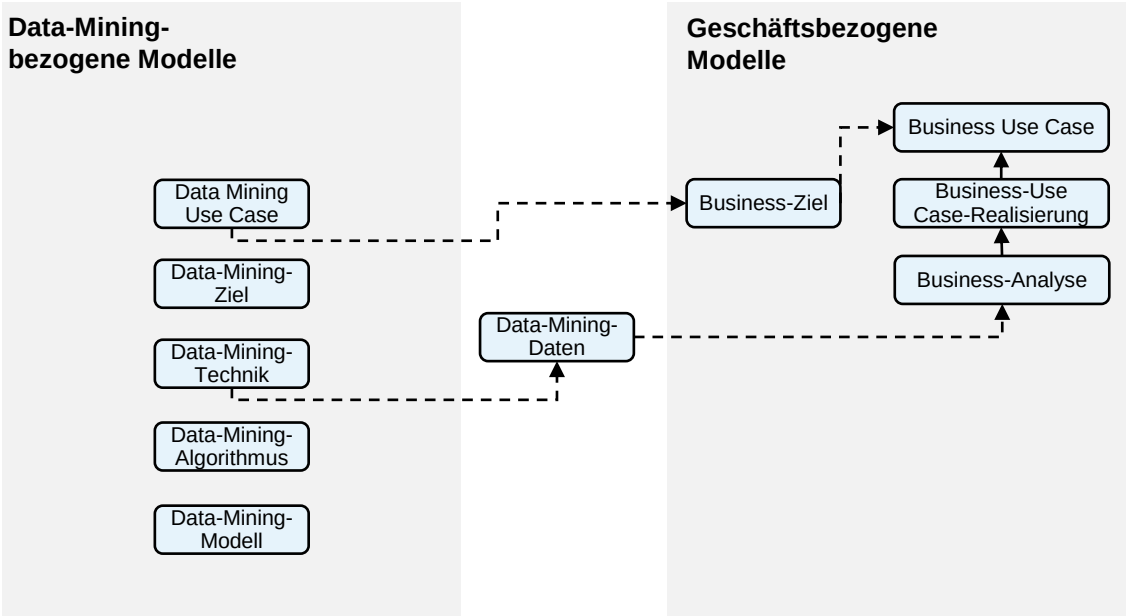

Bild 3-5: DM-UML-Modelle in Anlehnung an [SM13]

Für jedes Modell werden bestimmte Elemente definiert. Zunächst werden zur Analyse des Geschäfts die Business-Modelle erstellt:

- **Business-Use-Case-Modell**: In dieser Phase wird ermittelt, welche Teile des Unternehmens verbessert werden sollen (geschäftliche Anwendungsfälle, z. B. Besuch einer Webseite) und welche Elemente außerhalb des Unternehmens beteiligt sind (Geschäftsakteure).

- **Business-Ziel-Modell**: Die Business-Use-Cases müssen einem oder mehreren vom Unternehmen definierten Geschäftszielen (z. B. etwas aufbauen, den Gewinn steigern, die Produktivität verbessern) zugeordnet werden. Dabei sind Hierarchien vorgesehen, welche spezifische Ziele weiter in noch spezifischere Unterziele unterteilen.
- **Business-Analyse-Modell**: Die Business-Analyse zeigt, wie verschiedene Business-Elemente interagieren, um ein bestimmtes Ziel zu erreichen.
- **Business-Use-Case-Realisierungsmodell**: Das Modell verknüpft die Business-Analyse-Elemente mit dem entsprechenden Business Use Case.

Nach der Analyse des Geschäfts, ist der nächste Schritt herauszufinden, wo das Data-Mining-Projekt zur Verbesserung des Geschäfts beitragen kann. Die DM-UML Analysediagramme zeigen, welche Daten das Unternehmen für die Analyse sammelt und wo die Ergebnisse der Analyse verwendet werden können, um die Geschäftsziele zu erreichen.

- **Data-Mining-Daten-Modell**: stellt die Quellen der verfügbaren Daten für das Projekt dar (mit Tabellen, Spalten, Datentypen und Datenbeziehungen). Es wird von den Geschäftsentitäten abgeleitet, die die verfügbaren Daten repräsentieren.
- **Data-Mining-Use-Case-Modell**: beschreibt die vorgeschlagene Funktionalität des durch die Data-Mining-Aufgaben extrahierten Wissens aus Sicht des Nutzers, um ein bestimmtes Business-Ziel zu erreichen. Data Mining Use Cases werden aus Business Use Cases und Business-Zielen gebildet. Zu beachten ist, dass nicht alle Business Use Cases mit einem Data Mining Use Case verbunden sind. Ob ein Data-Mining Use Case existiert, hängt an dem Business Use Case und den verwendeten Daten, die analysiert werden können.
- **Data-Mining-Ziel-Modell**: legt zusammen mit den Data Mining Use Cases die Data-Mining-Ziele für jeden Use Case fest. Diese Ziele werden in Bezug auf die Geschäftsziele definiert und sind eine Übersetzung des Geschäftsproblems in Probleme, welche in Data-Mining-Begriffen ausgedrückt werden (z. B. Daten clustern, Vorhersagemodell erstellen etc.)
- **Data-Mining-Technik-Modell**: zeigt die Data-Mining-Techniken (z. B. Clustering), die zur Erreichung der Data-Mining-Ziele, die in den Data Mining Use Cases vorgeschlagen werden.
- **Data-Mining-Algorithmen-Modell**: unterstützt den Entwickler bei der Entscheidung, welche der Algorithmen (z. B. k-means) von dem Data-Mining-Technik-Modell genutzt werden sollte.
- **Data-Mining-Modell**: stellt die Data-Mining-Elemente innerhalb des zu verwendenden Data-Mining Tools dar (z. B. SPSS).

Bewertung: MARBAN und SEGOVIA stellen mit DM-UML ein umfassendes Werkzeug zur Projektmodellierung und -dokumentation vor, das alle Phasen eines Data-Mining-Projekts abdeckt. Durch die Verbindungen der einzelnen Modelle, werden die Abhängigkeiten der verschiedenen Aufgaben im Analytics-Projekt berücksichtigt. Vor allem zur Verbindung des Geschäftsverständnisses mit der Modellierung bietet DM-UML eine Lösung und damit wertvolle Anknüpfungspunkte für die Übersetzung der Business Use Cases in Data Analytics Use Cases. Aufgrund der Vielzahl an Modellen und Elementen ist diese Form der Modellierung allerdings sehr erklärungsbedürftig und damit für Nicht-Experten ungeeignet. Außerdem werden mögliche Ausprägungen, bzw. Szenarien (Gestaltungswissen) nicht mitgeliefert.

3.2.3 Methode zur Entwicklung sinnvoller/zweckmäßiger KI-Use-Cases nach HOFMANN ET AL.

HOFMANN ET AL. entwickelten eine Methode, um organisationsspezifische Anwendungsfälle für den Einsatz von KI aus der Problem- als auch aus der Chancenperspektive zu identifizieren [PJD+20]. Ziel der Methode sind KI-Anwendungsfälle, die den Kontext der Domäne und die Funktionen der KI berücksichtigen und damit zielgerichtet sind. Die Methode besteht aus fünf Schritten:

1) **Vorbereitung:** Im ersten Schritt werden relevante Informationen über den organisationsspezifischen Kontext gesammelt und strukturiert. Relevante Kontextgebiete sind dabei Technologie, Organisation und Umgebung. Zunächst müssen Unternehmen ein umfassendes Verständnis für KI-Technologien entwickeln. Im Hinblick auf die Organisation spielen Aspekte wie Strukturen, Ressourcen und Kultur eine entscheidende Rolle; es ist wichtig, diese in Form einer übergreifenden Unternehmensstrategie mit den neuen Technologien abzustimmen. Der Umweltkontext umfasst eine Bewertung der Anforderungen der Branche, der Wettbewerber, der Kunden und der Vorschriften. Dabei müssen Unternehmen förderliche und hinderliche Faktoren für die Entwicklung von KI-Anwendungsfällen identifizieren.

2) **Entdeckung:** Ziel des zweiten Schritts ist die Sammlung spezifischer Anwendungsdomänenprobleme und existierender KI-Lösungen. Folglich können Unternehmen KI-Use-Cases aus zwei Perspektiven entwickeln: (1) Problem-getrieben, indem KI genutzt wird, um existierende Probleme in der Anwendungsdomäne zu adressieren und (2) Chancen-getrieben, indem neue KI-Lösungen exploriert werden, um neue technologische Möglichkeiten in der Anwendungsdomäne zu eröffnen.

3) **Verständnis**: Der dritte Schritt zielt darauf ab, die identifizierten Domänenprobleme und KI-Lösungen weiter zu abstrahieren, um ihre zugrunde liegende Natur zu enthüllen. In Anbetracht des Domänenverständnisses müssen Organisationen einen Weg finden, die Prozesse, Aktivitäten und Fachkenntnisse der Domäne zu

strukturieren. Mit Blick auf das KI-Verständnis schlagen sie zur Unterscheidung der abstrakten Aufgaben KI-Funktionen wie z. B. Wahrnehmung, Vorhersage und Generierung vor. Darüber hinaus unterscheiden sie vier KI-Lösungstypen (regelbasiert, KI-aktiviert, KI-basiert und volle KI).

4) **Entwurf**: Im vierten Methodenschritt müssen die Probleme der Anwendungsdomäne und die KI-Lösungen durch Konsolidierung der gesammelten Informationen in einer Problem-Lösungs-Matrix einander zugeordnet werden. Hier sind drei Szenarien denkbar: (1) Ein Problem-Lösungs-Fit liegt vor, wenn die Matrixelemente sowohl Problem(e) als auch Lösung(en) enthalten. Die Organisation kann daraufhin den Mehrwert dieser Elemente bewerten. (2) Wenn in einem Matrixfeld nur ein Problem, jedoch keine Lösung zu finden ist, hat die Organisation die Möglichkeit, eine eingehendere Marktforschung zu betreiben und gezielt nach Lösungen zu suchen. (3) Falls sich im Matrixfeld nur eine Lösung befindet, jedoch kein Problem identifiziert wurde, sollte die Organisation prüfen, ob ein übersehenes Problem vorliegt oder ob die gefundene Lösung das Potenzial hat, die aktuellen Prozesse zu verbessern.

5) **Implementierung**: Im letzten Schritt müssen die Organisationen die theoretischen Überlegungen in die Praxis umsetzen und einen Plan für die erfolgreiche Implementierung ableiten. Dabei muss das Unternehmen entscheiden, ob es die KI-Lösung selbst entwickelt, die Entwicklung auslagert oder eine existierende Lösung von einem externen Anbieter erwirbt. Außerdem müssen die Unternehmen die drei Kontextfaktoren aus Schritt 1 berücksichtigen und so z. B. die Integration in die bestehende Infrastruktur bedenken.

Bewertung: Die Methode von HOFMANN ET AL. bietet einen vielversprechenden, systematischen Ansatz, um KI- oder Analytics-Use-Cases auf Basis von Problemen und Lösungen zu definieren. Sie adressiert damit die Übersetzung von Analytics-Use-Cases. Die zur Verfügung gestellten Domänenprobleme und KI-Funktionen sind jedoch zu allgemein, als dass sie für die betriebsdatengestützte Produktplanung einen Mehrwert liefern können.

3.3 Ansätze zum Datenverständnis

Abschnitt 3.3 stellt insgesamt fünf Ansätze im Kontext von Datenverständnis vor. Diese werden nochmal nach Ansätzen zur Datensammlung (Abschnitt 3.3.1) und Ansätzen zur Datenbeschreibung (Abschnitt 3.3.2) unterteilt.

3.3.1 Ansätze zur Datensammlung

Im Folgenden werden zwei Ansätze zur Datensammlung analysiert.

3.3.1.1 Datenlandkarte nach JOPPEN ET AL.

JOPPEN ET AL. stellen mit der Datenlandkarte eine semi-formale Spezifikationsmethode zur Visualisierung von Datenflüssen zwischen Datenquellen und Softwaresystemen in Herstellungsprozessen vor [JEK+19]. Sie berücksichtigt dabei die Geschäftsprozesse und die Anwendung und liefert somit wichtige Kontextinformationen für datengetriebene Use Cases. Die Datenlandkarte stellt insgesamt elf Objekte, abgeleitet von der OMEGA[14]-Notation, in vier Schichten zur Verfügung, welche typische Aspekte des Informationsaustauschs reflektieren: (1) Prozess, (2) Dokumente, (3) Ressourcen und (4) Informationsfluss. In der ersten Schicht wird der Geschäftskontext über *Prozessobjekte* abgebildet, welche in einer Prozesskette angeordnet sind. Die wesentlichen *Dokumente* (Papier und Digital) werden darunter in der zweiten Ebene einsortiert. Die dritte Schicht zeigt zugehörige Ressourcen. Dabei repräsentieren *IT-System-Objekte* Softwaresysteme, wie z. B. ERP, CAD, CRM oder Excel. Ein *Produktionsressourcen-Objekt* stellt Datenquellen nahe der Maschinenebene dar, z. B. Produktionsmaschinen selbst, Sensoren und Mensch-Maschine-Schnittstellen. Darüber hinaus bilden *Papierpuffer-Objekte* Datensenken für Papierdokumente und *implizite Informationsobjekte* Informationen, welche nicht explizit für die formale Informationsbeschaffung generiert wurden, wie z. B. Expertenwissen oder informelle Bedienereingaben. Der Informationsfluss zwischen diesen Objekten wird durch *Pfeile* visualisiert. Die Pfeile können zusätzlich durch ein Symbol eine *manuelle oder automatisierte Informationsübertragung* markieren. Mit Hilfe von *Konnektoren* werden kompliziertere Informationsflüsse entlang des Canvas aufgeteilt. Die vierte Schicht spezifiziert in den *Informationselementen* die Informationen, die ausgetauscht werden, im Detail z. B. über Datenpunkte und einzelne Dokumenteneinträge.

Um eine Datenlandkarte zu erstellen, empfehlen JOPPEN ET AL. fünf Schritte:

1) Ableitung der informationsgenerierenden Prozesse aus einer detaillierten Prozessdokumentation (z. B. in OMEGA)
2) Definition der wesentlichen Dokumente entlang der Prozessschritte
3) Platzierung der zugehörigen Ressourcen
4) Visualisierung des Informationsflusses zwischen Ressourcen und Dokumenten
5) Detaillierte Beschreibung der Informationselemente

Das Schema der Datenlandkarte ist in Bild 3-6 zu sehen.

[14] OMEGA steht für Objektorientierte Methode zur Geschäftsprozessmodellierung und -analyse [GPW09].

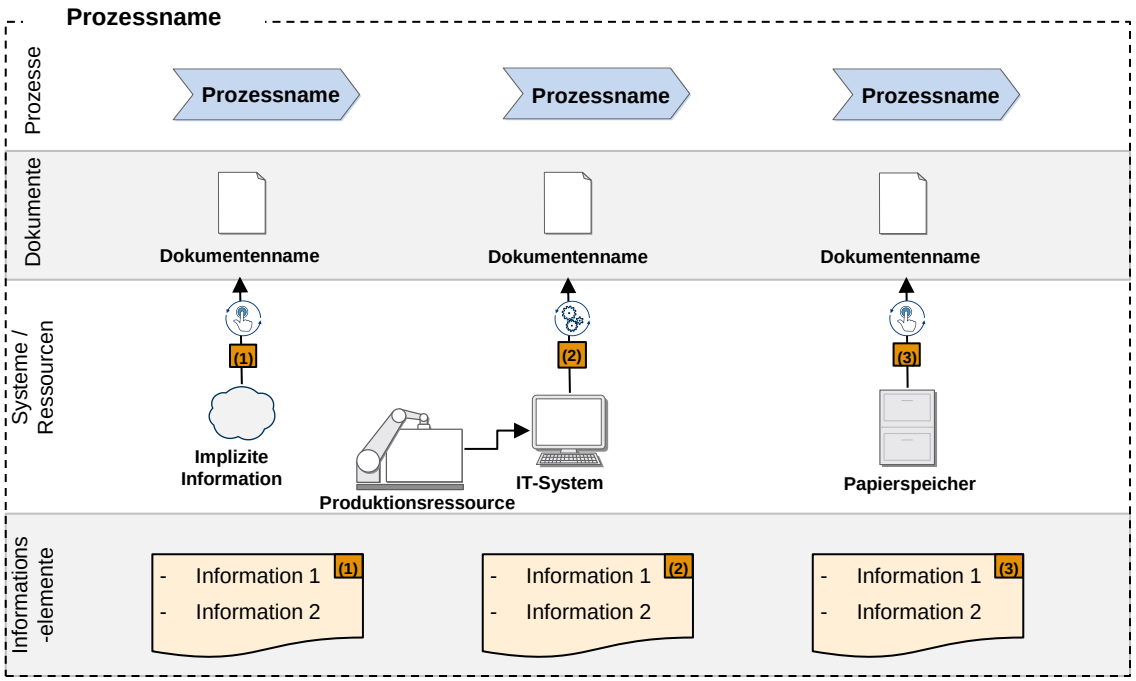

Bild 3-6: Schema der Datenlandkarte [JEK+19]

Bewertung: Die Methode von JOPPEN ET AL. ist ein guter Ansatz, um Datenflüsse zu visualisieren und Datenquellen im Prozesskontext besser zu verstehen. Daher eignet sich die Datenlandkarte als Hilfsmittel, wenn die Betriebsdaten im Prozess der Betriebsphase eines Produktes betrachtet werden sollen. Allerdings werden die anderen Dimensionen von Betriebsdaten aus Domänensicht nicht berücksichtigt.

3.3.1.2 Leitfaden für die Datenquellenauswahl nach STANULA ET AL.

STANULA ET AL. entwickelten einen systematischen Ansatz, um vielversprechende Datenquellen in der Produktion in Abhängigkeit der verfolgten Anwendung zu definieren [SZM18]. Dazu kombinieren sie die Prozess-Fehlermöglichkeits- und Einflussanalyse (PFMEA) und die Qualitätsfunktionseinsatz-Methode (QFD). Das Vorgehen orientiert sich an den ersten beiden Phasen des CRISP-DM-Modells und besteht aus vier Schritten:

1) **Geschäftsverständnis - Von Geschäftszielen zum Prozessverständnis:** In der datengetriebenen Produktion spielt der Systemrahmen eine große Rolle, der von den Geschäftszielen abgeleitet wird und aus dem Maschinenwerkzeug, dem Produkt und dem Prozess besteht. Dieser Systemrahmen bildet den Ausgangspunkt für eine Vorauswahl derjenigen Datenquellen, die potenziell nützlich für die Ziele sind. Um die Geschäftsziele in präzise Prozessziele zu übersetzen, wird eine PFMEA durchgeführt – eine Methode zur Analyse und Priorisierung möglicher Fehlerarten eines Prozesses. Im Rahmen dieser Methode werden Fehlerursachen, ihre Effekte auf die Performance der analysierten Einheiten (z. B. Hardware, Software, menschliche Aktionen) und ihrer Umgebung analysiert. Startpunkt für eine

PFMEA kann beispielsweise ein Prozessflussdiagramm sein, das die Schritte des Prozesses als Informationsbasis zusammenfasst. Ergebnis dieses Schrittes ist eine gewichtete Liste an (potenziellen) Fehlerarten entlang des Prozesses.

2) **Geschäftsverständnis – Datenvorauswahl:** Im Nachgang der PFMEA werden mit Hilfe der QFD bedeutsame Datenquellen ausgewählt. Die QFD ist ein Werkzeug zur Übersetzung von Kundenbedürfnissen (voice of customer – VoC) in technologische Anforderungen (voice of engineer – VoE). Zentraler Aspekt der QFD ist das Haus der Qualität (House of quality –HoQ) zur Korrelation der beiden Dimensionen in einer Matrixform. Die identifizierten Fehlerarten müssen innerhalb der QFD mit den potenziellen Datenquellen kombiniert werden. Dazu muss eine Vorauswahl an Quellen durch Experten stattfinden. Kreativitätstechniken wie die Delphi-Methode können hier unterstützen. Ergebnis sind potenzielle Datenquellen, die durch das zu messende Objekt und die Datenspezifika beschrieben werden.

3) **Geschäftsverständnis - Quality Function Deployment:** Um die gewichteten Fehlerarten mit den vorausgewählten Datenquellen zusammen zu bringen, wird im dritten Schritt die QFD durchgeführt. Der Input für die VoC wird dabei durch die PFMEA gebildet (Fehlerarten), der Input für die VoE durch den Schritt der Vorauswahl (vorausgewählte Quellen). Die Schritte der QFD sind in Bild 3-7 dargestellt.

4) **Datenverständnis:** Im vierten Schritt wird die Datenquellenauswahl validiert. Dazu werden ML-Workflows implementiert und überprüft, wie gut diese auf den ausgewählten Daten performen.

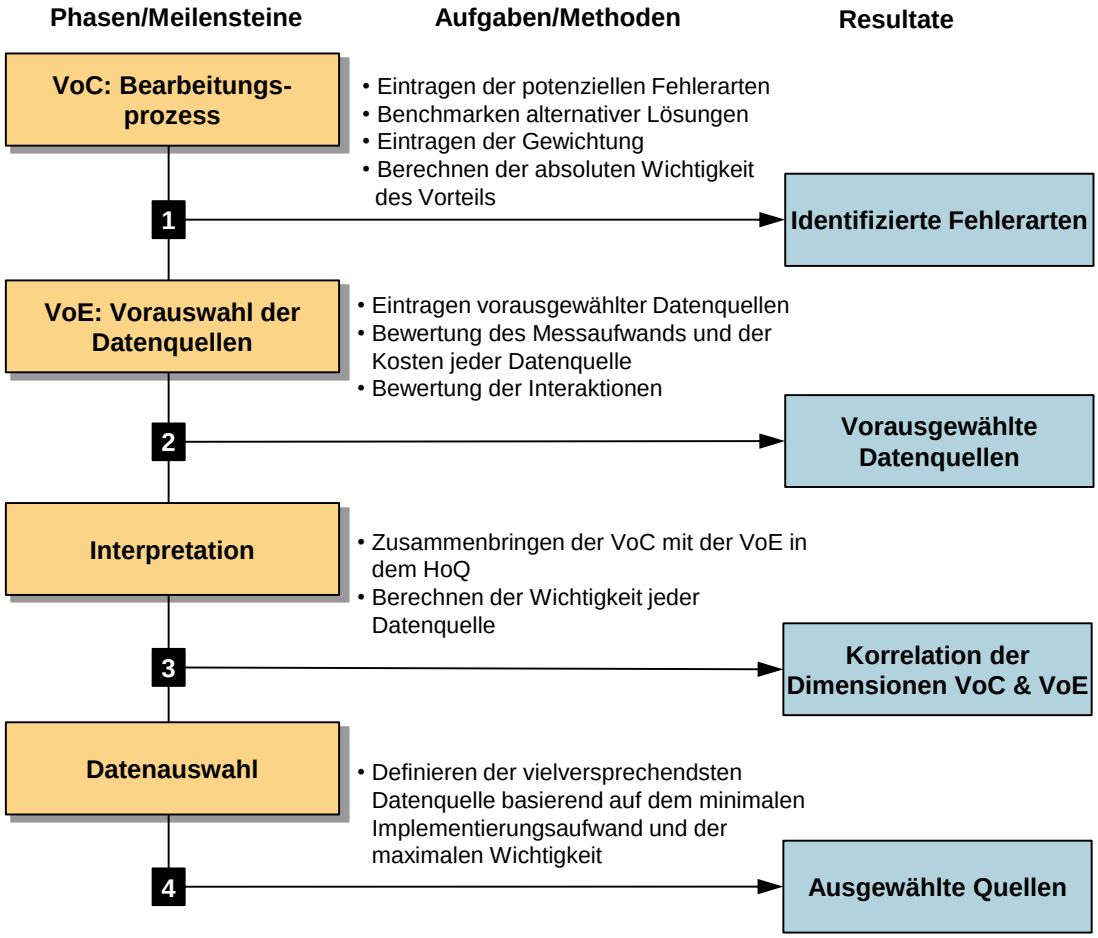

Bild 3-7: Vorgehensmodell der QFD in Anlehnung an STANULA ET AL. [SZM18]

Bewertung: STANULA ET AL. präsentieren mit ihrer Methode einen interessanten, sehr systematischen Ansatz, um vielversprechende Datenquellen für die anschließende Datenanalyse unter Berücksichtigung definierter Geschäftsziele auszuwählen. Damit stellen die Autoren den Prozess der Datensammlung und -auswahl direkt in Beziehung zu der Anwendung und der Datenanalyse. Insbesondere die Datenidentifikation durch den Ansatz potenzielle Fehlerarten mit durch Experten vorausgewählten Datenquellen in einer Matrix gegenüberzustellen ist vielversprechend, auch wenn dies schon früh Experten-Know-How erfordert. Zudem fokussiert der Einsatz einer PFMEA lediglich Geschäftsziele im Kontext von Produktionsprozessen.

3.3.2 Ansätze zur Datenbeschreibung

Im Kontext der Datenbeschreibung werden im Folgenden drei Ansätze analysiert.

3.3.2.1 Konzept zur Bestimmung der Nutzenpotenziale von Felddaten nach KREUTZER

KREUTZER präsentiert eine Methodik zur systematischen Identifizierung, Priorisierung und Umsetzung der Nutzenpotenziale von Felddaten cyber-physischer Systeme [Kre19]. Sie besteht aus vier Teilmodellen:

1) **Strukturierung des technologischen Nutzens cyber-physischer Systeme**: Zur Strukturierung des technologischen Nutzens erarbeitet KREUTZER initial relevante Stakeholder sowie Lebenszyklusphasen von CPS. Auf dieser Basis erfolgt eine Typologisierung des technologischen Nutzens cyber-physischer Systeme anhand eines Beschreibungsmodells. Mit Hilfe eines Erklärungsmodells können 24 Nutzenaspekte innerhalb der Typen abgeleitet werden.

2) **Charakterisierung von Felddaten cyber-physischer Systeme**: Im zweiten Teilmodell entwickelt KREUTZER ein Beschreibungsmodell, mit dem Felddaten auf einer inhaltlich-technologischen Ebene generisch beschrieben werden können (s. Bild 3-8). Es unterscheidet technische Messgrößen, Nutzer- und Systemdaten jeweils anhand von vier Gliederungsebenen.

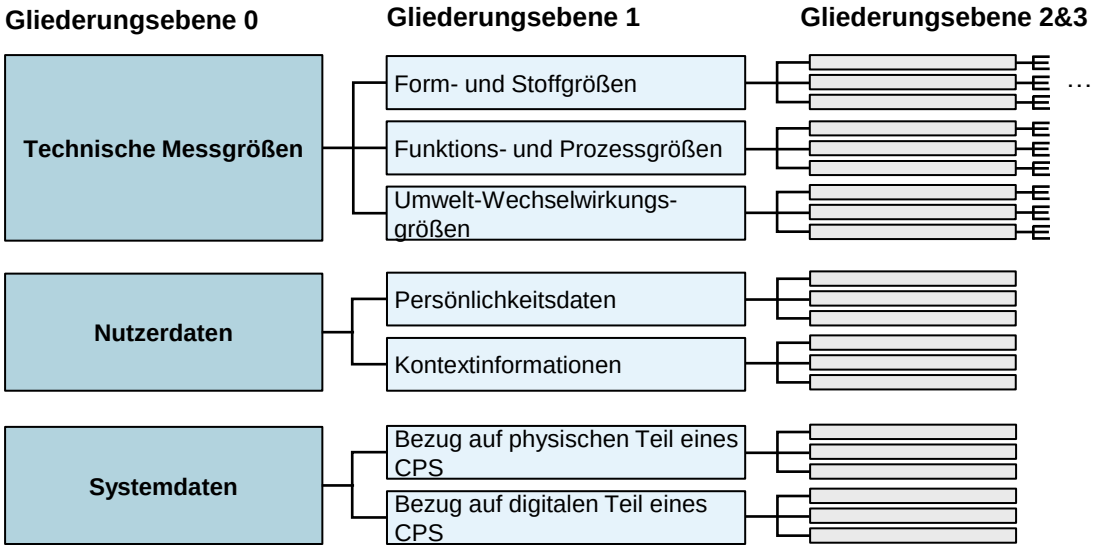

Bild 3-8: Beschreibungsmodell für Felddaten in Anlehnung an KREUTZER [Kre19]

3) **Untersuchung des Felddatenbedarfs von Nutzenaspekten**: KREUTZER untersucht im dritten Teilmodell, wie der Felddatenbedarf der in (1) abgeleiteten Nutzenaspekte bestimmt werden kann. Dafür erstellt er ein Erklärungsmodell zur Begründung der Wirkzusammenhänge zwischen Nutzenaspekten und Felddatensätzen anhand einer Untersuchung von 21 Use Cases.

4) **Bestimmung der Nutzenpotenziale von Felddaten cyber-physischer Systeme**: Zuletzt erarbeitet KREUTZER eine Vorgehensweise. Dafür definiert er drei Vorgehensalternativen (Market-Pull-Ansatz, Technology-Push-Ansatz und integrativer Ansatz), von denen er den letzteren weiterverfolgt. Zusätzlich stellt KREUTZER

Kriterien zur Bewertung von Nutzenaspekten und Felddatensätzen auf. Im Vorgehensmodell zeigt er die methodischen Schritte und das Zusammenwirken der einzelnen Bestandteile auf.

Bewertung: KREUTZER fokussiert mit seinem Ansatz zur Bestimmung der Nutzenpotenziale von Felddaten cyber-physischer Systeme vorgelagerte Schritte des Datenanalyseprozesses, welche zur Definition von nutzbringenden Use Cases notwendig sind. Sein vorgestelltes Beschreibungsmodell für Felddaten stellt jedoch sowohl domänen- als auch analyseseitige Charakterisierungsmerkmale zur Verfügung. Eine Integration ist daher zu prüfen.

3.3.2.2 OntoDT – Ontologie der Datentypen nach PANOV ET AL.

Ein Ansatz zur formalisierten Beschreibung von allgemeinen Daten liefern PANOV ET AL. mit ihrer Ontologie von Datentypen (OntoDT) [PSD16]. OntoDT definiert die Semantik, d.h. die Bedeutung der Schlüsselentitäten und repräsentiert das Wissen über Datentypen in einer maschinenfreundlichen Weise. Die Ontologie basiert auf dem ISO/ IEC 11404 Standard für Datentypen in Computersystemen. Nachfolgend werden die Schlüsselentitäten für die Repräsentation von Datentypen kurz beschrieben:

- **Datentyp und Werteraum**: Die *Datentyp-Klasse* legt fest, welchen Typ die Daten haben, welche verschiedenen Werte sie annehmen können, welche Eigenschaften diese Werte haben und welche Operationen auf ihnen durchgeführt werden können. Die *Datentyp-Klasse* hat eine „Member"-Verbindung zur *Werteraumspezifikations-Klasse* und eine „Operation"-Verbindung zu der *Charakterisierender-Vorgang-Klasse*. Die *Werteraumspezifikations-Klasse* spezifiziert die Sammlung an Werten für einen gegebenen Datentyp. Der Werteraum kann z. B. über Aufzählung der Werte, mit Axiomen unter Verwendung einer Reihe von Grundbegriffen oder als Teilmenge von Werten definiert werden.
- **Charakterisierender Vorgang**: Eine *charakterisierende Operation* ist definiert als direktive Informationseinheit, die diejenigen Operationen auf dem Datentyp spezifiziert, die ihn von den anderen Datentypen mit identischem Wertebereich unterscheiden. Es werden vier verschiedene Operationen unterschieden, darunter z. B. die monadische Operation, welche einen Wert eines gegebenen Datentyps auf einen Wert des gegebenen Datentyps oder auf einen Wert des booleschen Datentyps abbildet.
- **Datentyp-Eigenschaft**: Eine Datentyp-Eigenschaft ist definiert als eine Eigenschaft, die die inhärenten Eigenschaften der durch den Datentyp repräsentierten Dateneinheiten spezifiziert. Jeder Datentyp hat eindeutige Datentyp-Eigenschaften, wie Ordnung (definierte Ordnungsrelation in Wertebereich), Numerizität (Werte als Mengen, ausgedrückt in einem mathematischen Zahlensystem), Kardinalität (Kardinalität des Werteraums), Exaktheit (Unterscheidbare Werte im

Wertebereich), Gleichheit und Begrenztheit (Grenzen des Werteraums). Alle Dateneigenschafts-Klassen haben Unterklassen.

Innerhalb der OntoDT Ontologie wurde eine Taxonomie von Datentypen (s. Bild 3-9) basierend auf den Eigenschaften der Datentypen und ihrer Struktur definiert. Die obersten Ontologieklassen umfassen primitive Datentypen (Datentyp, dessen Wertebereich entweder explizit oder durch Aufzählung definiert ist), generierte Datentypen und nutzerdefinierte Datentypen (durch eine Typspezifikation definiert).

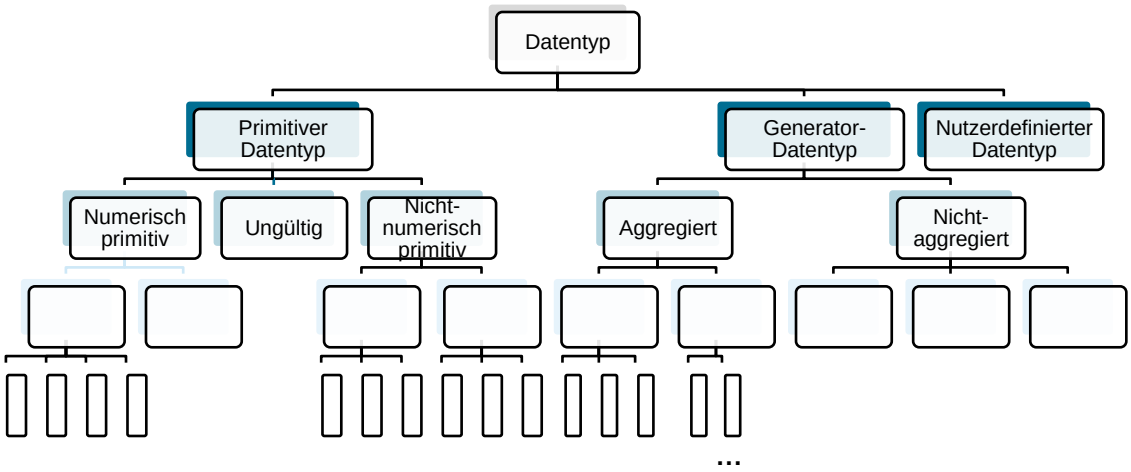

Bild 3-9: Die OntoDT Datentyp-Taxonomie in Anlehnung an PANOV ET AL. [PSD16]

Für die Data-Mining-Domäne stellt OntoDT ein Satz an Data-Mining-spezifischen Datentypen, die von den OntoDT Basisdatentypen durch Subklassifizierung abgeleitet werden. Dazu gehören z. B. Tupel von Primitiven, Menge von diskreten Datentypen, Sequenz von realen Zeitreihen und Baum von diskreten Datentypenknoten. Darüber hinaus bietet OntoDT die Flexibilität, einen beliebigen Datentyp zu definieren, der später zur Definition einer spezifischen Data-Mining-Aufgabe verwendet werden kann. Des Weiteren liefern PANOV ET AL. auf Basis der Datentypentaxonomie eine Taxonomie von Datensätzen. Hierbei werden auf der ersten Ebene ungelabelte und gelabelte Datensätze unterschieden.

Bewertung: PANOV ET AL. stellen mit ihrer OntoDT-Ontologie einen umfangreichen formalen Beschreibungsrahmen für Datentypen zur Verfügung. Die gemachten Unterscheidungen, definierten Eigenschaften und Datensätze liefern einen Anknüpfungspunkt für die Charakterisierung der Betriebsdaten. Für Nicht-Experten wirkt der Ansatz jedoch zu formal und nicht nachvollziehbar. Außerdem bleibt die Bestimmung als Aufgabe der Datentypen unklar.

3.3.2.3 Taxonomie zur Klassifizierung von Betriebsdaten nach MEYER ET AL.

MEYER ET AL. bieten eine Taxonomie zur Klassifizierung von Betriebsdaten anhand von sieben Kriterien (s. Bild 3-10) [MPK+22].

Kriterium	Ausprägungen			
Erzeuger	Mensch	Produkt	Umfeld	
Struktur	Strukturiert		Semistrukturiert	
Entstehung	Kontinuierlich		Event-getriggert	
Beschreibungsfokus	Produkt	Umfeld	Nutzer	Kunde
Zweck	Erfassen		Veranlassen	
Erzeugungsfrequenz	Hoch	Mittel	Niedrig	
Informationsdichte	Hoch	Mittel	Niedrig	

Bild 3-10: Taxonomie zur Klassifizierung von Betriebsdaten [MPK+22]

Darauf aufbauend präsentieren sie fünf Cluster innerhalb der Datenarten (s. Bild 3-11).

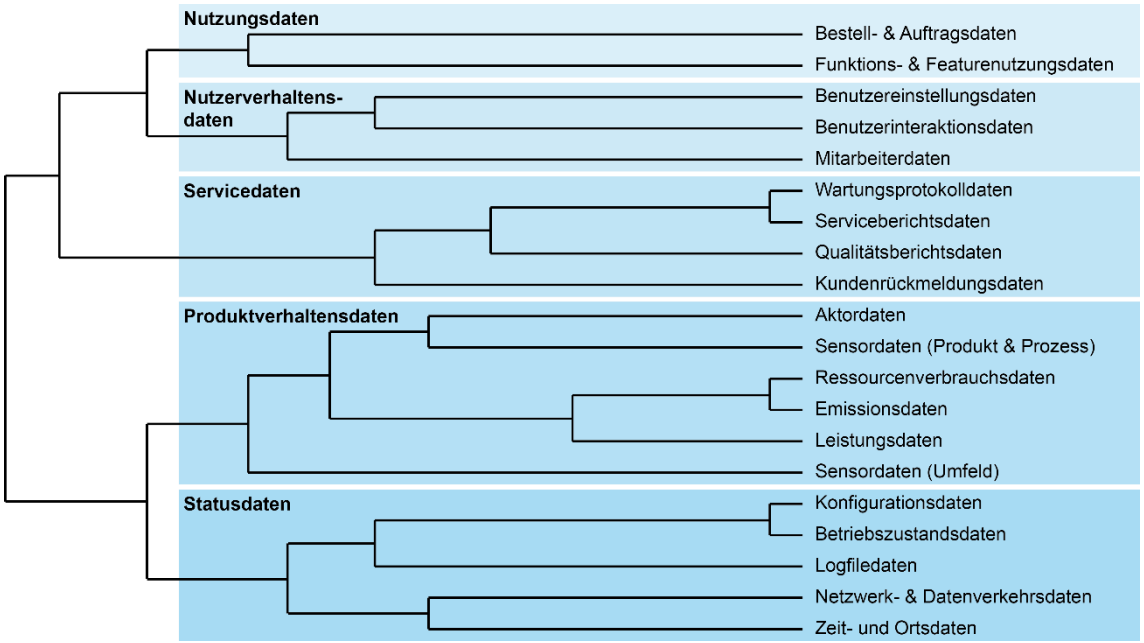

Bild 3-11: Dendrogramm der fünf identifizierten Betriebsdatencluster [MPK+22]

Nutzungsdaten beschreiben, wie ein Produkt von seinen Kunden und Nutzern verwendet wird. Diese Daten beziehen sich auf den Einsatz und die Anwendung des Produkts, nicht auf die Interaktion des Nutzers mit dem Produkt. Ein Beispiel für Nutzungsdaten sind Daten über die Nutzung von Funktionen und Merkmale eines Produkts.

Nutzerverhaltensdaten fassen zusammen, wie sich die Nutzer bei der Verwendung des Produkts verhalten. Im Gegensatz zu den Nutzungsdaten konzentrieren sich Nutzerverhaltensdaten auf die Interaktion des Nutzers mit dem Produkt. Benutzereinstellungsdaten sind ein Beispiel.

Servicedaten sind semistrukturierte Daten, die sich mit Problemen und der Qualität der Produkte befassen. Beispiele für diese Daten sind Serviceberichte und Informationen über Kundenreklamationen.

Produktverhaltensdaten zeigen die erbrachte Leistung eines Produktes im Betrieb. Sie zeichnen sich durch ihr großes Volumen und ihre kontinuierliche Erzeugung aus. Aktor-Daten sind ein Beispiel für diese Datenklasse.

Statusdaten beschreiben den Zustand des betrachteten Produkts. Diese Daten liefern den Kontext zu bestimmten Betriebssituationen. Beispiele für diese Daten sind Konfigurationsdaten und Logdateien.

Bewertung: Die Taxonomie zur Klassifizierung von Betriebsdaten nach MEYER ET AL. bietet eine gute Gliederung der relevanten Betriebsdaten. Die zur Klassifizierung eingesetzten Kriterien sind eher domänennah als analysenah. Daher eignet sich die Taxonomie für die Strukturierung der Betriebsdatenübersicht zur Identifikation der relevanten Daten.

3.4 Ansätze zur Methodenauswahl

In Abschnitt 3.4 folgen insgesamt vier Ansätze zur Data-Analytics-Methodenauswahl.

3.4.1 Evaluierungsmethode für ML-Algorithmen im Kontext der Produktentwicklung nach RIESENER ET AL.

RIESENER ET AL. stellen eine Bewertungsmethode vor, die Nutzern ohne tiefgreifende Kenntnisse von Algorithmen des maschinellen Lernens hilft, spezifische Stärken und Schwächen ausgewählter Algorithmen zu identifizieren. Damit soll eine mögliche Verwendung im Produktentwicklungsprozess ermittelt und eine geeignete Auswahl getroffen werden können [MCM+20]. Die Methode umfasst auf der einen Seite relevante Algorithmen für die Produktentwicklung. Dabei handelt es sich um die Algorithmen, die in der Fachliteratur dominierend eingesetzt werden. RIESENER ET AL. führen hier insgesamt 11 Algorithmen auf. Dazu gehören der k-nearest Neighbor Algorithmus (kNN), die Support Vector Machine (SVM), der Entscheidungsbaum, das künstliche neuronale Netz, Naive Bayes, Random Forest, lineare Regressionen, K-Means, der Apriori Algorithmus sowie das Expectation Maximization Clustering (EM) und der Q-learning Algorithmus. Auf der anderen Seite bietet die Methode neun verschiedene Bewertungskriterien, die für die Vorauswahl von ML-Algorithmen relevant sind: Lernaufgabe, Genauigkeit, Trainingsdauer, Rechenaufwand, Toleranz gegenüber fehlerhaften Werten, Online-Fähigkeit, Transparenz der Algorithmen, Toleranz gegenüber irrelevanten Werten und Lernformen. Anschließend wurden die Algorithmen mit Hilfe der Bewertungskriterien evaluiert und die Ergebnisse dazu in einer Matrix festgehalten (s. Tabelle 3-1).

Tabelle 3-1: Matrix zur Bewertung der Algorithmen in Anlehnung an RIESENER ET AL. [MCM+20]

○ = nicht erfüllt ◔ = knapp erfüllt ◑ = teilweise erfüllt ◕ = fast erfüllt ● = erfüllt	Lernform	Lernaufgabe	Genauigkeit	Dauer des Trainings	Berechnungsaufwand	Toleranz gegenüber fehlerhaften Werten	Toleranz gegenüber irrelevanten Werten	Online-Fähigkeit	Transparenz der Algorithmen
kNN	S	R/K	◔	●	○	○	◕	●	●
SVM	S	K	●	○	●	○	◕	○	◔
Entscheidungsbaum	S	K	◕	◑	◑	○	◔	●	◕
Künstliches neuronales Netz	S	R/K	●	○	○	◑	◕	●	○
Naive Bayes	S	K/(R)	◕	●	●	◕	○	●	◕
Random Forest	S	R/K	●	◑	●	●	●	○	○
Lineare Regression	S	R	●	●	●	○	○	○	○
K-Means	U	Cl	◑	◑	●	○	◔	●	◕
Apriori Algorithmus	U	A	◕	○	○	◑	◑	○	◕
Erwartungsmaximierung	U	Cl	◑	○	●	●	◔	●	●
Q-Lernen	R	-	◑	◔	○	◕	◕	●	○
R = Regression	K = Klassifikation		Cl = Clustering			A = Assoziation			
S = Supervised	U = Unsupervised				R = Reinforcement Learning				

Zur Auswahl eines Algorithmus aus der Matrix schlagen die Autoren vor, eine Problembeschreibung zu verwenden, die Informationen über die Aufgabe (z. B. Clustering, Klassifikation) mit Anforderungen an die Lösung enthält. Durch Bestimmung des minimalen Abstands zwischen der Problembeschreibung und der Algorithmenbewertung werden die am besten geeigneten Algorithmen ermittelt.

Bewertung: Der Ansatz von RIESENER ET AL. nimmt eine kriterienbasierte Bewertung relevanter Algorithmen für die Produktentwicklung vor und bietet damit einen guten und transparenten Überblick über Stärken und Schwächen von relevanten Algorithmen sowie ein Vorgehen zur Identifikation passender Algorithmen basierend auf einer Problembeschreibung und der Algorithmenbewertung. Damit werden auch diesbezüglich Kompetenzen geschult. Im Fokus des Ansatzes steht allerdings die Algorithmenauswahl auf Basis von einzelnen Kriterien, welche nicht alle Einflussfaktoren (Analyseziel, Nutzeranforderungen und Daten) abdecken. Das Vorgehen zur Auswahl berücksichtigt zwar das Problem, geht darauf aber nicht weiter ein; ein unterstützendes Gestaltungswissen fehlt.

3.4.2 Leitfaden für die effiziente Algorithmenauswahl in der Prozessoptimierung nach ZIEGENBEIN ET AL.

ZIEGENBEIN ET AL. stellen in Ergänzung zu ihrer Methode zur Datenquellenauswahl (s. Abschnitt 3.3.1.2) eine Methode zur systematischen Algorithmenauswahl im Produktionskontext vor [ZSM+19].

Sie bauen auf dem CRISP-DM-Verfahren auf und spezifizieren den Schritt Datenverständnis, indem sie vorausgewählte ML-Algorithmen sowie geeignete Datenquellen in einem QFD-Ansatz (Quality Function Deployment) kombinieren. Am Ende der Bewertung steht eine Eignungsbeurteilung für jedes der möglichen ML-Verfahren. Der Ablauf der Methode sieht wie folgt aus:

1) **Priorisierung der Datenquellen**: Datenquellen werden wie im Ansatz in Abschnitt 3.3.1.2 mit Hilfe einer QFD ausgewählt und anhand verschiedener Kriterien, wie Messaufwand und Potenzial für die Geschäftsziele, priorisiert. Anschließend werden die Quellen als VoC (Voice of Customer) in die QFD eingetragen.

2) **ML-Vorauswahl**: Da nicht alle Algorithmen in der QFD berücksichtigt werden können, wird eine strukturierte Vorauswahl durchgeführt. Als erster Filter dient das Ziel des Data-Mining-Projekts. Dies lässt sich realisieren, indem die übergeordneten Methoden Vorhersage, Clustering, Klassifikation und Ausreißeranalyse hinsichtlich ihrer Eignung für verschiedene Anwendungsziele bewertet werden. Die vorausgewählten Verfahren werden als VoE in die QFD aufgenommen.

3) **Nutzwertanalyse**: Anschließend werden die vorausgewählten Methoden nach ihren Stärken und Schwächen gewichtet. Zu diesem Zweck schlagen ZIEGENBEIN ET AL. eine Nutzwertanalyse basierend auf verschiedenen Kriterien vor. Eine Liste an typischen Kriterien wird zur Verfügung gestellt. Beispiele sind die Modellgenauigkeit, der Rechenaufwand, die Anzahl der Metaparameter und die notwendige statistische Kompetenz des Anwenders. Die individuellen Gewichte der Kriterien können über Experten oder systematisch, durch z. B. Paarvergleich, bestimmt werden. Die Gewichtungsfaktoren sollten je nach Anwendung und Anwender individuell festgelegt werden, während die Bewertung der maschinellen Lernverfahren allgemein erfolgen kann. Die berechneten Nutzwerte dienen als Gewichtsfaktoren in dem QFD-Prozess.

4) **Gesamtpriorität**: Anschließend werden die ML-Methoden dahingehend bewertet, wie gut sie zu den Eigenschaften der ausgewählten Datenquellen passen. Dazu werden die beiden Perspektiven VoC und VoE in der Kernmatrix des HoQ dargestellt (s. Bild 3-12). Die Eignung des Datensatzes für das ML-Verfahren und umgekehrt kann unabhängig von der Anwendung mit Hilfe einer Literaturrecherche und einer daraus resultierenden Methodenlandkarte festgestellt werden. Die Gewichtung, d.h. die Wichtigkeit der Datensatzmerkmale muss in Abhängigkeit des

Anwendungsfalls geschehen. Am Ende resultiert ein Nutzwert für jeden betrachteten Algorithmus.

	Prognose				
1 **Passende Datenquellen** / **Vorausgewählte ML-Algorithmen** 2	SVM	RF	ANN - MLP	....	Gewicht (1;3;9)
Spindeldrehzahl	6,40	7,05	6,00		3
Werkzeugposition (x-y-z-Achse)	6,40	7,05	6,00		9
Spindelstrom (c-Achse Spindel)	1,00	1,00	2,00		3
….				3	
ML Gewichtung	3,85	4,00	3,15		
Absolute Algorithmenbewertung	79,80	87.60	78.00		
Normalisierte Algorithmenbewertung	1,02	1,12	1,00	4	

Bild 3-12: Ausschnitt aus dem HoQ in Anlehnung an ZIEGENBEIN ET AL. [ZSM+19]

Bewertung: Der Ansatz von ZIEGENBEIN ET AL. adressiert die Algorithmenauswahl innerhalb des Pipeline-Design und bietet damit ihre eigenständige Bestimmung. Dabei berücksichtigt die Methode sowohl die Daten und deren Eigenschaften als auch die modellseitigen Einflussfaktoren, die zusammen mit den Nutzeranforderungen bewertet werden. Verständnis über die Zusammenhänge und Kompetenzaufbau werden dadurch gefördert. Die aufgestellten Kriterien, bzw. Datensatzmerkmale, entsprechen dem angestrebten Abstraktionsgrad für die Datenbeschreibung und der Methodenauswahl; eine Integration ist daher zu prüfen. Bei der Auswahl wird allerdings die wichtige Vorverarbeitung nicht miteinbezogen. Außerdem erfordert das Verfahren mit seinem Bewertungsschema ein zeitaufwändiges Vorgehen.

3.4.3 Lösungsmuster für Machine Learning nach NALCHIGAR und YU

NALCHIGAR und YU präsentieren einen Modellierungsrahmen für den Entwurf von Business-Analytics-Systemen [NY18]. Das Framework kombiniert drei Sichten: (1) die Geschäftssicht repräsentiert ein Unternehmen in Bezug auf Strategien, Aktoren, Entscheidungen und benötigte Erkenntnisse. Diese Sicht wird genutzt, um systematisch Analyseanforderungen zu ermitteln und die Arten von Analysen zu bestimmen, die der Nutzer benötigt. (2) Die Sicht Analytics-Design stellt das Kernkonzept eines Analytics-Systems

hinsichtlich Analytics-Ziel, ML-Algorithmus, Qualitätsanforderungen und Performance Metriken. Die Sicht identifiziert Design Abwägungen, erfasst die durchzuführenden Experimente und unterstützt die Algorithmenauswahl. (3) In der Datenaufbereitungssicht werden Datenverarbeitungsprozesse in Bezug auf Struktur und Inhalt von Datenquellen und die Gestaltung der Datenaufbereitungsaufgaben betrachtet. Diese Sichten ermöglichen zusammen die Verknüpfung der Unternehmensstrategien mit Analysealgorithmen, Datenspeichern und Aufbereitungsaktivitäten [NY18]. Bestandteil des Frameworks sind auch ein Set an Design-Katalogen, welche analytisches Design-Wissen strukturieren und generische Lösungen zur Verfügung stellen.

Auf dem Modellierungsrahmen bauen ihre Lösungsmuster für maschinelles Lernen auf [NYO+19]. Diese stellen generische ML-Designs für allgemein bekannte und wiederkehrende Business-Analytics-Probleme, wie Betrugserkennung, dar. Ihr Metamodell beschreibt die verschiedenen Elemente eines Lösungsmusters und zeigt ihre semantischen Beziehungen auf (s. Bild 3-13).

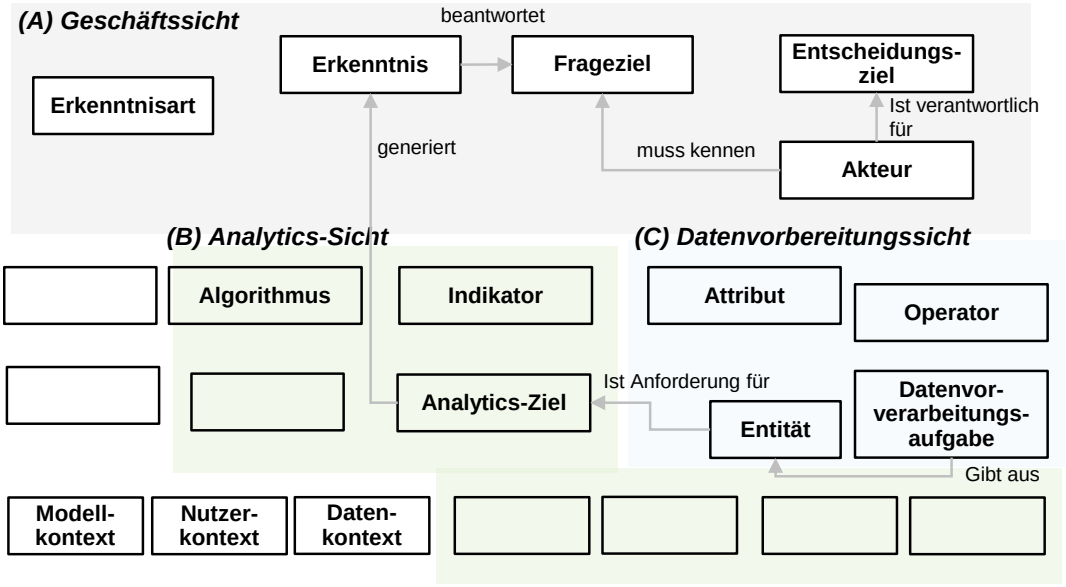

Bild 3-13: Ausschnitt des Metamodells zur Beschreibung der Lösungsmusterelemente in Anlehnung an NALCHIGAR ET AL. [NY18, NYO+19]

Ein Lösungsmuster beginnt mit einer Charakterisierung des Geschäftsproblems und der Bedürfnisse. Diese werden in Form von *Akteuren*, *Entscheidungszielen*, *Fragezielen* und *Erkenntnissen* dargestellt. Durch die Verknüpfung von Akteuren und Entscheidungen mit Fragen und Erkenntnissen übersetzt ein Lösungsmuster ein Geschäftsproblem in (eine Reihe von) wohldefinierte(n) ML-Probleme(n). Nach der Definition des ML-Problems liefert das Muster dann einen Lösungsentwurf für dieses Problem. Es beschreibt, welche Art von Analyse anwendbar ist und welche Algorithmen zu dieser Kategorie gehören. Dies wird in Form von *Analysezielen*, *Algorithmen* und *Verbindungen* dargestellt. Da jeder ML-Algorithmus bestimmte Annahmen hat, die seine Anwendbarkeit auf bestimmte Kontexte beschränken, ist ein wesentlicher Teil eines Lösungsmusters das Wissen

darüber, wann und wie verschiedene Algorithmen zu verwenden sind. Diese werden in Form von *Benutzer-*, *Daten-* und *Modellkontexten* dargestellt. Darüber hinaus beschreiben Lösungsmuster, welche Metriken für das jeweilige Problem anwendbar sind und wann welche Metrik zu verwenden ist. Diese werden in Form von *Indikatoren* und *Evaluierungsverbindungen* dargestellt. Ein Lösungsmuster zeigt darüber hinaus, welche nicht-funktionalen Anforderungen für die Domäne und das vorliegende Problem relevant sind. Es spiegelt auch das Wissen darüber wider, wie verschiedene Algorithmen in Bezug auf diese Anforderungen im Allgemeinen funktionieren. Diese werden in Form von „*Softgoals*“ und *Beitragsverknüpfungen* dargestellt. Lösungsmuster für die Datenbereinigung und -aufbereitung geben Hinweise darauf, welche Daten für das vorliegende Problem relevant sind und wie sie in die richtige Form für verschiedene Algorithmen gewandelt werden sollten. Diese werden in Form von *Entitäten*, *Operatoren* und *Datenflüssen* dargestellt.

NALCHIGAR und YU stellen zusätzlich eine Prototypenarchitektur zur Nutzung der Muster in realen Szenarien vor.

Bewertung: Die Lösungsmuster für Machine Learning von NALCHIGAR und YU stellen einen interessanten Lösungsansatz für die Konzeption von Data-Analytics-Pipelines dar. Der Ansatz geht von geschäftlichen Entscheidungen und Fragen aus und verknüpft diese mit ML-Algorithmen und Datenaufbereitungstechniken. Er berücksichtigt die Abhängigkeiten zwischen Aufgabe, Daten, Benutzeranforderungen und Modelleigenschaften. Das Metamodell liefert den Rahmen zur Generierung von Lösungsmustern zu individuellen Anwendungsfällen. Allerdings handelt es sich bei den Lösungsmustern aus Nutzersicht um vorbereitete Vorlagen, welche Nicht-Experten unterstützend einsetzen können, jedoch keine eigenständige und lernförderliche Erstellung von Pipelines unterstützen.

3.4.4 Domänenorientierte mehrstufige Ontologie nach TIANXING ET AL.

TIANXING ET AL. schlagen eine domänenorientierte mehrstufige Ontologie (DoMO) durch Zusammenführung und Verbesserung bestehender Data-Mining-Ontologien vor [TZ21]. Die Ontologie umfasst vier Ebenen: (1) Einschränkungen, die durch Datenmerkmale beschrieben werden, (2) Definition von Domänen-Datenmerkmalen, (3) Kern-Ontologie für eine bestimmte Domäne und (4) Benutzeranfragen und die Generierung eines Data-Mining-Prozesses. Als intelligenter Assistent soll DoMO vor allem Nicht-Experten helfen, Daten in Form von Ontologie-Entitäten zu beschreiben, die geeigneten Lösungen auf der Grundlage der Datenmerkmale und Aufgabenanforderungen auszuwählen und die Datenverarbeitungsprozesse der ausgewählten Lösungen zu erhalten. Die Architektur der Ontologie, bestehend aus vier Ebenen, ist in Bild 3-14 zu sehen.

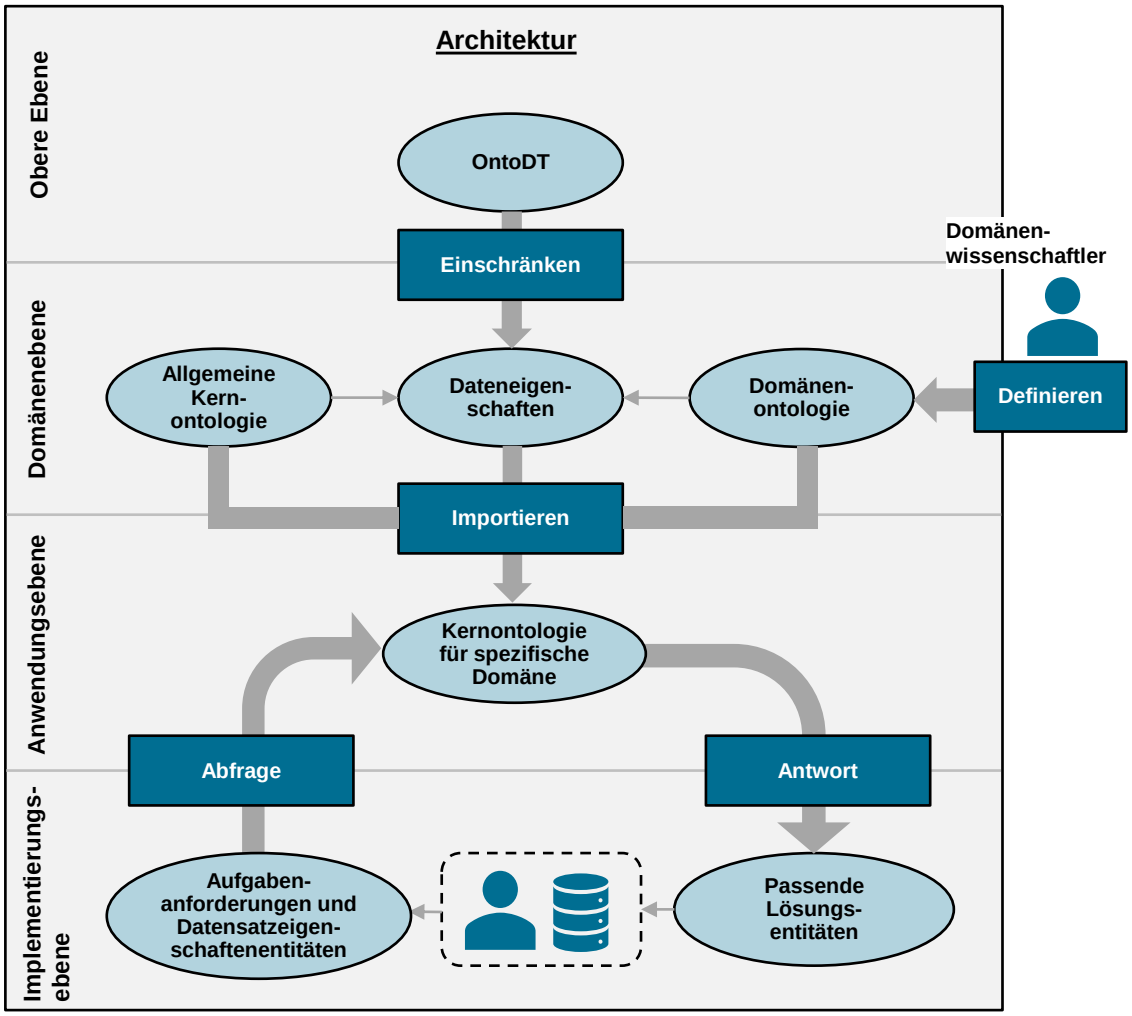

Bild 3-14: DoMO Architektur in Anlehnung an TIANXING ET AL. [TZ21]

Die Ebenen haben die folgenden Rollen:

1) Obere Ebene: In der obersten Ebene werden Entitäten zu allgemeinen Dateneigenschaften, wie z. B. Datentyp, Datentypeigenschaften und Datentyp-Wertebereich aus der OntoDT (s. Abschnitt 3.3.2.2), abgebildet. Die Charakteristika und Aufgabenanforderungen von Datensätzen sind die Grundlage für die Algorithmenauswahl.

2) Domänenebene: In der Domänenebene findet der Schaffensprozess von intelligenten Assistenten für spezifische Felder statt. Basierend auf den Einschränkungen der obersten Ebene definieren Experten die domänenspezifischen Daten.

3) Anwendungsebene: Experten definieren die Datencharakteristika des Feldes und importieren sie in die allgemeine Kernontologie. Damit wird eine Kernontologie für eine spezifische Domäne generiert. Nutzer können direkt auf der Ontologie Abfragen durchführen, um den Data-Mining-Prozess für eine spezifische Aufgabe zu bekommen.

4) Implementierungsebene: Die Generierung von Nutzerabfragen und Data-Mining-Prozessen erfolgt in dieser Ebene. Nutzer erhalten passende Lösungen gemäß der Datencharakteristika und Aufgabenanforderungen.

Die Kernontologie ist in der Initialisierungsphase eine allgemeine Ontologie und enthält eine „Input"-Ontologie sowie einige weitere existierende Data-Mining-Ontologien. Wenn Experten Domänenwissen in Form einer Domänenontologie importieren, entsteht eine Kernontologie für eine spezifische Domäne. Die Input-Ontologie hat als Input-Schnittstelle zum Ziel, Datencharakteristika-Entitäten entsprechend der Algorithmen-Charakteristika zu definieren und Anforderungen der Data-Mining-Aufgabe (Output des Algorithmus) zu beschreiben. Um passende Lösungen und Prozesse zu generieren, werden verschiedene Ontologien integriert. DoMO kommt so insgesamt auf 15 Klassen und sieben relevante Eigenschaften. Beispiele für Klassen sind Algorithmen-Charakteristika, welche die Performance von Data-Mining-Algorithmen einschließlich der Tolerierung einiger Datensatzfehler (z. B. fehlende Werte, Rauschen) beschreiben, und Datencharakteristika, wie die Anzahl an kategorischen und kontinuierlichen Variablen, Anzahl an Klassen und Anteil an fehlenden Werten.

Der Arbeitsablauf für Nutzer von DoMO für die Datenanalyse in einer bestimmten Domäne sieht wie folgt aus:

1) Definition der Eigenschaften der Domänendaten in Form einer Ontologie basierend auf den Einschränkungen von OntoDT
2) Zusammenführung der Domänen-Ontologie und der allgemeinen Kern-Ontologie, um die Kern-Ontologie für die spezifische Domäne zu erhalten
3) Manuelle Beschaffung von Aufgabenanforderungen und Datensätzen und deren Beschreibung in Form von Ontologie-Entitäten als Input
4) Ausführung des Auswahlprozesses auf dieser Kernontologie für eine bestimmte Domäne

Bewertung: Die OntoDM-Ontologie von TIANXING ET AL. stellt einen computer-verstehbaren Ansatz zur Generierung von Data-Mining-Prozessen dar. Sie berücksichtigt Aufgabenanforderungen und Datencharakteristika, um geeignete Lösungen zu finden und adressiert somit einen Großteil des CRISP-DM-Prozesses. Nicht-Experten können von der Ontologie in Form von Assistenten profitieren, indem sie geeignete Lösungen auf der Grundlage spezifischer Aufgabenanforderungen und Datencharakteristika abfragen. Ein solcher Assistent unterstützt sie jedoch nicht bei der eigenständigen Bestimmung der Anforderungen durch Übersetzung der Geschäftsziele und Dateneigenschaften. Definiertes Wissen wird nicht zugänglich gemacht.

3.5 Ansätze zur Toolauswahl

Zuletzt wird in diesem Abschnitt ein ausgewählter Ansatz zur Toolauswahl vorgestellt.

3.5.1 Methode zur Evaluierung und Auswahl von Data-Mining-Software nach COLLIER ET AL.

Einer der ersten Ansätze zur Auswahl von Data-Mining-Tools stammt von COLLIER ET AL. [CCS+99]. Er besteht aus einem Framework zur Evaluierung von Data-Mining-Tools sowie einer Methode zur Anwendung des Frameworks. Der Fokus dabei liegt auf kommerziellen Tools. Der Bewertungsrahmen wird aus vier Hauptkriterien geformt: Performance, Funktionalität, Benutzerfreundlichkeit und Unterstützung von Nebentätigkeiten. Darunter sind weitere Kriterien aufgeführt.

- **Performance**: Diese Kategorie fokussiert sich auf die qualitativen Aspekte der Fähigkeit des Tools, Daten unter verschiedenen Umständen zu verarbeiten. Kriterien sind hier die Plattformvielfalt, Softwarearchitektur, heterogener Datenzugang, Datengröße, Effizienz, Interoperabilität und Robustheit.
- **Funktionalität**: Damit ist die Einbeziehung einer Vielzahl von Fähigkeiten, Techniken und Methoden für Data Mining gemeint. Sie hilft anhand von Kriterien wie Algorithmenvielfalt, vorgeschriebene Methodik, Modellvalidierung, Datentypflexibilität, Algorithmenmodifizierbarkeit, Daten Sampling, Reporting und Modellexport zu beurteilen, wie gut sich das Tool an verschiedene Problemdomänen anpasst.
- **Benutzerfreundlichkeit**: Benutzerfreundlichkeit meint die Anpassung an verschiedenen Ebenen und Benutzertypen ohne Verlust der Funktionalität oder Nützlichkeit. Kriterien zur Beurteilung dessen sind die Nutzerschnittstelle, die Lernkurve, Nutzertypen, Datenvisualisierung, Fehler-Reporting, Aktionshistorie und Domänenvielfalt.
- **Unterstützung von Nebentätigkeiten**: Diese Kategorie untersucht, ob der Nutzer bei verschiedenen Vorverarbeitungs- und Nachbearbeitungsaufgaben wie Visualisierung und anderen Aufgaben unterstützt wird. Kriterien sind Datenbereinigung, Werteersetzung, Datenfilterung, Binning, Ableitung von Features, Randomisierung, Record, Löschen von Einträgen, Umgang mit leeren Stellen, Metadatenmanipulation und Ergebnisfeedback.

Die Methode zur Anwendung des Bewertungsrahmens nutzt Entscheidungsmatrixkonzepte und besteht aus den Phasen: (1) Tool-Vorauswahl, (2) Identifizierung zusätzlicher Auswahlkriterien, (3) Gewichtung der Auswahlkriterien, (4) Tool-Bewertung, (5) Ergebnisauswertung und (6) Tool-Auswahl. Die Tools werden im ersten Schritt auf Basis von starren Vorgaben, wie z. B. der Einschränkung auf ein bestimmtes Betriebssystem, auf eine handhabbare Anzahl reduziert. Im zweiten Schritt geht es darum, ergänzende Kriterien zu identifizieren, die spezifisch für die eigene Organisation sind, wie z. B. Kosten, Fähigkeiten der Endnutzer und Inhalte der Projekte. Irrelevante Kriterien aus dem Framework können verworfen werden. Anschließend werden den Kriterien jeder Kategorie Gewichte zugeordnet, sodass das Gesamtgewicht innerhalb der Kategorien 100% ergibt. Die

Gewichtung muss im Hinblick auf den Verwendungszweck und die Ziele der Software vorgenommen werden. Der vierte Schritt umfasst die vergleichende Bewertung der Tools. Die Bewertung wird dabei in Relation zu einem Referenztool vorgenommen, z. B. dem aus subjektiver Sicht bevorzugten Tool. Die Autoren schlagen hierzu eine diskrete Bewertungsskala von 1 (viel schlechter als das Referenztool) bis 5 (viel besser als das Referenztool) vor. Ein Gesamt-Score wird über die Scores der einzelnen Kategorien durch einen gewichteten Mittelwert gebildet. Im nächsten Schritt wird berücksichtigt, wenn die Bewertung mit der subjektiven Einschätzung nicht übereinstimmt. In einem solchen Fall werden die Gewichtungen aus Schritt 5 überprüft und ggf. angepasst. Zuletzt kann das am besten bewertete Tool ausgewählt werden.

Bewertung: Der Ansatz von COLLIER ET AL. zur Bewertung und Auswahl von Data-Mining-Software ist inzwischen über 20 Jahre alt. Da sich im Laufe dieser Zeit die Tools zur Datenanalyse sehr viel weiterentwickelt haben, müssen die Kriterien geprüft und ggf. angepasst werden. Die Anwendung einer Entscheidungsmatrix-Methode erscheint weiterhin als sehr sinnvoll.

3.6 Handlungsbedarf

Aufbauend auf der Beschreibung und Bewertung der ausgewählten Ansätze des Stands der Forschung wurde untersucht, inwieweit die vorgestellten Ansätze die Ziele an die Systematik aus Abschnitt 2.5 erfüllen. Bild 3-15 zeigt, dass kein einzelner Ansatz noch eine einfache Kombination mehrerer Ansätze die Ziele in vollem Umfang erreicht. Im Folgenden wird dies entlang der elf Ziele begründet. Außerdem wird der verbleibende Handlungsbedarf beschrieben.

Z1) Eignung für Nicht-Experten/ Citizen Data Scientists in der Industrie

Viele der vorgestellten Ansätze verfolgen eine Unterstützung von Anwendern und Nicht-Experten für einzelne verschiedene Aufgaben wie die Use-Case-Definition, Datenquellen- oder Data-Analytics-Methoden-Auswahl. Insbesondere der letzten Aufgabe widmen sich einige formalisierte und automatisierte Ansätze, welche zum Ziel haben, Data-Analytics und KI der breiten Masse zugänglich zu machen. Der datengetriebene Ansatz dieser Disziplinen ist jedoch nicht immer für die Ingenieursrolle in der Industrie geeignet, da sie häufig durch Black-Box-Modelle erforderliches Vertrauen und Transparenz nicht fördern. Mehr Transparenz bieten hingegen die systematischen Ansätze von ZIEGENBEIN ET AL. und STANULA ET AL., weshalb sie einige Impulse für die Systematik liefern.

Z2) Bereitstellung eines Strukturierungsrahmens für die Datenanalyse von Betriebsdaten in der strategischen Produktplanung

Keiner der vorgestellten Ansätze setzt auf einem Strukturierungsrahmen oder Prozess auf, der die Datenanalyse in Form von Anwendungsdefinition, Datenverständnis, Methodenauswahl und Umsetzung in der betriebsdatengestützten Produktplanung beschreibt und den Strukturierungsrahmen zur Ausgestaltung dieser Aufgaben liefert. Das Metamodell

für die Lösungsmuster für maschinelles Lernen von NALCHIGAR ET AL. beschreibt allerdings die verschiedenen Ebenen von Data Analytics im Allgemeinen und bietet damit eine gute Struktur. Auch das 4-Ebenen-Modell und die Data-Analytics-Canvas stellen eine Basis zur Entwicklung eines solchen domänenspezifischen Modells zur Verfügung. Gute Ansätze zur Automatisierung stellen sowohl NALCHIGAR ET AL. als auch TIANXING ET AL. zur Verfügung.

Z3) Bereitstellung von relevanten Data-Analytics-Gegenständen und -Zielen

Keiner der vorgestellten Ansätze erfüllt diese Anforderung vollständig oder teilweise. Die Ansätze zur Definition von Data-Analytics-Anwendungen unterstützen mehr die domänenunabhängige Definition von solchen Aspekten, stellen aber kein Gestaltungswissen in Form von potenziellen Data-Analytics-Gegenständen und Zielen bereit, welche zur Konkretisierung eines (Business-)Use-Cases genutzt werden können.

Z4) Übersetzung der Business Use Cases in Analytics-Aufgabenstellungen

Die Übersetzung der geschäftlichen Ziele in konkrete Analytics Use Cases wird in mehreren Ansätzen teilweise adressiert. Die DM-UML von MARBAN und SEGOVIA stellt vor allem die Verbindung zwischen dem Geschäftsverständnis und der Modellierung her und liefert damit die Verknüpfung dieser beiden Sichten. Da es sich um ein Modellierungswerkzeug handelt, wird jedoch mehr die Dokumentation der Verbindungen als die Identifikation der Verbindungsstellen fokussiert. Die Methode nach HOFMANN ET AL. bringt zwar Domänenprobleme und KI-Lösungen zusammen, bleibt dabei aber zu abstrakt, um dem (Citizen) Data Scientist damit einen guten Startpunkt für die weiteren Schritte zu bieten. Sowohl die Ansätze von NALCHIGAR ET AL., TIANXING und CHAPMAN ET AL. berücksichtigen die Verknüpfung zwischen geschäftlichen Entscheidungen und ML-Techniken, lassen den Prozess der Entstehung der Verknüpfungen aber außen vor, sodass andere Use Cases nicht so einfach übersetzt werden können. Die Modellentitäten bilden allerdings eine gute Basis als notwendige Informationen zur Übersetzung. Folglich soll im Rahmen der Systematik eine Lösung entwickelt werden, die die Anforderung vollständig erfüllt.

Z5) Bereitstellung einer strukturierten und detaillierten Betriebsdatenübersicht

Das Beschreibungsmodell nach KREUTZER stellt einen vier-gliedrigen Strukturierungsrahmen für Betriebsdaten bereit, welcher auf der untersten Ebene eine Vielzahl an konkreten Beispielen aufführt. Damit stellt er eine wichtige Quelle für die geplante Betriebsdatenübersicht dar. Die Taxonomie zur Klassifizierung von Betriebsdaten nach MEYER ET AL. liefert dazu eine sinnvolle Klassifizierung von Betriebsdaten. Die Übersicht der Systematik soll daher auf diesen Vorarbeiten aufbauen.

Z6) Bestimmung der relevanten Betriebsdaten: Eine Methode zur Nutzung einer solchen Übersicht, um sowohl erforderliche als auch existierende Daten zu identifizieren, stellt die Data Analytics Canvas nach KÜHN ET AL. dar. Ihr Einsatz soll daher im Rahmen der Systematik bedacht werden. Die Datenquellenauswahl nach STANULA ET AL. erfüllt

die Anforderung teilweise, indem sie einen guten Ansatz zum systematischen Abgleich von Datenquellen und benötigten Informationen präsentiert, sich jedoch bei den erforderlichen Daten auf die Fehlerarten beschränkt. Die Datenlandkarte erfüllt durch ihren Fokus auf Produktionsprozesse auch teilweise die Anforderungen. Eine prozessseitige Visualisierung und Bestimmung von Daten soll innerhalb der Systematik geprüft werden.

Z7) Bereitstellung eines Beschreibungsrahmens für Betriebsdaten

Das Beschreibungsmodell nach KREUTZER stellt auch relevante Charakterisierungsmerkmale für Betriebsdaten zur Verfügung. Der Fokus liegt hier jedoch nicht auf den Merkmalen, die für die Datenanalyse von Bedeutung sind. PANOV ET AL. zeigen mit ihrer OntoDT eine sehr formalisierte und damit wissensintensive Methode zur Beschreibung von allgemeinen Daten und erfüllen damit teilweise die Anforderung.

Z8) Bestimmung der Dateneigenschaften

Während einige Ansätze Merkmale und Methoden oder Tools zur Dokumentation dieser zur Verfügung stellen, adressiert nur ein Ansatz am Rand die Bestimmung der Eigenschaften durch den Nutzer. Im Rahmen der Algorithmenauswahl präsentieren ZIEGENBEIN ET AL. eine Tabellenvorlage mit Indikatoren für jede Datensatzeigenschaft. Diese soll für die Systematik passend adaptiert werden.

Z9) Bereitstellung von für die betriebsdatengestützten Produktplanung relevanten Vorverarbeitungsmethoden sowie Modellen und Algorithmen und Evaluierungsmetriken

Einige Ansätze im Bereich der Methodenauswahl erfüllen diese Anforderung teilweise, indem sie eine Auswahl an Algorithmen und Techniken für einen anderen Domänenfokus bereitstellen. RIESENER ET AL. zeigen 11 typische Algorithmen für die Produktentwicklung auf, NALCHIGAR ET AL. verschiedene Vorverarbeitungstechniken und Algorithmen für ausgewählte Use Cases wie die Betrugserkennung. Für die betriebsdatengestützte Produktplanung besteht also Handlungsbedarf.

Z10) Auswahl geeigneter Methoden unter Berücksichtigung der Abhängigkeiten

Diese Anforderung wird von den Ansätzen im Bereich der Methodenauswahl adressiert und meist teilweise oder vollständig erfüllt. Die technischeren wissensbasierten Ansätze, wie die Lösungsmuster und die DoMO-Ontologie, berücksichtigen beide die Abhängigkeiten der Methoden zur Anwendung und zu den Daten. Über Eingabe der relevanten Parameter werden passende Verfahren ausgewählt. Solche Rahmenkonzepte sollen insbesondere für ein digitales Tool der Systematik genutzt werden. Etablierte AutoML- oder verwandte datengetriebene Techniken setzen meist erst bei der Modellierung im Datenanalyseprozess an, bieten hier aber sehr gute Unterstützung beim Bauen und Evaluieren der Modelle. Daher soll die Systematik grundsätzlich eine Integration solcher Methoden unterstützen. Der konzeptionelle Ansatz von ZIEGENBEIN ET AL. unterstützt lediglich die Auswahl von Algorithmen und lässt die Vorverarbeitung außen vor. Vor diesem

Hintergrund soll die Systematik geeignete Ansätze der verschiedenen Verfahren aus dem Stand der Forschung nutzen und kombinieren, um den Nutzer der Systematik zu einer fundierten Auswahl zu befähigen.

Z11) Auswahl geeigneter Tools

Eine Auswahl von geeigneten Tools ermöglicht der Ansatz nach COLLIER ET AL. nur zum Teil, da nicht alle relevanten Entscheidungsfaktoren berücksichtigt werden. Um die Auswahl durchzuführen, soll geprüft werden, ob die Methode auf Basis einer Entscheidungsmatrix sinnvoll ist.

Fazit:

Die Analyse des Stands der Forschung ergibt, dass insbesondere produktplanungsspezifisches Lösungswissen fehlt, mit welchem Anwendungen vereinfacht definiert und Betriebsdaten identifiziert werden können. Zum Aufbau von Verständnis über die Betriebsdaten mangelt es an Hilfsmitteln, die eine Bestimmung der Dateneigenschaften unterstützen. Hingegen existieren einige Ansätze, welche sich für Nicht-Experten durch ein transparentes Vorgehen oder einzelnen Komponenten zum Wissensaufbau eignen; dies leisten sie jedoch nicht für den gesamten adressierten Prozess der Datenanalyse für die betriebsdatengestützte Produktplanung. Es besteht daher Handlungsbedarf für eine Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung.

Bewertung der untersuchten Ansätze hinsichtlich der gestellten Ziele **Fragestellung**: Wie gut erfüllen die untersuchten Ansätze (Zeile) die gestellten Ziele an eine Systematik zur Betriebsdaten-gestützten Datenanalyse in der Produktplanung (Spalte)? **Bewertungsskala**: ○ = nicht-erfüllt ◑ = teilweise erfüllt ● = vollständig erfüllt		Übergreifend		Definition von Anwendungen		Datenverständnis				Methoden- und Toolauswahl		
		Eignung für Nicht-Experten in der Industrie	Bereitstellung eines Strukturierungsrahmens	Bereitstellung Analytics-Gegenstände und Ziele	Übersetzung in Analytics Aufgabenstellungen	Bereitstellung einer Betriebsdatenübersicht	Bestimmung der relevanten Daten	Bereitstellung eines Beschreibungsrahmens	Bestimmung der Dateneigenschaften	Bereitstellung relevanter Methoden	Auswahl geeigneter Methoden	Auswahl geeigneter Tools
		Z1	**Z2**	**Z3**	**Z4**	**Z5**	**Z6**	**Z7**	**Z8**	**Z9**	**Z10**	**Z11**
Betriebsdatenanalyse in der PP und PE	Feedback-Assistenz-System nach DIENST	●	◑	○	○	○	○	○	○	○	○	○
	Taxonomie für Produktentwicklungsszenarien nach HOLLER ET AL.	◑	◑	○	○	○	○	○	○	○	○	○
	Anforderungerhebung durch explor. Analyse nach RIESENER ET AL.	◑	◑	○	◑	○	◑	○	○	○	○	○
	Konzept der techn. Vererbung nach LACHMEYER ET AL.	○	◑	○	○	○	○	○	○	○	○	○
	Data-Analytics-Baukasten für Produktdesigner nach EDDAHAB	◑	○	○	◑	○	○	○	○	◑	◑	○
Anwendungen	Data Analytics Canvas nach KÜHN ET AL.	◑	●	○	◑	○	●	○	○	○	○	○
	Use-Case-Modellierung nach MARBAN ET AL.	○	◑	○	◑	○	◑	○	○	○	◑	○
	Entwicklung von KI-Use-Cases nach HOFMANN ET AL.	◑	○	○	◑	○	○	○	○	○	○	○
Datenverständnis	Datenlandkarte nach JOPPEN ET AL.	◑	○	○	○	○	◑	○	○	○	○	○
	Datenquellenauswahl nach STANULA ET AL.	◑	○	○	◑	○	◑	◑	○	○	○	○
	Nutzenpotenziale von Felddaten nach KREUTZER	◑	○	○	○	◑	◑	○	○	○	○	○
	OntoDT nach PANOV ET AL.	○	○	○	○	○	○	●	○	○	○	○
	Klassifizierung von Betriebsdaten nach MEYER ET AL.	◑	○	○	○	◑	○	○	○	○	○	○
Methoden- und Toolauswahl	Evaluierungsmethode für ML-Algorithmen nach RIESENER ET AL.	●	○	○	○	○	○	○	○	◑	◑	○
	Algorithmenauswahl nach ZIEGENBEIN ET AL.	●	○	○	○	○	◑	●	◑	○	◑	○
	Lösungsmuster für ML nach NALCHIGAR ET AL.	◑	◑	○	◑	○	○	◑	○	◑	●	○
	DoMO nach TIANXING ET AL.	◑	◑	◑	◑	○	○	●	○	◑	●	○
	Auswahl von DM-Software nach COLLIER ET AL.	○	○	○	○	○	○	○	○	○	○	◑

Bild 3-15: Bewertung des untersuchten Stands der Forschung

4 Entwicklung der Systematik

Dieses Kapitel stellt das Ziel der vorliegenden Arbeit einschließlich ihrer Entwicklung vor: eine Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung. Diese soll die definierten Ziele in der Problemanalyse (Abschnitt 2.5) erfüllen. In Abschnitt 4.1 wird zunächst ein Überblick über die Systematik gegeben. Abschnitt 4.2 stellt den Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung vor. In Abschnitt 4.3 wird anschließend der Analytics-Baukasten für die betriebsdatengestützte Produktplanung beschrieben. In Abschnitt 4.4 wird das Vorgehen zur Datenanalyse in der betriebsdatengestützten Produktplanung erläutert. Abschnitt 4.5 präsentiert das darauf aufbauende digitale Assistenz- und Lerntool. Zum Abschluss gibt Abschnitt 4.6 die Ergebnisse der Kriterien-basierten Analyse wieder, in welcher die Systematik anhand der Ziele der Problemanalyse bewertet wird.

4.1 Überblick über die Systematik

Die Systematik besteht aus vier Elementen: 1) einem Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung, 2) einem Analytics-Baukasten für die betriebsdatengestützte Produktplanung, 3) einem Vorgehen zur Datenanalyse in der betriebsdatengestützten Produktplanung und 4) einem digitalen Assistenz- und Lerntool. Bild 4-1 gibt einen Überblick über die Systematik.

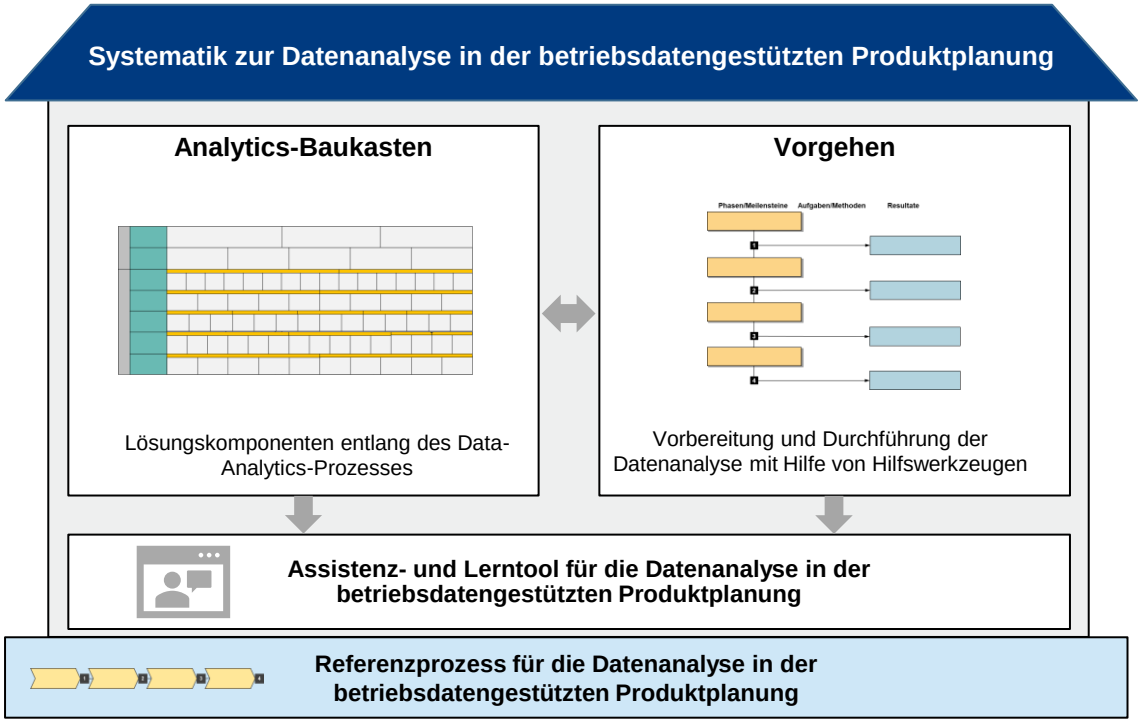

Bild 4-1: Überblick über die Systematik

1) Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung (Abschnitt 4.2): Der Referenzprozess gibt an, mit welchen Schritten Datenanalysen in der betriebsdatengestützten Produktplanung umgesetzt werden können. Damit stellt er die allgemeine Pipeline für Datenanalyseprojekte dar, welche durch verschiedene Komponenten in Form von Anwendungen, Datenquellen und Analyse-Techniken ausgestaltet werden kann. Der Referenzprozess basiert auf dem Referenzprozess für die betriebsdatengestützte Produktplanung und bettet sich damit in einen größeren Rahmen einschließlich der Planung und Verwertung von Betriebsdaten-Analysen ein.

2) Analytics-Baukasten für die betriebsdatengestützte Produktplanung (Abschnitt 4.3): Der Analytics-Baukasten strukturiert sich entlang des Referenzprozesses für die Datenanalyse in der betriebsdatengestützten Produktplanung und stellt entlang der dadurch resultierenden Dimensionen relevante Lösungskomponenten zur Use-Case-spezifischen Ausgestaltung und Konkretisierung einer allgemeinen Data Analytics Pipeline zur Verfügung.

3) Vorgehen zur Datenanalyse in der betriebsdatengestützten Produktplanung (Abschnitt 4.4): Dieses Vorgehen beschreibt ausführlich, wie Betriebsdaten-Analysen in der strategischen Produktplanung systematisch und erfolgreich vorbereitet und durchgeführt werden können. Es konkretisiert dabei den zuvor eingeführten Referenzprozess durch Einführung verschiedener Werkzeuge. Das Vorgehen befähigt ihre Anwender, die passenden Lösungskomponenten des Analytics-Baukastens zu bestimmen.

4) Assistenz- und Lerntool für die Datenanalyse in der betriebsdatengestützten Produktplanung (Abschnitt 4.5): Dieses Tool setzt die aus den zuvor beschriebenen Elementen bestehende Systematik in Form eines digitalen Lernassistenten um, welcher seinen Anwendern eine intuitive und geführte Entwicklung einer erfolgsversprechenden Data Analytics Pipeline ermöglicht.

4.2 Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung

Eine wesentliche Erkenntnis der Problemanalyse ist, dass Nicht-Experten einen Strukturierungsrahmen zur Konzipierung und Umsetzung von Data Analytics Projekten in der Produktplanung in Form einer Data-Analytics-Pipeline benötigen (vgl. Abschnitt 2.5). Daher wurde ein Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung entwickelt, welcher den Referenzprozess zur betriebsdatengestützten Produktplanung nach MEYER ET AL. [MWP+22] auf Basis der Problemanalyse (vgl. Abschnitt 2.3.1) verkürzt.

Der Referenzprozess nach MEYER ET AL. ist in Bild 4-2 abgebildet. Er setzt sich aus vier Hauptprozessen zusammen: 1) Planung von Betriebsdaten-Analysen, (2) Vorbereitung von Betriebsdaten-Analysen, (3) Durchführung von Betriebsdaten-Analysen und (4)

Verwertung von Betriebsdaten-Analysen in der strategischen Produktplanung. Jeder Hauptprozess besteht wiederum aus vier Subprozessen bzw. Phasen.

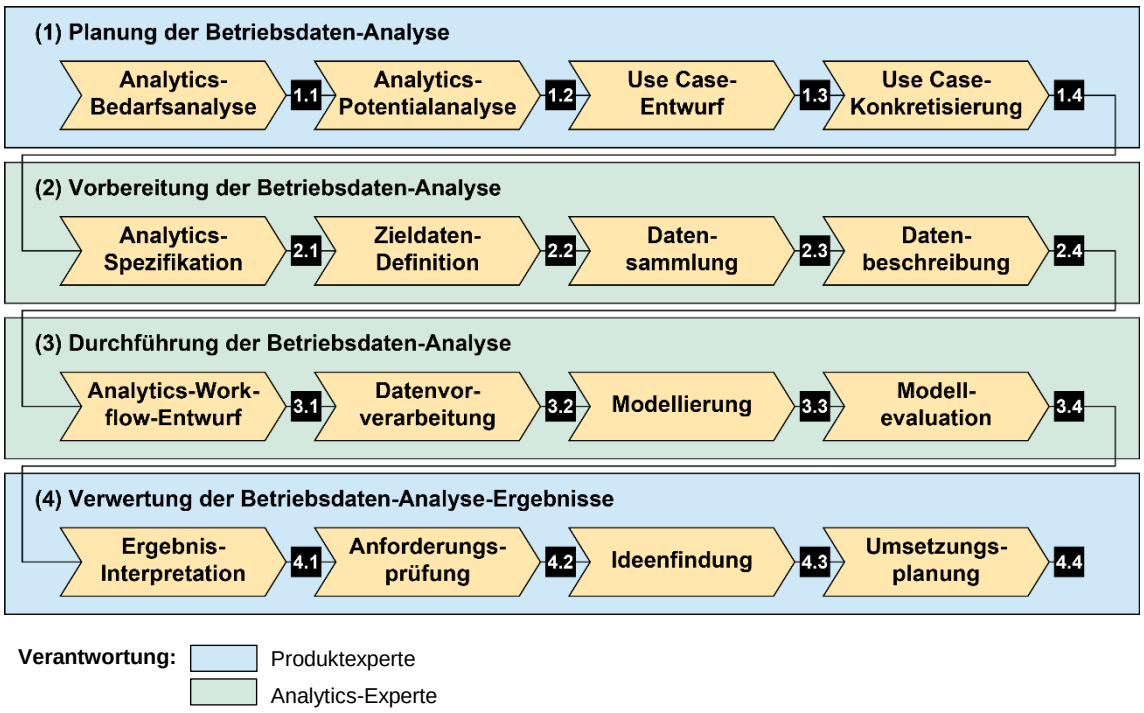

Bild 4-2: Referenzprozess für die betriebsdatengestützte Produktplanung [MWP+22]

Dieser Referenzprozess entstand aus mehreren kombinierten Ansätzen zur Entwicklung von Referenzprozessen, darunter das Prozessmodell für eine empirisch fundierte Referenzmodell-Konstruktion nach AHLEMANN und GASTL [AG07], das Vorgehen von FRANK ET AL. [FGW+20] sowie die Gestaltungsprinzipien für die Referenzmodellierung nach VOM BROCKE [vom07]. Daraus wurde ein **methodisches Vorgehen** aus vier Phasen entwickelt:

1) **Domänenanalyse:** In dieser Phase wurde das erforderliche Wissen für die betriebsdatengestützte Produktplanung gesammelt und analysiert. Dies umfasste literaturbasierte Inhalte sowie Ergebnisse einer Interviewstudie zu Potenzialen und Herausforderungen. Das gewonnene Domänenwissen diente als Grundlage zur Ableitung inhaltlich notwendiger Prozessschritte.

2) **Referenzprozess-Entwurf:** Hier erfolgte die Erstellung einer ersten Version des Referenzprozesses. Durch Analyse bestehender Referenzprozesse im Bereich Data Analytics wurden detaillierte Modelle erstellt, verglichen und zu einem umfassenden Referenzprozess aggregiert. Dieser wurde spezifisch auf die betriebsdatengestützte Produktplanung angepasst, indem notwendige Prozessschritte ergänzt und unnötige entfernt wurden.

3) **Theoretische Validierung:** Die theoretische Validierung erfolgte durch Experteninterviews mit drei Domänenexperten. Die erste Version des

Referenzprozesses wurde anhand vordefinierter Fragen diskutiert. Basierend auf den Rückmeldungen der Experten wurde der Referenzprozess überarbeitet und erneut vorgestellt. Nachdem die Experten zufrieden waren, galt der Referenzprozess als theoretisch validiert.

4) **Praktische Validierung:** Nach der theoretischen Validierung wurde der Referenzprozess in der Praxis bei vier produzierenden Unternehmen angewendet. Dabei wurden Verbesserungspotenziale identifiziert, welche zur weiteren Überarbeitung des Referenzprozesses genutzt wurden. Das Ergebnis dieser Phase ist der praktisch validierte Referenzprozess.

Die detaillierte Ausarbeitung sowie Beschreibung der Phasen ist der entsprechenden Veröffentlichung zu entnehmen [MWP+22].

Im Rahmen dieser Arbeit werden der zweite und dritte Hauptprozess des Referenzprozesses durch Zusammenführung einzelner Phasen vereinfacht und entsprechend der in der Problemanalyse identifizierten Phasen verschlankt, um die nachfolgenden Elemente der Systematik besser strukturieren zu können. Bild 4-3 zeigt den angepassten Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung. Die Phasen werden anschließend erläutert.

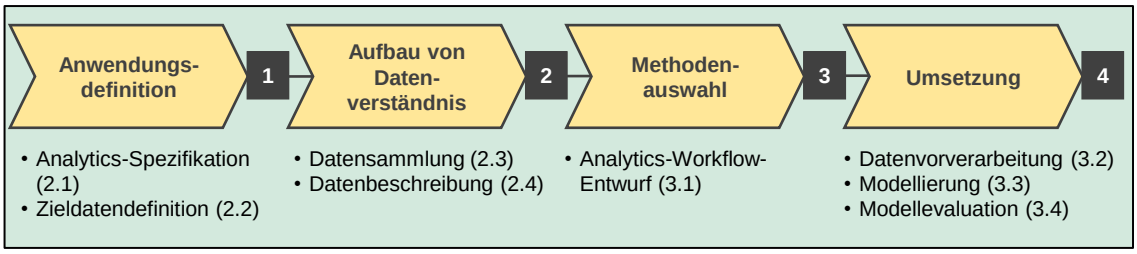

Bild 4-3: Verkürzter Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung und zugehörige Phasen des Referenzprozesses für die betriebsdatengestützte Produktplanung nach MEYER ET AL.

1) Definition der Anwendung: In dieser Phase findet ein Übersetzungsprozess des Use Cases vom Produkt- in den Analytics-Kontext statt. Dadurch werden die Analyseziele in Form einer technischen Analytics-Aufgabenstellung spezifiziert. Außerdem werden die Zieldaten grob definiert.

2) **Aufbau von Datenverständnis** (Datensammlung und -beschreibung): Diese Phase umfasst zum einen die Datenidentifikation und -sammlung. Der Datenbedarf muss weiter konkretisiert werden und mit den tatsächlich existierenden Daten im Unternehmen abgeglichen werden. Zum anderen werden die Daten und ihre Eigenschaften genau analysiert, um das notwendige Datenverständnis für die anschließende Methodenauswahl zu gewinnen.

3) **Methodenauswahl** (Vorverarbeitung, Modellierung und Evaluierung): In dieser Phase werden auf Basis der zuvor identifizierten Faktoren zur Anwendung und zu den Daten passende Vorverarbeitungsverfahren, Algorithmen und

Evaluierungsmetriken ausgewählt und in einer Analytics-Pipeline zusammengestellt. Diese dient als Konzept für die anschließende Umsetzung.

4) **Umsetzung**: Im Rahmen der Umsetzung findet zunächst eine Auswahl von Tools statt, die für die zusammengestellte Pipeline geeignet sind. Anschließend kann die Umsetzung mit den verschiedenen Hilfsmitteln starten. Ggf. sind weitere Durchläufe des gesamten Prozesses oder einzelne Sprünge zurück notwendig.

4.3 Analytics-Baukasten für die Datenanalyse in der betriebsdatengestützten Produktplanung

Die Ausführung der Problemanalyse in Abschnitt 2.3.2 bis 2.3.4 belegt den Bedarf für die Bereitstellung von relevanten Analytics-Zielen, Betriebsdaten und Vorverarbeitungsmethoden, Algorithmen und Evaluierungsmetriken. Im Folgenden wird ein Baukasten entwickelt, der diese Komponenten zur Verfügung stellt. Die Abschnitte 4.3.1 bis 4.3.3 gehen dabei in Anlehnung an den Referenzprozess (vgl. Abschnitt 4.2) für die Dimensionen Anwendung, Datenverständnis und die Methodenauswahl für Vorverarbeitung, Modellierung und Evaluierung jeweils auf das methodische Vorgehen sowie die Ergebnisse ein. In Abschnitt 4.3.4 werden die Ergebnisse in Form von Analytics-Lösungskomponenten in dem Analytics-Baukasten für die Datenanalyse in der betriebsdatengestützten Produktplanung zusammengeführt.

4.3.1 Anwendung

Die Problemanalyse zeigt, dass die Bereitstellung typischer Analyseziele ein erster wichtiger Schritt und Bestandteil bei der Definition der Anwendung, bzw. der Aufgabenstellung, ist und eine Brücke zu den Geschäftszielen schlägt (vgl. Abschnitt 2.3.2). Sie sind wichtig, um die adressierten Analytics-Probleme, z. B. Klassifikation, zu bestimmen. Daher wurden für den Baukasten der Systematik relevante Analyseziele und -probleme für die betriebsdatengestützte Produktplanung identifiziert. Dazu wurden zwei Forschungsansätze verfolgt: ein Experten-getriebener Ansatz zur Aufdeckung von interessanten, potenziell in Frage kommenden Zielen und ein Literatur-basierter Ansatz zur Bestimmung von Zielen („Produktplanungssicht“), welche bereits erfolgreich mit Hilfe von Data Analytics umgesetzt wurden („Analytics-Sicht“). In Abschnitt 4.3.1.1 wird zunächst das methodische Vorgehen zur Identifizierung dieser Ziele beschrieben. Anschließend werden die Analytics-Ziele (Ergebnisse) vorgestellt (vgl. Abschnitt 4.3.1.2).

4.3.1.1 Methodisches Vorgehen

Der erste Ansatz identifizierte Analytics-Ziele anhand der geschäftlichen Ziele der Produktplanung (Top-down-Ansatz). Um diese Ziele zu adressieren, wurden die Nutzen-elemente nach ALMQUIST ET AL. genutzt [ASB16, ACS18]. Diese Elemente beschreiben den

Nutzen, den ein Produkt dem Kunden bieten kann und erklären, warum Kunden ein Produkt kaufen und verwenden. 25 von insgesamt 40 verschiedenen Nutzenelementen wurden als passend für die Betriebsdatenanalyse bewertet. Um anhand dessen Analytics-Ziele zu identifizieren, wurden zwei Workshops von jeweils drei Stunden mit insgesamt 17 Experten aus der Produktplanung und dem Data Science durchgeführt. Details zum Vorgehen sind der dazugehörigen Veröffentlichung zu entnehmen [MPK+22]. Innerhalb von Gruppen sollten die Teilnehmer Anwendungsbeispiele (z. B. Analyse des Nutzerverhaltens), exemplarische Fragen (z. B. Wie navigieren Nutzer durch das Nutzerinterface?) und erforderliche Daten (z. B. Nutzerverhaltensdaten, Statusdaten) für die Analyse von Betriebsdaten für einige Nutzenelemente identifizieren. Nach dem Workshop wurden die die Ergebnisse analysiert und aggregiert, um Analytics-Ziele abzuleiten.

Der zweite Ansatz identifizierte Analytics-Ziele mit Hilfe einer strukturierten Literaturrecherche (Systematic Literature Review - SLR). Diese folgte den Leitlinien von KITCHENHAM ET AL. und KUHRMANN ET AL. und durchläuft drei Hauptphasen [KBB+09, KFD17]: 1. die Planung und Vorbereitung des Reviews, 2. die Durchführung des Reviews einschließlich der Datenerhebung und Studienauswahl und 3. die Analyse. Im Folgenden werden die einzelnen Phasen näher beschrieben.

Planung des Reviews

In dieser Phase wurden die Forschungsziele und die Art der Durchführung festgelegt. Dazu wurde die folgende Forschungsfrage formuliert:

RQ1: Für welche Anwendungen in der Produktplanung wird Data Analytics eingesetzt?

Wie in der Problemanalyse beschrieben (vgl. 2.3.2), bietet der Einsatz von Data Analytics in der strategischen Produktplanung viele Potenziale, aber eine große Herausforderung ist die Definition geeigneter Anwendungsfälle, die mit Data Analytics realisiert werden können. Selbst wenn die Geschäftsziele klar sind, kann es sein, dass die Datenperspektive und die Anwendungsfälle, die durch Data Analytics umgesetzt werden können, fehlen oder unklar sind. Daher sollte in der Literatur untersucht werden, welche spezifischen Probleme in der Produktplanung mit Hilfe von Data-Analytics-Techniken gelöst werden.

Um ein umfassendes Bild des Forschungsfeldes zu erhalten, wurden die Suchbegriffe auf Basis der Problemanalyse sorgfältig ausgewählt und getestet. Folgende Suchanfrage wurde durch Ergänzung von Begriffen und Synonymen aus den Feldern Data Analytics und Produktplanung erstellt:

1. ("Data Analytics" OR "Machine Learning" OR "Data Mining" OR "data-driven") AND ("product planning" OR "product design" OR "product management" OR "early phases of product development" OR "product optimization") AND ("review" OR "overview" OR "survey" OR “case study”)

2. ("Data Analytics" OR "Machine Learning" OR "Data Mining" OR "data-driven") AND ("data description" OR "classification" OR "regression" OR "clustering" OR

"dependency analysis" OR "causation analysis“ OR “causal discovery”) AND ("product planning" OR "product design" OR "product management" OR "early phases of product development" OR "product optimization")

Die Suchbegriffe wurden nur auf Titel, Zusammenfassungen und Schlüsselwörter angewendet. Um die Treffermenge weiter zu reduzieren, wurden Ein- und Ausschlusskriterien definiert (siehe Tabelle 4-1).

Tabelle 4-1: Ein- und Ausschlusskriterien

Einschlusskriterien	Ausschlusskriterien
Publikationen, die die definierten Forschungsfragen beantworten	Papiere, die für die definierten Forschungsfragen nicht relevant sind
Papiere, die sich auf die Produktplanung und frühe Phasen der Produktentwicklung im Kontext der Anforderungs- und Ideenextraktion beziehen	Papiere, die nicht in englischer Sprache verfasst sind
Alle Papiere, die Data Analytics/ ML/ KI nutzen, um Diagnosen im industriellen Kontext zu erstellen (falls Übertragbarkeit möglich)	Beiträge, die nicht unter die Definition der Produktplanung fallen
Beiträge, die zwischen 2005 und 2023 veröffentlicht wurden	Arbeiten, die keine konkreten datengetriebenen Lösungsansätze aufführen
Studien, Übersichtsarbeiten, Fallstudien	Arbeiten in medizinischen oder anderen branchenfremden Domänen

Durchführung des Reviews

In der Durchführungsphase wurde der Suchstring für die Suche in den Datenbanken verwendet: Eine automatisierte Suche in den Online-Bibliotheken IEEE Xplore, Scopus, SpringerLink und ScienceDirect ergab insgesamt ein Datensatz von 1603 Publikationen. Der detaillierte Ablauf ist in Bild 4-4 dargestellt. Der Datensatz wurde nach dem Vier-Augen-Prinzip gesichtet und auf Publikationen analysiert, die für die Forschungsfrage relevant sind. Dazu wurden die Titel und Zusammenfassungen der gesammelten Studien geprüft, um irrelevante Studien zu entfernen. Nach der Prüfung blieben 70 Publikationen übrig. Mit Hilfe eines „Backward Snowballing“ wurden weitere 22 relevante Publikationen ermittelt.

Eine zusätzliche Qualitätsprüfung und die Entfernung von Duplikaten resultiert in einem finalen Datensatz von 82 relevanten Arbeiten.

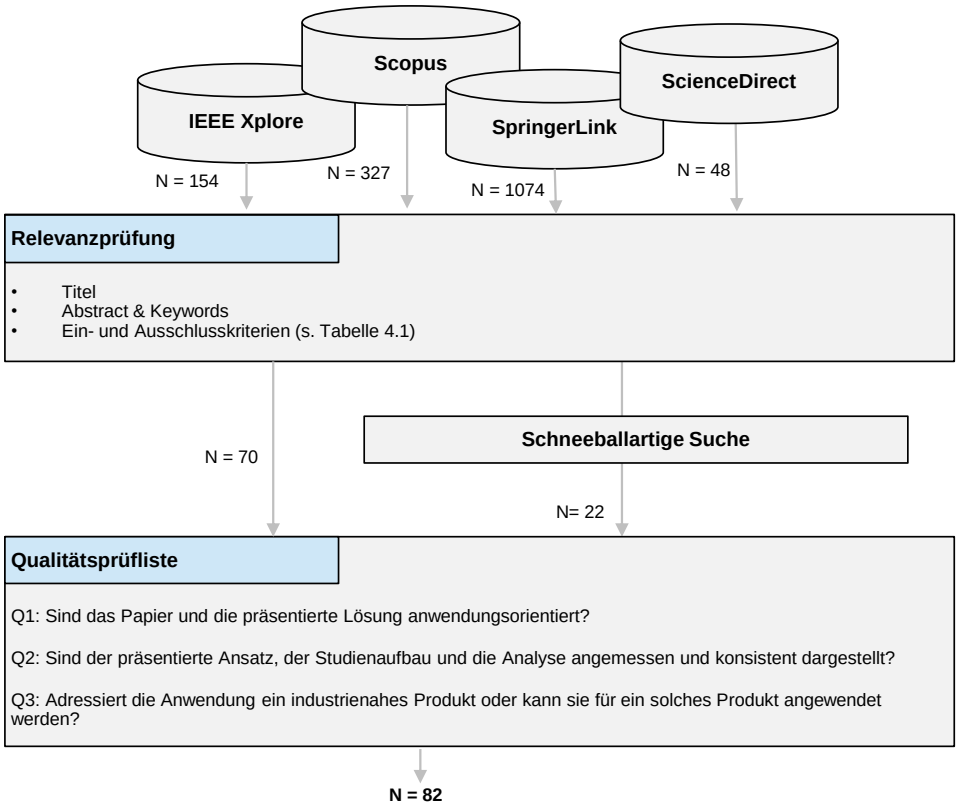

Bild 4-4: Ablauf der strukturierten Literaturstudie

Auswertung

Anschließend wurde der Datensatz sorgfältig analysiert und die Anwendungen, bzw. Ziele, sowie die Data-Analytics-Probleme extrahiert. Zwei Forscher aus dem Bereich Industrial Data Analytics einschließlich der Autorin der vorliegenden Arbeit ordneten unabhängig voneinander die extrahierten Anwendungen den aus dem ersten Ansatz resultierenden Ziel-Kategorien zu. Als Kodierungsbasis für die Probleme fungierten die Data-Analytics-Probleme aus der Problemanalyse (vgl. Abschnitt 2.3.2). Wenn die Zuordnungen nicht übereinstimmten, wurde dies diskutiert und ggf. eine andere Zuordnung vorgenommen oder eine besser passende Kategorie definiert. Anschließend wurde eine quantitative Auswertung durchgeführt, die Aufschluss darüber gab, wie oft die verschiedenen Kategorien von Anwendungen den Papieren zugeordnet werden konnten. Die Annahme dahinter ist, dass die Bedeutung von Analytics-Zielen höher ist, wenn die Häufigkeit der Erwähnung und die Ausführlichkeit in der Literatur hoch ist. Kategorien mit sehr wenigen Übereinstimmungen wurden nochmals daraufhin überprüft, ob sie in eine andere Kategorie integriert werden können. Alle erwähnten Methoden wurden jedoch geclustert, um größere Kategorien zu definieren, wie z. B. Skalierung und Normalisierung. Diese wurden dann zur Strukturierung der Vorverarbeitungsmethoden verwendet.

4.3.1.2 Ergebnisse

Ergebnis des ersten Ansatzes sind insgesamt 17 Analytics-Anwendungen, bzw. Ziele für die Analyse von Betriebsdaten in der Produktplanung. Sie lassen sich in drei Kategorien einordnen: (1) Ziele mit Produktfokus, (2) Ziele mit Prozess-Fokus und (3) Ziele mit Kunden- und Nutzer-Fokus. Eine detailliertere Beschreibung der einzelnen Ziele sind der entsprechenden Veröffentlichung zu entnehmen [MPK+22].

1) Ziele mit **Produktfokus**:
    - Verifizieren und Validieren von Anforderungen oder Identifizieren von neuen **Anforderungen**
    - Erkennen von **Zuverlässigkeitsproblemen** und damit verbundenen Auslösern
    - Untersuchen von bekannten **Fehlern**, ihren Ursachen und Effekten
    - Verstehen von **Abnutzung und Verschleiß**, ihren Ursachen und Effekten
    - Vergleichen von **Lösungsalternativen** für dieselbe Aufgabe
    - Verstehen der **Produktbedienung**
    - Erkennen von Komponenten, die nicht gut für die **Arbeitsbelastung** sind
2) Ziele mit **Prozessfokus**:
    - Identifizieren von Faktoren, die eine hohe **Produktivität** behindern oder fördern
    - Identifizieren von **Prozessmängeln** und Ineffizienzen
    - Aufdecken von verschwendeten **Ressourcen**
    - Bestimmen von **Emissionen** und beeinflussenden Faktoren
    - Sicherstellen, dass das Produkt allen **Vorschriften** entspricht
3) Ziele mit **Kunden- und Nutzerfokus**:
    - Bestimmen, wie Kunden das **Produkt nutzen**
    - Verstehen der Wichtigkeit von **Funktionen** und ihrer **Nutzung**
    - Ableiten von **Produktvarianten**
    - Verstehen, wie Nutzer sich **verhalten** und mit dem Produkt interagieren
    - Verstehen der Beschwerden und **Bedürfnisse** der Nutzer

Wie im letzten Abschnitt erläutert, dienten diese Zielkategorien zur initialen Kodierung der extrahierten Anwendungen aus dem zweiten, literaturbasierten Ansatz. Bei den

Fällen, wo die Forscher häufig zwei unterschiedliche Kategorien angaben, wurden die Zielkategorien zusammengelegt (z. B. Fehleranalyse und (Zuverlässigkeits-)Problemanalyse, Fehleranalyse und Prozessmängel, Kundenbedürfnisse und Kundenanforderungen). Kategorien, die häufiger neu hinzugefügt wurden, wurden ergänzt (z. B. Trendanalyse und Kundensegmentierung). Ziele, welche nicht oder nur sehr vereinzelt zugeordnet wurden (z. B. Produktvarianten, Vorschriften), wurden vernachlässigt. Zusätzlich wurden sehr häufig genannte Kategorien, wie die Fehleranalyse, im Anschluss noch einmal genauer betrachtet und eine Spezifizierung durch Aufteilung in mehrere Ziele vorgenommen (Fehler- und Problemdetektion und Fehlerdiagnose). Ergebnis dieses iterativen Prozesses sind sechs Data-Analytics-Ziele für die betriebsdatengestützte Produktplanung (sortiert nach Häufigkeit ihrer Nennung in den Publikationen): (1) Nutzerbedürfnisanalyse, (2) Fehlerdiagnose, (3) Fehler- und Problemdetektion, (4) Nutzerverhaltensanalyse, (5) Trendanalyse und (6) Nutzersegmentierungsanalyse.

Die Auswertung der Data-Analytics-Probleme ergab durch ähnliche Vorgänge eine Ergänzung der verwendeten Problemklassen um das Problem Text Mining, welches verschiedene Ansätze zur Themenaufdeckung, Textzusammenfassung und Attributidentifikation zusammenfasst. Auch der Ansatz der Assoziationsanalyse wurde hinzugefügt, da diese unter der Kategorie Abhängigkeitsanalyse sehr häufig genannt wurde. Damit ergaben sich sieben **Data-Analytics-Probleme**: (1) Beschreibung, (2) Klassifikation, (3) Regression, (4) Clustering, (5) Abhängigkeitsanalyse, (6) Assoziationsanalyse und (7) Text Mining.

Das häufigste Thema im Datensatz waren die **Nutzerbedürfnisse**. Im Mittelpunkt standen hierbei (1) die Extraktion von Zufriedenheit, bzw. Sentiment, über das Produkt und konkrete Produktattribute, um Handlungsbedarf auf der Entwicklungsseite zu identifizieren, (2) die Klärung von Kundenbedürfnissen, um Anhaltspunkte für Produktanpassungen und neue Anforderungen zu erhalten, und (3) die direkte Ermittlung von Kundenanforderungen, die an die Produktentwicklung weitergegeben werden können. Diese Ziele wurden am häufigsten mit Hilfe von Klassifikation gelöst, aber auch Text Mining war in diesem Zusammenhang sehr beliebt (insbesondere zur Verarbeitung von Review-Daten).

Mit nur einer Nennung weniger wurde die **Fehlerdiagnose** eingesetzt, um Informationen über das Produkt zu erhalten. Überwiegend wurden hier Einflussfaktoren auf Fehler aufgedeckt, um mögliche Schwachstellen und deren Ursachen zu identifizieren. Auch hier wurden Klassifikationsansätze mit elf Nennungen am häufigsten gezählt. Die Abhängigkeitsanalyse wurde achtmal zur Diagnose eingesetzt. Um Abhängigkeiten in Form von Assoziationsregeln abzuleiten, wurde an dritter Stelle die Assoziationsanalyse in Anspruch genommen.

An dritter Stelle steht die **Fehlererkennung**, die einen Schritt vor der Diagnose ansetzt und Fehler als auch Probleme und Anomalien aufspürt. Überwiegend sind hier Klassifikationsansätze zu finden. Auch das Clustering scheint ein beliebter Ansatz für die Fehlererkennung zu sein, insbesondere wenn keine Labels vorhanden sind.

Es folgen als weitere relevante Anwendungen die **Nutzerverhaltensanalyse**, die **Trendanalyse** und die **Nutzersegmentierungsanalyse**. Bei der Analyse des Nutzerverhaltens und der -segmentierung werden die Nutzer genauer betrachtet, um ihr Verhalten zu identifizieren und sie sinnvoll zu gruppieren. Dies kann mit weiteren Hinweisen für Produktanpassungen einhergehen. Clustering scheint hier ein populärer Ansatz zu sein. Die Trendanalyse konzentriert sich auf neue Markttrends und Veränderungen, aus denen sich neue Anforderungen an ein Produkt ergeben können.

Bild 4-5 zeigt die Verteilung der im Datensatz angesprochenen Anwendungen und zugrunde liegender Analytics-Probleme.

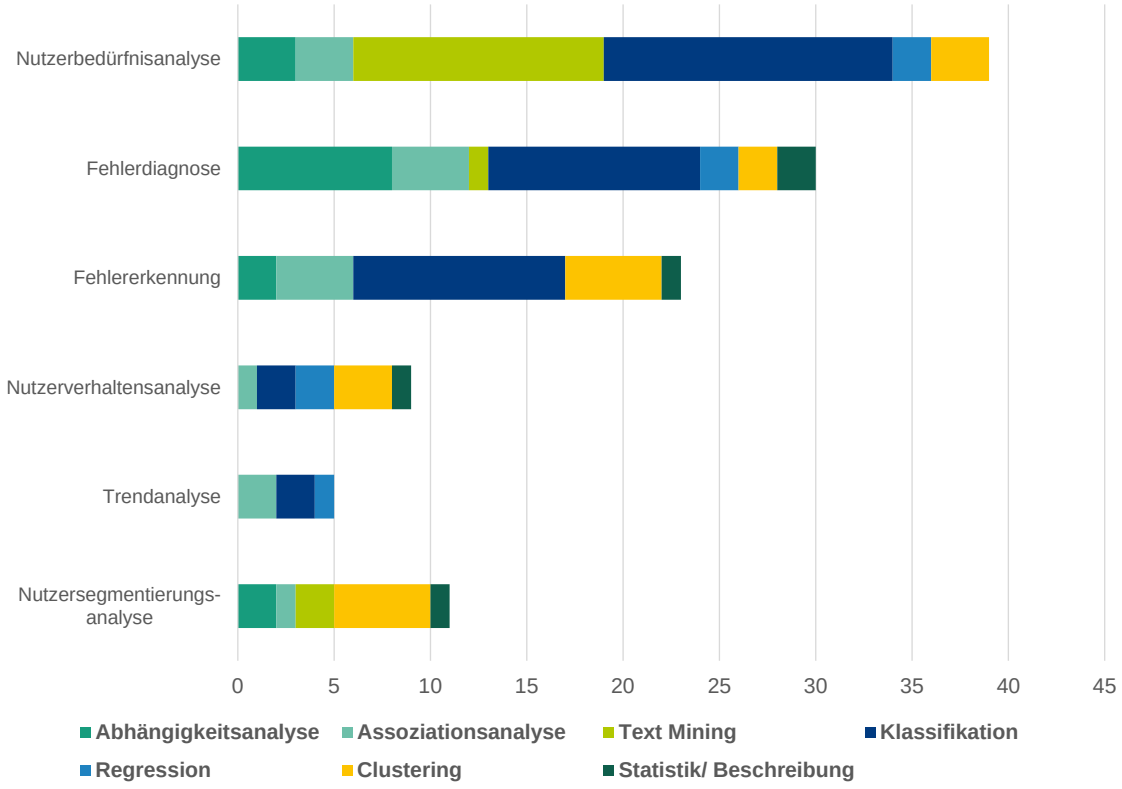

Bild 4-5: Data-Analytics-Anwendungen und Analytics-Probleme in der Literatur

4.3.2 Datenverständnis

Die Problemanalyse deckt auf, dass zum Aufbau von Datenverständnis zunächst die domänenseitige Auseinandersetzung mit den Datenquellen als Voraussetzung für die erfolgreiche Sammlung wichtig ist, bevor ihre Daten auf die analyserelevanten Eigenschaften geprüft werden können. Aus diesem Grund wurden als Bestandteile des Baukastens im ersten Schritt typische Betriebsdaten der Produktplanung identifiziert und im zweiten Schritt ein Beschreibungssystem für diese entwickelt. Diese bieten eine Orientierung für die Datenidentifikation und -beschreibung. Zum Beispiel können Fehlermeldungen als Statusdaten mit Hilfe des Beschreibungssystems anhand einer Auswahl relevanter Merkmale wie Datensatzgruppe, Dimensionalität und Qualität konkret beschrieben werden.

Zusätzlich bietet eine aggregierte Kombination von Merkmalen in Form von Clustern eine sehr schnelle Einordnung der identifizierten Daten. In Abschnitt 4.3.2.1 wird zunächst das methodische Vorgehen zur Entwicklung dieser Komponenten präsentiert, bevor Abschnitt 4.3.2.2 die Ergebnisse vorstellt.

4.3.2.1 Methodisches Vorgehen

Um im ersten Schritt den Blickwinkel aus der Produktplanung auf die Betriebsdaten zu erlangen, wird als Strukturierungsrahmen für die zu entwickelnde Übersicht über typische Betriebsdaten die Klassifizierung nach MEYER ET AL. (vgl. Abschnitt 3.3.2.3) genutzt [MPK+22]. Diese berücksichtigt u. a. Merkmale wie Erzeuger, Entstehung, Beschreibungsfokus und Zweck. Die Klassifikation, bestehend aus (1) Nutzungsdaten, (2) Nutzerverhaltensdaten, (3) Servicedaten, (4) Produktverhaltensdaten und (5) Statusdaten, wurde um typische Betriebsdatensatztypen ergänzt. Dies geschah mit Hilfe einer intensiven Analyse der Literatur, welche Beispiele aufführt (vgl. Literatur aus Abschnitt 2.3.3) sowie mehreren Forschungspartnerterminen im Rahmen des Forschungsprojekts *DizRuPt – Datengestützte Generationen- und Retrofitplanung*. An diesen Terminen nahmen bis zu sechs Experten aus Forschung und Industrie für die Bereiche Produktplanung, Product-Lifecycle-Management und Data Analytics teil.

Bild 4-6 liefert einen Überblick über das methodische Vorgehen zur Entwicklung des Beschreibungssystems für Betriebsdaten auf Basis der zuvor entwickelten Betriebsdatenübersicht.

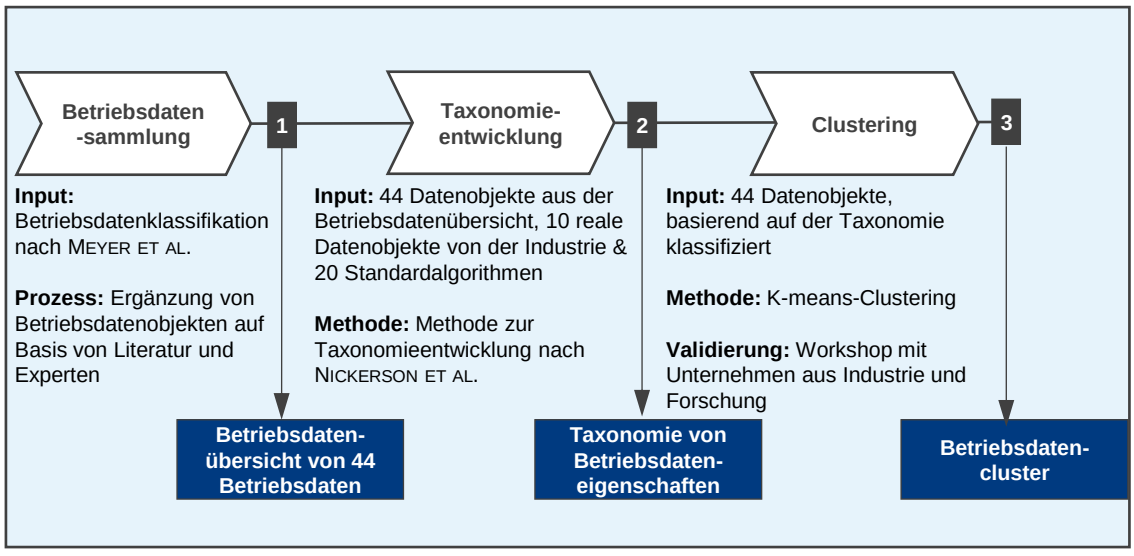

Bild 4-6: Methodisches Vorgehen für das Betriebsdatenbeschreibungssystem

Wie in Abschnitt 2.3.3 und 2.3.4 beschrieben, können Betriebsdaten eine Vielzahl an verschiedenen Eigenschaften aufweisen und all diese möglichen Eigenschaften beeinflussen die Methodenauswahl für die Vorverarbeitung und Modellierung. Daher war das methodische Vorgehen von der Forschungsfrage „Was sind die wesentlichen Eigenschaften von

Betriebsdaten, die für die Datenanalyse relevant sind?“ getrieben. Zur Beantwortung dieser Frage wurde die Methode zur Entwicklung einer Taxonomie nach NICKERSON ET AL. eingesetzt [NVM13]. Taxonomien werden oft synonym mit Begriffen wie Rahmen, Typologie oder Klassifikation verwendet und sind empirisch und/oder konzeptionell abgeleitete Gruppierungen in Bezug auf Dimensionen und Merkmale [PRB20]. Die Methode von NICKERSON ET AL. umfasst folgende Schritte: (1) Bestimmung eines Meta-Merkmals, (2) Bestimmung objektiver und subjektiver Endbedingungen und (3) die iterative Wahl des Ansatzes, bis alle Endbedingungen erfüllt sind. Für die eigentliche Taxonomie-Entwicklung unterscheiden NICKERSON ET AL. zwischen dem empirisch-konzeptionellen und dem konzeptionell-empirischen Ansatz. Beim empirisch-konzeptionellen Ansatz werden reale Objekte ausgewählt, Merkmale abgeleitet, mit konzeptionellen Bezeichnungen versehen und den Dimensionen zugeordnet. Beim konzeptionell-empirischen Ansatz schlagen die Forscher zunächst Dimensionen und Merkmale vor, bevor die Dimensionen und Merkmale durch Klassifizierung von Objekten untersucht werden. Dies führt zu einer ersten oder überarbeiteten Taxonomie. Bild 4-7 fasst die Methode zusammen.

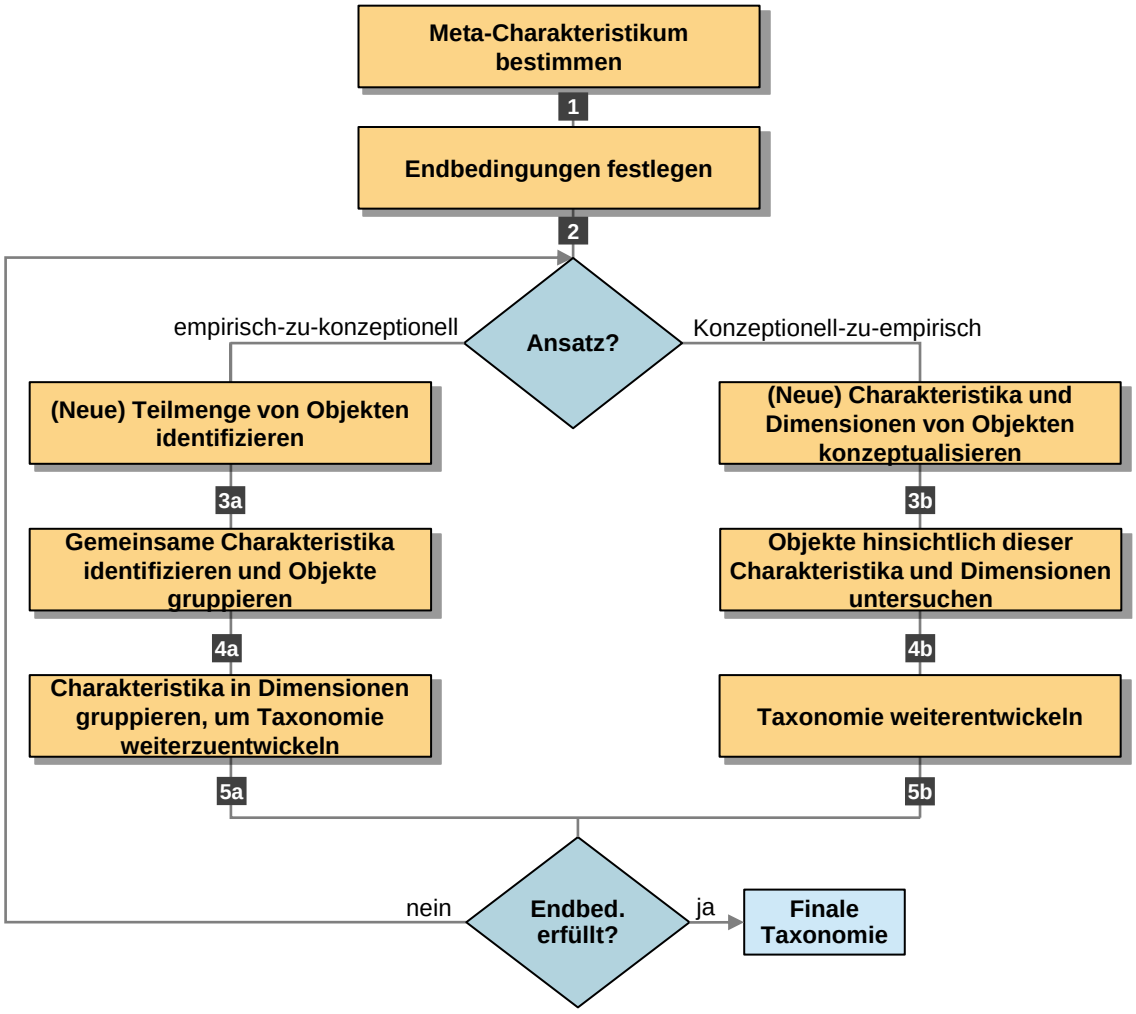

Bild 4-7: Methode zur Taxonomie-Entwicklung nach NICKERSON ET AL. [NVM13]

Im Einklang mit der Forschungsfrage stellten die analytisch relevanten Merkmale von Betriebsdaten das Meta-Charakteristikum dar. Dabei sollten allgemeine Merkmale, die für die gleichen Datensatztypen ähnlich ausgeprägt sein können, und individuelle Merkmale, die unternehmens- und infrastrukturabhängig sind, unterschieden werden. Im zweiten Schritt wurden die von NICKERSON ET AL. vorgeschlagenen objektiven Endbedingungen übernommen: jedes Merkmal ist in seiner Dimension eindeutig, jede Dimension ist eindeutig und wird nicht wiederholt, unter jedem Merkmal jeder Dimension ist mindestens ein Objekt klassifiziert und in der letzten Iteration sind keine neuen Dimensionen oder Merkmale hinzugekommen. Subjektiv sollte die Methode beendet werden, wenn die Taxonomie von allen beteiligten Forschern als prägnant, robust, umfassend, erweiterbar und erklärend eingestuft wird. Im dritten Schritt wurde in der ersten Iteration der konzeptionell-empirische Ansatz gewählt, um Dimensionen und Merkmale auf der Grundlage von Standardliteratur und Expertenwissen der beteiligten Forscher zu konzeptualisieren. Als Ausgangspunkt dienten die geläufigen Big-Data-Merkmale (vgl. Abschnitt 2.3.3), die im Hinblick auf das Metamerkmal gefiltert wurden. Zur Bewertung der initialen Taxonomie wurden die ersten 22 Datensatztypen aus der Betriebsdatenübersicht genutzt. In den nächsten Iterationen fand der empirisch-konzeptionelle Ansatz Anwendung. Zusammenfassend wurden weitere 22 Datenobjekte aus der Übersicht verwendet, um neue Merkmale oder andere Konstellationen abzuleiten sowie zehn reale Betriebsdatensätze aus der Industrie, um die einzelnen Dimensionen und Merkmale zu hinterfragen. Um die Perspektive der analytischen Seite noch besser abzudecken, wurden in der letzten Iteration Beschreibungen von 20 Algorithmen aus der Literatur verwendet, um herauszufinden, ob die Taxonomie final ist. Die Einzelheiten aller Iterationen sind der entsprechenden Veröffentlichung zu entnehmen [PEM+22].

Anschließend bestand das Ziel darin, Betriebsdaten mit ähnlichen allgemeinen Merkmalen zu identifizieren und potenzielle Merkmalskombinationen auf einen Satz üblicher Kombinationen zu begrenzen. Dies sollte dazu dienen, den Beschreibungsrahmen zu vereinfachen und durch eine solche Abstraktion eine zügigere Identifikation von Daten und ihren Eigenschaften zu ermöglichen. Verwendet wurden dazu die klassifizierten Datensätze der Übersicht gemäß der Taxonomie. Die dortigen Zuordnungen wurden wiederum mit einem Datenexperten mit langer Berufserfahrung in Forschung und Industrie hinterfragt, der häufig mit Nutzungsdaten arbeitet und daher deren Merkmale gut kennt. Daraus resultierte eine Matrix, die als Input für das automatische Clustering diente. Für das Clustering wurde ein prototypenbasierter Algorithmus, der populäre und weit verbreitete k-means-Algorithmus, verwendet. Dieser bestimmt einen Prototyp für jedes Cluster und bildet Cluster, indem er Datenobjekte dem nächstgelegenen Clusterprototyp zuordnet [Wu12]. Zur Bestimmung der optimalen Anzahl von Clustern k diente die grafische "Ellbogen"-Methode. Dies führte zu fünf Clustern. Die Interpretation dieser ergab, dass jedes Cluster sowohl für sich als auch im Verhältnis zu den anderen Clustern sinnvoll interpretiert werden konnte.

4.3.2.2 Ergebnisse

Bild 4-8 zeigt einen Ausschnitt der Betriebsdatenübersicht. Sie listet insgesamt 44 typische Daten der Betriebsphase entlang der fünf Kategorien nach MEYER ET AL. auf und befindet sich vollständig im Anhang A1. Es handelt sich dabei um Grundgrößen, welche durch zusätzliche Informationen (Variablen) wie die Zeit oder das System einen Datensatz bilden können. Auffällig ist, dass die meisten Daten den Statusdaten zuzuordnen sind. Darüber hinaus sind die Produktverhaltensdaten auch durch mehrere mögliche Datenobjekte vertreten.

Statusdaten	Produkt-verhaltensdaten	Nutzungsdaten	Nutzerverhaltens-daten	Servicedaten
• eingebaute physikalische Elemente (komplexes System, geringe Variationen) • eingebaute physikalische Elemente (einfaches	• Aktordaten für einen einzelnen Akteur (selektiv gesteuert) • Stellglieddaten für einen einzelnen Akteur (kontinuierlich gesteuert) • Aktordaten für	• Bestellung und Auftrag • Aggregierte Benutzer-aktivitäten (z. B. Nutzung von Funktionen)	• Benutzer-aktivitätsprotokoll/ Protokoll • Nutzungsprozess/ Interaktionspfad • Aktivitätsdaten über Benutzerschnitt-stellen • persönliche Mitarbeiterdaten utzeranmeldun	• Serviceberichte (automatisiert) • Reparaturprotokoll • Wartungsprotokoll • Garantiefall • Kunden-beschwerden • Kunden-rezensionen/ Bewertungen • Anregungen von Kunden

Bild 4-8: Betriebsdatenübersicht (Ausschnitt)

Die entwickelte Taxonomie besteht aus zwei Komponenten: einer Klassifikation für allgemeine Betriebsdatenmerkmale und einer weiteren für individuelle Merkmale. Im Gegensatz zu den allgemeinen Merkmalen wird bei den individuellen Merkmalen angenommen, dass ihre Ausprägungen für die verschiedenen Betriebsdatensätze von Unternehmen zu Unternehmen stark variieren. Bild 4-9 und Bild 4-10 zeigen die Komponenten der Taxonomie und mögliche Indikatoren zur leichteren Klassifizierung von Datenobjekten. Die vollständigen Taxonomien befinden sich im Anhang A2. Im Folgenden werden alle Dimensionen und Merkmale näher beschrieben.

<u>Allgemeine Dimensionen und Merkmale:</u>

- **Datensatzgruppe**: In dieser Dimension werden die Daten hinsichtlich ihrer Vielfalt, d.h. der Art des Datensatzes und des Strukturierungsgrades untersucht. Merkmale der ersten Ebene sind tabellarische Daten (strukturiert) und Textdaten. Diese können weiter untergliedert werden. Bei Datensatzdaten wird ein Datensatz als eine Sammlung von Datensätzen mit einem festen Satz von Datenfeldern (Variablen) angenommen. Üblich sind Tabellen- oder Matrixform. Im Allgemeinen gibt es keine explizite Beziehung zwischen den Datensätzen. Jeder Datensatz hat denselben Satz von Variablen. Bei Graphen basierten Daten handelt es sich um Daten mit Beziehungen zwischen Objekten oder um Daten mit Objekten, die Graphen sind (wenn Objekte Unterobjekte enthalten, die Beziehungen haben). Bei

geordneten Daten haben die Attribute Beziehungen mit einer zeitlichen oder räumlichen Reihenfolge. Geordnete Daten können weiter gruppiert werden in sequenzielle Transaktionsdaten (jede Transaktion ist mit einem Zeitpunkt verbunden), Sequenzdaten (Datensatz, der eine Folge von einzelnen Entitäten ist - Positionen anstelle von Zeitstempeln), Zeitreihendaten (jeder Datensatz ist eine Reihe von Messungen, die im Laufe der Zeit vorgenommen wurden) mit Signalen und ohne Signale und räumliche Daten (räumliche Attribute, wie Positionen oder Gebiete). Bei Textdaten kann zwischen strukturierten und halbstrukturierten Textdaten unterschieden werden. Bild-, Audio- und grafikbasierte Daten sind ausgegraut, da das Verfahren in Abschnitt 4.2.2.1 gezeigt hat, dass sie als Merkmal für Nutzungsdaten nicht relevant sind. Da diese Formate in Zukunft aber durchaus eine größere Rolle spielen können, werden sie ebenfalls aufgeführt.

- **Dimensionalität**: Die Dimensionalität ist ein weiterer wichtiger Faktor, der bei der Auswahl einer geeigneten Analysetechnik eine entscheidende Rolle spielen kann. Beispielsweise führen zu viele Dimensionen dazu, dass jede Beobachtung in einem Datensatz äquidistant zu allen anderen erscheint (Fluch der Dimensionalität), was ein großes Problem für Clustering-Algorithmen darstellt. Daher sind die Merkmale klein- und hochdimensional.

- **Spärlichkeit und Dichte:** Einige allgemeine Aspekte von Verteilungen haben oft einen starken Einfluss, was die Modellierung erschweren kann. Spärlichkeit (eng. Sparseness) ist ein solcher Sonderfall, bei dem die meisten Attribute eines Objekts den Wert 0 haben. Einige Data-Mining-Algorithmen, wie z. B. die Algorithmen für das Assoziationsregel-Mining, funktionieren nur bei spärlichen Daten gut [TSK16]. Andererseits funktionieren einige Algorithmen, wie z. B. Random Forests, am besten bei dichten Daten.

- **Komplexität:** Die Komplexität von Daten kann z. B. durch (Auto-)Korrelation ausgedrückt werden, was wichtig zu wissen ist, da z. B. eine der Annahmen der Regressionsanalyse darin besteht, dass die Daten keine Autokorrelation aufweisen. Daher müssen möglicherweise andere Methoden verwendet werden. Korrelation und Multikollinearität in den Daten können sich ebenfalls auf die Leistung des Modells auswirken. Algorithmen, wie die logistische oder die lineare Regression, sind in diesem Fall nicht gut geeignet, so dass dies vor dem Training behoben werden sollte.

- **Echtzeitverhalten** (Velocity): In der Datenanalyse oder beim maschinellen Lernen kann Echtzeit- oder Online-ML (Training eines Modells, indem Live-Daten durchlaufen werden, um das Modell kontinuierlich zu verbessern) von traditionellem Training unterschieden werden, bei dem ein Stapel historischer Daten verwendet wird. Ersteres erfordert andere Verfahren als letzteres.

- **Volumen**: Hinsichtlich des Umfangs kann ein Datenobjekt oder ein Datensatz klein, mittelgroß oder groß sein. Um dies zu beurteilen, ist die Menge der pro Tag

erzeugten Daten sicherlich wichtig. Das Volumen wirkt sich insofern auf die Analyse aus, als einige Methoden besser mit wenigen Trainingsstichproben umgehen können, z. B. Support Vector Machines, oder einige Algorithmen besser für die Verarbeitung großer Datenmengen geeignet sind.

	Dimension	Ausprägungen				Indikator (beispielhaft)
Datenmerkmale (allgemein)	Datensatz-gruppe (Typ)	Tabellarische Daten (strukturiert)	Relationale Daten (z. B. Tabelle, Datenmatrix)			Feste Anzahl an Datenfeldern, keine Verknüpfung zwischen Einträgen
			Graphenbasierte Daten			Verbindungen, Objektbeziehungen
			Geordnete Daten (zeitliche oder räumliche Ordnung)	Sequentielle Transaktionsdaten		Itemsets + Zeitangabe (keine bestimmte Frequenz)
				Sequenzdaten		Geordnet ohne Zeitstempel
				Zeitreihendaten	Signal	Messbare physikalische Parameter
					Kein Signal	Messungen über die Zeit im Allgemeinen
				Räumliche Daten		Positionen, Bereiche
		Text	Unstrukturiert			Fehlendes Format (keine Trennungen der Informationen)
			Halbstrukturiert			Tags, Metadaten
		Bild				Metadaten
		Video				Metadaten
	Dimensionalität	Niedrigdimensional				Deutlich weniger Features als Beobachtungen
		Hochdimensional				Mehr Features als Beobachtungen, > 100 Features
		Spärlich				Viele Lücken in den Daten,

Bild 4-9: Ausschnitt der Taxonomie für allgemeine Datencharakteristika

Individuelle Dimensionen und Merkmale:

Die individuellen Merkmale sind vor allem für die Auswahl der richtigen Vorverarbeitungstechniken wichtig, spielen aber auch eine Rolle bei den Modellierungsalgorithmen (vgl. Abschnitt 2.3.4).

- **Datenqualitätsprobleme**: Die Datenqualität hat einen großen Einfluss auf die Datenanalyse, z. B. sind einige Verfahren toleranter gegenüber fehlenden Werten, Ausreißern und ungewöhnlichen Datenverteilungen. Einige Datenvorverarbeitungsverfahren (z. B. Ausreißereliminierung, Normalisierung, Phasierung, Datenreduktion usw.) können erforderlich sein, um die Qualitätsprobleme anzugehen und die Daten für die Modellierung anzupassen. Merkmale sind zufälliges Rauschen, systematische Fehler, Ausreißer, Inkonsistenz, fehlende Werte und doppelte Daten.
- **Variablentyp**: Um einzelne Datenobjekte zu beschreiben, eignet sich der Variablentyp. Grundsätzlich wird hier zwischen kategorialen (qualitativen) und numerischen (quantitativen) Attributen unterschieden. Qualitativen Attributen fehlen die meisten Eigenschaften von Zahlen und sollten eher wie Symbole behandelt werden. Auch hier können nominale und ordinale Typen unterschieden werden. Quantitative Attribute werden durch Zahlen dargestellt und haben die meisten Eigenschaften von Zahlen. Binäre und Datumsvariablen können sowohl kategorisch als auch numerisch sein und sind Untermerkmale der Sonder-/Hybridform.

Zur Bewertung oder Bestimmung der Qualitätsmerkmale und zur besseren Einschätzung von Vorverarbeitungsmaßnahmen wird die Verwendung einer dreistufigen Skala "zu vernachlässigen", "zu berücksichtigen" und "dominant" vorgeschlagen. Qualitätseinschränkungen im Zusammenhang mit der Messqualität können vernachlässigbar sein, z. B., wenn der Datensatz konstante systematische Fehler enthält, aber nur die Relationen von Interesse sind, oder wenn zufällige Fehler vorhanden sind, die selten genug sind, um eine Auswirkung zu haben. Systematische Fehler, die korrigiert werden können, sind im Rahmen der Vorverarbeitung zu berücksichtigen. "Dominant" sind untragbare Sensorausfälle oder zufällige Fehler, die die Daten dominieren. Bei der abschließenden Bewertung muss auch der Anwendungsfall berücksichtigt werden.

	Dimension	Ausprägungen	Bewertung	Indikator (beispielhaft)
Datenmerkmale (individuell)	Daten-qualitäts-probleme	Systematische Fehler (Datenfehler z. B. Miskalibrierung verursacht)	Zu vernachlässigen	Konstante systematische Fehler (Lieferantenspezifikation oder Domänenwissen), wo nur Verhältnisse interessant sind
			Zu berücksichtigen	Systematische Fehler, die korrigiert werden können
			Dominant	untolerierbare Sensorfehler
		Zufälliges Rauschen (Rauschverteilung folgt keinem bekanntem Modell)	Zu vernachlässigen	Zufällige Fehler, die sehr selten sind
			Zu berücksichtigen	Zufällige Fehler
			Dominant	Zufällige Fehler, die die Daten dominieren
		Ausreißer	Zu vernachlässigen	Keine oder isolierte Ausreißer
			Zu berücksichtigen	Ausreißer können klar identifiziert werden
			Dominant	Ausreißer dominieren
		Inkonsistenz (Frequenzen, Einheit, Wertebereich)	Zu vernachlässigen	Frequenzen und Einheiten sind einheitlich
			Zu berücksichtigen	Variablen haben signifikant verschiedene Wertebereiche, Frequenzen variieren
			Dominant	Informationen zu Einheiten und Wertebereichen fehlen
		Fehlende Werte	Zu vernachlässigen	Keine oder nur gelegentlich fehlende Daten
			Zu berücksichtigen	Fehlende Da…

Bild 4-10: Ausschnitt der Taxonomie für individuelle Datencharakteristika

Das Ergebnis der anschließenden Clusteranalyse sind fünf Cluster bzw. Kategorien, die Kombinationen von allgemeinen Betriebsdatenmerkmalen abdecken, die typischerweise zusammen auftreten. Die Cluster sind in Bild 4-11 dargestellt, in der die häufigsten Merkmale pro Dimension hervorgehoben werden. Die Namen der Cluster sind durch die am stärksten ausgeprägten Merkmale geprägt.

	Cluster				
	Sequenzielle dünnbesetzte Echtzeitdaten	**Stark strukturierte historische Daten**	**Gemischt-strukturierte, hochdimensionale Echtzeitdaten**	**Echtzeit-Zeitreihendaten**	**Text Daten**
Beispiele	Sensordaten, Steuersignale für einzelne Aktoren, Hardware-Zustände, Software-Status, Warnmeldungen	Hardware-Konfigurationen, Werks-einstellungen, Warnungen, Beschwerden, Bewertungen	Aktordaten, Aktivitätsdaten	Vibrationsdaten, Hardwarestand, Softwarezustand, Produktionsmenge, Arbeitslast, Laufzeit, Energieverbrauch	Lizenzen, Protokolle (Wartung, Instandhaltung)
Datensatz-gruppe	Zeitreihen, sequenzielle Transaktionsdaten	Strukturiert (relationales Datenbankschema)	Gemischt	Zeitserien (Signale)	Semi-strukturierter bis unstrukturierter Text
Echt-zeitver-halten	Echtzeit	Nicht Echtzeit	Echtzeit	Echtzeit	Nicht Echtzeit
Menge	Kleine bis mittlere Datenmenge	Kleine Datenmengen	Mittlere bis große Datenmengen	Kleine Datenmengen	Kleine Datenmengen
Kom-plexi-tät	Unkorreliert	Unkorreliert	Unkorreliert	(Auto-)korreliert	Unkorreliert
Dimen-siona-lität	Niedrigdimensional	Niedrigdimensional	Hochdimensional	Niedrigdimensional	Niedrigdimensional
Vertei-lung	Dünnbesetzt	Dünnbesetzt	Dicht	Dicht	Dünnbesetzt

Bild 4-11: Betriebsdatencluster

- Cluster 1 - Sequenzielle, spärliche Echtzeitdaten: Dieses Cluster ist durch die Datengruppe der geordneten Daten gekennzeichnet, spezifischer durch Zeitreihen und sequenzielle Transaktionsdaten. Die Datenobjekte in diesem Cluster werden in den meisten Fällen in Echtzeit generiert, die Datengröße ist klein bis mittelgroß und es besteht meist keine offensichtliche Korrelation. Außerdem ist der Cluster durch geringe Dimensionalität und Sparsamkeit gekennzeichnet. Sensordaten, teilweise auch Aktordaten, Hardware- und Softwarezustände sowie Warn- und Fehlermeldungen lassen sich hier einordnen.

- Cluster 2 - hochstrukturierte, historische Daten: Dieses Cluster enthält überwiegend strukturierte Daten, die in relationalen Datenbanken gespeichert werden können. Die Datenmengen sind eher klein, auch weil die Daten eher spärlich sind. Beispiele für Datenobjekte sind Hardwarekonfigurationen, Werkseinstellungen, Warnmeldungen, Bewertungen und Anmeldedaten.

- Cluster 3 - Gemischt-strukturierte, hochdimensionale Echtzeitdaten: Diese Kategorie umfasst Daten aus verschiedenen Datensatzgruppen. Objekte, die hier eingeordnet werden können, haben oft ein halbstrukturiertes Format, können aber genauso gut sequentielle oder strukturierte Daten sein. Weitere Merkmale dieser

Daten sind ihr Echtzeitverhalten und eine mittlere bis große Datengröße. Oft sind sie auch hochdimensional und dicht.

- Cluster 4 - Echtzeit-Zeitreihendaten: Echtzeit-Zeitreihendaten sind durch ein Zeitreihenformat oder sogar Signalcharakteristika gekennzeichnet. Sie werden in Echtzeit erzeugt und sind meist kleine Daten. Signaldaten, wie z. B. Vibrationen, treten dagegen häufig in großen Datensätzen auf. Da der Schwerpunkt auf Zeitreihen liegt, zeichnen sie sich häufig durch Autokorrelation aus, sind aber eher niedrigdimensional und dicht. Vibrationsdaten, Hardware- und Softwarezustände, Laufzeit und Energieverbrauch können in dieses Cluster eingeordnet werden.
- Cluster 5 – Textdaten: Das letzte Cluster ist durch ein unstrukturiertes oder strukturiertes Textformat gekennzeichnet. Die Menge der Daten ist eher gering. Die Sparsamkeit ist durch das Format gegeben. Beispiele sind Lizenzen und verschiedene Protokolle.

4.3.3 Vorverarbeitung und Modellierung

Die Problemanalyse hat gezeigt, dass die Bereitstellung von relevanten Techniken für die Vorverarbeitung und Modellierung einen Mehrwert liefern kann, um den riesigen Lösungsraum für die Produktplanung einzugrenzen. Daher wurden für den Baukasten wichtige Vorverarbeitungstechniken, Modelle, bzw. Algorithmen und Evaluierungsmetriken identifiziert. Das Vorgehen basierte dabei auf der SLR (Abschnitt 4.3.1.1). Darüber hinaus wurde eine Umfrage mit Data Scientists durchgeführt, um den Blickwinkel der Praxis einzufangen.

4.3.3.1 Methodisches Vorgehen

Das Vorgehen setzt auf der in Abschnitt 4.3.1.1 beschriebenen systematischen Literaturrecherche auf und nutzt den dort generierten Datensatz. Es wurden folgende Forschungsfragen ergänzt:

RQ2: Welche Vorverarbeitungsmethoden (Bereinigung, Transformation, Feature Engineering) werden für Betriebsdaten eingesetzt?

Da die in der Betriebsphase von Produkten erzeugten und für die Produktplanung verwendeten Daten sehr heterogen und teilweise von schlechter Qualität sind, ist eine Vorverarbeitung unerlässlich. Anhand dieser Forschungsfrage wird untersucht, welche Techniken für die jeweiligen Anwendungsfälle und die dazugehörigen Algorithmen wichtig sind.

RQ3: Welche Algorithmen werden für die Modellierung verwendet?

Im Zentrum der Pipeline steht der Algorithmus, der die vorverarbeiteten Daten als Input nimmt und je nach Problemstellung z. B. Clusterzuordnungen, Klassen, etc. ausgibt. Wie

bereits erläutert, ist die Auswahl geeigneter Algorithmen nicht trivial. Es gibt eine große Anzahl unterschiedlicher Möglichkeiten, hinzu kommen die zu berücksichtigenden Faktoren (Ziel und Daten). Um den Lösungsraum einzugrenzen und die Auswahl zu vereinfachen, soll analysiert werden, welche Modelle in der Literatur zunehmend zur Lösung von Produktplanungsproblemen verwendet werden.

RQ4: Welche Evaluierungsmetriken werden verwendet?

Die Bewertung der Modelle ist wichtig, um die Ergebnisse zu beurteilen. Da in der Produktplanung sowohl überwachte als auch unüberwachte Methoden verwendet werden, ist die Zahl der möglichen Bewertungsmaßstäbe groß. Welche davon für Produktplanungsmodelle und -probleme relevant sind, wird durch diese Frage beantwortet.

Ähnlich wie bei den Anwendungen wurden alle Vorverarbeitungstechniken, Modelle und Evaluierungsmetriken aus den Veröffentlichungen extrahiert. Nach dem Vier-Augen-Prinzip wurden alle eindeutigen Verfahren identifiziert und ggf. zur Neukodierung verwendet. Im Anschluss wurde auch hier eine quantitative Auswertung durchgeführt. Da sich herausstellte, dass insgesamt nur wenige Vorverarbeitungsmethoden, bzw. viele Einzelnennungen, extrahiert werden konnten, wurde in dem Fall von einer solchen Auswertung abgesehen. Alle genannten Vorverarbeitungsmethoden wurden jedoch mittels einer thematischen Analyse nach BRAUN und CLARKE [BC06] geclustert und benannt, um so größere Kategorien, wie z. B. Skalierung und Normalisierung zu erhalten. Diese wurden dann zur Strukturierung der Vorverarbeitungsmethoden für die anschließende Umfrage und den Baukasten genutzt.

Um die Ergebnisse der SLR aus praktischer Sicht zu bewerten und mit praxisrelevanten Verfahren anzureichern, wurde anschließend eine Umfrage nach dem Leitfaden von LINAKER ET AL. [LSH+15] aufgesetzt und durchgeführt. Sie kann wie folgt zusammengefasst werden:

- Zielsetzung: Ziel der Umfrage war es, (1) zu evaluieren, ob die identifizierten Vorverarbeitungs- und Modellierungstechniken sowie -metriken aus der wissenschaftlichen Literatur auch in der Praxis angewandt werden, und (2) weitere relevante, in der Praxis verwendete Techniken zu identifizieren.
- Zielpublikum/Population: Zielgruppe waren Data-Science-Professionals in Deutschland, die in letzter Zeit in Industrieprojekten mit Betriebsdaten oder operativen Daten wie Sensor- und Log-Daten, gearbeitet haben.
- Stichprobendesign: die Teilnehmer wurden über das Netzwerk der Autoren rekrutiert, da die Bereitschaft zur Teilnahme durch die persönliche Verbindung erhöht wird.
- Gestaltung des Fragebogens: Die Umfrage bestand aus insgesamt 19 Fragen, darunter 2 einleitende Fragen zum Unternehmen und zu Erfahrungen mit der Analyse von Betriebsdaten. Es folgten mehrere Fragen zu Modellen in der Form:

„Welche der folgenden Modelle verwenden Sie regelmäßig?“ – jeweils zu (1) Modellen, (2) Vorverarbeitungsverfahren und (3) Metriken, strukturiert nach (1) dem Analyseproblem, (2) Datenqualitätsproblemen und Datentyp und (3) Lernform (unüberwacht vs. überwacht). Der Fragebogen wurde als selbstverwaltete, webbasierte Variante konzipiert, da er einfach zu verwalten ist und ein Einfluss des Forschers verhindert wird. Das Problem der geringen Rücklaufquote wurde durch entsprechende Einführungen und einen Testlauf mit externen Kollegen versucht zu minimieren.

- Analyse der Umfragedaten: Da es sich bei den Antworten um nominale Daten handelte, wurde eine Häufigkeitsanalyse der Nennungen durchgeführt.

4.3.3.2 Ergebnisse

Im Kontext der **Datenvorverarbeitung** (RQ2) werden Ausreißer in der Praxis regelmäßig und von den meisten Data Scientists mit Mittelwerten und Standardabweichungen sowie Grenzwerten (Boxplots) behandelt. Skalierung und Normalisierung ist auch ein Thema, wenn es um die Vorverarbeitung von Betriebsdaten geht. Der Min-Max-Skalierer scheint hier beliebt zu sein. Systematische Fehler werden eher selten angesprochen. Um die Daten in das richtige Format zu bringen, spielen One-Hot-Encoder eine große Rolle. Um Merkmale aus Zeitreihen zu extrahieren, werden die Fast-Fourier-Transformation und Zeitfenster am häufigsten verwendet. Bild 4-12 stellt die absoluten Häufigkeiten der Vorverarbeitungsverfahren in einer Treemap dar.

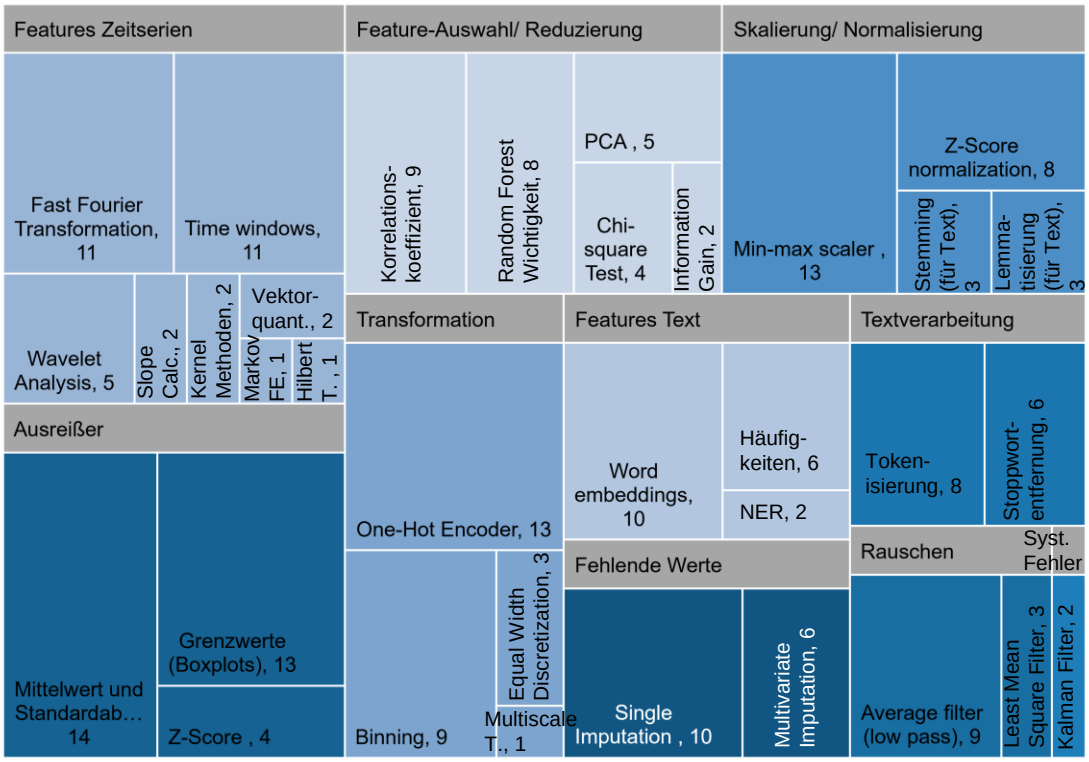

Bild 4-12: Anzahl der in der Praxis eingesetzten Vorverarbeitungsverfahren

In Bild 4-13 werden die Häufigkeiten der in der Literatur vorgestellten **Algorithmen** (RQ3) dargestellt. In den meisten Beiträgen werden überwachte Klassifizierungsverfahren eingesetzt. Mit 26 Nennungen ist das Modell der Support Vector Machine (SVM) sehr präsent. Mit größerem Abstand dahinter sind neuronale Netze, Convolutional Neural Networks (CNN) und Entscheidungsbäume (DT) zu nennen. Regressionsmethoden werden insgesamt eher wenig verwendet.

Aber auch unüberwachte Methoden werden häufig eingesetzt, da sie noch unbekannte Muster und Zusammenhänge aufdecken. Hier dominieren Modelle zur Abhängigkeits- und Assoziationsanalyse. Beliebt in der Forschungs-Community sind Bayes'sche Netzwerke (BN) und Assoziationsregeln (AR) (Apriori). K-Means ist der wichtigste Clustering-Algorithmus.

Es gibt auch eine Reihe von Einzelerwähnungen, die aus Gründen der Übersichtlichkeit nicht in das Diagramm aufgenommen wurden:

- Abhängigkeitsanalyse: Bayessche Netzwerke (K2), Bayessche Netzwerke (PC), (partieller) Korrelationskoeffizient
- Clustering: DBSCAN, hierarchisches Clustering, Biclustering (BCBimax)
- Beschreibung/Statistik: parallele Koordinaten, Häufigkeitsanalyse
- Text Mining: LLM, Word2Vec, BERT, NER, LSA
- Unüberwachte Klassifizierung: One-Class-SVM
- Überwachte Klassifizierung: Conditional Random Fields

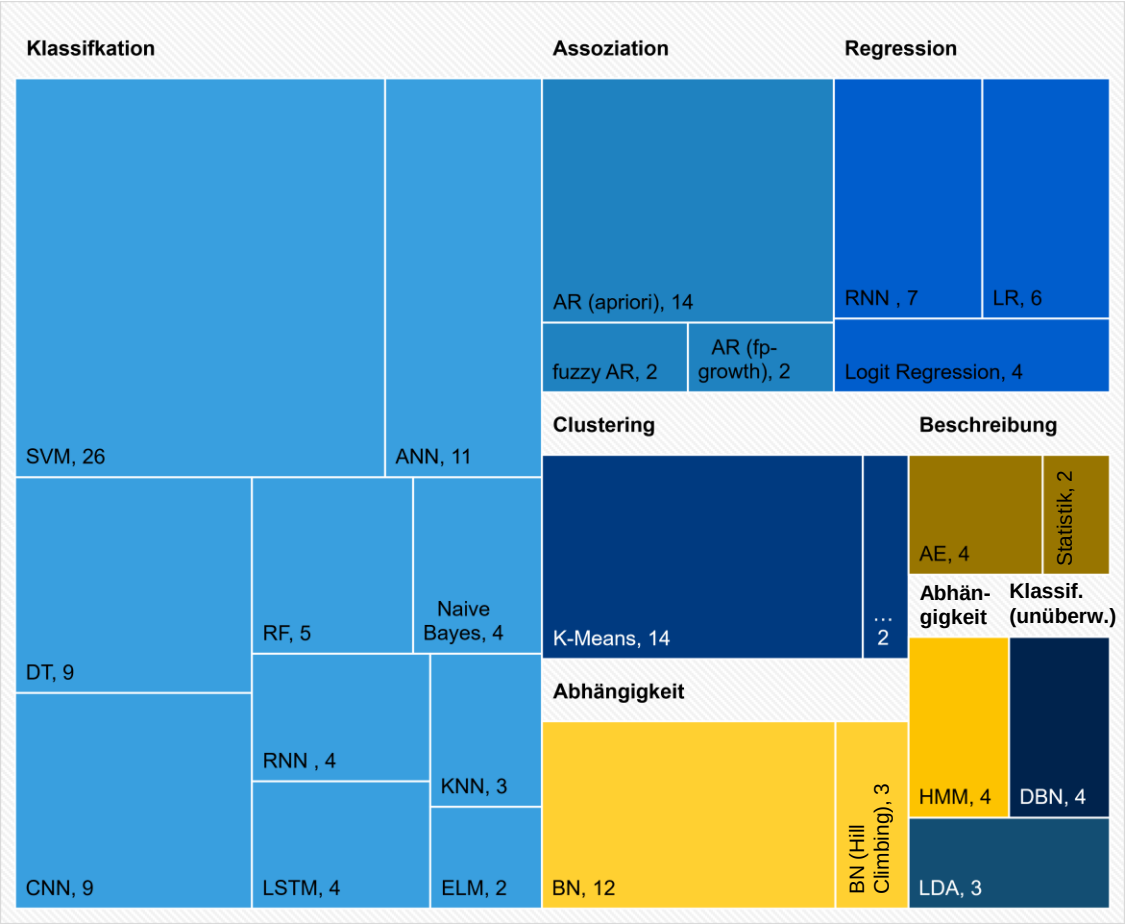

Bild 4-13: Überblick über in der Literatur eingesetzten Algorithmen im Kontext der Produktplanung

Ein Blick in die Praxis bestätigt die Beliebtheit von SVMs für die Klassifizierung (s. Bild 4-14). Ihre Verwendung ist jedoch fast gleichauf mit der von Random Forests und Entscheidungsbäumen. In der Praxis verwenden Datenwissenschaftler auch statistische Methoden und Regressionsmethoden mit ähnlicher Häufigkeit. An erster Stelle steht der K-means-Algorithmus, der mit 17 Nennungen der am häufigsten verwendete Algorithmus von allen ist. Die Abhängigkeitsanalyse und ihre Modelle werden dagegen nur von einem kleinen Teil der Data Scientists regelmäßig verwendet.

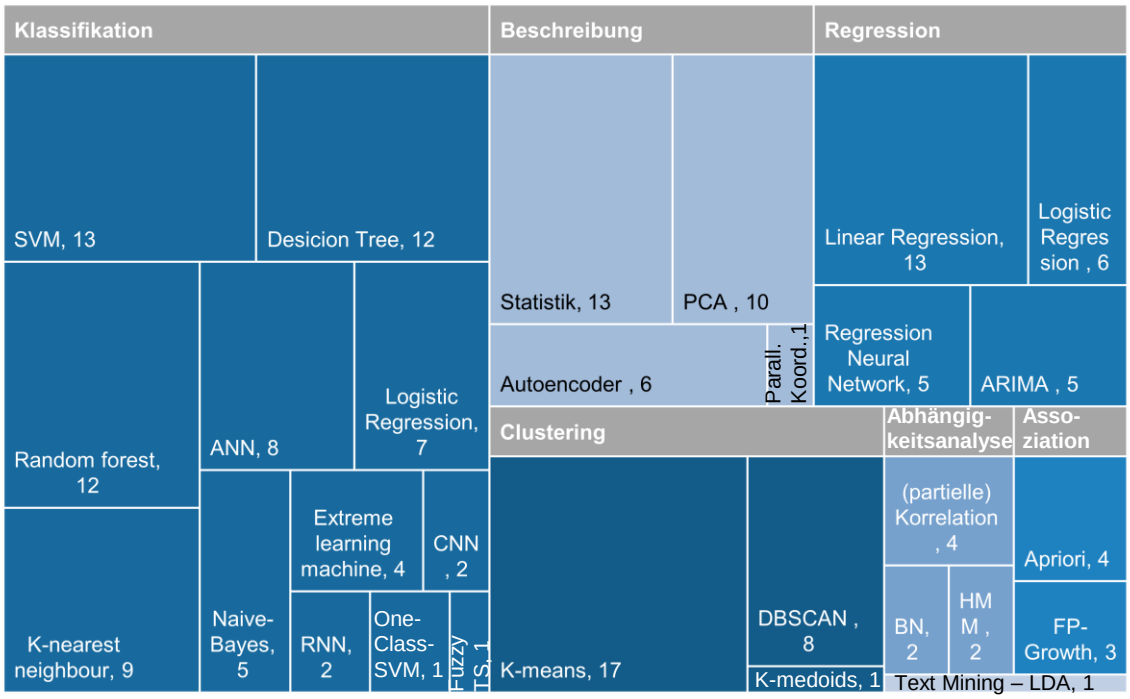

Bild 4-14: Anzahl der in der Praxis eingesetzten Algorithmen

Die meisten **Evaluierungsmetriken** (RQ4) können dem überwachten Lernen zugeordnet werden (s. Bild 4-15). Precision und Recall sind hier die am häufigsten verwendeten Metriken. Beim unüberwachten Lernen dominiert die externe Validierung, d.h. es wird ein Vergleich mit anderen (Experten-)Ergebnissen durchgeführt. Hierfür können wiederum Klassifikationsmetriken verwendet werden.

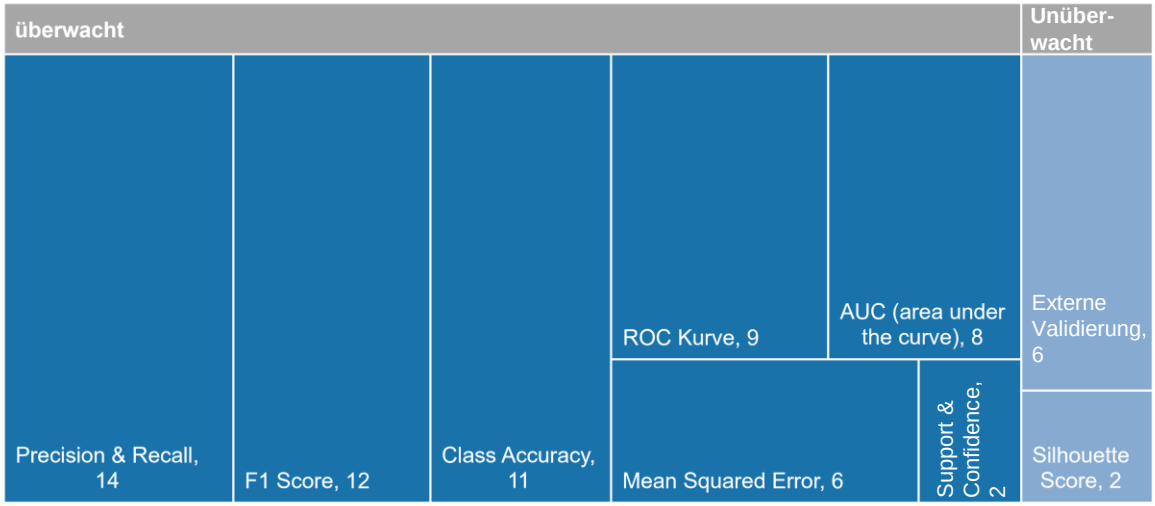

Bild 4-15: Anzahl der in der Praxis eingesetzten Evaluierungsmetriken

4.3.4 Zusammenfassung der Ergebnisse in einem Baukasten für die betriebsdatengestützte Produktplanung

Die Ergebnisse der Studien, die Komponenten zur Ausgestaltung des Referenzprozesses (vgl. Abschnitt 4.2), wurden in Form eines morphologischen Kastens strukturiert und zusammengefasst. Der Baukasten umfasst folglich fünf Dimensionen: Analytics-Anwendung, Datenverständnis, Vorverarbeitung, Modellierung und Evaluierung. Die Dimensionen Anwendungen, Datenverständnis und Vorverarbeitung unterteilen sich in die kleinteiligeren Aufgaben, welche in den Abschnitten 2.3.2 bis 2.3.4 hergeleitet und untersucht wurden. So untergliedert sich die Analytics Anwendung in das Analytics-Ziel und Analytics-Problem. Der Schritt des Datenverständnisses adressiert die Datensammlung durch die Betrachtung der Datenquellen und die Datenbeschreibung durch typische Kombinationen von Dateneigenschaften. Die Vorverarbeitung wird in die Aufgaben Bereinigung, Transformation und Feature Engineering, welches auch die Datenreduktion umfasst, aufgeteilt. Die Komponenten dieser Vorverarbeitungsaufgaben wurden wie bereits für den Fragebogen im Rahmen der Umfrage (vgl. Abschnitt 4.3.3.1) nach zuvor definierten Kategorien strukturiert. Sowohl die Problemanalyse als auch die aufgestellte Taxonomie für die individuellen Dateneigenschaften ermöglichen die Strukturierung der Datenvorverarbeitungsbausteine. Die Problemanalyse ergab ebenfalls die relevanten Abhängigkeiten zwischen Datenqualitätsproblemen aus dem Datenverständnis und der Datenbereinigung sowie zwischen dem Variablenformat und der Datentransformation (vgl. Abschnitt 2.3.4), sodass an dieser Stelle erste Zuordnungen stattfinden konnten. Komponenten des Feature Engineerings wurde zum einen in Merkmalsextraktionsansätze für die zwei relevanten und recht individuellen Datentypen Zeitreihen und Text und zum anderen in Ansätze zur Selektion, bzw. Reduktion unterteilt. Zur Strukturierung der Modelle dienten die in 4.3.1.2 ergänzten relevanten Analytics-Probleme für die Produktplanung. Der Baukasten ist in Bild 4-16 abgebildet. Insgesamt enthält der Baukasten 89 Komponenten, mit denen individuelle Pipelines für Use Cases der datengestützten Produktplanung ausgestaltet werden können. Der Übersicht halber wurden lediglich Komponenten mit Mehrfachnennung aufgenommen. Dabei ist zu betonen, dass es sich um eine initiale Wissensbasis handelt, die mit neuen gemachten Erfahrungen aus Forschung und Praxis erweitert werden soll, um eine hilfreiche Übersicht über relevante Komponenten zur Ausgestaltung einer Pipeline für die betriebsdatengestützte Produktplanung zu bieten. Wie eine individuelle Pipeline zusammengestellt werden kann, d. h. passende Komponenten für jede Dimension bestimmt, bzw. ausgewählt, werden können, wird im nächsten Abschnitt, im Vorgehen, beschrieben.

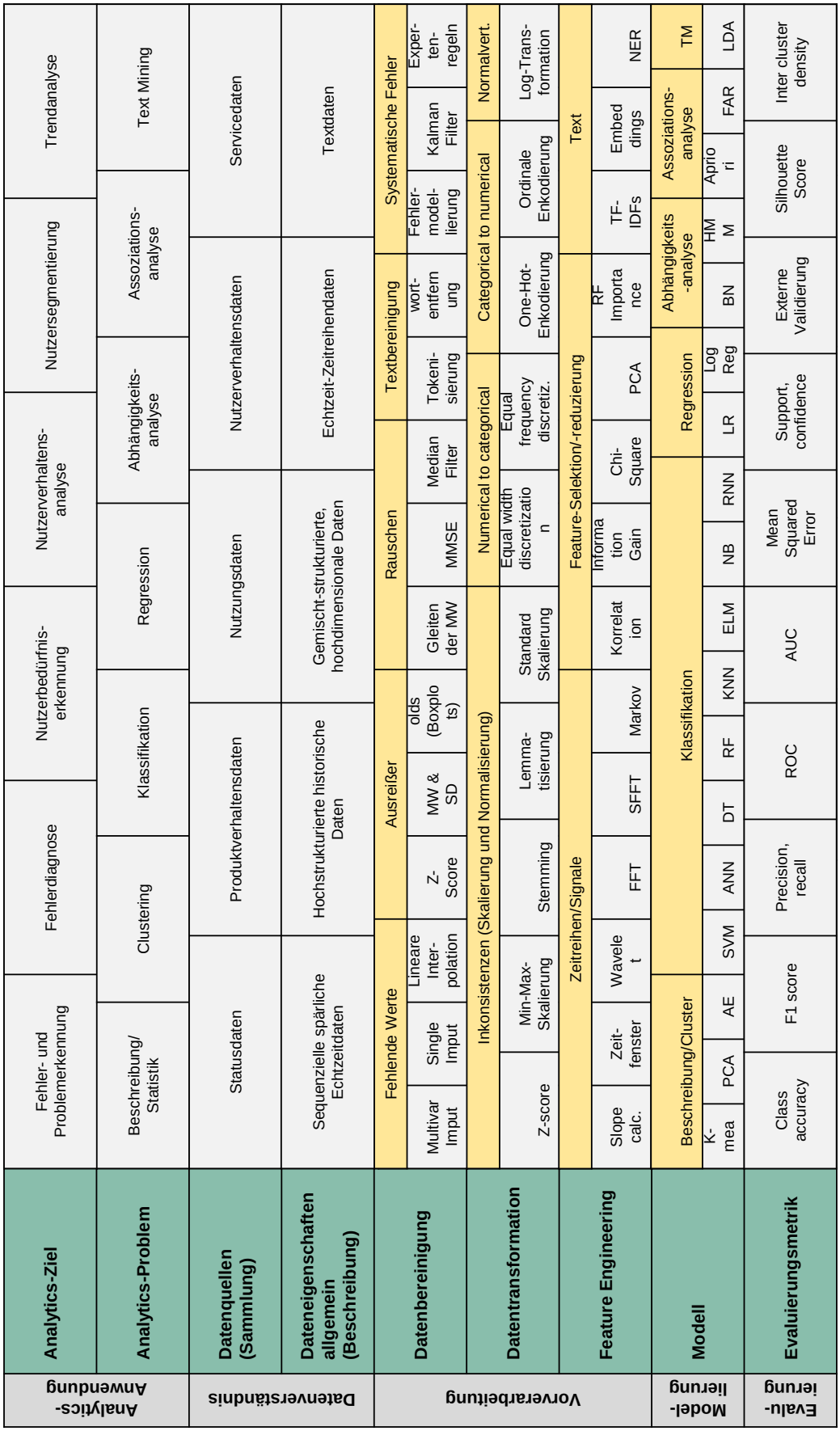

Bild 4-16: Data-Analytics-Toolbox für die betriebsdatengestützte Produktplanung

4.4 Vorgehen zur Datenanalyse in der betriebsdatengestützten Produktplanung

Während der Referenzprozess grundsätzlich zeigt, wie Anwender bei der Datenanalyse für die betriebsdatengestützte Produktplanung vorgehen können, benötigen Nicht-Experten ein systematisches Vorgehen und verschiedene Werkzeuge, um eine auf ihre Ziele und Bedürfnisse abgestimmte Analytics-Pipeline zu konzipieren und die passenden Lösungskomponenten für jeden Pipeline-Schritt zu bestimmen. Das Vorgehen ist in Bild 4-17 in Form eines Phasen-Meilenstein-Diagramms visualisiert. In den folgenden Abschnitten 4.4.1 bis 4.4.4 werden die einzelnen Phasen des Vorgehens detailliert beschrieben.

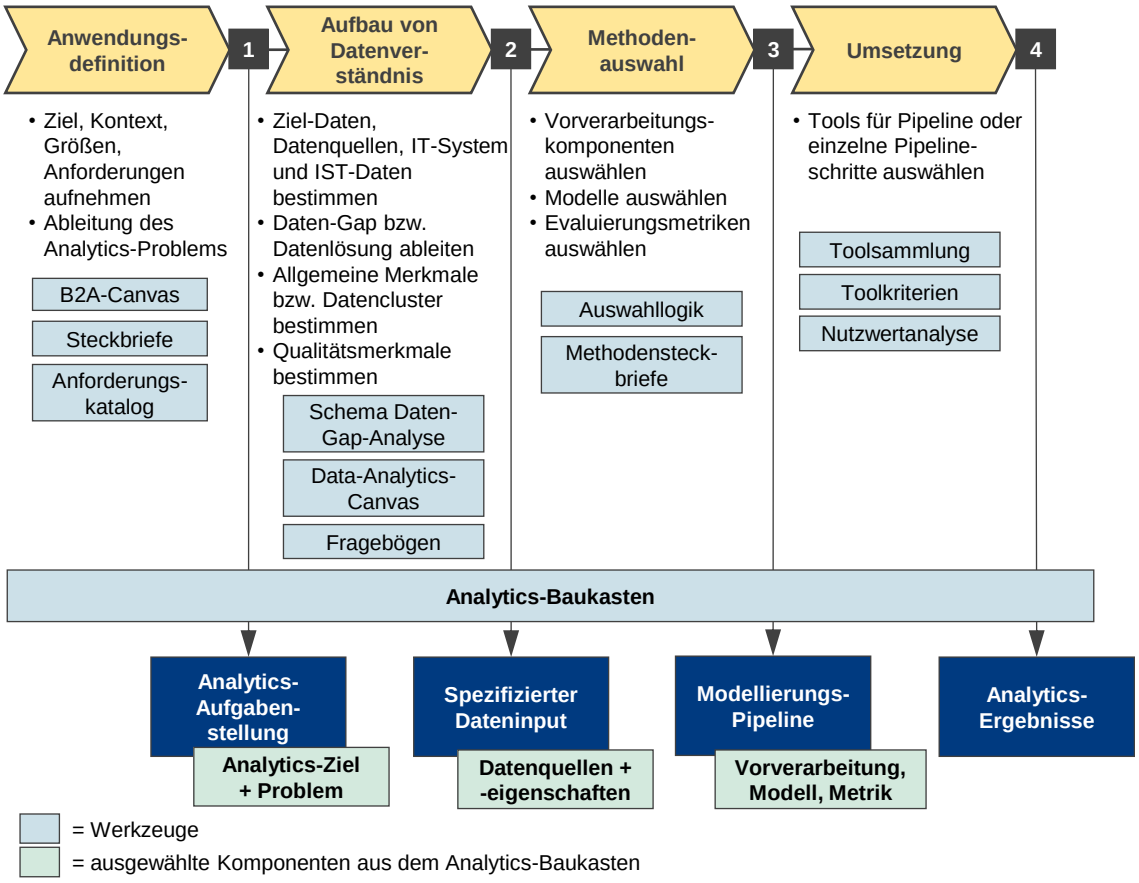

Bild 4-17: Vorgehen zur Datenanalyse in der betriebsdatengestützten Produktplanung

4.4.1 Anwendungsdefinition

Die erste Phase des Vorgehens zur Datenanalyse in der betriebsdatengestützten Produktplanung ist die Anwendungsdefinition. In dieser Phase wird die Analytics-Aufgabenstellung tiefer gelegt und mit der Bestimmung des Analytics-Ziels und -Problems wichtige Auswahlfaktoren festgelegt. Dies wird repräsentiert in der Leitfrage: *Wie lautet der Business-Use-Case in der „Analytics-Sprache“ und welche Anwendungsauswahlfaktoren sind zu berücksichtigen?*

Ergebnis des letzten Schritts des ersten Hauptprozesses *Planung von Betriebsdaten-Analysen* und damit Ausgangspunkt der Definition der Analytics-Anwendung sind konkrete Business Use Cases. Diese befinden sich noch in der "Sprache" des Produktmanagers und enthalten erste wichtige Informationen, wie bspw. das Ziel und die Anforderungen, die auch für den Data Scientist wichtig sind. Ziel dieses Schrittes ist, diese Use Cases in realisierbare Analytics-Aufgaben zu übersetzen, d. h. sie in eine Form zu bringen, die der Data Scientist mit Betriebsdaten umsetzen kann und seine nächsten Tätigkeiten strukturiert. Um diese wesentlichen Elemente der Analytics-Aufgabe bestimmen zu können, gilt es, die notwendigen Informationen und Parameter zusammenzutragen, die Aufschluss über die notwendigen Datenverarbeitungsschritte geben. Da der Stand der Forschung keine geeigneten Ansätze zur Übersetzung der Business-Use-Cases in Analytics-Anwendungen enthält, wurde ein visuelles Kollaborationstool, die Business-to-Analytics-Canvas (B2A-Canvas) entwickelt (s. Bild 4-18), mit dem der Data Scientist konzeptionell alle notwendigen Informationen sammeln und so systematisch Analyseansatz und Variablen definieren kann [PME+22].

Als **Methodisches Vorgehen z**ur Entwicklung dieses Hilfsmittels wurde ein Action-Design-Research-Ansatz genutzt, welcher die Möglichkeit bietet, ein Forschungsthema in einem realen Business-Kontext empirisch zu untersuchen [SHP+11, PTR+07]. Der Ansatz umfasste vier Stufen: 1) Problemformulierung, 2) Intervention, 3) Reflektion und Lernen und 4) Formalisierung des Gelernten.

1) Neben den Ergebnissen aus der Literatur, welche die Notwendigkeit für eine methodische Unterstützung für die Übersetzung begründen (vgl. Abschnitt 2.3.2), bestätigte auch eine frühere Studie zu der Frage „Wie können Data Analytics Use Cases in der Produktplanung spezifiziert und in konkrete Analytics-Aufgaben übersetzt werden?" [MPK+21] den Unterstützungsbedarf. So wurde beispielsweise die Herausforderung offengelegt, dass die Data-Analytics-Fragestellungen, die von Produktexperten formuliert werden, für Data Scientists nicht ausreichend Information bereitstellen, um einen Lösungsansatz zu bestimmen und somit viele Kommunikationsschleifen notwendig sind. Erste Lösungsansätze sollten Details wie Ziele, Probleme und standardisierte analyse-relevante Fragen bereitstellen sowie die Analytics-Probleme in den Fokus der Betrachtungen setzen.
2) Auf Basis dieser Erkenntnisse wurde im zweiten Schritt ein neues Workshop-Konzept erarbeitet, um noch besser zu verstehen, wie Data Scientists vorgehen, wenn sie eine Aufgabe mit einer zu beantwortenden Fragestellung im Produktplanugskontext erhalten und anschließend die Aufgabenstellung für sich spezifizieren. Dazu wurde ein Workshop mit sieben Experten aus dem Bereich Industrial Data Science durchgeführt. Diese hatten zur Aufgabe, zunächst Ziele und Fragestellungen von Produktmanagern einem vordefinierten Data-Analytics-Problem (z. B. Clustering) (vgl. Abschnitt 4.2.1.2) zuzuordnen. Parallel sollten Annahmen, Fragen und Umformulierungen aufgenommen werden, die für die Zuordnung notwendig waren. Im zweiten Schritt sammelten die Workshopteilnehmer weitere

Aspekte, die zur Tieferlegung der Use Cases und für einen erfolgreichen Start in die Datenverarbeitung für sie wichtig sind. Da sich herausstellte, dass die Teilnehmer beim zweiten Schritt anfangs ziemlich verloren waren, sich mit der Zeit aber drei Hauptkategorien von zu klärenden Aspekten herauskristallisierten (Produkt, Zielinformation und Anforderungen), konnten diese in einer zweiten Workshopiteration berücksichtigt werden.

3) Durch die Beobachtungen der Teilnehmer während der Workshops, der aufgenommenen Ergebnisse und zusätzlich gestellten Fragen am Ende der Workshops, wie z. B. „Hat die Übersetzung Ihrer Meinung nach im Workshop funktioniert?“, „Welchen Problemen sind Sie dabei begegnet?“, konnten mehrere Erkenntnisse abgleitet werden. Beispielsweise konnten Analytics-Probleme nur zugeordnet werden, wenn die Teilnehmer gewisse Fragen für sich beantwortet haben oder gewissen Annahmen getroffen haben. Alle „Learnings“ sind der entsprechenden Veröffentlichung zu entnehmen [PME+22].
4) Die vierte Stufe formalisierte das Gelernte, indem die zuvor gemachten Beobachtungen und Erkenntnisse generalisiert und abstrahiert wurden und die Lernprinzipien aus (3) als Elemente in eine Canvas zur Übersetzung von Business Use Cases in Data Analytics Use Cases überführt wurden.

Das **Ergebnis** ist in Bild 4-18 dargestellt. Sie zeigt ein exemplarisch ausgefülltes Canvas. Im Fokus der Canvas stehen die potenziellen Data-Analytics-Probleme. Um das passende Problem zu bestimmen, werden die Felder des Canvas genutzt. Diese helfen, den Use Case technisch weiter zu spezifizieren. Als Brücke zum Business Use Case kann das Kernziel des Anwendungsfalls genutzt werden, welches als zusammenfassender Titel des Canvas dient. Im Kontext der betriebsdatengestützten Produktplanung stehen im Kern sieben verschiedene Ziele zur Auswahl (s. Analytics-Ziele im Baukasten - Kap. 4.3.4).

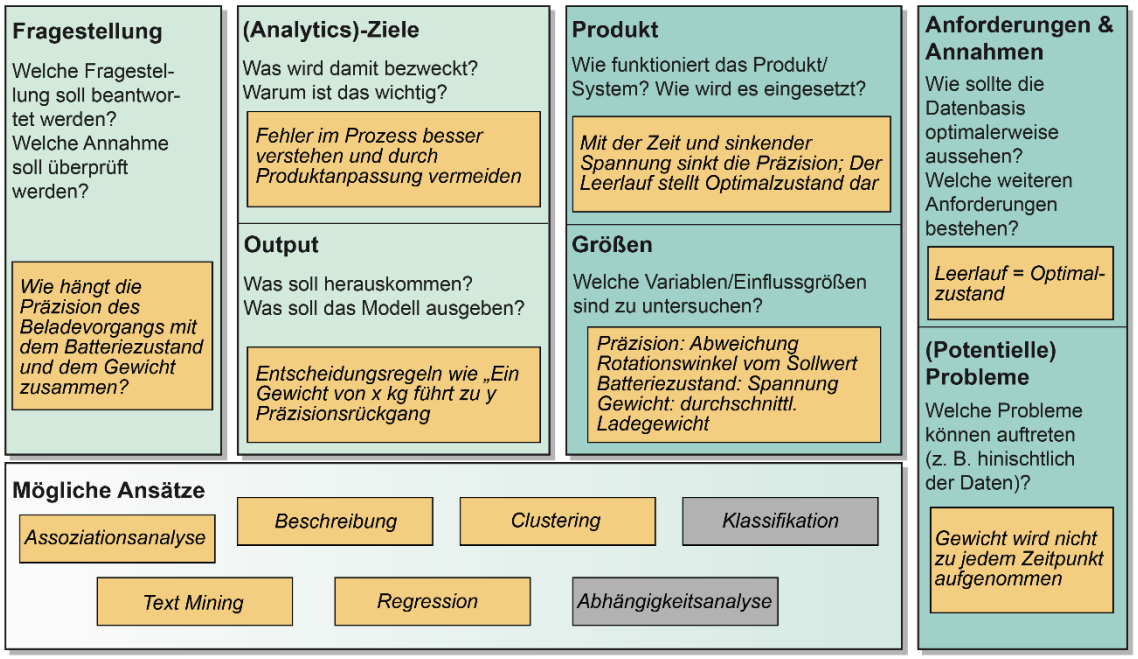

Bild 4-18: Business-to-Analytics-Canvas (B2A-Canvas) für die Übersetzung von Business-Use-Cases in Analytics-Aufgabenstellungen

Das gewünschte Analytics-Ziel lässt sich mit Hilfe von kurzen Analytics-Ziel-Steckbriefen bestimmen (s. Bild 4-19).

Analytics-Ziel: Nutzerbedürfnisanalyse

Zusammenfassung

Die **Nutzerbedürfnisanalyse** untersucht welche Interessen, Meinungen, Wünsche und Erwartungen und Anforderungen die Zielgruppe und Nutzer eines Produkts oder einer Dienstleistung haben. Ziel ist es, ein tiefes Verständnis für die Nutzer zu entwickeln, um Produkte oder Dienstleistungen so zu gestalten, dass sie den Bedürfnissen der Nutzer optimal entsprechen.

Untersuchungsaspekte

1) **Bedürfnisse und Wünsche**: Erfassen, was Nutzer wirklich benötigen und wünschen, auch jene Bedürfnisse, die sie möglicherweise nicht direkt artikulieren können.
2) **Schmerzpunkte**: Identifizieren von Problemen und Herausforderungen, die Nutzer mit bestehenden Produkten oder Dienstleistungen erleben.
3) **Präferenzen und Prioritäten**: Ermitteln, welche Merkmale oder Funktionen für die Nutzer am wichtigsten sind.
4) **Marktlücken**: Aufdecken von Bereichen, in denen bestehende Produkte oder Dienstleistungen die Bedürfnisse der Nutzer nicht vollständig erfüllen.
5) **Demografische und psychografische Merkmale**: Sammeln von Informationen über die demografische Zusammensetzung der Nutzergruppe und ihre Einstellungen, Werte und Lebensstile.

Ziele innerhalb der Produktplanung

Neben dem Ziel der **Entwicklung** von neuen Produkten, die **Anpassung** bestehender Produkte oder die Schaffung **personalisierter Nutzererfahrungen**, hat die Nutzerbedürfnisanalyse zum Ziel, die Kundenzufriedenheit zu steigern, das Angebot auf dem Markt zu differenzieren, das Risiko von Produktfehlschlägen zu minimieren und Innovationen zu fördern.
Die Nutzerbedürfnisanalyse ist ein zentraler Bestandteil des nutzerzentrierten Designs und der **strategischen Produktplanung**, da sie sicherstellt, dass Produkte und Dienstleistungen wirklich wertvoll und relevant für die Zielgruppe sind.

Bild 4-19: Beispiel für ein Analytics-Ziel-Steckbrief

Die Fragen des Business Use Cases werden wieder aufgegriffen und im Falle einer technisch unscharfen Formulierung in eine analysierbare Form umformuliert (z. B. „Ist die Prozessorleistung ausreichend groß dimensioniert?“ in „Wie viele Leistungspeaks liegen über der Lastgrenze?“). Dabei ist zu beachten, dass zur Beantwortung mancher Fragen oder Use Cases auch mehrere Analytics-Aufgaben notwendig sein können. Zur Gewährleistung der Vollständigkeit wird auch das wesentliche Ziel des Use Cases übernommen und ggf. nochmal geschärft. Data Scientist und Produktmanager stimmen den Output der Analyse genau ab. Hierzu gilt es, transparent aufzuzeigen, welche Outputs mit welchen Problemklassen möglich sind (typische Outputs von Segmentierungsverfahren sind z. B. ähnliche Objekte oder Gruppenzuweisungen). Steckbriefe, die typische Fragestellungen und Outputs der Analytics-Probleme aufzeigen, unterstützen dabei als Entscheidungsgrundlage. Bild 4-20 zeigt exemplarisch einen Analytics-Problem-Steckbrief.

Analytics-Problem: Klassifikation

Zusammenfassung

Die **Klassifikation** ist die Suche nach Mustern anhand eines Klassifikationsmerkmals. Primäres Ziel der Klassifikation ist das Erlernen der Zuordnung von Beobachtungen zu vorgegebenen Klassen (z.B. Bauteil defekt/nicht defekt). Das Klassenlabel ist ein kategorischer Wert und bekannt für jede Beobachtung. Ein Klassifikationsmodell dient zwei wesentlichen Rollen: 1. als prädiktives Modell, um bisher ungelabelte Instanzen zu klassifizieren und 2. als deskriptives Modell, um Eigenschaften zu identifizieren, die Instanzen von mehreren Klassen unterscheiden (vor allem dort sinnvoll, wo es nicht ausreicht ein Modell zu haben, das eine Vorhersage macht, ohne zu begründen, wie es zu einer solchen Entscheidung kommt. Viele Data Analytics Probleme können in Klassifikationsprobleme transformiert werden.

Typische Fragestellungen

- Welcher Kategorie gehört ein bestimmtes Objekt oder Ereignis an?
- Kann auf Basis vorhandener Merkmale vorhergesagt werden, zu welcher Gruppe oder Kategorie ein neues Objekt oder Ereignis gehören wird?
- Wie kann man Standardfälle von Ausnahmen unterscheiden?

Output

Zuordnung zu einer Klasse → Klassenlabels, Wahrscheinlichkeiten, Rangfolge der Klassen; Entscheidungsregeln; Konfidenzintervalle (Maß für die Zuverlässigkeit für die Klassifikation)

Algorithmen

z. B. Support Vector Machines, K-Nearest Neighbors, Entscheidungsbäume, Naive Bayes, Logistische Regression

Bild 4-20: Beispiel für ein Analytics-Problem-Steckbrief

Die Bestimmung des geeigneten Analytics-Problems wird erreicht, indem die Fragestellung, das Ziel und der gewünschte Output zusammen betrachtet werden. Mit den Feldern *Produkt, Größen, Anforderungen & Annahmen* sowie *Probleme* sammelt der Data Scientist weitere wichtige Informationen. Dies dient einem besseren Produktverständnis, der Vertiefung einzelner Variablen sowie der Definition von wichtigen Störfaktoren und Anforderungen. Um das notwendige Domänenwissen aufzunehmen, ist insbesondere das Produktverständnis zu stärken. Dies kann durch Nutzung weiterer Methoden, wie einer FMEA oder einem hierarchischen Funktionsmodell, die auch Ansätze aus dem Stand der

Forschung nutzen, unterstützt werden. Weitere Methoden sind z. B. CONSENS [ADG+09] und Ishikawa-Diagramme [Sys06].

Ein Anforderungskatalog (s. Bild 4-21) gibt einen Überblick über potenzielle Nutzer-Anforderungen an eine Datenanalyse, welche bei der späteren Methodenauswahl der Verfahren aus dem Analytics-Baukasten berücksichtigt werden müssen. Dieser stellt verschiedene Kriterien aus dem Stand der Forschung zusammen [NY18, ZSM+19, MCM+20]. Zutreffende Anforderungen sind an dieser Stelle festzuhalten.

Algorithmusmerkmal	Erklärung	Anforderung
Lernart	Überwachtes (Label notenwendig) oder unüberwachtes Lernen	Überwacht
Benötigte Rechenleistung	Erforderliche Rechenkapazität für die Anwendung der Methode	-
Implementierungsaufwand	Vorhandensein einer Implementierung der Methode in Bibliotheken gängiger Analysesoftware	Bibliothek vorhanden
Transparenz der Methode	Nachvollziehbarer Entscheidungsprozess oder Black-Box	Nachvollziehbares Modell
Schwierigkeitsgrad	Erforderliche statistische Kompetenz des Nutzers	Geringe Kompetenz erforderlich
Andere Nachteile (z. B. Overfitting-Risiko)	Berücksichtigung weiterer Aspekte, wie z. B. das Risiko, dass das Modell die Daten zu präzise beschreibt	-

Bild 4-21: Nutzeranforderungskatalog

Nach dieser Phase können die definierten Komponenten im Baukasten für die betriebsdatengestützte Produktplanung markiert werden. Er dient somit auch der Dokumentation der Ergebnisse.

4.4.2 Aufbau von Datenverständnis (Datensammlung und Beschreibung)

Die zweite Phase des Vorgehens dient dem Aufbau eines Datenverständnisses. Dazu werden im ersten Schritt im Zuge einer Datensammlung und Daten-Gap-Analyse die Zieldaten und zugehörigen Datenquellen bestimmt sowie mit den tatsächlich vorliegenden Daten abgeglichen. Der resultierende Dateninput wird anschließend hinsichtlich seiner Eigenschaften untersucht und beschrieben, um relevante Datenfaktoren für die nachfolgende Methodenauswahl zu bestimmen. Die Leitfrage lautet daher: *Welche Daten erlauben die Umsetzung der Analytics-Aufgabenstellung und welche Eigenschaften, bzw. Daten-Faktoren, sind bei der Methodenauswahl zu berücksichtigen?*

Wenn der Datenbedarf grob über die notwendigen Größen definiert ist, muss geprüft werden, ob dieser im Unternehmen abgedeckt ist, über welche Datenquellen und Systeme dies erfolgt oder ob die notwendigen Daten fehlen (Daten-Gap-Analyse). Der Data Scientist übernimmt diese Aufgabe zusammen mit den Domänenexperten und ggf. Datenverantwortlichen, um den notwendigen Überblick über die Datenlandschaft im Kontext

des Produkts zu erhalten. Methodisch wird dies durch das Schema der Daten-Gap-Analyse (s. Bild 4-22) unterstützt.

(1) Ziel-Daten	(2) Datenquelle	(3) IT-System	(4) IST-Daten	(5) Gap	(6) Fazit/Lösung
Präzision	Status-Daten – Hardware-Status	IoT	Abweichung Rotations-winkel (RW) Ist von RW Soll	-	Präzision durch Abweichung des RW gegeben
Gewicht auf Bagger-schaufel	Produktver-haltensdaten - Sensor	IoT	Durch-schnittliches Gewicht des LKWs für 3 Zyklen	Nur stark aggregierte Werte verfügbar	Regelmäßige Messung des Gewichts auf Schaufel notwendig
Batterie-zustand	Status-Daten – Hardware-Status	IoT	Spannung	-	Batterie-zustand durch Spannung gegeben

Bild 4-22: Schema für die Daten-Gap-Analyse

Dieses strukturiert die gedanklichen Schritte eines Ziel-/Ist-Datenabgleichs in einer chronologischen Reihenfolge: (1) Bestimmung der benötigten Zieldaten aus der B2A-Canvas, (2) Bestimmung potenzieller relevanter Datenquellen aus der Datenübersicht (ggf. zum schnellen Finden von Alternativen), (3) Bestimmung des IT-Systems als Ort der Datenhaltung, falls vorhanden, (4) Bestimmung der vorliegenden „Ist-Daten“, (5) Ableitung der Lücke zwischen Ziel- und Ist-Daten und (6) Ermittlung der Lösung (zu verwendender Dateninput). Je nach Kenntnissen der beteiligten Personen können einzelne Schritte wie die Datenquelle übersprungen werden. Die Bestimmung der Ist-Daten (4) kann zusätzlich durch weitere Hilfsmittel zur Modellierung von Datenflüssen unterstützt werden, um auch über die Zieldaten hinaus die Datenlandschaft des Unternehmens zu dokumentieren. Der Stand der Forschung hat gezeigt, dass Daten im Kontext des Produkts beispielsweise durch die Data-Analytics-Canvas modelliert werden können. Bild 4-23 zeigt ein exemplarisch ausgefülltes Data-Analytics-Canvas.

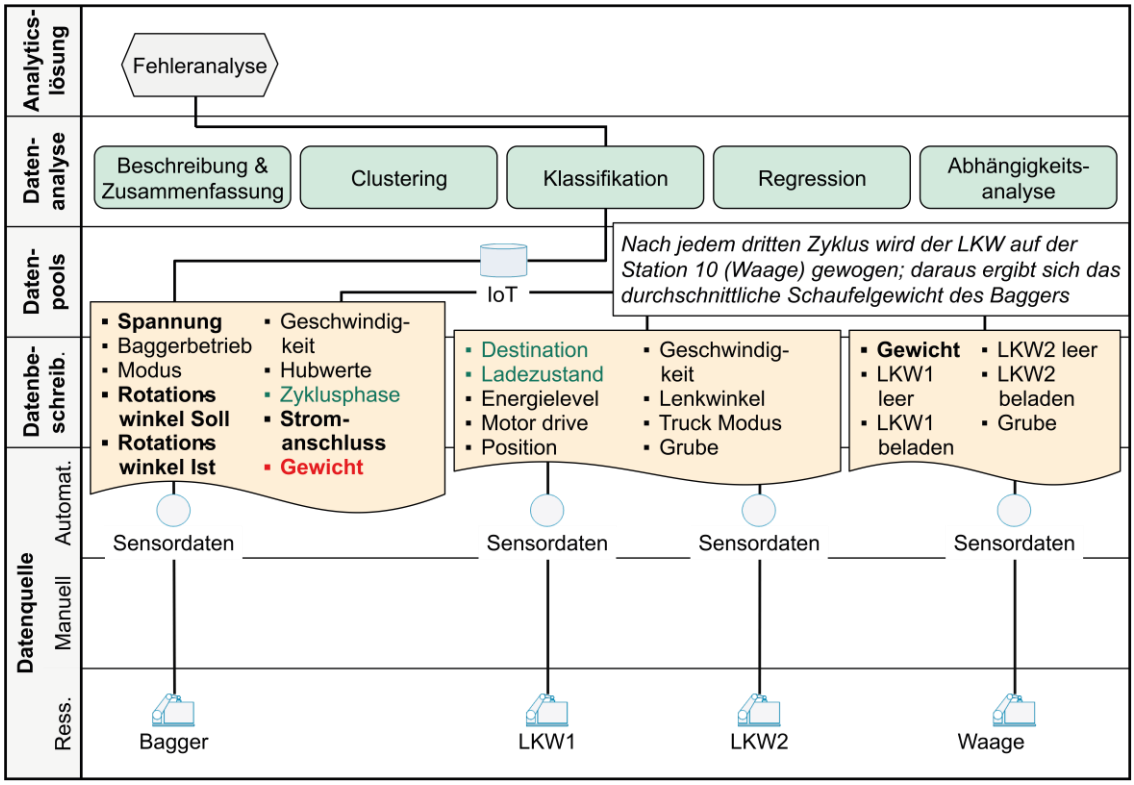

Bild 4-23: Beispiel einer ausgefüllten Data-Analytics-Canvas

Stellt sich heraus, dass wichtige Zieldaten fehlen und auch keine alternativen Datenobjekte oder eine Kombination von verschiedenen Datenpunkten die Zieldaten ersetzen können, müssen passende Daten akquiriert werden. Das DMME-Vorgehen von HUBER ET AL. bietet hier erste Ansätze, wie das im Kontext von technischen Anwendungen umgesetzt werden kann (vgl. Abschnitt 2.3.1). Im Rahmen dieser Arbeit wird die Datenakquirierung jedoch nicht betrachtet. Auch die ggf. notwendige Integration von Datenobjekten aus verschiedenen Datenquellen wird an dieser Stelle nicht weiter thematisiert.

Sobald alle notwendigen Daten gesammelt wurden, kann ein tieferes Verständnis über die analyserelevante Beschreibung der Daten aufgebaut werden, mit dem Ziel allgemeine Daten- und Qualitätsmerkmale für die Methodenauswahl zu bestimmen. Für die Bestimmung der allgemeinen Merkmale (vgl. Taxonomie in Abschnitt 4.3.2.2) stehen zwei Möglichkeiten zur Verfügung: (1) die schnelle und einfache Einordnung der Inputdaten in die Datencluster über Zuordnung zu den passenden Beispieldaten (z. B. können Spannungsmessungen über Sensor- und Statusdaten in das Cluster *sequenzielle dünnbesetzte Echtzeitdaten* eingeordnet werden) und (2) die selbstständige Bestimmung der Merkmale mit Hilfe eines Fragebogens (s. Bild 4-24). Dabei ist zu beachten, dass die erste Möglichkeit nur eine sehr grobe Einschätzung liefert. Bei der Bestimmung der individuellen Merkmale, wie Qualität und Variablentyp, unterstützt ein weiterer Fragebogen (s. Bild 4-25). Die Fragebögen basieren auf der Taxonomie für allgemeine und individuelle Betriebsdatenmerkmale (s. Kapitel 4.3.2.2) und ermöglichen eine simple Bestimmung der zutreffenden Eigenschaften über die Indikatoren. Bei den individuellen Eigenschaften

unterstützt die Angabe von Werkzeugen bei der Exploration der Daten und damit der korrekten Ermittlung entsprechender Merkmale.

Charakteristik/ Tags	Frage	Ja	Bemer-kung
(Sehr) strukturierte Daten	Liegen die Daten als Tabelle oder Matrix mit einer feststehenden Anzahl an Datenfelder vor, ohne Verbindung zwischen den Einträgen?	☐	
Sequentielle Transaktionsdaten	Liegen die Daten als Transaktionen mit einem Zeitbezug vor, die z. B. Events zu einem Zeitpunkt beschreiben?	☐	
Sequenzdaten	Sind die Daten geordnet, aber ohne Zeitstempel?	☐	
Zeitserien	Liegen die Daten als Messungen über die Zeit vor?	☐	
Signale	Liegen die Daten als gemessene physikalische Parameter über die Zeit vor?	☐	
Räumliche Daten	Enthalten die Daten eine Information über einen Standort?	☐	
Unstrukturierter Text	Liegen die Daten im Textformat vor und enthalten sie Tags und Metadaten?	☐	
Semistrukturierter Text	Liegen die Daten im Textformat vor und enthalten sie Tags und Metadaten?	☐	
Niedrig dimensional	Haben die Daten deutlich weniger Features/Attribute als Beobachtungen?	☐	
Hoch dimensional	Haben die Daten mehr Features als Beobachtungen?	☐	
Dünnbesetzte Daten	Haben die Daten viele Lücken in Form von Nullen, sind z. B. nur Veränderungen sichtbar?	☐	
Dichte Daten	Enthalten die Daten Werte, die sich kontinuierlich ändern?	☐	
(Auto-)Korrelation	Haben die Datenpunkte eine hohe Abhängigkeit zu ihren Nachbarpunkten?	☐	
Geringe Daten-mengen	Haben die Daten eher 100 Zeilen (pro Tag)?	☐	
Normale Datenmengen	Haben die Daten eher 1000 Zeilen?	☐	
Riesige Daten-mengen	Haben die Daten eher 100.000 Zeilen?	☐	

Bild 4-24: Fragebogen zur Bestimmung der allgemeinen Merkmale der Betriebsdaten

Eigenschaft (Tag)	Indikator	Werkzeuge	Ja	Bemerkung
Systematische Fehler	Daten liefern nicht reproduzierbare Ergebnisse, stark abweichende Ergebnisse im Vergleich zu anderen Sensoren, abnormale Datenverteilung, werden von Umgebung stark beeinflusst	Kontroll- und Vergleichs-messungen, Verteilung prüfen	☐	Neu-akquirierung der Daten ggf. notwendig
Zufälliges Rauschen	Daten enthalten unerwartete Muster, Abweichungen oder hohe Variation.	Visualisierung (Graphen, Plots), statistische Maße (z. B-Varianz, Std)	☐	
Ausreißer	Daten enthalten Datenpunkte, die außerhalb des erwarteten Bereichs liegen oder sich deutlich von anderen unterscheiden	Visualisierung (Plots. Histogramme, Boxplots), statistische Maße, Clustering	☐	
Inkonsistenzen	Daten enthalten unterschiedliche Wertebereiche, Abtastungen variieren	Visualisierung	☐	
Fehlende Werte	Daten enthalten stellenweise fehlende Daten	Visualisierung, Zusammenfass ungsstatistiken	☐	Bei sehr vielen fehlenden Daten, Neu-akquirierung ggf. notwendig
Duplikate	Daten enthalten doppelte Werte, die nicht relevant sind		☐	
Numerisch	Variablen nehmen Zahlenwerte an, entweder jeden beliebigen Wert innerhalb eines bestimmten Bereichs (kontinuierlich) oder nur bestimmte Werte (diskret)	„dtype“-Funktion	☐	
Kategorisch	Variablen nehmen eine begrenzte Anzahl von Werten an, i.d.R. nicht numerische		☐	

Bild 4-25: Fragebogen zur Bestimmung der individuellen Merkmale der Betriebsdaten

4.4.3 Methodenauswahl

In der dritten Phase werden die Methoden der Vorverarbeitung, die Algorithmen und Evaluierungsmetriken ausgewählt, mit dem Ziel die Kern-Data-Analytics-Pipeline zu konzipieren. Dies geschieht durch einen Abgleich der potenziellen Komponenten des Baukastens mit den bereits definierten Auswahlfaktoren. Folgende Leitfrage steht daher im Fokus: *Welche Pipeline-Komponenten eignen sich gegeben des Analytics-Ziels, Analytics-Problems, der Nutzeranforderungen und der Eigenschaften der Inputdaten?*

Da die Datenvorverarbeitung, wie in der Problemanalyse gezeigt, stark im Zusammenhang mit dem Algorithmus steht, ist es sinnvoll, im ersten Schritt die in Frage kommenden Algorithmen auszuwählen. Zur Unterstützung dieser Auswahl dienen neben dem vorgefilterten Lösungsraum an potenziell wichtigen Modellen des Baukastens von Experten

generierte Modellsteckbriefe (s. Bild 4-26). Die Struktur orientiert sich an den zu berücksichtigen Auswahlfaktoren (vgl. Abschnitt 2.3.4). Die Steckbriefe stellen neben einer kurzen Beschreibung und Subformen bzw. Algorithmen des Modells, Bezüge zu der Anwendung in der Produktplanung, den Nutzeranforderungen, den Daten und zu den Evaluierungsmetriken her. Für die Anwendung, Daten und Evaluierung werden direkt passende Komponenten aus dem Baukasten in Form von „Tags“ verknüpft. Darüber hinaus greift der Steckbrief weiteren Kontext zum Modell auf, wie z. B. wichtige Annahmen, Vor- und Nachteile. Auf Basis all dieser Informationen lässt sich der Lösungsraum weiter eingrenzen.

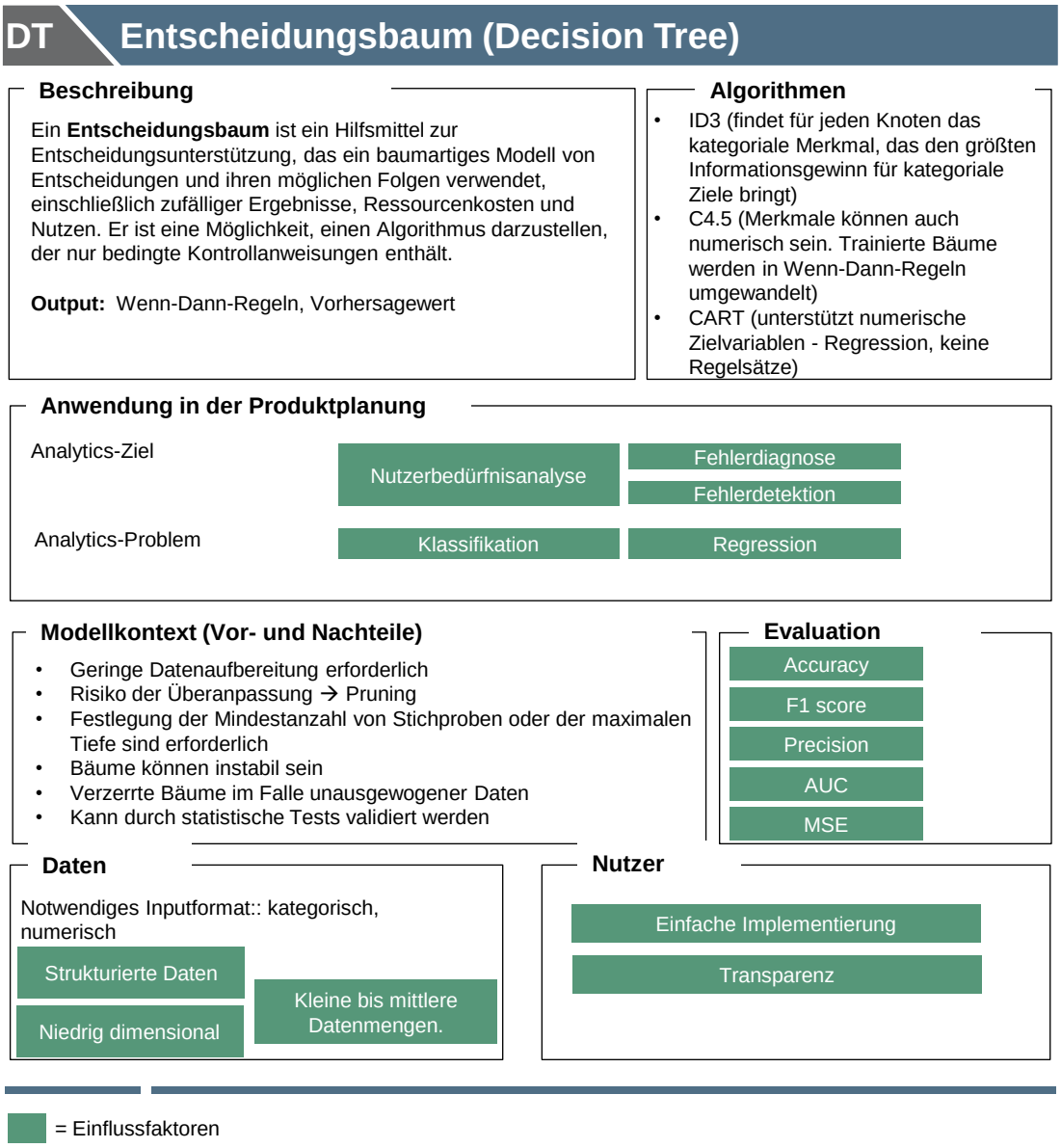

Bild 4-26: Beispiel für einen Modellsteckbrief

Der Entscheidungsprozess, bzw. die Auswahllogik, ist in Bild 4-27 veranschaulicht und lässt sich wie folgt darstellen:

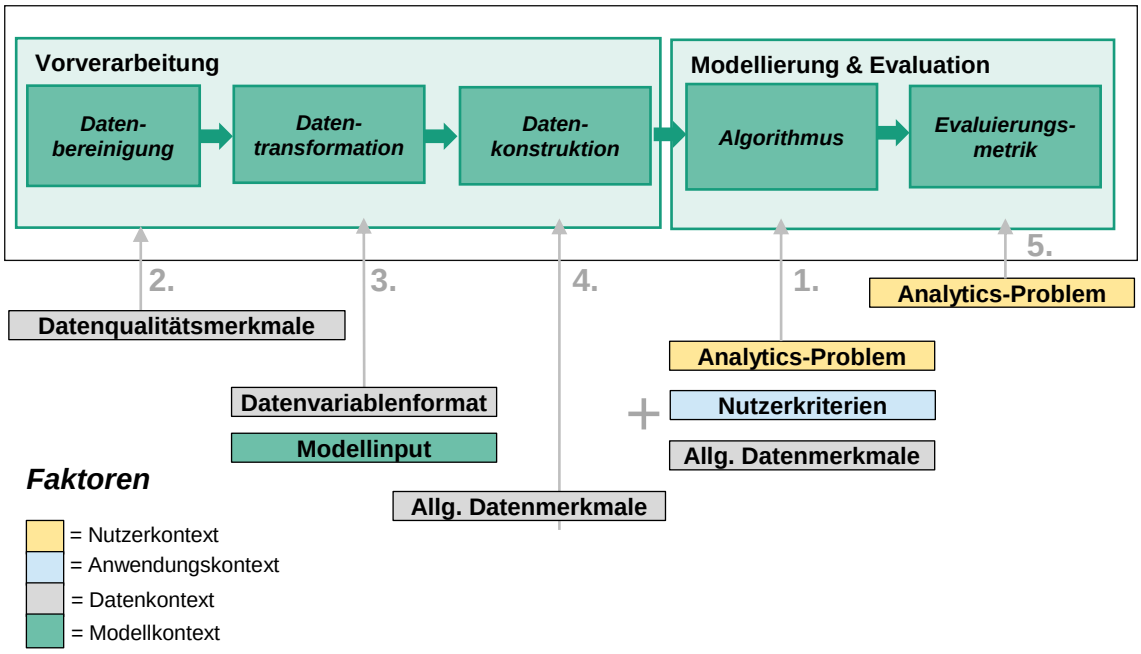

Bild 4-27: Auswahllogik für die Vorverarbeitungsmethoden, Algorithmen und Evaluierungsmetriken

1) Als erstes werden passende Algorithmen bestimmt. Dies geschieht mittels der Auswahlfaktoren *Analytics-Problem*, *Nutzerkriterien* und *allg. Datenmerkmale*. Das definierte **Analytics-Problem** in der ersten Phase (vgl. Abschnitt 4.3.1) trifft bereits im Baukasten durch die Strukturierung eine Vorauswahl einiger weniger Modelle. Zusätzlich können die Steckbriefe nach den entsprechenden Tags zu Analytics-Ziel und Analytics-Problem durchsucht werden, da die Expertensteckbriefe weitere Verknüpfungen aufzeigen können, die über den größtenteils literatur-basierten Baukasten hinausgehen. Die festgehaltenen **Nutzeranforderungen** aus der ersten Phase werden schließlich mit den wichtigen Aspekten aus Nutzersicht des Steckbriefs abgeglichen und die Modelle mit den meisten Übereinstimmungen gewählt. Zusammen mit den Vor- und Nachteilen der jeweiligen Modelle muss hier im Ermessen des Data Scientists häufig eine Abwägung getroffen werden, welcher Aspekt im vorliegenden Fall schwerer wiegt. Es ist jedoch auch immer sinnvoll, im Fall von mehreren vielversprechenden Ansätzen, sie einfach auszuprobieren und Ergebnisse zu vergleichen. Die in der zweiten Phase ermittelten **allgemeinen Dateneigenschaften** können einen zusätzlichen Filter darstellen, indem sie mit den sich für das Modell eignenden Datentypen auf dem Steckbrief abgeglichen werden.

2) Die ausgewählten Modelle können nun mit passenden Vorverarbeitungskomponenten kombiniert werden. Die Auswahl der Datenbereinigungskomponenten erfolgt im ersten Schritt anhand der definierten **Qualitätseigenschaften**. Der Baukasten greift diese auf, um schnell einige wenige passende Verfahren

vorzuschlagen. Eine weitere Eingrenzung der Verfahren kann wiederrum durch Expertenerklärungen unterstützt werden.

3) Die **Modelle** selbst bestimmen in erster Linie die notwendige Datentransformation. Die benötigten Inputvariablen gehen aus dem Steckbrief hervor und werden im Fall eines abweichenden **Ist-Variablenformats** der Daten (s. individuelle Eigenschaften Abschnitt 4.3.2) durch Verfahren der entsprechenden Kategorie im Baukasten umgewandelt.

4) Auch für die Feature-Engineering-Verfahren können die Modelle einen Ausschlag geben, wenn die Modelle beispielsweise nicht mit hoch-dimensionalen Daten umgehen können, dies aber einer Eigenschaft der vorliegenden Daten entspricht. Durch die Kategorisierung im Baukasten bietet sich vor allem auch eine Auswahl nach den vorliegenden **Datengruppen** an. Darüber hinaus ist das Konstruieren von Merkmalen stark domänengetrieben und erfordert zusätzliches Wissen und manuelle Arbeit.

5) Passende **Evaluierungsmetriken** werden für die Modelle auf den Steckbriefen vorgeschlagen. Sie hängen insbesondere von dem betrachteten Analytics-Problem ab, da sie häufig nach überwachten und unüberwachten Problemen kategorisiert werden (vgl. Kap. 2.3.4.3).

4.4.4 Umsetzung (Toolauswahl)

Ziel des vierten Schrittes ist die Umsetzung der konzipierten Data-Analytics-Pipelines mit Hilfe passender Tools, um Ergebnisse zu erhalten. Es gibt eine Vielzahl an unterschiedlichen Tools, welche verschiedene Einsatzschwerpunkte und Vorteile mitbringen (vgl. Abschnitt 2.3.5). Diese Systematik setzt den Fokus auf Python-Bibliotheken, da sie im Data Science sehr beliebt und frei verfügbar sind und gleichzeitig viel Flexibilität erlauben [Rog18]. Um das Ziel zu erreichen, wählen die Nutzer passende Bibliotheken aus. Dazu wird zunächst überprüft, welche Bibliotheken die relevanten Pipeline-Schritte umsetzen können. Im Anschluss wird (optional) eine Nutzwertanalyse durchgeführt, um anhand der wichtigsten Kriterien zwischen alternativen Bibliotheken die geeignetste auszuwählen. Folgende Leitfrage ergibt sich: *Mit welchen Python-Tools kann ich die Data-Analytics-Pipelines mit meinen Anforderungen am besten umsetzen?*

Als Grundlage zu diesem Vorgehen wurde eine initiale Sammlung an beliebten Python-Bibliotheken durch das Zusammentragen der favorisierten Tools von insgesamt fünf Data Scientists erstellt. Dabei wurde berücksichtigt, dass möglichst jeder der Pipeline-Schritte umgesetzt werden kann. Die Bewertung der Bibliotheken hinsichtlich der Erfüllung dieser Anforderungen nahmen zwei Personen, darunter die Autorin der Arbeit, unabhängig voneinander und mit Hilfe der Dokumentationen und der Tools selbst vor. Bei Uneinigkeit wurde nochmal die Dokumentation geprüft oder eine dritte unabhängige Person dazu befragt. Bild 4-28 zeigt einen Ausschnitt des Ergebnisses dieses Vorgehens. Diese

Tabelle dient den Nutzern als erster Filter zur Einschränkung der in Frage kommenden Tools. Sie ist vollständig im Anhang A3 aufgeführt.

	Datenbereinigung					Datentransformation	
	Fehlende Werte	Ausreißer	Noise	Textbereinigung	Systematische Fehler	Skalierung und Normalisierung	Transformation
Scikit-Learn	Ja	Ja	Ja	Ja	Ja	Ja	Ja
XGBoost	Nein	Nein	Nein	Nein	Nein	Nein	Nein
LightGBM	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Catboost	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Annoy	Nein	Nein	Nein	Nein	Nein	Nein	Nein
H20ai	Ja	Ja	Ja	Ja	Ja	Ja	Ja
StatsModels	Nein	Nein	Nein	Nein	Ja	Nein	Nein
Pattern	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Prophet	Nein	Nein	Nein	Nein	Nein	Nein	Nein
TPOT	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Auto-sklearn	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Numpy	Nein	Nein	Nein	Nein	Nein	Ja	Ja
Pandas	Ja	Ja	Ja	Ja	Ja	Ja	Ja
Matplotlib	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Feature Engine	Ja	Ja	Ja	Ja	Ja	Ja	Ja
SciPy	Nein	Ja	Ja	Nein	Nein	Nein	Nein
Dtreeviz	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Category_Encoders	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Tslearn	Nein	Nein	Nein	Nein	Nein	Ja	Ja
Sktime	Ja	Ja	Ja	Nein	Ja	Ja	Ja
TensorFlow	Ja	Ja	Ja	Nein	Nein	Ja	Ja
Keras	Nein	Nein	Nein	Nein	Nein	Ja	Ja
mljar	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Layzpredict	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Nltk	Nein	Nein	Nein	Ja	Nein	Nein	Nein
TextBlob	Nein	Nein	Nein	Ja	Nein	Nein	Nein
PM4PY	Ja	Ja	Nein	Nein	Nein	Nein	Ja

Bild 4-28: Ausschnitt aus der Python-Tool-Übersicht

An diese Vorauswahl schließt sich die systematische Bewertung der Bibliotheken mittels einer Nutzwertanalyse an. Diese Methode bietet einen strukturierten Ansatz, um aus einer Reihe von Alternativen die bestmögliche Lösung zu identifizieren, insbesondere, wenn viele Aspekte und Kriterien zu berücksichtigen sind und viele Personen am Entscheidungsprozess beteiligt sind [Küh19]. Die Entscheidungsalternativen sind durch die vorausgewählten Tools bereits vorgegeben. Vor der Durchführung der Nutzwertanalyse sind noch Entscheidungskriterien auszuwählen und zu gewichten. Tabelle 4-2 schlägt verschiedene Kriterien vor, die sich aus den Faktoren nach NADI und SAKR [NS23] ergeben und zur Auswahl von Data Science Tools herangezogen werden können (vgl. Abschnitt 2.3.5).

Tabelle 4-2: Kriterien zur Auswahl von Data Science Tools

	Kriterium	Beschreibung
Technische Faktoren	Nutzbarkeit/Benutzerfreundlichkeit	Wie einfach es ist, die Bibliothek zu nutzen
	Dokumentation	Vorhandensein von (offizieller) Dokumentation und Informationen für die Nutzung der Bibliothek
	Zweckmäßigkeit	Inwieweit der Zweck der Bibliothek mit den erforderlichen Anforderungen übereinstimmt
	Reife und Stabilität	Ob die Bibliothek für die Nutzung stabil ist
	Performance	Die allgemeine Performance der Bibliothek
Menschliche Faktoren	Aktivitätsgrad	Wie aktiv die Community der Bibliothek ist (d.h. Größe, Reaktionsfähigkeit)
	Erfahrung	Die kollektive Erfahrung der Community (Bibliotheksnutzer) mit der Bibliothek.
Sonstige	Statistische Stichhaltigkeit	Ob Vertrauen in den statistischen Hintergrund des Entwicklers besteht
	Konsistenz und Skalierbarkeit	Ob Bibliothek konsistente Ergebnisse zu anderen Bibliotheken liefert und große Datenmengen bewältigen kann
	Individualisierung	Das Maß, in dem eine Bibliothek dem Benutzer ermöglicht, an benutzerdefinierten Anwendungsfällen zu arbeiten

Die Gewichtung kann nach verschiedenen Methoden erfolgen (z. B. 10er-Skala oder Schulnoten). KÜHNAPFEL stellt hierzu einige alternative Methoden vor [Küh19]. Anschließend werden die Entscheidungsalternativen und die Bewertungskriterien in einer Matrix gegenübergestellt. Jede Entscheidungsalternative wird für jedes Bewertungskriterium bewertet, so dass am Ende die Summe aller Einzelbewertungen für jede Entscheidungsalternative gebildet werden kann. Ein Beispiel für eine Nutzwertanalyse zeigt Tabelle 4-3.

Tabelle 4-3: Beispiel für eine Nutzwertanalyse

Bewertungs-kriterien	Gew.	Tensorflow		Keras	
		Bewertung	Punktewert	Bewertung	Punktewert
Nutzbarkeit	**0,4**	7	2,8	6	2,4
Aktivitätsgrad	**0,35**	9	3,15	8	2,8
Dokumenta-tion	**0,25**	8	2	7	1,75
Summe	**100**		**7,95**		**6,95**

Die Kriterien lassen sich auch für Low-Code oder No-Code-Tools, wie Knime oder Rapidminer, anwenden, welche auch Anwendern ohne Programmierkenntnisse die Umsetzung von Pipelines erlauben.

Die final konzipierten Pipelines können nun mit den entsprechenden Tools in Experimenten umgesetzt und ihre Leistung mit Hilfe der Evaluierungsmetriken verglichen werden. Dabei können wiederrum andere Tools, wie AutoML und Programmierassistenten, unterstützen (vgl. Kap. 2.4).

Nach erfolgreicher Umsetzung der Pipelines können die gefilterten Baukästen anderen Nutzern als Muster und Orientierungshilfe dienen.

4.5 Prototyp für ein digitales Assistenz- und Lerntool

Um die Systematik einfacher anwendbar zu gestalten und sie weiterhin mehr im Sinne einer naturalistischen Evaluierung, also einer Leistungsbewertung des Lösungsartefaktes in seiner realen Umgebung, validieren und evaluieren zu können, wurde die Systematik als digitaler Assistent in Form einer Webanwendung implementiert [VPB16, PRT+12]. Dieser soll Anwendern den Nutzen und die Eignung der Systematik demonstrieren. Ein Tool wie dieses ermöglicht in besonderem Maße die Konsolidierung der Bestandteile der Systematik sowie ihre Manifestierung. Dadurch wird die Anwendung für Nicht-Experten weiter vereinfacht. Abschnitt 4.5.1 erläutert das Vorgehen zur Entwicklung des Tools, bevor in Abschnitt 4.5.2 der Prototyp vorgestellt wird.

4.5.1 Methodisches Vorgehen zur Entwicklung des Tools

Das konkrete Vorgehen zur Entwicklung des Tool-Prototypens setzt sich aus drei Phasen zusammen und ist in Bild 4-29 dargestellt. Im Folgenden werden die Phasen kurz erläutert.

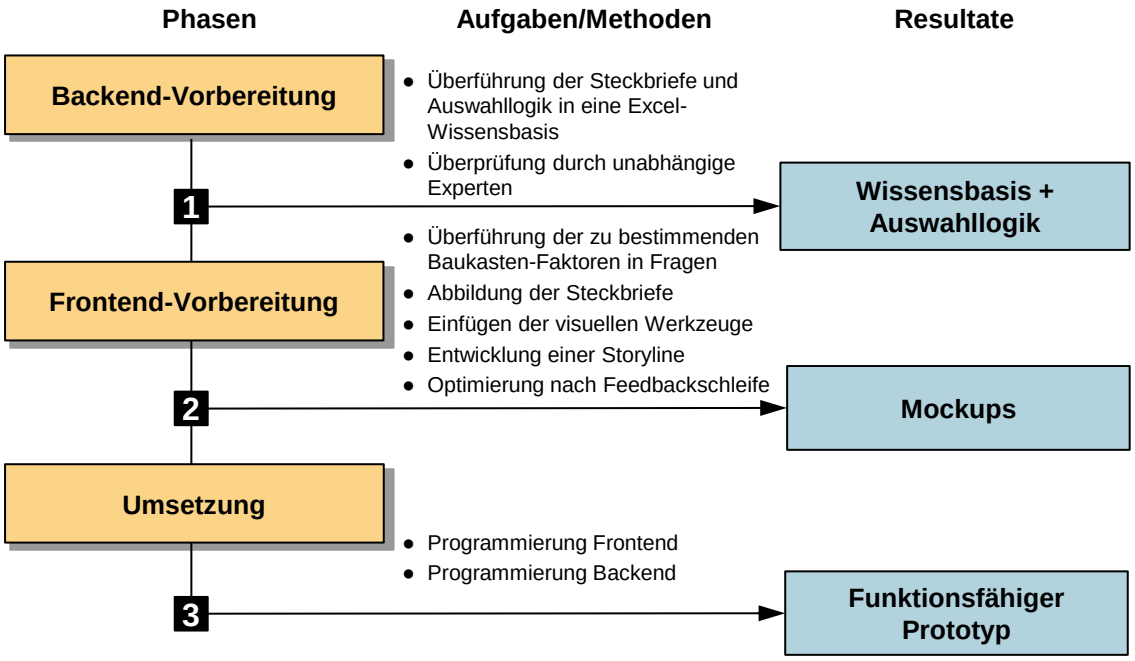

Bild 4-29: Vorgehen zur Entwicklung des Tool-Prototypen

1) **Backend-Vorbereitung**: Zuerst wurden alle Steckbriefe zu den Algorithmen und Analytics-Zielen und -Problemen aus dem Analytics-Baukasten sowie die Expertenerklärungen zu den verschiedenen Vorverarbeitungsverfahren und Evaluierungsmetriken in eine Exceltabelle erfüllt, die als Wissensbasis dienen soll. Die Tabelle besteht aus zwölf Tabellenblättern (s. Anhang A4): vier Blätter, welche die Kern-Pipeline bestehend aus Datenbereinigung, Transformation, Feature Engineering und Algorithmen, abdecken, sowie sechs Blätter, welche die Einflussfaktoren zur Auswahl der Kern-Pipeline-Verfahren repräsentieren. Neben den Faktoren aus dem Analytics-Baukasten Analytics-Ziel und Analytics-Problem wurden zusätzlich die Nutzerkriterien aus dem Anforderungskatalog (s. Abschnitt 4.4.1), die allgemeinen Datencharakteristika, die Datenqualitätsaspekte und das Datenformat aus den Fragebögen (s. Abschnitt 4.4.2) aufgenommen. Diese Einflussfaktoren werden als Drop-Down-Liste in den Blättern für die Methoden der Kern-Pipeline verwendet, um die notwendigen Verknüpfungen herstellen zu können. Die Evaluationsmetriken aus dem Baukasten werden auf einem weiteren Blatt aufgeführt und im Blatt Algorithmen aufgerufen. Zuletzt ist die Toolübersicht (s. Abschnitt 4.4.4) zu nennen, die auf dem zwölften Tabellenblatt abgebildet ist. Die Algorithmen sind gemäß der Algorithmensteckbriefe (s. Bild 4-26) in insgesamt zwölf Spalten organisiert: Kurzbeschreibung, Output, Trainingsalgorithmen, Modellvarianten, Modellkontext (Vor- und Nachteile), Analytics-Ziel, Analytics-Problem, Eignung für allgemeine Dateneigenschaften, notwendiges Inputformat, Nutzeranforderungen, Evaluationsmetriken und Kommentar. Bild 4-30 zeigt das Format exemplarisch für einen Algorithmus. Die Auswahllogik aus Bild 4-27 ist durch die Farbgebung angedeutet. Die farblich hervorgehobenen

Einflussfaktoren sind mit der Nutzereingabe abzugleichen und Algorithmen mit übereinstimmenden Faktoren auszugeben. Die Informationen zu den Vorverarbeitungsverfahren sind gemäß den geringeren relevanten Faktoren schlanker organisiert und umfassen jeweils die Spalten Erklärung, Voraussetzung (, dass die Verfahren gut funktionieren), sowie Vor- und Nachteile. Für die Datenbereinigungsverfahren wird der Faktor *Datenqualität* ergänzt; für die Transformationsverfahren *Modellinput* und *Datenvariablenformat* und für die Feature-Engineering-Verfahren *allgemeine Datenmerkmale*. Als letzte Aufgabe dieser Phase wurde die gesamte Wissensbasis von einem unabhängigen Data-Science-Experten mit mehrjähriger Berufserfahrung sowohl in der Forschung als auch der Industrie geprüft, um eine gute Informationsqualität und Faktizität zu gewährleisten.

Modell	Kurzbeschreibung	Output	Trainings-algorithmen	Modell-varianten
K-means	K-means ist ein unüberwachter ML-Algorithmus, der zur Gruppierung von Daten in k Cluster verwendet wird. Dabei versucht der Algorithmus, die Datenpunkte so in Cluster zu unterteilen, dass die Abstände zwischen den Datenpunkten innerhalb eines Clusters möglichst klein sind, während die Abstände zwischen den Clustern möglichst groß sind.	Gruppen-zugehörigkeit	Näherungsmaß benötigt: Euklidische Distanz oder Manhattan Distanz für Datenpunkte im euklidischen Raum; Kosinus Ähnlichkeit und Jaccard geeigneter für Dokumente, Varianten durch unterschiedliche Distanzmaße: Euklidische Distanz oder Manhattan Distanz für Datenpunkte im euklidischen Raum; Kosinus Ähnlichkeit und Jaccard geeigneter für Dokumente	

Modellkontext	Analytics-Ziel	Analytics-Problem	Daten-eigenschaften	Nutzer-anforderungen
Zu den Vorteilen von k-means gehören seine Einfachheit, Schnelligkeit und Skalierbarkeit sowie die Möglichkeit, auch große Datensätze zu verarbeiten. Ein Nachteil ist, dass der Algorithmus sensibel auf die ….	Fehler- und Problemerkennung Nutzer-segmentierung, Nutzerverhaltens-analyse	Clustering	strukturierte Daten, dichte Daten, Normale Datenmenge, Geringe Datenmenge	Einfache Implementierung (durch Low-Code-Funktionen)

Evaluationsmetriken	Inputformat	Kommentar
silhouette score, external validation….	numerisch	-

Bild 4-30: Format der Wissensbasis für Algorithmen

2) **Frontend-Vorbereitung**: In der zweiten Phase wurde die grafische Benutzeroberfläche der Anwendung geplant, mit dem Ziel Mockups zu erstellen. Dazu

wurden die Screens der Anwendung zunächst nach den Dimensionen des Analytics-Baukastens strukturiert: 1) Anwendung, 2) Datenverständnis, 3) Vorverarbeitung und 4) Modellierung. Das systematische Vorgehen aus Abschnitt 4.4 wurde entsprechend der Schritte in diese Struktur integriert. Die Dimensionen 3) und 4) entsprechen der Methoden- und Toolauswahl im Vorgehen. Während 1) und 2) Benutzereingaben erfordern, um die relevanten Faktoren zu bestimmen, sollten 3) und 4) keine aktiven Eingaben mehr nötig machen, sondern gemäß der Auswahllogik im Backend die passenden Methoden einschließlich ihrer Steckbriefinformationen aus der Wissensbasis anzeigen. Um die Auswahl und die Zusammenhänge für den Nutzer möglichst transparent und nachvollziehbar zu gestalten, auch zum Zweck des Lernens, sollten die ausschlaggebenden Faktoren und damit die vorgenommenen Nutzereingaben in den Informationen hervorgehoben werden (s. Bild 4-31). Die Nutzereingaben selbst wurden in Form von Fragen und Anzeige der dazugehörigen Informationen gestaltet. Neben den Fragen wurden auch die visuellen Werkzeuge des Vorgehens (Business-to-Analytics-Canvas, Schema zur Daten-Gap-Analyse und Data-Analytics-Canvas) integriert, um den Nutzern die Konkretisierung und Dokumentation der Anwendung und Betriebsdaten zu ermöglichen. Da diese jedoch zur vereinfachten Implementierung keine Eingaben für die Auswahl entgegennehmen sollten, wurden insbesondere beim Schritt der Anwendungsdefinition Fragestellungen zum Analytics-Ziel und -Problem ausgelagert (s. Bild 4-32). Somit werden diese Angaben, bzw. Faktoren, nicht ausschließlich in der Business-to-Analytics-Canvas erarbeitet.

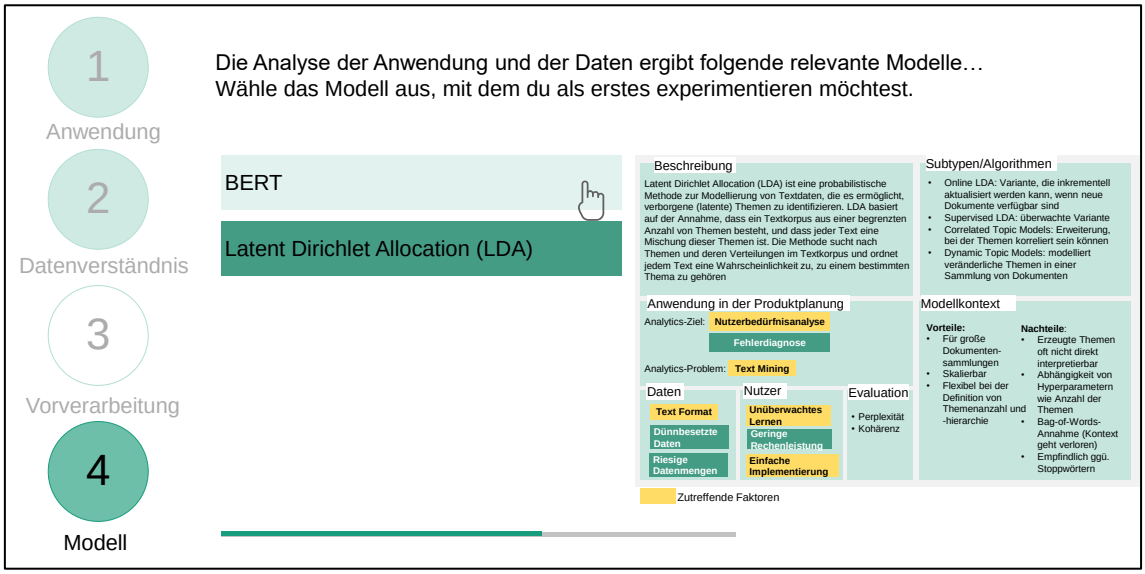

Bild 4-31: Mockup-Screen – Übersicht über vorgeschlagene Algorithmen

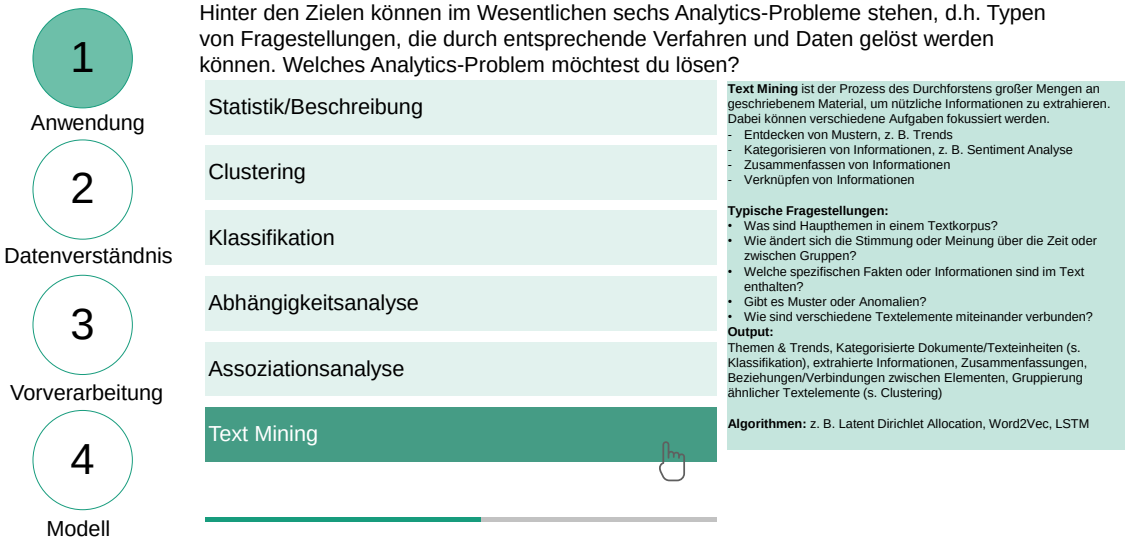

Bild 4-32: Mockup-Screen: Auswahl des Analytics-Problems

Darüber hinaus wurde zur Förderung eines Verständnisses über die Nutzung und Funktionsweise des Tools eine Storyline (s. Einleitungstexte in den Mockups) konzipiert, über welche der Nutzer anschaulich durch den Prozess geleitet werden soll. Die Mockups wurden am Ende der Phase im Rahmen einer Feedbackschleife mit potenziellen Nutzern einmalig optimiert.

3) **Umsetzung**: Die Anwendung wurde exemplarisch und vereinfacht als Webapplikation entwickelt. Das Frontend besteht aus einzelnen serverseitig gerenderten Screens, die mit der Komponentenbibliothek TailWind (https://tailwindcss.com/) umgesetzt wurden. Ablauf und Design richten sich hierbei nach den im vorhergehenden Abschnitt aufgeplanten Mockups. Das Backend ist auf Basis von Micronaut als Serverapplikation umgesetzt. Neben dem Ausliefern des Frontends und dem Session-Handling ist hier auch die Kernlogik zur Empfehlung passender Modelle und Techniken implementiert. Zur besseren technischen Handhabbarkeit wurde die Excel basierte Wissensbasis zunächst in ein JSON Format überführt. Die Wissensbasis wurde statisch auf Vollständigkeit geprüft, indem zunächst sämtliche Kombinationen von Antworten (Analytics Ziele, Dateneigenschaften, etc.) generiert und für jede dieser Kombinationen sichergestellt wurde, dass es passende Vorschläge von Modellen und Vorverarbeitungstechniken gibt. Das Vorschlagssystem kommt im Userflow zum ersten Mal zum Einsatz, wenn aus den Zielen und Dateneigenschaften mögliche Modellkandidaten angezeigt werden. Die Wissensbasis hält für jedes Modell in der jeweiligen Sektion eine Liste der kompatiblen Antworten aus vorhergehenden Screens bereit. Als Modellkandidat gelten diejenigen Modelle, für die die Nutzerantwort in jeder Kategorie als kompatible Antwort aufgeführt ist. Da die Wissensbasis so gestaltet ist, dass sie einen ausmaterialisierten Regelsatz zur Algorithmen- und Verfahrensauswahl darstellt, muss im Backend kein regelbasiertes System mehr implementiert

werden. Die Kandidatenauswahl lässt sich auf eine Reduktion der Gesamtmenge der Verfahren abbilden, für die in jeder Kategorie vom Nutzer eine als (gemäß der Wissensbasis) kompatibel gelistete Antwort gegeben wurde.

4.5.2 Ergebnis

Der entstandene Prototyp ist unter dem Link dpipe.eu-central-1.elasticbeanstalk.com aufrufbar. Die Webanwendung umfasst insgesamt 14 verschiedene Ansichten entlang des Vorgehens zur Datenanalyse in der betriebsdatengestützten Produktplanung. Sie enthält die notwendigen Nutzereingaben sowie die Empfehlung der Modelle als auch der Vorverarbeitungsverfahren. Die visuellen Werkzeuge werden über einen Link auf einem digitalen Whiteboard zur Verfügung gestellt, wo die Nutzer auch gemeinsam mit anderen Teammitgliedern die Templates bearbeiten können. Die Informationen zu den jeweiligen Analytics-Komponenten wurden lediglich exemplarisch in Form der Steckbriefe dargestellt (s. Bild 4-33). In den anderen Fällen kommt eine einfache Textanzeige zum Einsatz (s. Bild 4-34).

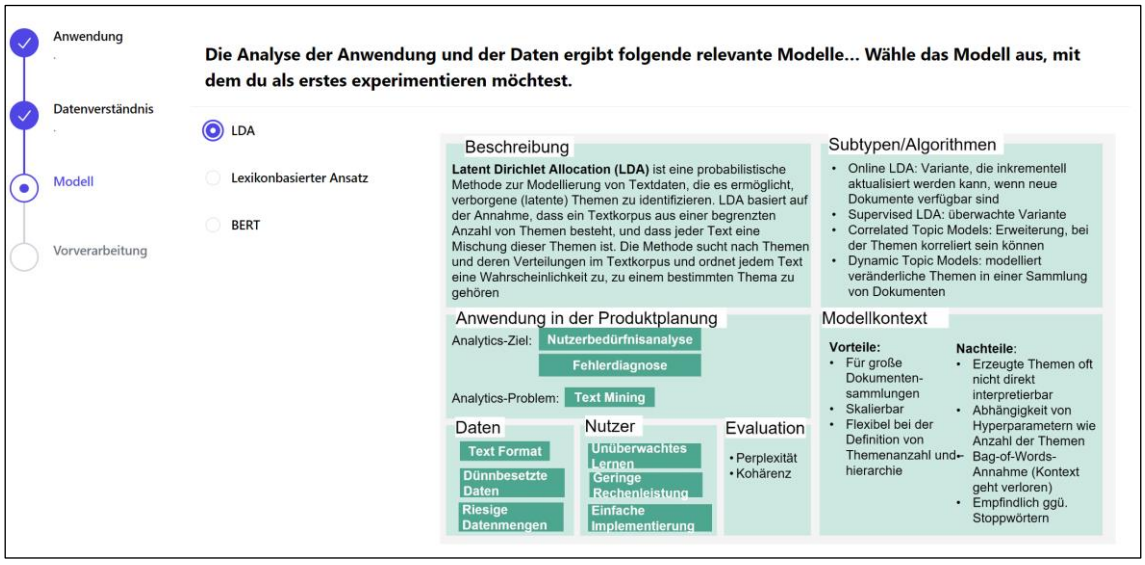

Bild 4-33: Tool-Ansicht der Modellempfehlung

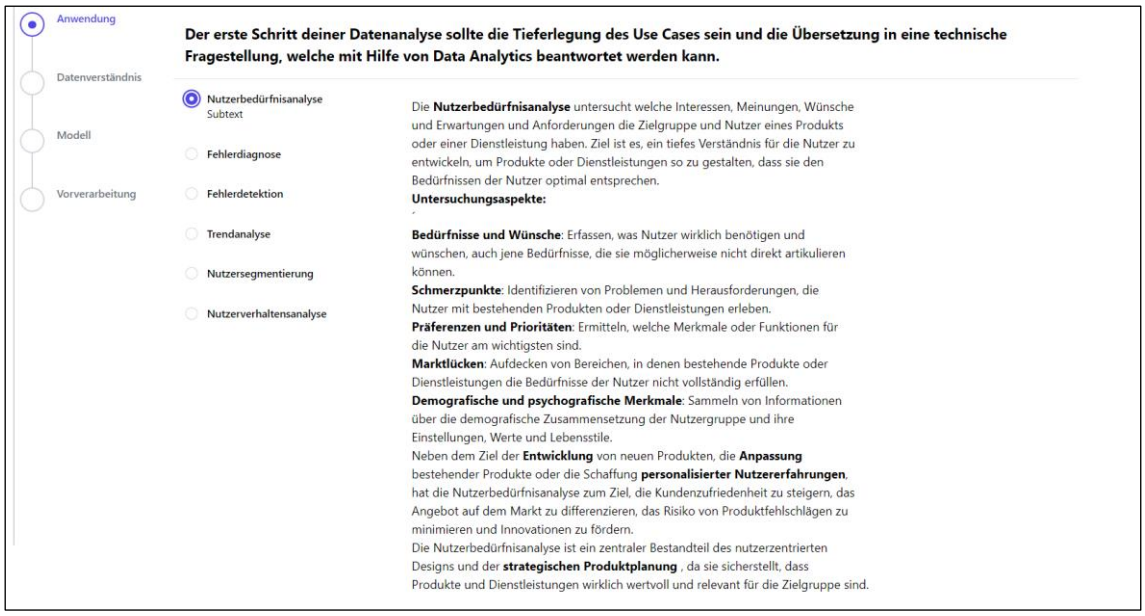

Bild 4-34: Tool-Ansicht der Auswahl des Analytics-Ziels

4.6 Kriterien-basierte Analyse der Systematik

In diesem Abschnitt wird die Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung anhand der an sie gestellten Ziele (Abschnitt 2.5) bewertet. Dies entspricht der künstlichen Evaluierung, insbesondere der Kriterien-basierten Analyse, in der DSR [VPB16].

DA-Prozess übergreifende Ziele

Z1) Befähigung (und Training) von Nicht-Experten/Citizen Data Scientists in der Industrie: Die Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung befähigt Nicht-Experten bzw. Produktexperten (Citizen Data Scientists) und BerufseinsteigerInnen im Data Science durch ihre Strukturierung und diversen Hilfsmittel zur Konzipierung von erfolgsversprechenden Data-Analytics-Pipelines. Insbesondere das detaillierte Vorgehen und die erklärungsliefernden, Experten-basierten Artefakte gewährleisten einen vertrauenswürdigen Aufbau von Wissen und Verständnis für das Zusammenspiel der Schritte und vielen möglichen Komponenten im Data-Analytics-Prozess für die Produktplanung. Durch das Tool wird die praktische Anwendbarkeit und damit die Trainingseffizienz gesteigert

Z2) Bereitstellung eines Strukturierungsrahmens für die Datenanalyse von Betriebsdaten in der strategischen Produktplanung

Die Systematik stellt ein Strukturierungsrahmen in Form eines Referenzprozesses für die Datenanalyse von Betriebsdaten in der strategischen Produktplanung bereit. Dieses stellt einen Handlungsrahmen dar und beschreibt die Schritte, die es zu erledigen gilt. Außerdem ermöglicht er die Ausgestaltung anhand von für die Produktplanung relevanten Analytics-Komponenten. Der Strukturierungsrahmen baut auf bewährten Data-Analytics-

Referenzprozessen auf und überführt diese in einen einfachen und verständlichen Kontext der strategischen Produktplanung.

Ziele an/für die Definition von Analytics-Anwendungen für die strategische Produktplanung

Z3) Bereitstellung von relevanten Data-Analytics-Zielen und -Problemen

Die Systematik stellt über den Strukturierungsrahmen relevante Data-Analytics-Ziele und -Probleme für die betriebsdatengestützte Produktplanung bereit. Diese wurden mittels eines DSR-Ansatzes und einer umfangreichen Literaturstudie identifiziert. Die Ziele und Probleme stellen die Verbindung zwischen dem Business-Use-Case und der Analytics-Aufgabenstellung dar und bilden die Grundlage für eine Übersetzung.

Z4) Übersetzung der Business Use Cases in Analytics-Aufgabenstellungen

Dieses Ziel wird mit der Business-to-Analytics-Canvas im Schritt der Anwendungsdefinition erfüllt. In der Canvas wird zunächst der zuvor definierte Business Use Case aufgegriffen und tiefer gelegt. Sie verhilft den Nutzern so dazu, dass sie einen technischeren Blickwinkel einnehmen können. Zuletzt findet mit Hilfe der aufgenommenen Informationen eine Übersetzung in konkrete Analytics-Probleme statt. Die B2A-Canvas wurde mit Hilfe eines mehrstufigen Action-Design-Research-Ansatzes entwickelt.

Ziele an das Datenverständnis

Z5) Bereitstellung einer strukturierten und detaillierten Betriebsdatenübersicht

Die Systematik stellt eine Betriebsdatenübersicht als Hilfsmittel zur Sammlung und Identifizierung relevanter Daten für einen Use Case der betriebsdatengestützten Produktplanung bereit. Sie strukturiert eine Vielzahl an verschiedenen Betriebsdaten entlang von fünf Kategorien, welche unterschiedliche Aspekte des Produktes, wie z. B. Zweck, Nutzerinteraktion und Qualität repräsentieren.

Z6) Bestimmung der relevanten Betriebsdaten: Über die Daten-Gap-Analyse und Data-Analytics-Canvas ermöglicht die Systematik den Nutzern die Bestimmung der für einen Use Case relevanten Daten. Dafür werden im Rahmen des Vorgehens die definierten Zieldaten mit den vorhandenen Ist-Daten abgeglichen und dokumentiert.

Z7) Bereitstellung eines Beschreibungsrahmens für Betriebsdaten

Die Systematik stellt einen Beschreibungsrahmen für Betriebsdaten zur Verfügung. Dieser besteht aus acht Dimensionen mit allgemeinen und individuellen Merkmalen. Er bildet alle notwendigen Eigenschaften für ein umfassendes Datenverständnis ab. Der Rahmen wurde als Taxonomie mit Hilfe der Methode nach NICKERSON ET AL. entwickelt.

Z8) Bestimmung der Dateneigenschaften

Das Ziel, dass Nutzer die Dateneigenschaften ihrer Daten mit Hilfe des Beschreibungsrahmens einfach bestimmen können, wird durch die Fragebögen in der zweiten Phase des

Vorgehens erfüllt. Die Fragebögen basieren auf der Taxonomie und stellen einfache Fragen, um die zutreffenden Eigenschaften zu erkennen.

Ziele an die Methodenauswahl

Z9) Bereitstellung von für die betriebsdatengestützten Produktplanung relevanten Vorverarbeitungsmethoden sowie Algorithmen und Evaluierungsmetriken

Die Systematik stellt über den Strukturierungsrahmen Vorverarbeitungsmethoden sowie Algorithmen und Evaluierungsmetriken bereit, die in der betriebsdatengestützten Produktplanung relevant sind. Damit wird der riesige Lösungsraum an potenziellen Verfahren eingeschränkt und eine anschließende Auswahl geeigneter Methoden vereinfacht. Die Verfahren wurden über eine strukturierte Literaturanalyse in Verbindung mit einer Umfrage ermittelt und bilden somit eine wissenschaftliche und praktische Sichtweise ab.

Z10) Auswahl geeigneter Methoden unter Berücksichtigung der Abhängigkeiten

Der Entscheidungsprozess in der Phase Methodenauswahl ermöglicht die Auswahl weniger geeigneter Vorverarbeitungsmethoden, Algorithmen und Evaluierungsmetriken, mit denen eine Umsetzung starten kann. Dieser Prozess berücksichtigt die Einflusskontexte Anwendung, Nutzer, Daten und Modell, die im Rahmen der Problemanalyse identifiziert wurden, und stellt somit eine fundierte Auswahl sicher.

Ziele an die Umsetzung

Z11) Auswahl geeigneter Tools

Die Auswahl geeigneter Tools wird über die Toolübersicht und die Nutzwertanalyse im letzten Schritt des Vorgehens zur Datenanalyse in der betriebsdatengestützten Produktplanung gewährleistet. Die Toolübersicht bietet eine Experten-basierte Bewertung von beliebten Tools hinsichtlich der notwendigen Pipeline-Schritte, während die Nutzwertanalyse eine sehr spezifische Auswahl anhand von Kriterien, die für Data-Science-Tools relevant sind, zulässt.

5 Anwendung der Systematik als Prototyp und Evaluierung

In diesem Kapitel wird die Anwendung der Systematik in der Praxis in Form von zwei verschiedenen Fallbeispielen (vgl. Abschnitt 5.1) sowie eine darauf aufbauende Nutzenevaluation (vgl. Abschnitt 5.2) beschrieben. Die Fallbeispiele dienen nach PEFFERS ET AL. dazu, die entwickelte Lösung in der realen Welt zu demonstrieren und ihre Funktionalität zu validieren, während die Nutzenevaluation die Bewertung der Lösung hinsichtlich ihrer Effektivität und Nützlichkeit beinhaltet (vgl. Kap. 1.3.1). Aufgrund der komprimierten Darstellung und der einfacheren Demonstration und Kommunikation wurden die Fallbeispiele als auch die Nutzenevaluation auf die Systematik zum Teil in Form des Prototypens (vgl. Abschnitt 4.5) angewendet.

5.1 Vorstellung der Fallbeispiele

Die entwickelte Systematik wurde mit zwei produzierenden Unternehmen am Beispiel eines ausgewählten Produkts angewendet. Nachfolgend wird die Anwendung der Systematik anhand der zwei Beispiele gezeigt:

1) Fallbeispiel 1: **Fehlerdiagnose** (Zusammenhang von Kassettenwechsel und Servicefällen bei Geldautomat Systemen)

2) Fallbeispiel 2: **Nutzerbedürfnisanalyse** (Stimmungen und Themen im Kontext einer elektronischen Verbindungsklemme)

5.1.1 Fallbeispiel „Geldautomat“

Das erste Anwendungsbeispiel beschreibt die Anwendung der Systematik mit einem Unternehmen, das Geldautomaten für Banken und Kassensysteme für den Einzelhandel produziert. Im Fokus steht eine Produktfamilie der Geldautomaten.

Im ersten Schritt der **Definition der Anwendung** wurde mit Hilfe der ersten Frage nach dem Ziel der Analyse und der Business-to-Analytics Canvas (s. **Fehler! Verweisquelle konnte nicht gefunden werden.**) ein zuvor definierter Business Use Case tiefer gelegt und hinsichtlich Analytics-relevanter Fragen ausspezifiziert. Ursprünglich hatte das Produktmanagement die Vermutung, dass ein Teil der Gerätefehler und Performance-Beeinträchtigungen auf eine unsachgemäße Operator-Bedienung (z. B. bei der Notenbefüllung durch externe Dienstleister) zurückgeführt werden kann. Übergreifendes Ziel des Fallbeispiels „Geldautomat“ war die Identifikation von Fehlerursachen, um die Systemperformance zu erhöhen. Damit konnte das Analytics-Ziel *Fehlerdiagnose* bestimmt werden. Die sich daraus ergebende, analysierbare Kernfrage lautet: Gibt es einen (zeitlichen) Zusammenhang zwischen Kassettenwechseln und Fehlermeldungen? Es gilt damit, die vermutete Hypothese, dass die Fehlermeldungen häufig durch eine unsachgemäße Bedienung von externen Dienstleistern verursacht werden, zu validieren. Im Falle einer Bestätigung sollte der Vorgang durch Anpassungen am Produkt optimiert werden. Für eine

weitere Spezifizierung wurde als gewünschter Output der Datenanalyse die Häufigkeit von Fehlermeldungen nach Kassettenwechseln sowie die Zeit zwischen Kassettenwechseln und Fehlern definiert. Für das Produktverständnis wurden u.a. die verschiedenen Probleme im Kontext der Hardware herausgestellt, wie z. B. ein Stau nach Befüllung oder eine Abnutzung, welche auch zu Stau führen kann. Als wichtige Größen wurden die System-ID, die Fehlercodes, die Fehler-Events und die Kassettenwechsel-Events eingetragen. Als weitere Hinweise wurde mit aufgenommen, dass teilweise auch Gründe für Störmeldungen in Form von Textdaten vorliegen.

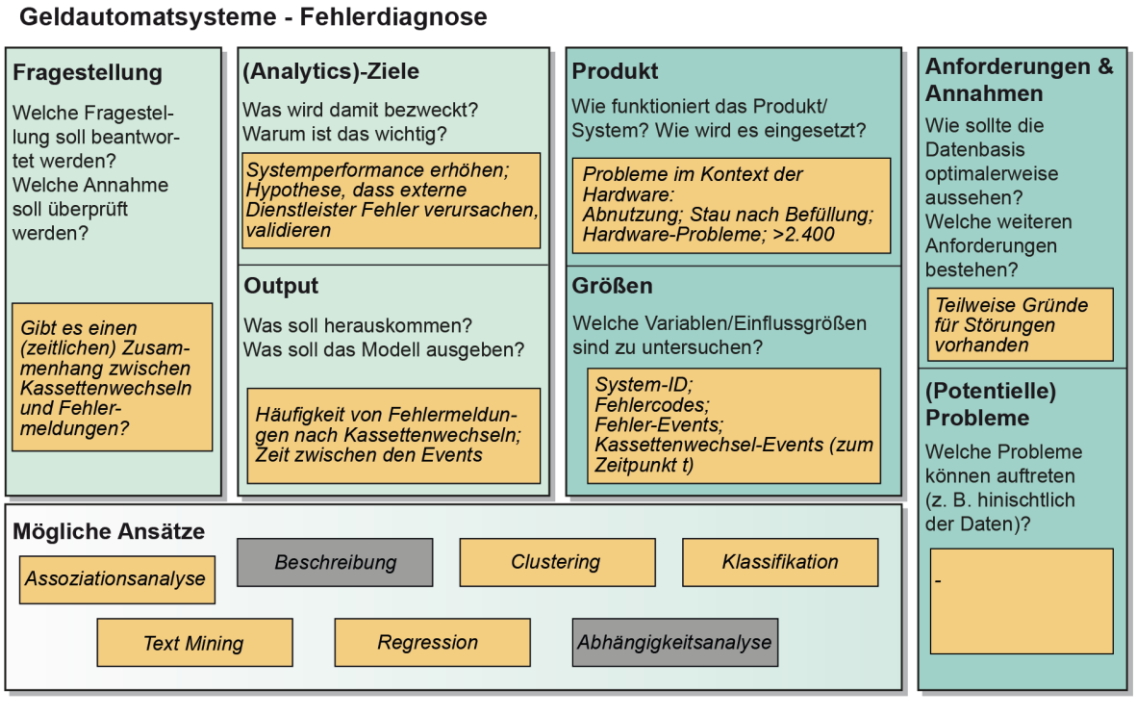

Bild 5-1: Ausgefüllte B2A-Canvas für die Fehlerdiagnose

Anschließend konnten vereinfacht durch die zwei Leitfragen und Anwählen von Auswahlmöglichkeiten die notwendigen Einflussfaktoren des Analytics-Problems (*Beschreibung/Statistik*, *Abhängigkeitsanalyse*) und der Nutzeranforderungen (*Transparenz*, *einfache Implementierung*) für die spätere Methodenauswahl ermittelt werden.

Im Schritt **Aufbau von Datenverständnis** zeigte sich durch Ausfüllen der Data Analytics Canvas (s. **Fehler! Verweisquelle konnte nicht gefunden werden.**) ein klareres Bild über die vorhandenen Daten. Das Unternehmen konnte zur Beantwortung der Fragestellung drei relevante Datensätze aus einem Datenpool nutzen: 1) System-Log-Daten mit den Variablen System-Seriennummer, Zeitpunkt des Events (Kassettenwechsel) und die betroffene Kassetteneinheit. 2) Ticketdaten mit der Seriennummer, Zeitpunkt der Fehlermeldung und Fehlercode (geschätzte Fehlerursache). 3) Servicedaten mit Fehlercode und mögliche Fehler. So wurde auch ersichtlich, über welche Variablen sich die Datensätze

verknüpfen lassen. Mithilfe dieser Methode konnten die Datenquellen *Status*- und *Servicedaten* (Events) bestimmt werden.

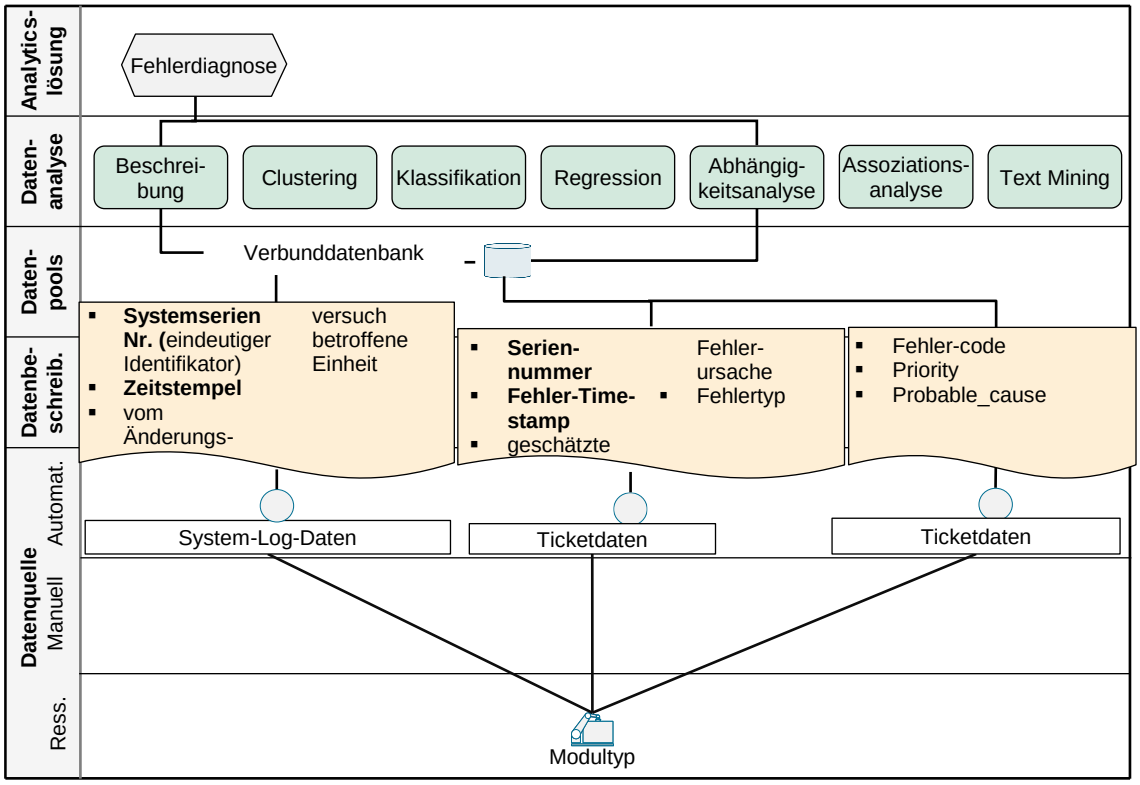

Bild 5-2: Ausgefüllte Data-Analytics-Canvas für die Fehlerdiagnose

Anschließend wurden die Daten mit Hilfe der Fragebogen beschrieben. Ergebnis dieser Beschreibung sind die Merkmale *sequenzielle Transaktionsdaten*, *niedrig dimensional*, *normale Datenmenge*, (überwiegend) *kategorisch*. Auffällig dabei ist, dass keine nennenswerten Qualitätsprobleme bei Exploration der Datensätze aufgedeckt wurden.

Aufgrund der Faktoren aus den Schritten 1 und 2 hat der digitale Assistent im dritten Schritt (**Modellauswahl**), gestützt auf sein Wissen über die Analytics-Probleme, passende Verfahren für die Beschreibung/Statistik und die Abhängigkeitsanalyse ausgewählt. Dabei wurden deskriptive Statistik für das erste Analytics-Problem und Algorithmen wie Hidden Markov Models (HMMs) und Bayesian Networks (BNs) für das zweite Problem identifiziert. Im Rahmen dieser Case Study wurden die Verfahren der Statistik zur sorgfältigen Exploration gewählt. Die Umsetzung einer Abhängigkeitsanalyse durch HMMs oder BNs kann nachgelagert stattfinden. Eine geeignete Vorverarbeitungsmethode wurde aufgrund der fehlenden Qualitätsprobleme und notwendigen Transformationen nicht ausgegeben. Das Ergebnis – die ausgewählten Komponenten – ist in **Fehler! Verweisquelle konnte nicht gefunden werden.** dargestellt.

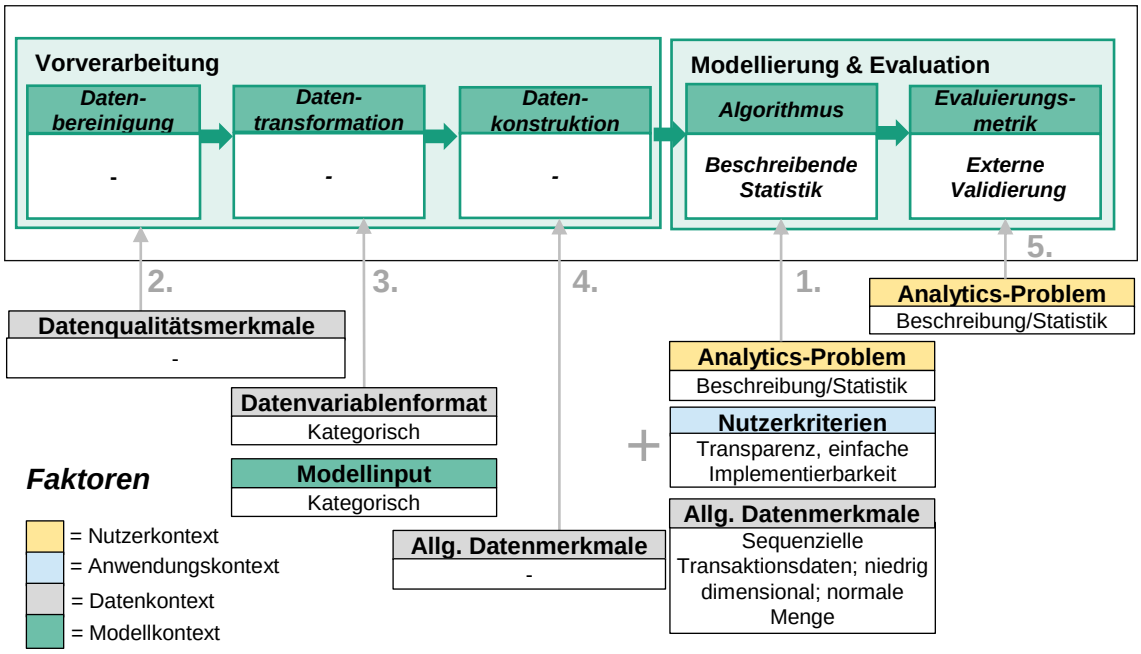

Bild 5-3: Ergebnis der Methodenauswahl für das Fallbeispiel „Geldautomat“

Im letzten Schritt der Systematik, der **Umsetzung**, waren das Ergebnis der Vorausauswahl insgesamt drei verschiedene Bibliotheken. Nach genauerer Prüfung dieser Tools erwies sich die Process-Mining-Bibliothek *PM4PY* als die erste Wahl, da diese u.a. statistische Auswertungen in Form von Aktivitätenfolgen umsetzt, Event-Logs ohne viel Vorverarbeitung verarbeitet und sehr wenig Code erfordert. Die Bibliothek *pandas* ermöglicht zusätzlich die notwendigen Datenmanipulationen, um deskriptive Analysen wie Häufigkeitsauswertungen, Anteilsberechnungen und Filterungen vorzunehmen.

Anschließend konnte der konzipierte Use Case mit Hilfe der Bibliotheken umgesetzt werden. Nach dem Einlesen der kombinierten Daten waren nur wenige Bereinigungen nötig. Eine dieser Bereinigungen war z. B. die Umformatierung der Variable Zeitstempel in eine „Datetime-Variable“. Um die Daten noch besser zu beschreiben und erste Muster zu erkennen, wurden neben statistischen Kennzahlen und Aktivitätenfolgen Plots über verschiedene Häufigkeitsverteilungen erstellt, welche die Zusammenhänge zwischen zwei Variablen darstellten. Es konnten folgende Kernergebnisse für die drei definierten Fragen festgehalten werden:

Welcher Zusammenhang besteht zwischen den Operator-Bedienungen und den Fehlern am Gerät?

Tabelle 5-1: Statistische Kennzahlen für das Fallbeispiel „Geldautomat“

Metrik	**Wert**
Gesamtzahl der Ereignisse (Count)	*289696*
Anzahl Seriennummern (distinct)	*1135*

Anzahl Seriennummern (distinct) (nur System mit Kassettenwechseln)	*148*
Anteil von Systemen mit Kassettenwechseln	*13% (148/1135)*
Anzahl Tickets (distinct)	*236*
Anzahl Kassettenwechsel Events (Count)	*20903*
Durchschnittl. Anzahl Tickets pro Seriennummer	*9,36*
Durchschnittl. Anzahl Kassettenwechsel-Events pro System	*141,24*
Durchschnittl. Zeit zwischen Kassettenwechsel-Events	*1 Tag 05:12:39*
Anteil von Systemen mit auftretenden Fehlern nach einem Kassettenwechsel	*82% (122/148)*

Fehler! Verweisquelle konnte nicht gefunden werden. führt verschiedene statistische Kennzahlen auf, die auf der einen Seite einen Überblick über die Systeme und die Events geben, auf der anderen Seite durch Anteilsberechnungen einen ersten Eindruck über mögliche Zusammenhänge aufdecken. Ein Hinweis auf einen Zusammenhang liegt besonders darin, dass bei den meisten Systemen, in denen ein Kassettenwechsel in den Daten erfolgt, Fehler nach einem solchen Wechsel auftreten.

Welche Arten von Fehlern werden durch unsachgemäße Bedienung verursacht?

Fehler! Verweisquelle konnte nicht gefunden werden. zeigt die Häufigkeit des Auftretens von verschiedenen Fehlercodes nach einem Kassettenwechsel auf. Der mit großem Abstand am häufigsten auftretende Fehler ist A100091 („Init auf Reject Transport hat den Stau nicht beseitigt“).

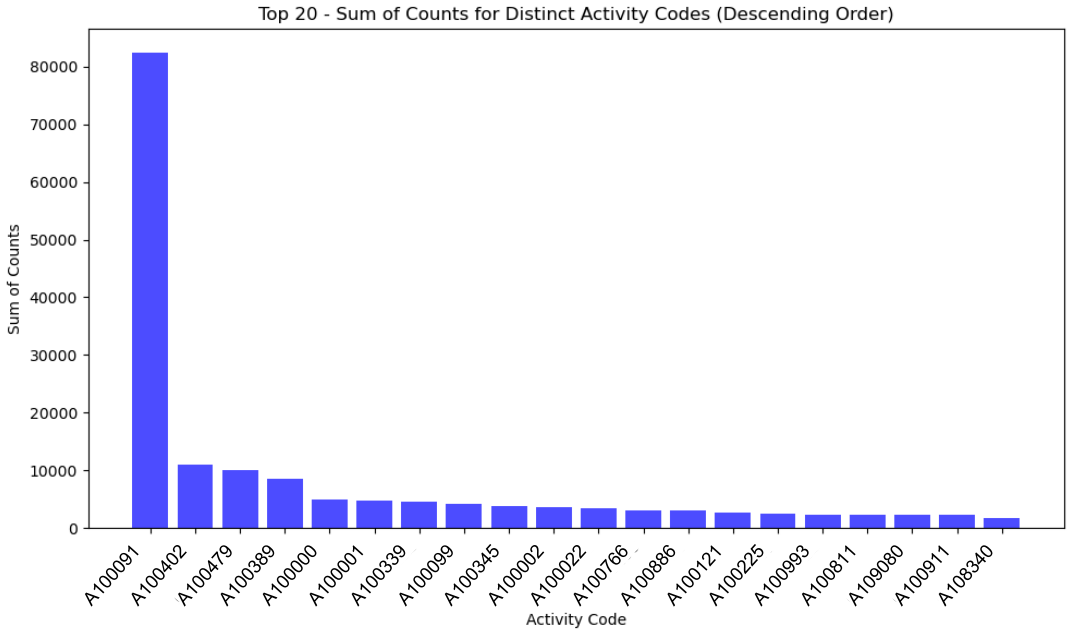

Bild 5-4: Anzahl verschiedener Fehlercode-Events nach einem Kassettenwechsel (Top 20)

Das Petri-Netz in **Fehler! Verweisquelle konnte nicht gefunden werden.** visualisiert den Fluss an Aktivitäten und Events in Systemen. Dabei wird ersichtlich, dass der Fehlertyp A10 am häufigsten nach einem Kassettenwechsel auftritt. Außerdem finden meist mehrere Kassettenwechsel statt, bevor es zu einem Fehler kommt.

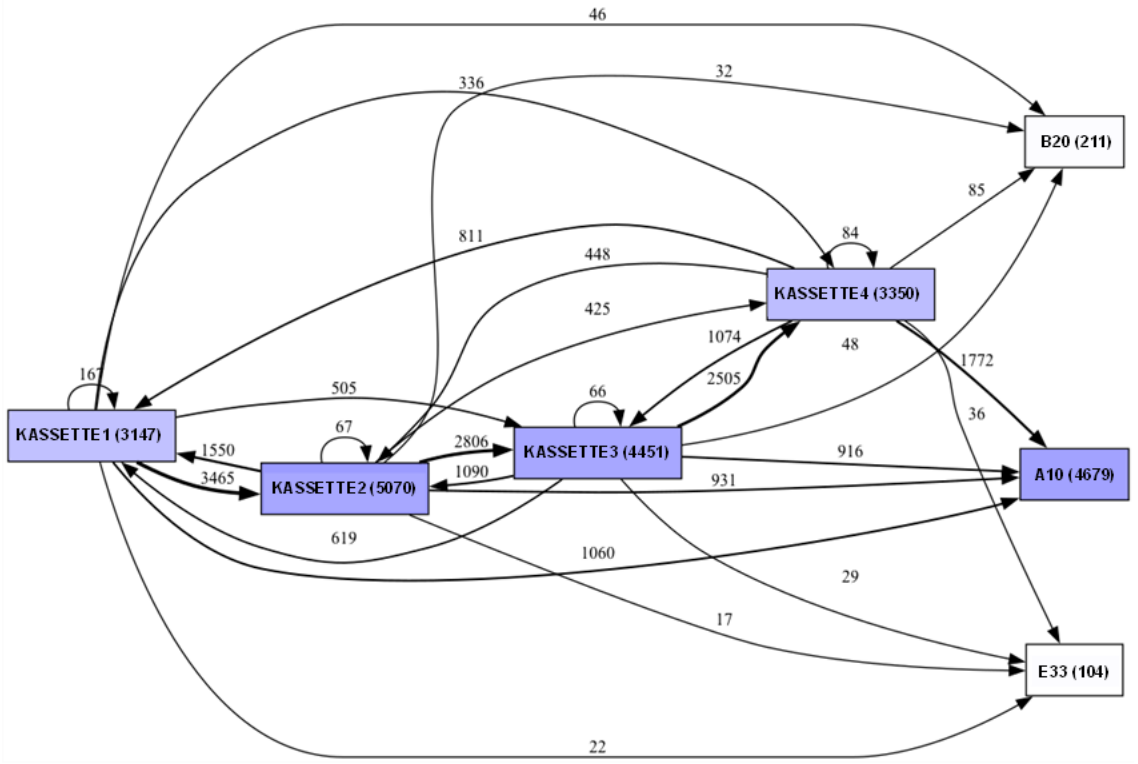

Bild 5-5: Petri-Netz zur Visualisierung der Kassettenwechsel- und Fehlerevents

Wie häufig treten Störungen durch (unsachgemäße) Bedienung auf und nach welchem zeitlichen Abstand?

Zur Beantwortung dieser Frage wurde eine Verteilung darüber aufgestellt, wie häufig Fehler nach 0 bis 250 Stunden auftreten (s. **Fehler! Verweisquelle konnte nicht gefunden werden.**). Am häufigsten treten Fehler zwischen 50 und 180 Stunden nach einem Kassettenwechsel auf.

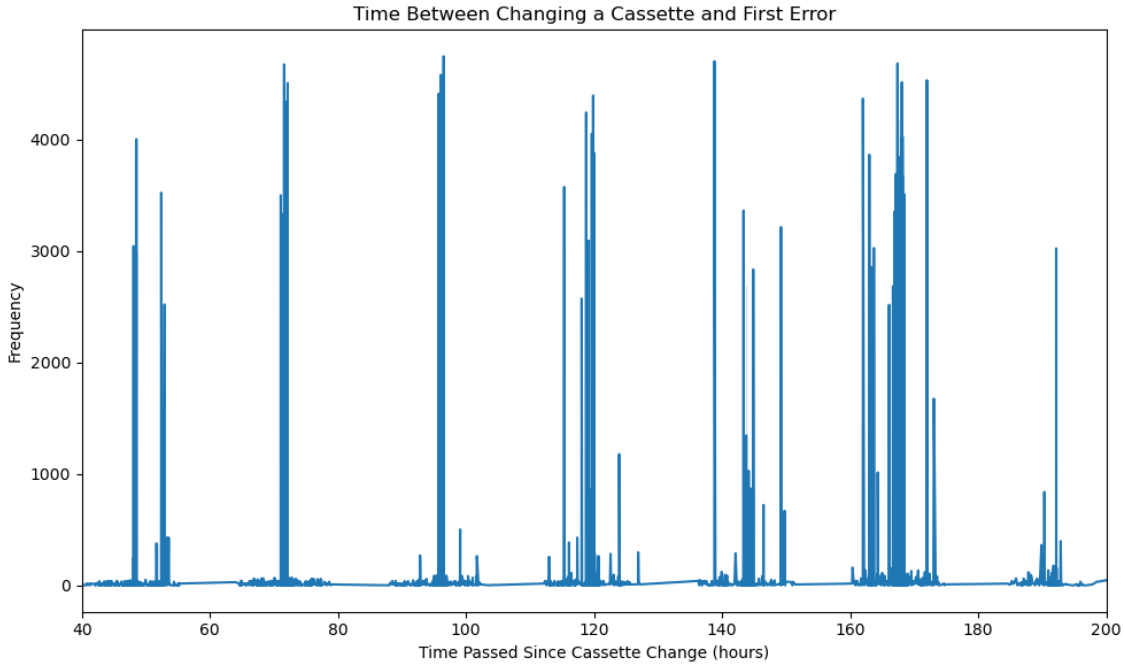

Bild 5-6: Zeit zwischen Kassettenwechseln

Nach der Umsetzung steht die Interpretation der Ergebnisse aus, die allerdings nicht Teil dieser Systematik ist. Es wird aber deutlich, dass die deskriptive Beschreibung der Daten die untersuchte Hypothese, dass eine (unsachgemäße) Operator-Bedienung häufig Fehler an den Systemen verursacht, noch nicht klar bestätigt werden kann, aber ein Zusammenhang zwischen den Kassettenwechseln und auftretenden Fehlern zu bestehen scheint. So ist z. B. die Aktivitätenfolge „Kassettenwechsel – Fehler“ recht stark ausgeprägt. Im nächsten Schritt könnte nun geprüft werden, ob signifikante Korrelationen bestehen und welche weiteren Faktoren, wie z. B. Service-Verträge und Währung, eine Rolle spielen. Die Sinnhaftigkeit von Anpassungen hinsichtlich der Systembedienung ist folglich weiter zu prüfen.

5.1.2 Fallbeispiel „Elektroklemmen“

Das zweite Fallbeispiel skizziert die Anwendung der Systematik in einem Unternehmen, welches Komponenten für die elektrische Verbindungstechnik sowie elektronische Bauteile für die Automatisierungstechnik herstellt. Als Produkt werden die Verbindungsklemmen näher betrachtet.

Die Produktmanager identifizierten folgenden Use Case: Identifikation der Bedürfnisse von Nutzern der Verbindungsklemmen, um Hinweise für die Produktplanung zu gewinnen. Im Schritt der **Definition der Anwendung** wurde der Use Case mit Hilfe der Leitfrage nach dem Analytics-Ziel und der Business-to-Analytics-Canvas (s. Bild 5-7) tiefer gelegt. Das übergreifende Ziel bestand in dem Verstehen der Nutzerbedürfnisse und -probleme, die nicht unbedingt direkt an das Unternehmen kommuniziert werden, sondern in den Communities geteilt werden. Die Definition des Analytics-Ziels *Nutzerbedürfnisanalyse* konnte zügig erfolgen. Eine Datenanalyse sollte die Frage beantworten, ob die Klemmen überwiegend positiv oder negativ bewertet werden und welche konkreten Aspekte (z. B. Features) im Zusammenhang mit der Bewertung genannt werden. Die Auswertung sollte dazu zum einen den Anteil positiver und negativer Produktmeinungen ausgeben, zum anderen die verschiedenen behandelten Themen, bzw. Produktaspekte, mit ihrer jeweiligen Bewertung. Als relevante Größen wurden die Themenschwerpunkte und Stimmungen (negativ, positiv) in den Nutzerkommentaren festgehalten. Unter den Anforderungen, bzw. Problemen, wurden lediglich die aufwändige Datenvorverarbeitung und die Sicherstellung der Betrachtung von Verbindungsklemmen genannt.

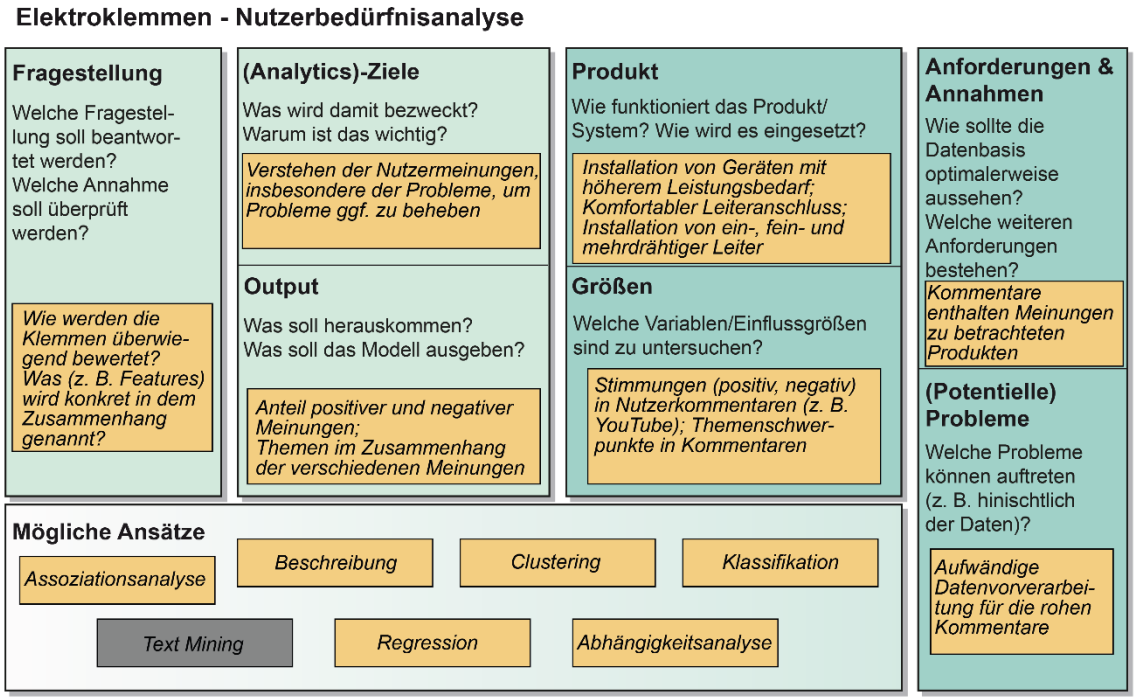

Bild 5-7: Ausgefüllte B2A-Canvas für die Nutzerbedürfnisanalyse

Nach der Beantwortung der Fragen nach dem Analytics-Problem und den persönlichen Präferenzen konnten die entscheidenden Faktoren für die spätere Auswahl der Analytics-Methoden bestimmt werden: *Text Mining* (Analytics-Problem) sowie *unüberwachtes Lernen* und *Einfache Implementierung* (Nutzeranforderungen).

Anschließend konnte im Schritt **Aufbau von Datenverständnis** durch das Ausfüllen des Data Analytics Canvas ein klareres Bild über die vorhandenen Daten gewonnen werden.

Das Unternehmen konnte vor allem das Internet als relevante Datenressource zur Beantwortung der Fragestellung identifizieren. Insbesondere drei Formate wurden als wertvolle Lieferanten von Nutzermeinungen erkannt: 1) Instagram, 2) Youtube und 3) Community-Seiten wie myDealz.de. Aufgrund der Vielzahl an Kommentaren und Meinungen zu Verbindungsklemmen wurde Youtube als erste Datenquelle definiert.

Im nächsten Schritt wurden die Daten anhand der Fragebögen beschrieben. Ergebnis dieser Beschreibung sind die Eigenschaften *unstrukturierter Text* und *normale Datenmenge*. Durch das Datenformat Text ergaben sich hinsichtlich der Qualität die üblichen Probleme wie Fehler und Inkonsistenten (z. B. Abkürzungen, Rechtschreibfehler, unvollständige Sätze).

Auf Basis der Faktoren aus Schritt 1 und 2 konnte der digitale Assistent im dritten Schritt, der **Modellauswahl**, für das Analytics-Problem Text Mining die relevanten Verfahren Latent Dirichlet Allocation (LDA), den lexikonbasierten Ansatz und BERT identifizieren. Im Rahmen dieser Fallstudie wurde entschieden, die ersten beiden Verfahren umzusetzen, da im ersten Schritt die möglichen Schwächen des BERT-Verfahrens, die Modellkomplexität und Rechenintensität, als Argument gegen dessen Nutzung verwendet wurden. Als Vorverarbeitungsmethoden wurden die vorgeschlagenen Textbereinigungstechniken übernommen sowie zur Normalisierung der Wörter die Methode Lemmatisierung. Eine externe Validierung war die vorgeschlagene Evaluierungskomponente. Die resultierende Pipeline ist in Bild 5-8 visualisiert.

Der letzte Schritt, die **Umsetzung**, brachte als Ergebnis der Vorauswahl insgesamt vier verschiedene Bibliotheken hervor. Nach näherer Betrachtung dieser Werkzeuge erwiesen sich die Bibliotheken nltk, sklearn und TextBlob als erste Wahl, da sie die benötigten Komponenten Textbereinigung, LDA und lexikonbasierter Ansatz einfach umsetzen.

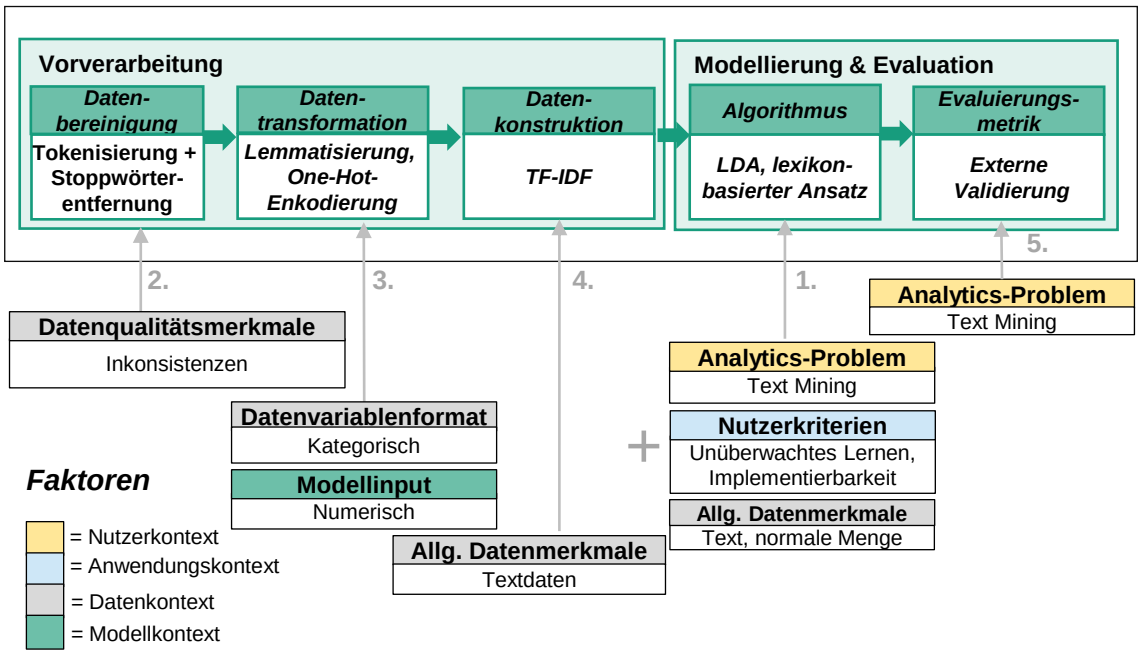

Bild 5-8: Ergebnis der Methodenauswahl für das Fallbeispiel „Elektroklemmen“

Darauf aufbauend wurde die Umsetzung der Analyse gestartet. Bevor die ausgewählten Pipeline-Komponenten mit Hilfe der Python-Bibliotheken implementiert wurden, mussten die Daten aus Youtube extrahiert werden. Für diesen Case wurde ein Video von dem Youtube-Kanal „Silver Cymbal“ gewählt, das eine neue Klemme (Inline 221) vorstellt. Über eine API wurden insgesamt ca. 800 Kommentare zu dem Video extrahiert und in einer CSV-Datei gespeichert. Danach wurden die Daten über die Bibliothek *nltk* entsprechend der ausgegebenen Komponenten vorverarbeitet: Tokenisierung, Lemmatisierung und Entfernung der Stoppwörter (mit Hilfe einer Standard-Stoppwörter-Liste). Mit nur wenigen Zeilen Code konnte mit Hilfe der Bibliothek *sklearn* ein LDA-Topic-Modell über die Daten berechnet werden. Die Anzahl der resultierenden Topics ist ein zu definierender Parameter und wurde in dem Fall auf 5, 10 und 15 gesetzt. Dabei erwiesen sich für den menschlichen Betrachter die Themen unter n=*5* als qualitativsten. Bild 5-9 zeigt die Themencluster mit ihren zehn relevantesten Schlüsselwörtern.

	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights
0	house	19.3	resistance	20.2	twist	19.5	work	13.9	circuit	18.1
1	type	19.2	test	13.4	lever	12.4	brand	11.3	push	12.8
2	price	13.1	spring	11.8	block	12.1	cable	9.3	heat	10.8
3	installation	9.2	conductor	9.0	terminal	11.6	copper	9.2	contact	9.3
4	cost	8.6	version	7.2	problem	9.5	people	7.3	power	8.8
5	wall	7.8	home	6.8	screw	7.8	guy	7.1	love	8.8
6	lot	6.1	bit	5.9	style	7.6	area	6.7	tape	7.9
7	standard	6.1	day	5.2	kind	7.1	space	6.7	look	7.5
8	point	5.8	end	5.2	switch	6.8	size	6.3	loss	6.6
9	outlet	5.2	insulation	5.0	case	4.7	rating	6.3	application	6.5

Bild 5-9: resultierende Themencluster aus dem Topic Modelling

Die Themen können anhand dieser Schlüsselwörter wie folgt benannt werden:

1) Kosten und Standards von Wand- und Anschlussinstallationen
2) Leiterwiderstand und Isolation
3) Drehhebeln und Schraubklemmen in verschiedenen Schalterarten
4) Kabeleigenschaften, Kupfer und Marke
5) Wärmeentwicklung und Leistung in Schaltkreisen

Anschließend wurde mit Hilfe von *TextBlob* über den lexikonbasierten Ansatz für jeden Kommentar ein Sentiment-Score berechnet, welcher die Größen Polarität und Subjektivität repräsentiert.

- Der Polaritätswert ist eine Gleitkommazahl innerhalb der Spanne [-1.0, 1.0], wobei -1 eine sehr negative Stimmung, +1 eine sehr positive Stimmung und 0 eine neutrale Stimmung anzeigt.
- Die Subjektivität ist eine Gleitkommazahl innerhalb der Spanne [0.0, 1.0], wobei 0.0 sehr objektiv und 1.0 sehr subjektiv ist.

Bild 5-10 zeigt einen Scatterplot, welcher die Kommentare entlang der Achsen Subjektivität und Polarität einordnet. Dabei fällt auf, dass der Großteil der Kommentare positiv eingestuft wird. Insgesamt werden 387 Kommentare positiv eingeschätzt (>0.0), 107 Kommentare negativ und 155 Kommentare neutral. Nur wenige erreichen einen sehr hohen, bzw. sehr niedrigen Score.

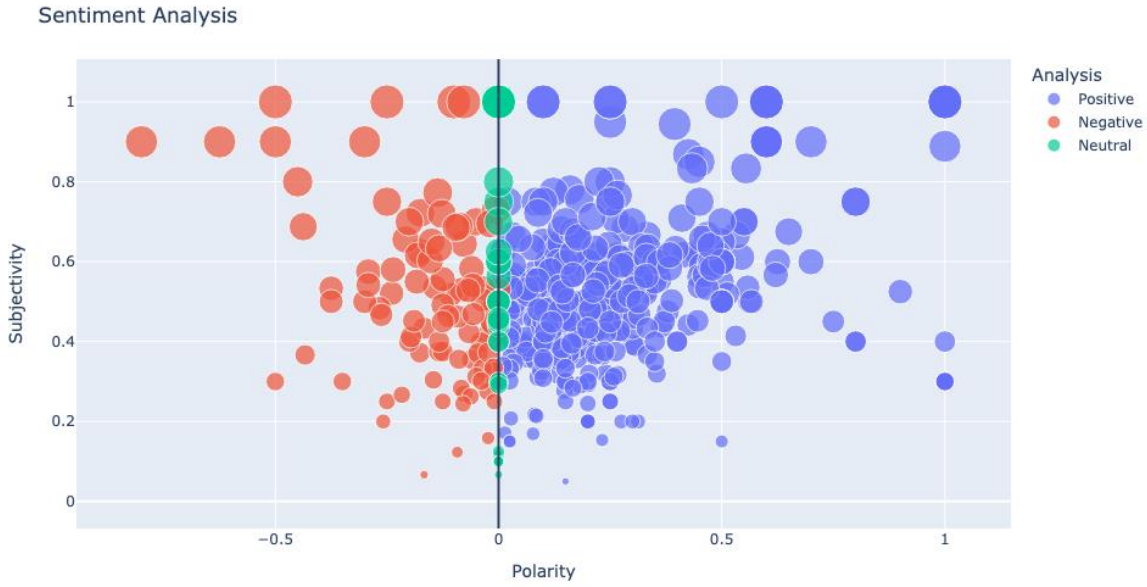

Bild 5-10: Scatterplot über die Sentiments der Kommentare

Ein weiterer Plot (s. Bild 5-11) zeigt die Verteilung der Kommentare im Zusammenhang mit ihren zugeordneten Hauptthemen auf. Dabei lässt sich kein Muster feststellen, welches auf eine themenabhängige Bewertung hindeutet. Bei genauerer Betrachtung der positiven Kommentare sticht das Thema 1 (Kosten und Standards von Wand- und Anschlussinstallationen) im Durchschnitt und absolut am deutlichsten hervor. Bei den negativen Kommentaren ist Thema 3 (Drehhebeln und Schraubklemmen in verschiedenen Schalterarten) relativ gesehen am häufigsten das führende Thema.

Bild 5-11: Scatterplot über die Sentiments und Topics der Kommentare

Bei genauerer Betrachtung der Kommentare anhand ihrer Schlüsselbegriffe sind in den Wortwolken (s. Bild 5-12), welche häufig auftretende Begriffe stärker hervorheben, nur teilweise Unterschiede zwischen den am negativsten (>-0.6) und am positivsten (>0.6) bewerteten Kommentaren erkennbar. Zum Beispiel sind Begriffe wie "bread", "ad" und "builder" sowohl bei den positiven als auch bei den negativen Kommentaren häufig anzutreffen. Darüber hinaus fallen offensichtlich positiv annotierte Begriffe wie „ease“ bei den als negativ eingestuften Kommentaren auf. Bei stichprobenartiger Prüfung der Kommentare stellt sich auch heraus, dass der Begriff meist in Verbindung mit einem Lob steht; andere Sätze in demselben Kommentar enthalten jedoch negative Andeutungen. Das deutet darauf, dass die lexikonbasierte Sentiment Analyse in dem Anwendungsfall zwischen verschiedenen Sentiment-Ausprägungen nicht gut differenzieren kann. Als nächster Schritt könnte somit das Testen eines anderen Verfahrens aus dem Baukasten (z. B. BERT) sinnvoll sein oder auch ein Anpassen der Vorverarbeitungsschritte.

Bild 5-12: Wortwolken der positiven und negativen Kommentare

5.1.3 Fazit

Anhand der beiden Anwendungsbeispiele konnte gezeigt werden, dass die Systematik in Form des digitalen Assistenten in der Praxis und für unterschiedliche Ziele und Probleme erfolgreich angewendet werden kann. Das heißt, dass das Durchlaufen der Schritte *Definition der Anwendung* bis *Umsetzung* mit Hilfe der Hilfsmittel die Planung eines passenden Datenanalyseprojektes ermöglicht und auch zu ersten Datenanalysenergebnissen und Erkenntnissen führt, welche anschließend im Kontext von neuen Produktanforderungen interpretiert werden oder als Ausgangspunkt für weitere Analysen dienen können. Die erzielten Funktionalitäten werden folglich erfüllt. Im ersten Fall konnte die untersuchte Hypothese noch nicht direkt widerlegt werden, sodass der vermutete Zusammenhang zwischen Kassettenwechseln und Fehlermeldungen zwar naheliegt, jedoch weitere Untersuchungen und ggf. ein Durchlaufen des Prozesses mit weiteren Fragestellungen erforderlich sind. Im zweiten Fall konnte die untersuchte Fragestellung, ob die Klemmen überwiegend positiv oder negativ bewertet werden, beantwortet und auch relevante Themen identifiziert werden. Dies kann bereits ein Ansatzpunkt für die Produktplanung und neue Anforderungen darstellen. Eine Validierung durch den Menschen ergab jedoch, dass die konkreten Aspekte, die im Zusammenhang mit den positiven und negativen Kommentaren genannt werden, noch nicht hinreichend signifikant sind, sodass sich auch hier weitere Iterationen als sinnvoll herausstellen. Ein Vorgehen zur Interpretation der Ergebnisse und Entwicklung von Ideen für neue Produktgenerationen bieten MEYER ET AL. [MFK+22a, DK23] und wird an dieser Stelle nicht weiter behandelt.

Insgesamt wird deutlich, dass die Systematik den Prozess von der Anwendungsdefinition bis zur Modell- und Toolauswahl erfolgreich unterstützt, indem mögliche Pipeline-Komponenten vorgeschlagen werden. Diese ermöglichen, schnell in die Umsetzung zu starten, Trends sowie Tendenzen zu erkennen und Anpassungen bei erneuten Durchläufen des Prozesses vorzunehmen. Die Phase der Umsetzung selbst wird nur geringfügig unterstützt, steht allerdings auch nicht im Fokus der Systematik. Die Anwendungsbeispiele verdeutlichen, dass die Modellauswahl nicht alle erforderlichen Schritte begleitet. Hierzu gehören beispielsweise grundlegende Bereinigungen und Datenmanipulationen, um Rohdaten einzulesen, zu integrieren und zu formatieren. Ebenso fehlen Aspekte wie die Erstellung von Visualisierungen zur Präsentation der Ergebnisse. Dazu ist zusätzliches Wissen oder Erfahrung notwendig. Nutzer bedürfen daher an dieser Stelle zusätzliche Unterstützung. Auch die Optimierung der Ergebnisse durch automatisierte Ansätze, wie beispielsweise *GridSearch* zur Bestimmung der besten Hyperparameter, wird nicht adressiert. Diese Arbeit verweist in diesem Zusammenhang auf Data-Science-demokratisierende Tools wie AutoML und No-Code-Tools (s. Kapitel 2.4).

5.2 Nutzenevaluation

Um den Nutzen der Systematik für potenzielle Anwender nach ihrer erfolgreichen Anwendung in zwei Fallbeispielen zu bewerten, wurden Interviews mit Produktexperten und

Data-Analytics-Anfängern geführt und ausgewertet. Die Interviews ermöglichen das Erfassen der fundierten Meinungen der adressierten Anwender.

In Abschnitt 5.2.1 wird das Vorgehen der Nutzenevaluation beschrieben. Folgend stellt der Abschnitt 5.2.2 die Ergebnisse dieser Evaluation dar. Im abschließenden Abschnitt 5.2.3 wird ein Fazit zur Evaluation gezogen.

5.2.1 Methodisches Vorgehen der Nutzenevaluation

Der Ansatz basierte auf einem semi-strukturierten Interviewleitfaden. Dieser umfasste eine Einleitung zur Erfassung grundlegender Hintergrundinformationen, eine Erläuterung des Zwecks des Interviews, Schlüsselfragen und einen Abschluss, der eine Nachbesprechung ermöglicht [MN07].

Jedem Interviewteilnehmer wurden nach einer ausführlichen Vorstellung des Prototypen-Tools anhand des Tools selbst, der Mockups und der exemplarischen Fallbeispiele Fragen gestellt zu (1) dem allgemeinen Nutzen der Systematik bzw. des Tools; (2) der Befähigungsleistung des Tools und der einzelnen Artefakte (Hilfsmittel); (3) den Stärken und Schwächen und möglichen Verbesserungen und (4) der Vertrauenswürdigkeit des Tools. Die Fragen und ihre Strukturierung sind in Tabelle 5-2 aufgeführt.

Tabelle 5-2: Interviewfragen zur Nutzenevaluation

Kernaspekt	**Fragen**
Nutzen	Wie hilfreich ist das vorgeschlagene Tool Ihrer Meinung nach als Lernmittel für Data Science in der Produktplanung und der Entwicklung geeigneter Pipelines für solche Use Cases? *1) gar nicht hilfreich bis 5) sehr hilfreich*
Befähigung	Sind Sie der Meinung, dass das Tool Sie zu erfolgreichen Analytics-Projekten für die PP befähigt? Wenn ja, wie schafft es einen Mehrwert?
	Trägt das Tool mit seinen Bestandteilen (Fragen, Templates und Wissensbasis) zu einem besseren Verständnis des gesamten Data-Analytics-Prozesses und seiner Abhängigkeiten bei? Bitte erklären Sie es.
Stärken/Schwächen	Was sind Ihrer Meinung nach Stärken und Schwächen des Tools?
Vertrauen	Wie vertrauenswürdig finden Sie das Tool und die daraus resultierenden Ergebnisse und Informationen (im Gegensatz zu anderen Tools, wie z. B. ChatGPT oder AutoML-Tools)?

Befragt wurden insgesamt fünf Experten und Expertinnen aus der Industrie aus zwei unterschiedlichen Gruppen: 1) datenaffine Produktplanungs-, bzw. Produktmanagementexperten und 2) Data Scientists mit max. 1-2 Jahren Berufserfahrung.

Tabelle 5-3 gibt einen Überblick über die Interviewpartner und ihre Selbsteinschätzung ihrer Data-Science-Kenntnisse.

Tabelle 5-3: Übersicht über die Interviewpartner

Nr.	Funktion	Data-Science-Kenntnisse (1 – keine Kenntnisse bis 5 – langjährige Erfahrung)
1	Produktmanager Steuerungssoftware	4
4	Produktmanager Elektronikverbindungen	2 +
5	Produktmanager Software	3
2	Data Scientist mit Fokus auf Servicedaten	3
3	(Junior) Data Scientist in der Forschung	2 +

Die Interviews wurden über ein Webkonferenz-Tool oder persönlich durchgeführt; jedes davon dauerte etwa 45 Minuten. Um die Authentizität und Integrität der Daten zu gewährleisten, wurde jedes Interview aufgezeichnet und wortwörtlich transkribiert.

Anschließend wurde eine strukturierende Inhaltsanalyse nach MAYRING [May94, May19] vorgenommen. Dieser Ansatz hat das Ziel, bestimmte Aspekte aus dem Material herauszufiltern und unter vorher festgelegten Ordnungskriterien einen Querschnitt durch das Material zu legen. Die Strukturierungsdimensionen orientieren sich an den Fragen und ihren Kernaspekten aus dem Interviewleitfaden. Unterhalb dieser Dimensionen wurden weitere Kategorien deduktiv aus dem Interviewmaterial herausgefiltert.

5.2.2 Ergebnisse der Nutzenevaluation

Die strukturierende Analyse der Interviewinhalte ist in Tabelle 5-4 zusammengefasst. Die vollständige Tabelle mit Zitaten befindet sich im Anhang A5. Nachfolgend wird auf die einzelnen Kernaspekte näher eingegangen.

Kernaspekt 1: **Nutzen**

Alle fünf Interviewpartner bestätigten den Nutzen der Systematik, bzw. des Tools, durch eine Bewertung von vier (hilfreich) bis fünf (sehr hilfreich). Interviewpartner 1,2 und 4 schätzten das Tool jeweils als hilfreiches Lernmittel ein, während Interviewpartner 3 und 5 dem Tool ein „sehr hilfreich“ attestierten. Dabei schränkte Interviewpartner 3 diese

höchste Bewertung auf das Lernen von Data Science in der Produktplanung ein, wohingegen das Tool als Lernmittel für das Entwickeln von Pipelines als „hilfreich“ angesehen wurde. Als Argument wurde am häufigsten (3 von 5 Interviewteilnehmer) genannt, dass keine Vorkenntnisse über Data-Science-Methoden erforderlich sind. Folgend mit zwei Nennungen ist eine gute Übersicht über die möglichen Anwendungen und Verfahren zu erwähnen. Vereinzelt sind Aspekte wie das Stellen der richtigen Fragestellungen, gute Dokumentation und Nachvollziehbarkeit des Prozesses sowie „learning on the job“ hervorgehoben worden.

Kernaspekt 2: **Befähigung**

Dass das Tool zu erfolgreichen Analytics-Projekten befähigt, wurde von allen Interviewpartnern bestätigt. 3 von 5 Interviewten nannten zum einen den Überblick über die Techniken und den Lösungsraum, zum anderen den systematischen Ansatz und Strukturierung als klare Mehrwerte. Mit jeweils zwei Nennungen kamen die Interviewpartner zu dem Schluss, dass die Einschränkung der Verfahren und die schnellere Umsetzung von Projekten einen Mehrwert lieferten. Interviewpartner 2 erwähnte zusätzlich die Dokumentation und Kommunikation, welche durch das Tool ermöglicht werden. Darüber hinaus stellte Interviewpartner 4 fest, dass das Tool die Beantwortung der wichtigen Fragestellungen sicherstellt und somit zur Befähigung zu erfolgreichen Analytics-Projekten beiträgt.

Auch die Frage, ob das Tool mit seinen verschiedenen Bestandteilen ein besseres Verständnis des Data-Analytics-Prozesses fördert, wurde von allen Interviewten bejaht. Mehr als einmal wurden hierbei die Einfachheit, z. B. bei den Erklärungen, und die Visualisierung der Kausalkette, d.h. die Nachvollziehbarkeit der Verknüpfungen, die zu einer Methodenauswahl führen, erwähnt. Weitere zu nennende Aspekte in dem Kontext sind die Förderung eines Verständnisses über die Data-Science-Logik, Nachvollziehbarkeit bei den Fragen, eine leichtere Einordnung in das Gesamtvorgehen und die Bereitstellung von Werkzeugen zur gemeinsamen Erarbeitung von Projektartefakten.

Kernaspekt 3: **Stärken und Schwächen**

Als Stärken wurden häufig bereits genannte Aspekte nochmal hervorgehoben, darunter die einfache Sprache, welche vor allem für Nicht-Experten die Inhalte im richtigen Detailgrad vermittelt, die Arbeitsersparnis durch den eingeschränkten Lösungsraum und die Abnahme von Aufgaben wie der Methodenauswahl. Als weitere Stärken wurden das Lernen während der Arbeit, der geführte Prozess, die intuitive und nachvollziehbare Anwendung und die Transparenz aufgefasst.

Neben den Stärken wurden auch einige Schwächen aufgedeckt. Interviewpartner 2 und 3 nannten hier beide, dass das Tool noch keine absolute Erklärbarkeit bietet, in dem Sinne, dass das Tool nicht offenlegt, wie es (im Backend) konkret die Vorschläge erarbeitet, sondern nur die Faktoren aufzeigt, die die Entscheidung beeinflussen. Darüber hinaus gab es weitere Einzelnennungen; beispielsweise fehlte Interviewpartner 1 ein umfassenderer

Überblick über die relevante Datenwelt in Form von Beschreibungen zu Beginn der Phase des Datenverständnisses. Interviewpartner 3 machte auf die Abhängigkeit des Tools von der Wissensbasis aufmerksam und mit welcher Sorgfalt sie erstellt wurde. Interviewpartner 4 erwähnte als potenzielle Schwäche den Medienbruch, der sich durch den Sprung auf ein digitales Whiteboard ergibt, sowie ein teilweise zu schneller, bzw. unerwarteter, Ansichtenwechsel, insbesondere die Pipeline-Übersicht am Ende. Interviewpartner 5 gab an, dass das Tool die Unschärfen des Data-Analytics-Prozesses nicht abbildet. So würden viele Entscheidungen im Prozess nicht immer allein von klar definierten Parametern, wie der Datenmenge oder der Qualität, abhängen. Dies sei aber für ein Lern-Tool nicht in erster Linie relevant.

Kernaspekt 4: **Vertrauen**

Zuletzt schätzten alle Interviewten das vorgestellte Tool als vertrauenswürdig ein. Insbesondere im Vergleich zu anderen Tools, wie ChatGPT, wurde das Tool als vertrauenswürdiger bewertet. Als Gründe wurde mehrfach die fundierte Wissensbasis genannt, auf die anfangs aufmerksam gemacht wird, und das dahinterliegende Expertenwissen. Diese stünden der statistischen, wahrscheinlichkeitsbasierten Generierung von Ergebnissen gegenüber und erweckten deutlich mehr Vertrauen. Außerdem würde das Vertrauen gesteigert dadurch, dass Entscheidungen transparent dargestellt werden und der Nutzer befähigt wird, präzise Angaben zu machen, sodass das Tool valide Vorschläge aufzeigen kann.

Tabelle 5-4: Strukturierende Analyse der Interviews

Thema	Subthema	Interviewte
Nutzen	Gute Übersicht (über mögliche Anwendungen und Verfahren)	1,3
	Richtige Fragestellungen	3
	Dokumentation des Prozesses auch für Seniorexperten	2
	Keine Vorkenntnisse über Methoden erforderlich	1, (2,5)
	Nachvollziehbarkeit des Prozesses	5
	"Learning on the job"	5
Befähigung	Einschränkung der Modelle/des Lösungsraums	1, 5
	Dokumentation & Kommunikation	2
	Überblick über Techniken/Lösungsraum	1,2, 3
	Schnellere Umsetzung	3, 5
	Sicherstellung über Beantwortung wichtiger Fragen	4
	Systematischer Ansatz & Struktur	2, 4, 5
	Verständnis über Großteil der Logik	1
	Einfachheit (z. B. Erklärungen)	2, 5
	Verknüpfung/Visualisierung der Kausalkette	1, 3, 5
	Erleichterung des Mappings	3
	Nachvollziehbare Fragen	4
	Einordnung in Gesamtvorgehen möglich	4
	Werkzeug zur gemeinsamen Bearbeitung durch Templates	5
Stärken	Fokussiertes Lernen durch Einschränkung der relevanten Verfahren	1
	Arbeitsersparnis	1, 2
	Learning on the job	1
	Geführter Prozess	3
	Einfache Sprache (richtiger Detailgrad für Nicht-Experten)	3, 5
	Intuitive und nachvollziehbare Anwendung	4
	Visual Tools für Nicht-Experten	4
	Transparenz	4
Schwächen	Einstieg in die Datenwelt fehlt	1

	Noch transparentere Zusammenfassung, keine absolute Explainability	2, 3
	Abhängigkeit von (Stand) d. Knowledge Base	3
	Medienbruch	4
	Vorschlagsansichten teilweise zu schnell	4
	Tool bildet Unschärfen des DS-Prozesses nicht ab	5
Vertrauen	Logisch, nicht zufällig	1
	Befähigung der Nutzer (gute Angaben zu machen)	1
	Fundierte Wissensbasis	1, 4
	Expertenwissen	2, 5
	Entscheidungen werden transparent	5

5.2.3 Fazit zur Nutzenevaluation

Die Nutzenevaluation der Systematik in Form des Lernassistenten liefert positive Ergebnisse. Alle Interviewpartner, trotz unterschiedlicher Data-Science-Kenntnisse, stuften das Tool als Lernmittel mindestens als „hilfreich“ ein und bestätigten somit einen Nutzen. Auch die Befähigungseigenschaft wurde von allen Befragten bekräftigt. Dabei hoben sie hervor, dass das Tool zu erfolgsversprechenden Analytics-Projekten qualifiziert, indem es den Nutzer sehr systematisch und strukturiert durch den Prozess leitet. Gleichzeitig bietet es einen guten Überblick über einen relevanten, eingeschränkten Lösungsraum, was nicht zuletzt ein stärker fokussiertes Lernen und eine schnellere Umsetzung ermöglicht. Den Bestandteilen des Tools bescheinigten die Befragten, dass sie zu einem besseren Verständnis des Prozesses und seiner Abhängigkeiten beitragen. Die Fragen seien sehr nachvollziehbar und ermöglichten an jeder Stelle eine Einordnung in das Gesamtvorgehen. Die Templates förderten eine gemeinsame Erarbeitung von relevanten Data-Analytics-Faktoren innerhalb des Projektteams. Verknüpfungen würden durch die zusammenfassenden Visualisierungen sichtbar.

Neben der Vielzahl an genannten Stärken, wurden Schwächen meist nur einmalig genannt, was darauf hindeutet, dass diese Schwächen eher individuell und spezifisch für einzelne Nutzer sind. Auffällig war hier am ehesten der Wunsch nach noch stärkerer Transparenz. Hieraus ergibt sich weiterer Forschungsbedarf.

Die abschließende Frage hinsichtlich der Vertrauenswürdigkeit des Tools wurde von allen Interviewpartnern positiv beantwortet. Demnach sei das Tool insbesondere im Vergleich zu anderen automatisierten Tools durch die Experten-basierte Wissensbasis deutlich vertrauenswürdiger.

6 Zusammenfassung, Limitationen und Ausblick

Aufgrund der neusten technologischen Entwicklungen haben produzierende Unternehmen einen wachsenden Zugang zu umfangreichen Betriebsdaten ihrer Produkte. In der strategischen Produktplanung besteht die Möglichkeit, diese Daten mithilfe verschiedener Data-Analytics-Lösungen zu analysieren. So können Hersteller besser verstehen, wie ihre Produkte verwendet werden und wie sie funktionieren. Die gewonnenen Erkenntnisse können in neue Produktanforderungen oder neue Produktideen überführt werden. Der Prozess der Datenanalyse stellt für produzierende Unternehmen, denen es häufig an Data-Science-Experten fehlt, jedoch eine bedeutende Herausforderung dar. Das Thema dieser Arbeit ist daher die Datenanalyse in der betriebsdatengestützten Produktplanung.

Nach der thematischen und forschungsmethodischen Einführung der Arbeit in **Kapitel 1** wird in **Kapitel 2** eine detaillierte Problemanalyse präsentiert. In dieser zeigt sich, dass die betriebsdatengestützte Produktplanung als Schnittpunkt zwischen strategischer Produktplanung, Data Analytics und Cyber-physischen Systemen eine Vielzahl an Potenzialen mit sich bringt, darunter z. B. den Erhalt von objektiven Informationen über Zustand, Störungen und Bedienung der Produkte während des normalen Betriebs und der Möglichkeit, das Produkt nutzstiftend zu optimieren. Allerdings belegen Studien, dass die Komplexität des Datenanalyseprozesses Unternehmen, insbesondere kleine und mittelständische, vor große Herausforderungen stellt. Ein wesentlicher Grund dafür sind das Fehlen von passenden Kompetenzen und Know-How im Unternehmen. Um diese Herausforderungen zu adressieren, ist die Demokratisierung von Data Science durch AutoML, No-Code-Tools, aber auch Schulungen und Lernmittel, ein wichtiger Hebel. Um die Herausforderungen des Data-Analytics-Prozesses im Detail zu verstehen, wurde anschließend der Prozess genau beleuchtet. Dabei stellte sich heraus, dass sich ein generischer Prozess bestehend aus den Schritten 1) Anwendungsdefinition, 2) Datenverständnisaufbau, 3) Modellauswahl und 4) Umsetzung für den Datenanalyseprozess in der betriebsdatengestützten Produktplanung eignet. Eine umfassende Literaturstudie offenbart diverse Herausforderungen und Ausgestaltungsfaktoren für jeden dieser Schritte. Diese werden abschließend in Ziele an die zu entwickelnde Systematik überführt.

In **Kapitel 3** wird auf Basis der identifizierten Ziele der Stand der Forschung untersucht. Zunächst werden dabei allgemeine Ansätze zur Betriebsdatenanalyse in der Produktplanung und -entwicklung analysiert. Dabei zeigt sich, dass kein bestehender Ansatz die Ziele aus Kapitel 2 realisiert. Im Anschluss werden ergänzende Ansätze für die jeweiligen Datenanalyseprozessschritte Anwendungsdefinition, Datenverständnis sowie Methoden- und Toolauswahl betrachtet. Diese Ansätze beanspruchen nicht, sämtliche Anforderungen zu erfüllen, jedoch zielen sie darauf ab, Lücken in den allgemeinen Ansätzen zu behandeln. Die Analyse überprüft daher, ob durch die Kombination von allgemeinen und ergänzenden Ansätzen die Ziele erfüllt werden können. Dabei wird festgestellt, dass auch

dies nicht der Fall ist. In Anbetracht dessen wird am Ende des Kapitels der noch ausstehende Handlungsbedarf festgelegt.

Die entwickelte Systematik wird in **Kapitel 4** vorgestellt. Sie besteht aus vier Bestandteilen:

1) Der **Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung** gibt an, mit welchen Schritten Datenanalysen in der betriebsdatengestützten Produktplanung umgesetzt werden können. Damit stellt er die allgemeine Pipeline für Datenanalyseprojekte dar, welche durch verschiedene Komponenten in Form von Anwendungen, Datenquellen und Analyse-Techniken ausgestaltet werden können. Der Referenzprozess basiert auf dem Referenzprozess für die betriebsdatengestützte Produktplanung und bettet sich damit in einen größeren Rahmen einschließlich der Planung und Verwertung von Betriebsdaten-Analysen ein.

2) Der **Analytics-Baukasten für die betriebsdatengestützte Produktplanung** strukturiert sich entlang des Referenzprozesses für die Datenanalyse in der betriebsdatengestützten Produktplanung und stellt entlang der dadurch resultierenden Dimensionen relevante Lösungskomponenten zur Use-Case-spezifischen Ausgestaltung und Konkretisierung einer allgemeinen Data Analytics Pipeline zur Verfügung.

3) Das **Vorgehen zur Datenanalyse in der betriebsdatengestützten Produktplanung** beschreibt ausführlich, wie Betriebsdaten-Analysen in der strategischen Produktplanung systematisch und erfolgreich vorbereitet und durchgeführt werden können und konkretisiert dabei den zuvor eingeführten Referenzprozess durch Einführung verschiedener Werkzeuge. Das Vorgehen befähigt die Anwender, die passenden Lösungskomponenten des Analytics-Baukastens zu identifizieren.

4) Das **Assistenz- und Lerntool für die Datenanalyse in der betriebsdatengestützten Produktplanung** setzt die aus den zuvor beschriebenen Elementen bestehende Systematik in Form eines digitalen Lernassistenten um, welcher seinen Anwendern eine intuitive und geführte Entwicklung einer erfolgsversprechenden Data Analytics Pipeline ermöglicht.

In der abschließenden Kriterien-basierten Analyse wird festgestellt, dass die Systematik die an sie gestellten Ziele erfüllt.

In **Kapitel 5** werden die Anwendung und Evaluation der Systematik beschrieben. Die Systematik in Form des Tools wurde in zwei verschiedenen Fallbeispielen mit unterschiedlichen Produkten und Zielen erfolgreich angewendet. Die Nutzenevaluation mit insgesamt fünf potenziellen Nutzern der Systematik ergab einen durchschnittlichen Nutzen von 4,4 auf einer Skala von 1 *gar nicht hilfreich* bis 5 *sehr hilfreich*. Darüber hinaus

bestätigten alle Interviewpartner, dass das Tool sie zu erfolgreichen Analytics-Projekten befähigt und zu einem besseren Verständnis des Prozesses beiträgt. Auch die gesteigerte Vertrauenswürdigkeit des Tools wurde bekräftigt, vor allem im Vergleich zu modernen Tools wie ChatGPT. Neben Stärken des Tools wurden auch Schwächen offensichtlich, die jedoch nur vereinzelt genannt wurden, allerdings gute Anhaltspunkte für die Limitationen, bzw. Weiterentwicklung des Tools, bieten.

Zusammengefasst erfüllt die entwickelte Systematik die an sie gestellten Ziele und kann in Folge der Anwendungs- und Nutzenevaluation in ersten Ansätzen als validiert erklärt werden.

Die wesentliche **Limitation** ist der geringe Umfang der Anwendungsbeispiele mit lediglich zwei verschiedenen Zielstellungen, was eine allgemeingültige Validität der Systematik nicht versichern kann. Die Evaluation liefert einige glaubwürdige Hinweise, dass die Systematik zur Datenanalyse in der betriebsdatengestützten Produktplanung befähigen kann. Allerdings ist die Interviewstudie mit fünf Teilnehmern weit von einer repräsentativen Studie entfernt und kann daher nur Signale für die Nützlichkeit und den praktischen Wert der Systematik geben. Darüber hinaus wurden die potenziellen Nutzer anhand eines konkreten Anwendungsbeispiels durch das Tool geführt und auf Basis dessen nach ihrer Einschätzung zum Nutzen gefragt. Noch realistischer können Nutzer die Systematik sicherlich nach Durcharbeit eines eigenen Anwendungsfalls mit Hilfe des Tools bewerten. Eine derartig aufwändige Evaluation konnte im Rahmen der vorliegenden Arbeit nicht durchgeführt werden. Des Weiteren besitzen auch die einzelnen Elemente der Systematik ihre jeweils eigenen Limitationen, welche zum Teil den entsprechenden Veröffentlichungen im Detail entnommen werden können. Nachstehend werden daher nur ausgewählte Limitationen erwähnt:

1) **Referenzprozess für die Datenanalyse in der betriebsdatengestützten Produktplanung:** Der Prozess ist auf den ersten Blick ein generischer Data-Analytics-Prozess und wird erst durch das Vorgehen für die Datenanalyse in der betriebsdatengestützten Produktplanung ausspezifiziert. Zudem berücksichtigt er nicht wichtige Phasen wie die Datenakquise und das Deployment. Die Umsetzung wird auch nur zu einem gewissen Teil unterstützt. Hier setzen jedoch sehr gut andere, bereits existierende Tools wie AutoML-Werkzeuge an.

2) **Analytics-Baukasten für die betriebsdatengestützte Produktplanung**: Die im Baukasten aufgeführten Komponenten sind das Ergebnis einer zwar umfangreichen, aber thematisch eingeschränkten Literaturrecherche und einer Umfrage mit einer relativ kleinen Stichprobe. Das bedeutet, dass möglicherweise ganz neue Use Cases und eher spezielle Verfahren fehlen. Jedoch ist der Baukasten als initiale Wissensbasis gedacht, welche durch weitere Ziele, Algorithmen und Verfahren leicht zu ergänzen ist. Außerdem ist der manuelle Aufwand zu erwähnen, der mit einer solchen Identifikation von relevanten Techniken verbunden ist.

3) **Vorgehen zur Datenanalyse in der betriebsdatengestützten Produktplanung**: Die Auswahllogik im Schritt der Methodenauswahl kann auf den ersten Blick recht komplex wirken und wurde daher auch im Rahmen des digitalen Tools automatisiert. Darüber hinaus kann sie sicherlich nicht alle Unschärfen des Data-Science-Prozesses abbilden, d.h. erfolgreiche Pipelines sind nicht nur von den berücksichtigten Einflussfaktoren abhängig, sondern teilweise von vielen Weiteren, welche oftmals nur durch die Erfahrung eines geübten Data Scientists oder komplexen Algorithmen (z. B. auf dem Meta-Lernen) erkannt werden können. Da die Systematik jedoch mehr als Befähigungs- und Lerninstrument und weniger als ergebnisorientiertes Hilfstool zu verstehen ist, sei diese Limitation hier nur der Vollständigkeit halber aufgeführt.

4) **Assistenz- und Lerntool für die Datenanalyse in der betriebsdatengestützten Produktplanung**: Das Tool ist stark abhängig von der dahinterliegenden Wissensbasis und deren Aktualitäts- und Qualitätsstand sowie Umfang. Somit sind die zu planenden Use Cases und die möglichen Ergebnisse in Form von Pipelines begrenzt.

Vor diesem Hintergrund ergibt sich **weiterer Forschungsbedarf**. Um die Limitationen hinsichtlich des Forschungsdesigns der vorliegenden Arbeit zu adressieren, bedarf es weiterer Studien, in denen die Anwendung und Evaluation vertieft werden. Die Systematik sollte dabei bei möglichst vielen verschiedenen Use Cases aus der Produktsicht eingesetzt werden, um Unterschiede in den Ergebnissen und mögliche fehlende Komponenten festzustellen. Zudem sollte eine große Nutzerstudie durchgeführt werden, bei der Nutzer beim Einsatz der Systematik, bzw. des Tools, beobachtet und befragt werden. Zusätzlich könnte eine vergleichende Studie sinnvoll sein, in welcher die Nutzung des Assistenztools der von anderen existierenden Tools, wie beispielweise ChatGPT-basierenden Werkzeugen, gegenübergestellt wird.

Inhaltlich sind folgende Forschungsbedarfe besonders hervorzuheben: (1) Es sollte eine Nutzung, bzw. Integration, von modernen Texttechnologien wie große Sprachmodelle (large language models) geprüft werden. Dabei sind sowohl die Einsatzmöglichkeiten innerhalb des Tools, um die Assistenzfunktionen zu erweitern, als auch die zur Entwicklungsunterstützung der Wissensbasis interessant. Solche Technologien können letztendlich auch dazu genutzt werden, um die Aktualität des bereitgestellten Wissens zu gewährleisten. (2) Eine Limitation der Systematik besteht darin, dass nicht der gesamte Data-Analytics-Prozess einschließlich Datenakquise, Umsetzung und Deployment adressiert wird. Da es sich hierbei auch um unterstützungsbedürfende Schritte handelt, sollte die Systematik um diese Schritte zur besseren Prozessdurchgängigkeit erweitert werden. (3) Um die Demokratisierung von (Industrial) Data Analytics voranzutreiben, spielen Trainings und verschiedene Lernkonzepte eine große Rolle. Daher sollte geprüft werden, in welche Form von Konzept und in welcher Art und Weise die Systematik integriert werden kann, um den größtmöglichen Lernerfolg für die sogenannten Citizen Data Scientists zu ermöglichen. (4) Die Evaluation hat hervorgebracht, dass die Systematik nicht nur

Anfängern einen Nutzen stiftet, sondern durchaus auch Seniorexperten, in Bezug auf Dokumentation beispielsweise. Aus diesem Punkt erwächst der Forschungsbedarf, die Systematik, bzw. das Tool, an verschiedene Bedürfnisse und Lernstile anzupassen. Denn je individueller ein Tool auf seine Nutzer eingehen kann, desto motivierter und lernfähiger sind diese.

Abkürzungsverzeichnis

AutoML	Automated Machine Learning
BMBF	Bundesministerium für Bildung und Forschung
BMWi	Bundesministerium für Wirtschaft und Klimaschutz
CPS	Cyber Physical Systems
CRM	Customer Relationship Management
CRISP-DM	Cross Industry Standard Process for Data Mining
DIKW	Daten – Information – Wissen – Weisheit
DS	Design Science
DS-Pipeline	Data Science Pipeline
DSR	Design Science Research
DSRM	Design Science Research Methodology
IoT	Internet of Things
IS	Informationssysteme
KDD	Knowledge Discovery in Databases
KI	Künstliche Intelligenz
KMU	Kleine und mittelständische Unternehmen
LMS	Least Mean Squares
LSA	Latent Semantic Analysis
MDS	Multidimensionale Skalierung
MES	Manufacturing Execution System
ML	Computer Aided Design
NLP	Natural Language Processing
NSF	National Science Foundation
OECD	Organisation für wirtschaftliche Zusammenarbeit und Entwicklung
PCA	Principal Component Analysis
SDM	Semantic Data Mining

SMM Semantic Meta Mining

SVM Support Vector Machines

Literaturverzeichnis

VDI/VDE 3714 Blatt 4 - Implementation and operation of big data applications in the manufacturing industry - Analysis process classes, 2021

[AAA+14] ABDUL RAWOOF PINJARI; AKBAR BAKHSHI ZANJANI; AAYUSH THAKUR; ANISSA NUR IRMANIA; M. KAMALI; JEFFREY BRADFORD SHORT; DAVE PIERCE; LISA PARK: Using Truck Fleet Data in Combination with Other Data Sources for Freight Modeling and Planning, 2014

[ABB+11] AGRAWAL, D.; BERNSTEIN, P.; BERTINO, E.; DAVIDSON, S.; DAYAL, U.; FRANKLIN, M.; GEHRKE, J.; HAAS, L.; HALEVY, A.; HAN, J.: Challenges and opportunities with Big Data 2011-1, 2011

[ABB+19] AMERSHI, S.; BEGEL, A.; BIRD, C.; DELINE, R.; GALL, H.; KAMAR, E.; NAGAPPAN, N.; NUSHI, B.; ZIMMERMANN, T.: Software engineering for machine learning: A case study. In: IEEE (Hrsg.): 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 2019, S. 291–300

[Abo20] ABOU EDDAHAB, F.: Exemplifying smart functions for a next generation data analytics toolbox, Delft University of Technology, 2020

[ABU+14] ALBERS, A.; BURSAC, N.; URBANEC, J.; LÜDCKE, R.; RACHENKOVA, G.: Knowledge Management in Product Generation Development – an empirical study. DFX 2014: Proceedings of the 24th Symposium Design for X: Bamburg, Germany 1-2 Oktober 2014, 2014, S. 13–24

[ACK+19] AHUJA, R.; CHUG, A.; KOHLI, S.; GUPTA, S.; AHUJA, P.: The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science, (152), 2019, S. 341–348

[ACP22] ASHMORE, R.; CALINESCU, R.; PATERSON, C.: Assuring the Machine Learning Lifecycle. ACM Computing Surveys, (54)5, 2022, S. 1–39

[ACS18] ALMQUIST, E.; CLEGHORN, J.; SHERER, L.: The B2B elements of value. Harvard business review, (96)3, 2018, S. 18

[ADG+09] ADELT, P.; DONOTH, J.; GAUSEMEIER, J.; GEISLER, J.; HENKLER, S.; KAHL, S.; KLÖPPER, B.; KRUPP, A.; MÜNCH, E.; OBERTHÜR, S.; PAIZ, C.; PODLOGAR, H.; PORRMANN, M.; RADKOWSKI, R.; ROMAUS, C.; SCHMIDT, A.; SCHULZ, B.; VÖCKING, H.; WITKOWSKI, U.; WITTING, K.; ZNAMENSHCHYKOV, O.: Selbstoptimierende Systeme des Maschinenbaus - Definitionen, Anwendungen, Konzepte. HNI-Verlagsschriftenreihe, 2009

[AEF+12] ARTIKIS, A.; ETZION, O.; FELDMAN, Z.; FOURNIER, F.: Event processing under uncertainty. In: Behrend, A. (Ed.): Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems. the 6th ACM International Conference, 7/16/2012 - 7/20/2012, Berlin, Germany, ACM Conferences, ACM, New York, NY, 2012, pp. 32–43

[AG07] AHLEMANN, F.; GASTL, H.: Process model for an empiracally grounded reference model construction: Reference modeling for business systems analysis. IGI Global, 2007, S. 77–97

[AG99] ANDRÉ BOURDREAU; GUY COUILLARD: Systems Integration and Knowledge Management. Information Systems Management, (16)4, 1999, S. 24–32

[AHK01] AGGARWAL, C. C.; HINNEBURG, A.; KEIM, D. A.: On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: van Bussche, J. den; Vianu, V. (Hrsg.): Database theory - ICDT 2001 00 – 8th international conference, London, UK, January 4-6, 2001. Proceedings. Lecture Notes in Computer Science, Band 1973, Springer, Berlin, 2001, S. 420–434

[Ale13] ALEXANDRU ADRIAN ȚOLE: Big Data Challenges. Database Systems Journal, (4), 2013, S. 31–40

[ALW+18] ATAT, R.; LIU, L.; WU, J.; LI, G.; YE, C.; YANG, Y.: Big Data Meet Cyber-Physical Systems: A Panoramic Survey. IEEE Access, (6), 2018, S. 73603–73636

[AN00] ALPAR, P.; NIEDEREICHHOLZ, J. (Hrsg.): Data Mining im praktischen Einsatz – Verfahren und Anwendungsfälle für Marketing, Vertrieb, Controlling und Kundenunterstützung. Business Computing, Vieweg+Teubner Verlag, Wiesbaden, 2000

[Ao09] ASHTON, K.; OTHERS: That ‘internet of things’ thing. RFID journal, (22)7, 2009, S. 97–114

[ARB+17] ALBERS, A.; RAPP, S.; BIRK, C.; BURSAC, N.: Die Frühe Phase der PGE - Produktgenerationsentwicklung. Stuttgarter Symposium für Produktentwicklung, SSP 2017, Fraunhofer Verlag, 2017, S. 345–354

[AS94] AGRAWAL, R.; SRIKANT, R.: Fast Algorithms for Mining Association Rules in Large Databases: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB ’94, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1994, S. 487–499

[ASB16] ALMQUIST, E.; SENIOR, J.; BLOCH, N.: The elements of value. Harvard business review, (94)9, 2016, S. 47–53

[Aue04] AUER, C.: Performance Measurement Für das Customer Relationship Management – Controlling des IKT-Basierten Kundenbeziehungsmanagements. Wirtschaftswissenschaften Ser, Deutscher Universitäts Verlag, Wiesbaden, 2004

[BAK18] BZDOK, D.; ALTMAN, N.; KRZYWINSKI, M.: Statistics versus machine learning. Nature methods, (15)4, 2018, S. 233–234

[BAT+16] BESIM BILALLI; ALBERTO ABELLÓ; TOMÀS ALUJA-BANET; ROBERT WREMBEL: Towards Intelligent Data Analysis: The Metadata Challenge. undefined, 2016

[Bat05] BATES, M. J.: Information and knowledge: An evolutionary framework for information science. Information Research: An international electronic journal, (10)4, 2005, n4

[BB15] BOSCH-SIJTSEMA, P.; BOSCH, J.: User Involvement throughout the Innovation Process in High-Tech Industries. Journal of Product Innovation Management, (32)5, 2015, S. 793–807

[BBM13] BALCAN, N.; BLUM, A.; MANSOUR, Y.: Exploiting Ontology Structures and Unlabeled Data for Learning. In: Dasgupta, S.; McAllester, D. (Hrsg.): Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, Atlanta, Georgia, USA, 2013, S. 1112–1120

[BC06] BRAUN, V.; CLARKE, V.: Using thematic analysis in psychology. Qualitative Research in Psychology, (3)2, 2006, S. 77–101

[Bel16] BELLMAN, R. E.: Adaptive Control Processes – A Guided Tour. Princeton University Press, Princeton, NJ, 2016, [2016

[BFG+07] BELLANDI, A.; FURLETTI, B.; GROSSI, V.; ROMEI, A.: Ontology-Driven Association Rule Extraction: A Case Study, 2007

[BG19] BAHETI, R.; GILL, H.: Cyber-Physical Systems: ICM – 2019 IEEE International Conference on Mechatronics proceedings Humboldt Building, TU Ilmenau, Ilmenau, Germany, 18 - 20 March, 2019. 2019 IEEE International Conference on Mechatronics (ICM), 3/18/2019 - 3/20/2019, Ilmenau, Germany, Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 2019, S. 430–432

[BGP+21] BENDER, B.; GERICKE, K.; PAHL, G.; BEITZ, W. (Hrsg.): Pahl/Beitz Konstruktionslehre – Methoden und Anwendung erfolgreicher Produktentwicklung. 9. Auflage, Springer eBook Collection, Springer Vieweg, Berlin, 2021

[BH09] BISSANTZ, N.; HAGEDORN, J.: Data Mining (Datenmustererkennung). Wirtschaftsinformatik, (51)1, 2009, S. 139–144

[BJS19] BAIER, L.; JÖHREN, F.; SEEBACHER, S.: CHALLENGES IN THE DEPLOYMENT AND OPERATION OF MACHINE LEARNING IN PRACTICE, 2019

[BK11] BAE, J.; KIM, J.: Product development with data mining techniques: A case on design of digital camera. Expert Syst. Appl., (38), 2011, S. 9274–9280

[BPD+22] BRUNHEROTO, P. H.; PEPINO, A. L. G.; DESCHAMPS, F.; LOURES, EDUARDO DE FREITAS ROCHA: Data analytics in fleet operations: A systematic literature review and workflow proposal. Procedia CIRP, (107), 2022, S. 1192–1197

[Bra19] BRASCHLER, M.: Applied Data Science – Lessons Learned for the Data-Driven Business. Springer International Publishing AG, Cham, 2019

[BRH+18] BERMAN, F.; RUTENBAR, R.; HAILPERN, B.; CHRISTENSEN, H.; DAVIDSON, S.; ESTRIN, D.; FRANKLIN, M.; MARTONOSI, M.; RAGHAVAN, P.; STODDEN, V.; SZALAY, A. S.: Realizing the potential of data science. Commun. ACM, (61)4, 2018, S. 67–72

[BRH+22] BIE, T. DE; RAEDT, L. DE; HERNÁNDEZ-ORALLO, J.; HOOS, H. H.; SMYTH, P.; WILLIAMS, C. K. I.: Automating data science. Commun. ACM, (65)3, 2022, S. 76–87

[BRM19] BARTSCHAT, A.; REISCHL, M.; MIKUT, R.: Data mining tools. WIREs Data Mining and Knowledge Discovery, (9)4, 2019

[Bro10] BROY, M. (Hrsg.): Cyber-Physical Systems – Innovation durch softwareintensive eingebettete Systeme ; [acatech Symposium …]. acatech DISKUTIERT, Springer, Berlin [u.a.], 2010

[BS06] BATINI, C.; SCANNAPIECA, M.: Data quality: concepts, methodologies and techniques – Concepts, methodologies and techniques. Scholars Portal, Berlin, Heidelberg, 2006

[BS22-ol] BLACKMAN, R.; SIPES, T.: The Risks of Empowering "Citizen Data Scientists". Unter: https://hbr.org/2022/12/the-risks-of-empowering-citizen-data-scientists

[BS95] BRODLEY, C.; SMYTH, P.: The process of applying machine learning algorithms. In: NRL, Navy Center for Applied Research in AI Washington, DC (Hrsg.): Working notes for applying machine learning in practice: a workshop at the twelfth international conference on machine learning, 1995, S. 7–13

[BSJ+18] BAUER, N.; STANKIEWICZ, L.; JASTROW, M.; HORN, D.; TEUBNER, J.; KERSTING, K.; DEUSE, J.; WEIHS, C.: Industrial Data Science: Developing a Qualification Concept for Machine Learning in Industrial Production, 2018

[BTG12] BHATT, N.; THAKKAR, A.; GANATRA, A.: A survey and current research challenges in meta learning approaches based on dataset characteristics. International Journal of soft computing and Engineering, (2)10, 2012, S. 234–247

[BUB16] BECKER, W.; ULRICH, P.; BOTZKOWSKI, T.: Data Analytics im Mittelstand. Springer-Verlag, 2016

[BWR22] BISWAS, S.; WARDAT, M.; RAJAN, H.: The art and practice of data science pipelines. In: McIntosh, S.; Nguyen, T. V. (Hrsg.): 2022 ACM/IEEE 44th International Conference on Software Engineering – ICSE 2022 22-27 May 2022, Pittsburgh, Pennsylvania proceedings. ICSE '22: 44th International Conference on Software Engineering, 21 05 2022 29 05 2022, Pittsburgh Pennsylvania, IEEE, [Piscataway, NJ], 2022, S. 2091–2103

[CCK+00] CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R.: CRISP-DM 1.0: Step-by-step data mining guide, 2000

[CCL18] CORRALES, D.; CORRALES, J.; LEDEZMA, A.: How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. Symmetry, (10)4, 2018, S. 99

[CCM+18] CHEN, C.-Y.; CHOU, T.-Y.; MU, C.-Y.; LEE, B.-J.; CHANDRAMOULI, M.; CHAO, H.: Using Data Mining Techniques on Fleet Management System: Proceedings of the ESRI International User Conference, 2018

[CCS+99] COLLIER, K.; CAREY, B.; SAUTTER, D.; MARJANIEMI, C.: A methodology for evaluating and selecting data mining software. In: IEEE (Hrsg.): Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers, 1999, 11-pp

[CEK+16] CHAÂRI, R.; ELLOUZE, F.; KOUBÂA, A.; QURESHI, B.; PEREIRA, N.; YOUSSEF, H.; TOVAR, E.: Cyber-physical systems clouds: A survey. Computer Networks, (108), 2016, S. 260–278

[CEK04] COOPER, R. G.; EDGETT, S. J.; KLEINSCHMIDT, E. J.: Benchmarking Best NPD Practices—II. Research-Technology Management, (47)3, 2004, S. 50–59

[CGM+16] COLEMAN, S.; GOEB, R.; MANCO, G.; PIEVATOLO, A.; TORT-MARTORELL, X.; REIS, M.: How Can SMEs Benefit from Big Data? Challenges and a Path Forward: S. Coleman et al. Quality and Reliability Engineering International, (32), 2016

[CHT09] CHOUDHARY, A. K.; HARDING, J. A.; TIWARI, M. K.: Data mining in manufacturing: a review based on the kind of knowledge. Journal of Intelligent Manufacturing, (20)5, 2009, S. 501

[CL04] CALLAHAN, J.; LASRY, E.: The importance of customer input in the development of very new products. R and D Management, (34)2, 2004, S. 107–120

[CLC15] CORRALES, D. C.; LEDEZMA, A.; CORRALES, J. C.: A conceptual framework for data quality in knowledge discovery tasks (FDQ-KDT): A Proposal. JCP, (10)6, 2015, S. 396–405

[CMX+21] CHENG FAN; MEILING CHEN; XINGHUA WANG; JIAYUAN WANG; BUFU HUANG: A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data: Frontiers in Energy Research, 2021

[Dat23-ol] DATA SCIENCE ASSOCIATION: Data Science Code of Professional Conduct - Terminology. Unter: https://www.datascienceassn.org/code-of-conduct.html

[DDS+17] DUŠANKA, D.; DARKO, S.; SRDJAN, S.; MARKO, A.; TEODORA, L.: A comparison of contemporary data mining tools: XVII International Scientific Conference on Industrial Systems, 2017, S. 150–155

[DGK+15] DUMITRESCU, R.; GAUSEMEIER, J.; KÜHN, A.; LUCKEY, M.; PLASS, C.; SCHNEIDER, M.; WESTERMANN, T.: Auf dem Weg zur Industrie 4.0: Erfolgsfaktor Referenzarchitektur. It's OWL Clustermanagement, 2015

[DGL+13a] DEMCHENKO, Y.; GROSSO, P.; LAAT, C. DE; MEMBREY, P.: Addressing big data issues in scientific data infrastructure. In: IEEE (Hrsg.): 2013 International conference on collaboration technologies and systems (CTS), 2013, S. 48–55

[DGL+13b] DEMCHENKO, Y.; GROSSO, P.; LAAT, C. DE; MEMBREY, P.: Addressing big data issues in Scientific Data Infrastructure. In: Smari, W. W. (Ed.): 2013 International Conference on Collaboration Technologies and Systems (CTS 2013) – San Diego, California, USA, 20 - 24 May 2013 ; [including symposia and workshops. 2013 International Conference on Collaboration Technologies and Systems (CTS), 5/20/2013 - 5/24/2013, San Diego, CA, USA, IEEE, Piscataway, NJ, 2013, pp. 48–55

[Die14] DIENST, S.: Analyse von Maschinendaten zur Entscheidungsunterstützung bei der Produktverbesserung durch die Anwendung eines Feedback Assistenz Systems. Universitätsbibliothek der Universität Siegen, 2014

[DK23] DUMITRESCU, R.; KOLDEWEY, C. (Hrsg.): Datengestützte Produktplanung – Mit Betriebsdaten und Data Analytics zur faktenbasierten Planung zukünftiger Produktgenerationen im produzierenden Gewerbe. Verlagsschriftenreihe des Heinz Nixdorf Instituts Band 408, Heinz Nixdorf Institut, Paderborn, 2023

[DM17] DENKENA, B.; MÖRKE, T.: Cyber-physical and gentelligent systems in manufacturing and life cycle – Genetics and intelligence -- keys to industry 4.0. Academic Press, an imprint of Elsevier, London [etc.], op. 2017

[DMQ+16] DEMMINGER, C.; MOZGOVA, I.; QUIRICO, M.; UHLICH, F.; DENKENA, B.; LACHMAYER, R.; NYHUIS, P.: The Concept of Technical Inheritance in Operation: Analysis of the

Information Flow in the Life Cycle of Smart Products. Procedia Technology, (26), 2016, S. 79–88

[DMS+05] DIPPOLD, R.; MEIER, A.; SCHNIDER, W.; SCHWINN, K.: Unternehmensweites Datenmanagement – Von der Datenbankadministration bis zum Informationsmanagement ; mit 19 Tabellen. 4. Auflage, Zielorientiertes Business-Computing, Vieweg, Braunschweig, 2005

[DS19] DZISEVIC, R.; SESOK, D.: Text Classification using Different Feature Extraction Approaches: 2019 Open Conference of Electrical, Electronic and Information Sciences – 25 April 2019, Vilnius, Lithuania. 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), 4/25/2019 - 4/25/2019, Vilnius, Lithuania, Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 2019, S. 1–4

[DSD+19] DORIS JUNG LIN LEE; STEPHEN MACKE; DORIS XIN; ANGELA LEE; SILU HUANG; ADITYA G. PARAMESWARAN: A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. IEEE Data Eng. Bull., (42), 2019, S. 59–70

[DSV+21] DOGRA, V.; SINGH, A.; VERMA, S.; KAVITA; JHANJHI, N. Z.; TALIB, M. N.: Understanding of Data Preprocessing for Dimensionality Reduction Using Feature Selection Techniques in Text Classification. In: Peng, S.-L.; Hsieh, S.-Y.; Gopalakrishnan, S.; Duraisamy, B. (Eds.): Intelligent Computing and Innovation on Data Science – Proceedings of ICTIDS 2021. Lecture Notes in Networks and Systems, 248, Springer Singapore; Imprint Springer, Singapore, 2021, pp. 455–464

[DT13] DOGAN, N.; TANRIKULU, Z.: A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. Information Technology and Management, (14)2, 2013, S. 105–124

[Dud-ol] DUDEN: Daten. Unter: https://www.duden.de/rechtschreibung/Daten, 7. Januar 2024

[DWL15] DOU, D.; WANG, H.; LIU, H.: Semantic data mining: A survey of ontology-based approaches. In: IEEE (Hrsg.): Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015), 2015, S. 244–251

[DZS20] DEHKORDI, M. B.; ZARAKI, A.; SETCHI, R.: Feature extraction and feature selection in smartphone-based activity recognition. Procedia Computer Science, (176), 2020, S. 2655–2664

[Edl01] EDLER, A.: Nutzung von Felddaten in der qualitätsgetriebenen Produktentwicklung und im Service, 2001

[EH04] E. AWAD; H. GHAZIRI: Knowledge management / Elias M. Awad, Hassan Ghaziri, 2004

[EIO+22] EZUGWU, A. E.; IKOTUN, A. M.; OYELADE, O. O.; ABUALIGAH, L.; AGUSHAKA, J. O.; EKE, C. I.; AKINYELU, A. A.: A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, (110), 2022, S. 104743

[EM13] EHRLENSPIEL, K.; MEERKAMM, H.: Integrierte produktentwicklung: Denkabläufe, methodeneinsatz, zusammenarbeit. Carl Hanser Verlag GmbH Co KG, 2013

[Eng99] ENGLISH, L. P.: Improving data warehouse and business information quality – Methods for reducing costs and increasing profits. J. Wiley & Sons, New York, 1999

[FGM+11] FLORES, M. J.; GÁMEZ, J. A.; MARTÍNEZ, A. M.; PUERTA, J. M.: Handling numeric attributes when comparing Bayesian network classifiers: does the discretization method matter? Applied Intelligence, (34)3, 2011, S. 372–385

[FGW+20] FRANK, M.; GAUSEMEIER, J.; WIDDERN, N. H.-C. VON; KOLDEWEY, C.; MENZEFRICKE, J. S.; REINHOLD, J.: A reference process for the Smart Service business: development and practical implications. In: The International Society for Professional Innovation Management (ISPIM) (Hrsg.): ISPIM Conference Proceedings, 2020, S. 1–19

[FKP07] FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W.: A Novel Framework for Imputation of Missing Values in Databases. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, (37)5, 2007, S. 692–709

[FMS19] FOUNTAINE, T.; MCCARTHY, B.; SALEH, T.: Building the AI-powered organization. Harvard business review, (97)4, 2019, S. 62–73

[FPS96a] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.: From data mining to knowledge discovery in databases. AI magazine, (17)3, 1996, S. 37

[FPS96b] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Commun. ACM, (39)11, 1996, S. 27–34

[FVP22] FLORENCE JACOB; VIRGINIE PEZ; PIERRE VOLLE: Principles, methods, contributions, and limitations of design science research in marketing: Illustrative application to customer journey management. Recherche et Applications en Marketing (English Edition), (37)2, 2022, S. 2–29

[FXY15] FAN, C.; XIAO, F.; YAN, C.: A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automation in Construction, (50), 2015, S. 81–90

[Gar00] GARLAN, D.: Software architecture: a roadmap: Proceedings of the Conference on the Future of Software Engineering, 2000, S. 91–101

[Gau00] GAUSEMEIER, J.: Kooperatives Produktengineering – Ein neues Selbstverständnis des ingenieurmässigen Wirkens. HNI-Verlagsschriftenreihe Band 79, Heinz Nixdorf Institut, Universität Paderborn, Paderborn, 2000

[GDE+19] GAUSEMEIER, J.; DUMITRESCU, R.; ECHTERFELD, J.; PFÄNDER, T.; STEFFEN, D.; THIELEMANN, F.: Produktinnovation – Strategische Planung von Produkten, Dienstleistungen und Geschäftsmodellen. Hanser, München, 2019

[GHG+19] GIL, Y.; HONAKER, J.; GUPTA, S.; MA, Y.; D'ORAZIO, V.; GARIJO, D.; GADEWAR, S.; YANG, Q.; JAHANSHAD, N.: Towards human-guided machine learning. In: Fu, W.-T. (Ed.): Proceedings of the 24th International Conference on Intelligent User Interfaces. IUI '19: 24th International Conference on Intelligent User Interfaces, 17 03 2019 20 03 2019, Marina del Ray California, ACM Conferences, ACM, New York, NY, 2019, pp. 614–624

[GJM+16] GAVRILOVSKI, A.; JIMENEZ, H.; MAVRIS, D.; RAO, A.; SHIN, S.-H.; HWANG, I.; MARAIS, K.: Challenges and Opportunities in Flight Data Mining: A Review of the State of the Art, 2016

[GK12] GAUSEMEIER, J.; KOKOSCHKA, M.: Schutzmaßnahmen vor Produktpiraterie. Gausemeier, J.; Glatz, R.; Lindemann, U.: Präventiver Produktschutz–Leitfäden und Anwendungsbeispiele, Carl Hanser Verlag, München, 2012

[GLD00] GAMBERGER, D.; LAVRAC, N.; DZEROSKI, S.: Noise detection and elimination in data preprocessing: Experiments in medical domains. Applied Artificial Intelligence, (14)2, 2000, S. 205–223

[GLH15] GARCIA, S.; LUENGO, J.; HERRERA, F.: Data preprocessing in data mining. Springer, 2015

[Gom-ol] GOMEZ, N.: Data Source: Definition, Types, & Common Examples. Unter: https://analystanswers.com/data-source-definition-types-common-examples/?utm_content=cmp-true, 7. Januar 2024

[Gör00] GÖRZ, G. (Hrsg.): Handbuch der künstlichen Intelligenz. 3. Auflage, Oldenbourg, München, 2000

[GPW+14] GAUSEMEIER, J.; PLASS, C.; WENZELMANN, C.; UNTERNEHMENSGESTALTUNG, Z.: Strategien, Geschäftsprozesse und IT-Systeme für die Produktion von morgen. Munich/Vienna, 2014

[GPW09] GAUSEMEIER, J.; PLASS, C.; WENZELMANN, C.: Zukunftsorientierte Unternehmensgestaltung-Strategien. Geschäftsprozesse und IT-Systeme für die Produktion von morgen, (2), 2009

[GTD13] GAUSEMEIER, J.; TSCHIRNER, C.; DUMITRESCU, R.: Der Weg zu Intelligenten Technischen Systemen. Industrie Management, (29)1, 2013

[Gug23] GUGGENHEIM, D.: Empowering the Citizen Data Scientist in Everyone: The no-code, low-math toolkit for data analytic excellence. CitizenAnalytics LLC, 2023

[HA19] HORVATH, I.; ABOU EDDAHAB-BURKE, F.-Z.: Using Data Analytics To Extract Knowledge From Middle-Of-Life Product Data. International Journal of Advanced Research and Publications, (4), 2019, S. 1–21

[Ham16-ol] HAMMOND, K.: The Periodic Table of AI. Unter: https://periodensystem-ki.de/Mit-Legosteinen-die-Kuenstliche-Intelligenz-bauen

[Har18-ol] HARVARD BUSINESS REVIEW: Closing the Data Gap in Product Development. Unter: https://hbr.org/sponsored/2018/11/closing-the-data-gap-in-product-development

[HB21] HOPKINS, A.; BOOTH, S.: Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. In: Fourcade, M. (Ed.): Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, 19 05 2021 21 05 2021, Virtual Event USA, ACM Digital Library, Association for Computing Machinery, New York,NY,United States, 2021, pp. 134–145

[HC05] HUANG, T.-S.; CHANG, C.-F.: The role of data mining in the product design and development process: Proc. 6th Int. Comput.-Aided Ind. Design Concept Design, 2005, S. 198–203

[HDG+20] HAN, Y.; DING, N.; GENG, Z.; WANG, Z.; CHU, C.: An optimized long short-term memory network based fault diagnosis model for chemical processes. Journal of Process Control, (92), 2020, S. 161–168

[HGH+15] HILDEBRAND, K.; GEBAUER, M.; HINRICHS, H.; MIELKE, M. (Hrsg.): Daten- und Informationsqualität – Auf dem Weg zur Information Excellence. 3. Auflage, Springer Vieweg, Wiesbaden, 2015

[HJ20] HOU, L.; JIAO, R. J.: Data-informed inverse design by product usage information: a review, framework and outlook. Journal of Intelligent Manufacturing, (31)3, 2020, S. 529–552

[HKN+09] HILARIO, M.; KALOUSIS, A.; NGUYEN, P.; WOZNICA, A.: A data mining ontology for algorithm selection and meta-mining, 2009, S. 76–87

[HLM18] HENKE, N.; LEVINE, J.; MCINERNEY, P.: Analytics translator: The new must-have role. Harvard business review, 2018

[HN20] HAPKE, H. M.; NELSON, C.: Building machine learning pipelines – Automating model life cycles with TensorFlow. O'Reilly, Beijing, 2020

[HNU+17] HOLLER, M.; NEIDITSCH, G.; UEBERNICKEL, F.; BRENNER, W.: Digital product innovation in manufacturing industries-towards a taxonomy for feedback-driven product development scenarios, 2017

[HSW+18] HOLLAUER, C.; SHALUMOV, B.; WILBERG, J.; OMER, M.: GRAPH DATABASES FOR EXPLOITING USE PHASE DATA IN PRODUCT-SERVICE-SYSTEM DEVELOPMENT: A METHODOLOGY TO SUPPORT IMPLEMENTATION: Proceedings of the DESIGN 2018 15th International Design Conference. 15th International Design Conference, May, 21-24, 2018, Design Conference Proceedings, Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia; The Design Society, Glasgow, UK, 2018, S. 1571–1582

[HV14] HERSTATT, C.; VERWORN, B.: The ‘Fuzzy Front End’ of Innovation. In: Management, E. I. F. T. A. I. (Hrsg.): Bringing technology and innovation into the boardroom – Strategy, innovation and competences … for business value. Palgrave Macmillan, [Place of publication not identified], 2014, S. 347–372

[HWS+19] HUBER, S.; WIEMER, H.; SCHNEIDER, D.; IHLENFELDT, S.: DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. Procedia CIRP, (79), 2019, S. 403–408

[IAG+15] IGBA, J.; ALEMZADEH, K.; GIBBONS, P. M.; HENNINGSEN, K.: A framework for optimising product performance through feedback and reuse of in-service experience. Robotics and Computer-Integrated Manufacturing, (36), 2015, S. 2–12

[Int17] INTERNATIONAL, DAMA: DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition). Technics Publications, LLC, Denville, NJ, USA, 2017

[Ise08] ISERMANN, R.: Mechatronische Systeme – Grundlagen. 2. Auflage, Springer eBook Collection Computer Science & Engineering, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008

[JDG11] JANKOWSKI, N.; DUCH, W.; GRĄBCZEWSKI, K.: Meta-learning in computational intelligence. Band 358, Springer, 2011

[JEK+19] JOPPEN, R.; ENZBERG, S.; KÜHN, A.; DUMITRESCU, R.: Data map – method for the specification of data flows within production. Procedia CIRP, (79), 2019, S. 461–465

[JF14] JAN MACHÁ\VC; FRANTISEK STEINER: Risk Management in Early Product Lifecycle Phases. International Review of Management and Business Research, (3), 2014, S. 1151

[JGL+14] JAGADISH, H. V.; GEHRKE, J.; LABRINIDIS, A.; PAPAKONSTANTINOU, Y.; PATEL, J. M.; RAMAKRISHNAN, R.; SHAHABI, C.: Big data and its technical challenges. Commun. ACM, (57)7, 2014, S. 86–94

[Jos20] JOSHI, A. V.: Machine learning and artificial intelligence, 2020

[Juo17] JUODYT\.E, M.: Overview: Data Mining Pipeline, 2017

[JZL+22] JIAFENG, Y.; ZHUKOVA, N.; LEBEDEV, S.; TIANXING, M.: An Architecture of the Semantic Meta Mining Assistant for Adaptive Domain-Oriented Data Processing. International Journal of Embedded and Real-Time Communication Systems, (13)1, 2022, S. 1–38

[Kan19] KANTARDZIC, M.: Data Mining: Concepts, Models, Methods, and Algorithms. Wiley, 2019

[KBB+09] KITCHENHAM, B.; BRERETON, O. P.; BUDGEN, D.; TURNER, M.; BAILEY, J.; LINKMAN, S.: Systematic literature reviews in software engineering-a systematic literature review. Information and software technology, (51)1, 2009, S. 7–15

[KBZ14] KIND, C.; BENTLAGE, A.; ZENKER, M.: Analyse von Lebenszyklusdaten. ERP Management, (10 (2014)), 2014, S. 50–52

[KEA+15] KAISLER, S.; ESPINOSA, J. A.; ARMOUR, F.; MONEY, W.: Advanced Analytics for Big Data. In: Khosrowpour, M. (Hrsg.): Encyclopedia of information science and technology. 3. Auflage, Advances in Information Quality and Management, Information Science Reference, Hershey, Pa., 2015, S. 7584–7593

[KFD17] KUHRMANN, M.; FERNÁNDEZ, D. M.; DANEVA, M.: On the pragmatic design of literature studies in software engineering: an experience-based guideline. Empirical software engineering, (22)6, 2017, S. 2852–2891

[KGM+14] KEPNER, J.; GADEPALLY, V.; MICHALEAS, P.; SCHEAR, N.; VARIA, M.; YERUKHIMOVICH, A.; CUNNINGHAM, R. K.: Computing on masked data: a high performance method for improving big data veracity, (20), 2014, S. 1–6

[KGM+15] KASSNER, L.; GRÖGER, C.; MITSCHANG, B.; WESTKÄMPER, E.: Product Life Cycle Analytics – Next Generation Data Analytics on Structured and Unstructured Data. Procedia CIRP, (33), 2015, S. 35–40

[Kit13] KITCHIN, R.: Big data and human geography. Dialogues in Human Geography, (3)3, 2013, S. 262–267

[Kit17] KITCHIN, R.: The data revolution – Big data, open data, data infrastructures & their consequences. Sage, Los Angeles [etc.], dr. 2017

[KJR+18] KÜHN, A.; JOPPEN, R.; REINHART, F.; RÖLTGEN, D.; ENZBERG, S. VON; DUMITRESCU, R.: Analytics Canvas – A Framework for the Design and Specification of Data Analytics Projects. Procedia CIRP, (70), 2018, S. 162–167

[KKM+20] KULIN, M.; KAZAZ, T.; MOERMAN, I.; POORTER, E. DE: A survey on Machine Learning-based Performance Improvement of Wireless Networks: PHY, MAC and network layer, 2020

[KM06] KURGAN, L. A.; MUSILEK, P.: A survey of knowledge discovery and data mining process models. Knowledge Engineering Review, (21)1, 2006, S. 1–24

[KM16] KITCHIN, R.; MCARDLE, G.: What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, (3)1, 2016, 2053951716631130

[KNH+16] KAMMERL, D.; NOVAK, G.; HOLLAUER, C.; MÖRTL, M.: Integrating usage data into the planning of Product-Service Systems: 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2016, S. 375–379

[Kre19] KREUTZER, R.: Methodik zur Bestimmung der Nutzenpotenziale von Felddaten cyber-physischer Systeme. Dissertation, RWTH Aachen; IIF - Institut für Industriekommunikation und Fachmedien GmbH, 2019

[Küh19] KÜHNAPFEL: Nutzwertanalysen in Marketing und Vertrieb. Springer Fachmedien Wiesbaden, Wiesbaden, 2019

[Kur05] KURBEL, K.: Produktionsplanung und-steuerung im Enterprise Resource Planning und Supply Chain Management. Oldenbourg Verlag, 2005

[KW17] KOLERUS, J.; WASSERMANN, J.: Zustandsüberwachung von Maschinen – Das Lehr- und Arbeitsbuch für den Praktiker. 7. Auflage, utb-studi-e-book Band 5181, UTB GmbH; UVK, Stuttgart, 2017

[Lan01] LANEY, D.: 3D data management: Controlling data volume, velocity and variety. META group research note, (6)70, 2001, S. 1

[LAT+20] LARIOS VARGAS, E.; ANICHE, M.; TREUDE, C.; BRUNTINK, M.; GOUSIOS, G.: Selecting third-party libraries: the practitioners' perspective. In: Devanbu, P. (Hrsg.): Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE '20: 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 08 11 2020 13 11 2020, Virtual Event USA, ACM Digital Library, Association for Computing Machinery, New York,NY,United States, 2020, S. 245–256

[LCL09] LIBRALON, G. L.; CARVALHO, A. C. P. L. F. DE; LORENA, A. C.: Pre-processing for noise detection in gene expression classification data. Journal of the Brazilian Computer Society, (15)1, 2009, S. 3–11

[LCS+06] LARRAÑAGA, P.; CALVO, B.; SANTANA, R.; BIELZA, C.; GALDIANO, J.; INZA, I.; LOZANO, J. A.; ARMAÑANZAS, R.; SANTAFÉ, G.; PÉREZ, A.; ROBLES, V.: Machine learning in bioinformatics. Briefings in bioinformatics, (7)1, 2006, S. 86–112

[Lew92] LEWIS, D. D.: Feature Selection and Feature Extraction for Text Categorization: Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992, 1992

[Li19] LI, C.: Preprocessing Methods and Pipelines of Data Mining: An Overview, 2019

[LLS00] LIM, T.-S.; LOH, W.-Y.; SHIH, Y.-S.: A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning, (40)3, 2000, S. 203–228

[LM22] LACHMAYER, R.; MOZGOVA, I.: Technical Inheritance as an Approach to Data-Driven Product Development. Design Methodology for Future Products: Data Driven, Agile and Flexible, 2022, S. 47–64

[LM98] LIU, H.; MOTODA, H.: Feature Extraction, Construction and Selection – A Data Mining Perspective. Springer US, Vol. 453 Boston, MA, 1998

[Lov83] LOVELL, M. C.: Data Mining. The Review of Economics and Statistics, (65), 1983, S. 1

[LPW+16] LUETH, L.; PATSIOURA, C.; WILLIAMS, D.; OTHERS: The current state of data analytics usage in industrial companies. Industrial Analytics, Digital Analytics Association Germany, 2016, S. 38–49

[LSH+15] LINÅKER, J.; SULAMAN, S.; HOST, M.; MELLO, R. DE: Guidelines for Conducting Surveys in Software Engineering, 2015

[LZW+14] LIU, Y.; ZHOU, Y.; WEN, S.; TANG, C.: A Strategy on Selecting Performance Metrics for Classifier Evaluation. International Journal of Mobile Computing and Multimedia Communications, (6)4, 2014, S. 20–35

[Mar14] MARR, B.: Big data: The 5 vs everyone must know. LinkedIn Pulse, (6), 2014

[Mat19-ol] MATHESON, R.: Democratizing data science – Tool for nonstatisticians automatically generates models that glean insights from complex datasets. Unter: https://news.mit.edu/2019/nonprogrammers-data-science-0115

[May19] MAYRING, P.: Qualitative content analysis: Demarcation, varieties, developments. In: Freie Universität Berlin (Hrsg.): Forum: Qualitative Social Research, 2019

[May94] MAYRING, P.: Qualitative inhaltsanalyse. Band 14, UVK Univ.-Verl. Konstanz, 1994

[MBS+14] MURTHY, P.; BHARADWAJ, A.; SUBRAHMANYAM, P.; ROY, A.; RAJAN, S.: Cloud Security Alliance report on Big Data Taxonomy, 2014

[MCM+20] MICHAEL RIESENER; CHRISTIAN DOELLE; MICHAEL MENDL-HEINISCH; NICLAS KLUMPEN: Identification of evaluation criteria for algorithms used within the context of product development. Procedia CIRP, (91), 2020, S. 508–515

[MEK+22] MERKELBACH, S.; ENZBERG, S. VON; KÜHN, A.; DUMITRESCU, R.: Towards a Process Model to Enable Domain Experts to Become Citizen Data Scientists for Industrial

Applications. In: IEEE (Hrsg.): 2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems (ICPS), 2022, S. 1–6

[MFK+22a] MEYER, M.; FICHTLER, T.; KOLDEWEY, C.; DUMITRESCU, R.: How can Data Analytics Results be Exploited in the Early Phase of Product Development? 13 Design Principles for Data-Driven Product Planning, 2022

[MFK+22b] MEYER, M.; FICHTLER, T.; KOLDEWEY, C.; DUMITRESCU, R.: Potentials and challenges of analyzing use phase data in product planning of manufacturing companies. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, (36), 2022

[MG10] MARINICA, C.; GUILLET, F.: Knowledge-Based Interactive Postmining of Association Rules Using Ontologies. IEEE Transactions on Knowledge and Data Engineering, (22)6, 2010, S. 784–797

[ML02] MOTODA, H.; LIU, H.: Feature selection, extraction and construction. Communication of IICM (Institute of Information and Computing Machinery, Taiwan), (5)67-72, 2002, S. 2

[MMR+06] MCCARTHY, J.; MINSKY, M. L.; ROCHESTER, N.; SHANNON, C. E.: A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine, (27)4, 2006, S. 12

[MN07] MYERS, M. D.; NEWMAN, M.: The qualitative interview in IS research: Examining the craft. Information and Organization, (17)1, 2007, S. 2–26

[MNS19] MOUSTAFA REDA, M.; NASSEF, M.; SALAH, A.: Categorization of Factors Affecting Classification Algorithms Selection. International Journal of Data Mining & Knowledge Management Process, (9)4, 2019, S. 1–19

[Moo17-ol] MOORE, S.: Gartner says more than 40 percent of data science tasks will be automated by 2020. Unter: https://www.gartner.com/en/newsroom/press-releases/2017–01-16-gartner-says-more-than-40-percent-of-data-science-tasks-will-be-automated-by-2020.

[MOR11] MANSINGH, G.; OSEI-BRYSON, K.-M.; REICHGELT, H.: Using ontologies to facilitate post-processing of association rules by domain experts. Information Sciences, (181)3, 2011, S. 419–434

[Moz17] MOZGOVA, I.: Intelligente Datenanalyse für die Entwicklung neuer Produktgenerationen. Proceedings of the Wissenschaftsforum Intelligente Technische Systeme (WInTeSys), 11. und 12.05. 2017, Paderborn, Germany. Heinz Nixdorf Institut, (369), 2017, S. 335–346

[MPK+21] MEYER, M.; PANZNER, M.; KOLDEWEY, C.; DUMITRESCU, R.: Towards Identifying Data Analytics Use Cases in Product Planning. Procedia CIRP, (104), 2021, S. 1179–1184

[MPK+22] MEYER, M.; PANZNER, M.; KOLDEWEY, C.; DUMITRESCU, R.: 17 Use Cases for Analyzing Use Phase Data in Product Planning of Manufacturing Companies. Procedia CIRP, (107), 2022, S. 1053–1058

[MS21] MASOOD, A.; SHERIF, A.: Automated Machine Learning. Packt Publishing; Safari, Erscheinungsort nicht ermittelbar, 2021

[Muk15] MUKHAMEDIEV, R.: Machine learning methods: An overview. CMNT, (19), 2015, S. 14–29

[MWK+21] MEYER, M.; WIEDERKEHR, I.; KOLDEWEY, C.; DUMITRESCU, R.: UNDERSTANDING USAGE DATA-DRIVEN PRODUCT PLANNING: A SYSTEMATIC LITERATURE REVIEW. Proceedings of the Design Society, (1), 2021, S. 3289–3298

[MWP+22] MEYER, M.; WIEDERKEHR, I.; PANZNER, M.; KOLDEWEY, C.; DUMITRESCU, R.: A Reference Process Model for Usage Data-Driven Product Planning, 2022

[NAL21] NIETO, F. J.; AGUILERA, U.; LÓPEZ-DE-IPIÑA, D.: Analyzing Particularities of Sensor Datasets for Supporting Data Understanding and Preparation. Sensors (Basel, Switzerland), (21)18, 2021

[NDB+19] NGUYEN, G.; DLUGOLINSKY, S.; BOBÁK, M.; TRAN, V.; LÓPEZ GARCÍA, Á.; HEREDIA, I.; MALÍK, P.; HLUCHÝ, L.: Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, (52)1, 2019, S. 77–124

[NM18] NORTH, K.; MAIER, R.: Wissen 4.0-Wissensmanagement im digitalen Wandel. HMD Praxis der Wirtschaftsinformatik: Vol. 55, No. 4, 2018

[NP19] NAGARAJAH, T.; PORAVI, G.: A Review on Automated Machine Learning (AutoML) Systems. In: Institute of Electrical and Electronics Engineers (Ed.): IEEE 5th International Conference for Convergence in Technology (I2CT). 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 3/29/2019 - 3/31/2019, Bombay, India, IEEE, Piscataway, NJ, 2019, pp. 1–6

[NS23] NADI, S.; SAKR, N.: Selecting third-party libraries: the data scientist's perspective. Empirical Software Engineering, (28)1, 2023

[NT17] NABATI, E. G.; THOBEN, K. D.: Big Data Analytics in the Maintenance of Off-Shore Wind Turbines: A Study on Data Characteristics: Dynamics in Logistics. Springer, Cham, 2017, pp. 131–140

[NVM13] NICKERSON, R. C.; VARSHNEY, U.; MUNTERMANN, J.: A method for taxonomy development and its application in information systems. European Journal of Information Systems, (22)3, 2013, S. 336–359

[NY18] NALCHIGAR, S.; YU, E.: Business-driven data analytics: A conceptual modeling framework. Data & Knowledge Engineering, (117), 2018

[NYO+19] NALCHIGAR, S.; YU, E.; OBEIDI, Y.; CARBAJALES, S.; GREEN, J.; CHAN, A.: Solution Patterns for Machine Learning. In: Giorgini, P.; Weber, B. (Hrsg.): Advanced Information Systems Engineering. Springer International Publishing, Cham, 2019, S. 627–642

[OB13] OLSSON, H. H.; BOSCH, J.: Towards data-driven product development: A multiple case study on post-deployment data usage in software-intensive embedded systems. In: Springer (Hrsg.): International Conference on Lean Enterprise Software and Systems, 2013, S. 152–164

[OBU+16] OLSON, R. S.; BARTLEY, N.; URBANOWICZ, R. J.; MOORE, J. H.: Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science, 2016

[OP06] OGAWA, S.; PILLER, F. T.: Reducing the risks of new product development. MIT Sloan management review, 2006

[PA16] PATGIRI, R.; AHMED, A.: Big Data: The V's of the Game Changer Paradigm. In: Chen, J.; Yang, L. T. (Eds.): The Eighteenth IEEE International Conference on High Performance Computing and Communications, the Fourteenth IEEE International Conference on Smart City, the Second IEEE International Conference on Data Science and Systems – HPCC/SmartCity/DSS 2016 12-14 December 2016, Sydney, Australia proceedings. 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 12/12/2016 - 12/14/2016, Sydney, Australia, IEEE, Piscataway, NJ, 2016, pp. 17–24

[PC19] PIMENTEL, B. A.; CARVALHO, A. C. DE: A new data characterization for selecting clustering algorithms using meta-learning. Information Sciences, (477), 2019, S. 203–219

[PC20] PIMENTEL, B. A.; CARVALHO, A. C. DE: A Meta-learning approach for recommending the number of clusters for clustering algorithms. Knowledge-Based Systems, (195), 2020, S. 105682

[Pei19] PEI WANG: On Defining Artificial Intelligence. Journal of Artificial General Intelligence, (10), 2019, S. 1–37

[PEM+22] PANZNER, M.; ENZBERG, S. VON; MEYER, M.; DUMITRESCU, R.: Characterization of Usage Data with the Help of Data Classifications. Journal of the Knowledge Economy, 2022

[Pfa01] PFAHRINGER, B.: Meta-Learning by Landmarking Various Learning Algorithms, 2001

[PGK+09] PREECE, S. J.; GOULERMAS, J. Y.; KENNEY, L. P. J.; HOWARD, D.: A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE transactions on bio-medical engineering, (56)3, 2009, S. 871–879

[PGP+22] POZUECO, L.; GUPTA, N.; PAÑEDA, X. G.; CORCOBA, V\'\IA, ROBERTO; RIONDA, A.: Data Analytics to Support a Smart Fleet Management Strategy. IEEE Intelligent Transportation Systems Magazine, 2022

[PH15] PORTER, M. E.; HEPPELMANN, J. E.: How Smart, Connected Products Are Transforming Companies (Harvard Business Review). Online verfügbar unter https://hbr.org/2015/10/how-smart-connected-products-are-transforming-companies. pdf, 2015

[Phi20] PHILIP BOUCHER: Artificial intelligence: How does it work, why does it matter, and what can we do about it?, 2020

[PJ20] PESSOA, M.; JAUREGUI-BECKER, J.: Smart design engineering: a literature review of the impact of the 4th industrial revolution on product design and development. Research in Engineering Design, 2020, S. 1–21

[PJD+20] PETER HOFMANN; JAN JÖHNK; DOMINIK PROTSCHKY; NILS URBACH: Developing Purposeful AI Use Cases - A Structured Method and Its Application in Project Management. 15th International Conference on Wirtschaftsinformatik (WI), 2020

[PLA19] PABLO MARTÍ; LETICIA SERRANO-ESTRADA; ALMUDENA NOLASCO-CIRUGEDA: Social Media data: Challenges, opportunities and limitations in urban studies. Computers, Environment and Urban Systems, (74), 2019, S. 161–174

[PME+22] PANZNER, M.; MEYER, M.; ENZBERG, S. VON; DUMITRESCU, R.: Business-to-Analytics Canvas - Translation of Product Planning-Related Business Use Cases into Concrete Data Analytics Tasks. Procedia CIRP, (109), 2022, S. 580–585

[PRB20] PUSCHEL, L. C.; ROGLINGER, M.; BRANDT, R.: Unblackboxing Smart Things—A Multilayer Taxonomy and Clusters of Nontechnical Smart Thing Characteristics. IEEE Transactions on Engineering Management, 2020, S. 1–15

[PRT+12] PEFFERS, K.; ROTHENBERGER, M.; TUUNANEN, T.; VAEZI, R.: Design science research evaluation. In: Springer (Hrsg.): Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings 7, 2012, S. 398–410

[PSD16] PANOV, P.; SOLDATOVA, L. N.; DŽEROSKI, S.: Generic ontology of datatypes. Information Sciences, (329), 2016, S. 900–920

[PTR+07] PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S.: A design science research methodology for information systems research. Journal of management information systems, (24)3, 2007, S. 45–77

[Pyl99] PYLE, D.: Data preparation for data mining. morgan kaufmann, 1999

[QLZ+23] QUAN, H.; LI, S.; ZENG, C.; WEI, H.; HU, J.: Big Data and AI-Driven Product Design: A Survey. Applied Sciences, (13)16, 2023, S. 9433

[QMH+18] QUANMING YAO; MENGSHUO WANG; HUGO JAIR ESCALANTE; ISABELLE GUYON; YI-QI HU; YU-FENG LI; WEI-WEI TU; QIANG YANG; YANG YU: Taking Human out of Learning Applications: A Survey on Automated Machine Learning. ArXiv, (abs/1810.13306), 2018

[Raf19-ol] RAFFEINER, M.: Erkunden Sie Ihre Datenlandschaft

[Ras18] RASCHKA, S.: Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, 2018

[RDL+21] RIESENER, M.; DÖLLE, C.; LENDER, B.; SCHUH, G.: Requirements Engineering through Exploratory Analysis of Usage Data. In: IEEE (Hrsg.): 2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR), 2021, S. 1–6

[RJ00] RÜDIGER WIRTH; JOCHEN HIPP: Crisp-dm: towards a standard process modell for data mining, 2000

[RKD17] REINHART, F.; KÜHN, A.; DUMITRESCU, R.: Schichtenmodell für die Entwicklung von Data Science Anwendungen im Maschinen- und Anlagenbau: Wissenschaftsforum Intelligente Technische Systeme (WInTeSys). Heinz Nixdorf MuseumsForum, 2017, S. 321–334

[RLS05] RAKESH MENON; LOH HAN TONG; S. SATHIYAKEERTHI: Analyzing textual databases using data mining to enable fast product development processes. Reliability Engineering & System Safety, (88)2, 2005, S. 171–180

[Ro11] RUSSOM, P.; OTHERS: Big data analytics. TDWI best practices report, fourth quarter, (19)4, 2011, S. 1–34

[Rob19] ROBERT G. COOPER: The drivers of success in new-product development. Industrial Marketing Management, (76), 2019, S. 36–47

[Rog18] ROGEL-SALAZAR, J.: Data science and analytics with python. CRC Press, 2018

[Row07] ROWLEY, J.: The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science, (33)2, 2007, S. 163–180

[RRC19] REBALA, G.; RAVI, A.; CHURIWALA, S.: Machine learning definition and basics. An introduction to machine learning, 2019, S. 1–17

[RRL+20] REDDY, G. T.; REDDY, M. P. K.; LAKSHMANNA, K.; KALURI, R.; RAJPUT, D. S.; SRIVASTAVA, G.; BAKER, T.: Analysis of Dimensionality Reduction Techniques on Big Data. IEEE Access, (8), 2020, S. 54776–54788

[RSG+14] REIF, M.; SHAFAIT, F.; GOLDSTEIN, M.; BREUEL, T.; DENGEL, A.: Automatic classifier selection for non-experts. Pattern Analysis and Applications, (17), 2014, S. 83–96

[RSW10] RENEAR, A. H.; SACCHI, S.; WICKETT, K. M.: Definitions of dataset in the scientific and technical literature. Proceedings of the American Society for Information Science and Technology, (47)1, 2010, S. 1–4

[Run12] RUNKLER, T. A.: Data Analytics: Models and Algorithms for Intelligent Data Analysis. Lehrbuch - Springer, Vieweg+Teubner Verlag, 2012

[Run20] RUNKLER, T. A.: Data analytics. Springer, 2020

[SA19] SHIVANI GUPTA; ATUL GUPTA: Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. Procedia Computer Science, (161), 2019, S. 466–474

[SA98] SARABJOT S. ANAND; ALEX G. BÜCHNER: Decision support using data mining, 1998

[SAB19] SASSI, I.; ANTER, S.; BEKKHOUCHA, A.: An Overview of Big Data and Machine Learning Paradigms. In: Ditzinger; Ezziyyani (Hrsg.): Advanced Intelligent Systems for Sustainable Development (AI2SD'2018). Advances in Intelligent Systems and Computing, Springer International Publishing, Cham, 2019, S. 237–251

[Sal94] SALZBERG, S. L.: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning, (16)3, 1994, S. 235–240

[Sar21a] SARKER, I. H.: Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN computer science, (2)6, 2021, S. 420

[Sar21b] SARKER, I. H.: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN computer science, (2)3, 2021, S. 160

[SB08] SPITTA, T.; BICK, M.: Informationswirtschaft: Eine Einführung. Springer-Verlag, 2008

[SE18] SOROOSH NALCHIGAR; ERIC YU: Business-driven data analytics: A conceptual modeling framework. Data & Knowledge Engineering, (117), 2018, S. 359–372

[Sha19-ol] SHAHAB D. MOHAGHEDH: Traditional Statistics vs. Artificial Intelligence and Machine Learning. Unter: https://jpt.spe.org/traditional-statistics-vs-artificial-intelligence-and-machine-learning

[SHB19] SHABESTARI, S. S.; HERZOG, M.; BENDER, B.: A Survey on the Applications of Machine Learning in the Early Phases of Product Development. Proceedings of the Design Society: International Conference on Engineering Design, (1), 2019, S. 2437–2446

[She00] SHEARER, C.: The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, (5)4, 2000, S. 13–22

[SHP+11] SEIN; HENFRIDSSON; PURAO; ROSSI; LINDGREN: Action Design Research. MIS Quarterly, (35)1, 2011, S. 37

[Sim96] SIMON, H.: The Sciences of Artificial. 3. Auflage, MIT Press, Cambridge, MA, 1996

[SKM+12] SCHÄFER, A.; KNAPP, M.; MAY, M.; VOß, A.; FÜR INTELLIGENTE ANALYSE UND INFORMATIONSSYSTEME IAIS, FRAUNHOFER INSTITUT: Big Data – Vorsprung durch Wissen – Innovationspotenzialanalyse, 2012

[SM13] SEGOVIA, J.; MARBÁN, O.: Extending UML for Modeling Data Mining Projects (DM-UML). Journal of Information Technology & Software Engineering, (3), 2013

[SM17] SHOBANADEVI, A.; MARAGATHAM, G.: Data mining techniques for IoT and big data — A survey. In: Institute of Electrical and Electronics Engineers (Ed.): Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) – 7-8 December 2017. 2017 International Conference on Intelligent Sustainable Systems (ICISS), 12/7/2017 - 12/8/2017, Palladam, IEEE, Piscataway, NJ, 2017, pp. 607–610

[SMB+18] STEFAN STIEGLITZ; MILAD MIRBABAIE; BJÖRN ROSS; CHRISTOPH NEUBERGER: Social media analytics – Challenges in topic discovery, data collection, and data preparation. International Journal of Information Management, (39), 2018, S. 156–168

[SRF+09] SUBRAMANIAM, L. V.; ROY, S.; FARUQUIE, T. A.; NEGI, S.: A survey of types of text noise and techniques to handle noisy text. In: Lopresti, D. (Ed.): Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data. The Third Workshop, 7/23/2009 - 7/24/2009, Barcelona, Spain, ACM Other conferences, ACM, New York, NY, 2009, p. 115

[SS14] SAHA, B.; SRIVASTAVA, D.: Data quality: The other face of Big Data. In: Institute of Electrical and Electronics Engineers; IEEE Computer Society (Eds.): 2014 IEEE 30th International Conference on Data Engineering (ICDE 2014) – Chicago, Illinois, USA, 31 March - 4 April 2014. 2014 IEEE 30th International Conference on Data Engineering (ICDE), 3/31/2014 - 4/4/2014, Chicago, IL, USA, IEEE, Piscataway, NJ, 2014, pp. 1294–1297

[SSB+13] STADELMANN, T.; STOCKINGER, K.; BRASCHLER, M.; CIELIEBAK, M.; BAUDINOT, G.; DÜRR, O.; RUCKSTUHL, A.: Applied data science in Europe: Challenges for academia in keeping up with a highly demanded topic: 9th European Computer Science Summit, Amsterdam, Niederlande, 8-9 October 2013, 2013

[SSE+14] STEENSTRUP, K.; SALLAM, R. L.; ERIKSEN, L.; JACOBSON, S. F.: Industrial analytics revolutionizes big data in the digital business. Gartner Research, 2014

[Sut21] SUTMÖLLER, N.: Big Data im Spannungsfeld von Wirtschaft und Gerechtigkeit. Dissertation, 2021

[Sys06] SYSKA, A.: Ishikawa-Diagramm. Produktionsmanagement: Das A—Z wichtiger Methoden und Konzepte für die Produktion von heute, 2006, S. 63–65

[SZ[21] SINGH, A.; ZAMAN, N.; [NACHNAME NICHT VORHANDEN], K.: Understanding of Data Preprocessing for Dimensionality Reduction Using Feature Selection Techniques in Text Classification A Systematic Literature Review of Machine Learning Estimation Approaches used in Scrum Projects View project Clustering Schemes in Big Data using Image Processing View project, 2021

[SZM18] STANULA, P.; ZIEGENBEIN, A.; METTERNICH, J.: Machine learning algorithms in production: A guideline for efficient data source selection. Procedia CIRP, (78), 2018, S. 261–266

[SZY+16] STEINER, S.; ZENG, Y.; YOUNG, T. M.; EDWARDS, D. J.; GUESS, F. M.; CHEN, C.-H.: A Study of Missing Data Imputation in Predictive Modeling of a Wood-Composite Manufacturing Process. Journal of Quality Technology, (48)3, 2016, S. 284–296

[TAR+19] TUGGENER, L.; AMIRIAN, M.; ROMBACH, K.; LORWALD, S.; VARLET, A.; WESTERMANN, C.; STADELMANN, T.: Automated Machine Learning in Practice: State of the Art and Recent Results, (70), 2019, S. 31–36

[Ten01] TENG, C.-M.: A Comparison of Noise Handling Techniques: FLAIRS Conference, 2001, S. 269–273

[TH02] THOMKE, S.; HIPPEL, E. VON: Innovators. Harvard business review, (80)4, 2002, S. 74–81

[TP17] TRIPATHY, M.; PANDA, A.: A Study of Algorithm Selection in Data Mining using Meta-Learning. Journal of Engineering Science and Technology Review, (10)2, 2017, S. 51–64

[TQL18] TAO, F.; QI, Q.; LIU, A.: Data-driven smart manufacturing. Journal of Manufacturing Systems, (48), 2018, S. 157–169

[TSK16] TAN, P.-N.; STEINBACH, M.; KUMAR, V.: Introduction to data mining. Pearson Education India, 2016

[TT23] TJADEN, J.; TJADEN, B.: MLpronto: A tool for democratizing machine learning. PloS one, (18)11, 2023, e0294924

[Tya03] TYAGI, S.: Using data analytics for greater profits. Journal of Business Strategy, (24), 2003, S. 12–14

[TZ21] TIANXING, M.; ZHUKOVA, N.: The Data Mining Dataset Characterization Ontology, 2021, S. 231–238

[UE16] ULRICH, K. T.; EPPINGER, S. D.: Product design and development. 6th Edition, McGraw-Hill, New York, NY, 2016

[UW01] UNNEBRINK, K.; WINDELER, J.: Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Statistics in medicine, (20)24, 2001, S. 3931–3946

[VA16] VERMA, J. P.; AGRAWAL, S.: Big Data Analytics: Challenges And Applications For Text, Audio, Video, And Social Media Data. International Journal on Soft Computing, Artificial Intelligence and Applications, (5)1, 2016, S. 41–51

[VAE+19] VOET, H.; ALTENHOF, M.; ELLERICH, M.; SCHMITT, R. H.; LINKE, B.: A Framework for the Capture and Analysis of Product Usage Data for Continuous Product Improvement. Journal of Manufacturing Science and Engineering, (141)2, 2019

[Van19] VANSCHOREN, J.: Meta-Learning. In: Hutter, F.; Kotthoff, L.; Vanschoren, J. (Hrsg.): Automated Machine Learning: Methods, Systems, Challenges. Springer International Publishing, Cham, 2019, S. 35–61

[Vas96] VASEGHI, S. V.: Advanced Signal Processing and Digital Noise Reduction. Vieweg+Teubner Verlag, Wiesbaden, 1996

[VDI22] VDI/VDE 3714 Blatt 1 - Implementation and operation of big data applications in the manufacturing industry - Analysis process, 2022

[vom07] VOM BROCKE, J.: Design principles for reference modeling: reusing information models by means of aggregation, specialisation, instantiation, and analogy: Reference modeling for business systems analysis. IGI Global, 2007, S. 47–76

[VPB16] VENABLE, J.; PRIES-HEJE, J.; BASKERVILLE, R.: FEDS: a Framework for Evaluation in Design Science Research. European Journal of Information Systems, (25)1, 2016, S. 77–89

[VPB17] VENABLE, J. R.; PRIES-HEJE, J.; BASKERVILLE, R. L.: Choosing a design science research methodology, 2017

[VRW+23] VILLANUEVA ZACARIAS, A. G.; REIMANN, P.; WEBER, C.; MITSCHANG, B.: AssistML: an approach to manage, recommend and reuse ML solutions. International Journal of Data Science and Analytics, (16)4, 2023, S. 455–479

[VSV+21] VENKATA VARA PRASAD, D.; SENTHIL KUMAR, P.; VENKATARAMANA, L. Y.; PRASANNAMEDHA, G.; HARSHANA, S.; JAHNAVI SRIVIDYA, S.; HARRINEI, K.; INDRAGANTI, S.: Automating water quality analysis using ML and auto ML techniques. Environmental research, (202), 2021, S. 111720

[vV15] VAN RIJN, J. N.; VANSCHOREN, J.: Sharing RapidMiner Workflows and Experiments with OpenML: MetaSel@ PKDD/ECML, 2015, S. 93–103

[WD10] WIMALASURIYA, D. C.; DOU, D.: Components for information extraction. In: Huang, X. J. (Hrsg.): CIKM'10 – Proceedings of the 19th International Conference on Information & Knowledge Management and Co-Located Workshops; Oktober 26-30, 2010, Toronto, Ontario, Canada. CIKM '10: International Conference on Information and Knowledge Management, 26 10 2010 30 10 2010, Toronto ON Canada, ACM, New York, NY, 2010, S. 9–18

[WE17] WEBSTER, J. G.; EREN, H.: Measurement, Instrumentation, and Sensors Handbook: Electromagnetic, Optical, Radiation, Chemical, and Biomedical Measurement. CRC Press, 2017

[WFH+18] WILBERG, J.; FAHRMEIER, L.; HOLLAUER, C.; OMER, M.: DERIVING A USE PHASE DATA STRATEGY FOR CONNECTED PRODUCTS: A PROCESS MODEL: Proceedings of the DESIGN 2018 15th International Design Conference. 15th International Design Conference, May, 21-24, 2018, Design Conference Proceedings, Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia; The Design Society, Glasgow, UK, 2018, S. 1441–1452

[Win92] WINSTON, P. H.: Artificial intelligence. Addison-Wesley Longman Publishing Co., Inc, 1992

[Wol96] WOLPERT, D. H.: The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation, (8)7, 1996, S. 1341–1390

[Wor11-ol] WORLD ECONOMIC FORUM: Personal Data – The Emergence of a New Asset Class. Unter: https://www.weforum.org/reports/personal-data-emergence-new-asset-class/

[WRW+20] WANG, D.; RAM, P.; WEIDELE, D. K. I.; LIU, S.; MULLER, M.; WEISZ, J. D.; VALENTE, A.; CHAUDHARY, A.; TORRES, D.; SAMULOWITZ, H.; AMINI, L.: AutoAI. In: ACM Special Interest Group on Artificial Intelligence (Ed.): Proceedings of the 25th International Conference on Intelligent User Interfaces Companion. IUI '20: 25th International Conference on Intelligent User Interfaces, 17 03 2020 20 03 2020, Cagliari Italy, ACM Digital Library, Association for Computing Machinery, New York,NY,United States, 2020, pp. 77–78

[WS96] WANG, R. Y.; STRONG, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of management information systems, (12)4, 1996, S. 5–33

[WTH+17] WILBERG, J.; TRIEP, I.; HOLLAUER, C.; OMER, M.: Big Data in product development: Need for a data strategy. In: IEEE (Hrsg.): 2017 Portland International Conference on Management of Engineering and Technology (PICMET), 2017, S. 1–10

[Wu09] WU, X. (Ed.): The top ten algorithms in data mining – … the IEEE International Conference on Data Mining identified the top 10 algorithms in data mining for presentation at ICMD '06 in Hong Kong. Chapman & Hall, London, 2009

[Wu12] WU, J. (Ed.): Advances in K-means clustering – A data mining thinking. Zugl: Tsinghua Univ., Diss., 2010. Springer, Heidelberg, 2012

[WWO+20] WEIDELE, D. K. I.; WEISZ, J. D.; ODUOR, E.; MULLER, M.; ANDRES, J.; GRAY, A.; WANG, D.: AutoAIViz. In: Paternò, F. (Ed.): Proceedings of the 25th International Conference on Intelligent User Interfaces. IUI '20: 25th International Conference on Intelligent User Interfaces, 17 03 2020 20 03 2020, Cagliari Italy, ACM Digital Library, Association for Computing Machinery, New York,NY,United States, 2020, pp. 308–312

[Zah20-ol] ZAHARIA, M., LITZEL, N.: Die Evolution der Data Teams - Data Engineers und Data Scientists werden sich annähern. Unter: https://www.bigdata-insider.de/die-evolution-der-data-teams-data-engineers-und-data-scientists-werden-sich-annaehern-a-922569/

[Zim20] ZIMMERMANN, A.: Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey. WIREs Data Mining and Knowledge Discovery, (10)2, 2020

[Zio19-ol] ZIONTS, M.: Empowering The Citizen Data Scientist. Unter: https://www.forbes.com/sites/forbestechcouncil/2019/02/20/empowering-the-citizen-data-scientist/

[ZRF18] ZHANG, H.; RAO, H.; FENG, J.: Product innovation based on online review data mining: a case study of Huawei phones. Electronic Commerce Research, (18)1, 2018, S. 3–22

[ZSB15] ZAKIR, J.; SEYMOUR, T.; BERG, K.: Big Data Analytics. Issues in Information Systems, (16)2, 2015

[ZSM+19] ZIEGENBEIN, A.; STANULA, P.; METTERNICH, J.; ABELE, E.: Machine Learning Algorithms in Machining: A Guideline for Efficient Algorithm Selection. In: Schmitt, R.; Schuh, G. (Eds.): Advances in Production Research – Proceedings of the 8th Congress of the German Academic Association for Production Technology (WGP), Aachen, November 19-20, 2018. 2019, Cham, Springer International Publishing, Cham, 2019, pp. 288–299

Anhang

Inhaltsverzeichnis

A1 Betriebsdatenübersicht

Statusdaten	Produktverhaltensdaten	Nutzungsdaten	Nutzerverhaltensdaten	Servicedaten
• eingebaute physikalische Elemente (komplexes System, geringe Variationen) • eingebaute physikalische Elemente (einfaches System) • Hardware-Konfiguration • Hardwarestand (lokales Protokoll über Sensorik) • Hardwarestatus (übertragenes Protokoll) • Hardwarestatus (lokale Speicherung von Zuständen) • Hardwarestatus (über Mensch/ Protokoll) • Werkseinstellungen • Versionsnummern • aktuelle Lizenzen • installierte Updates (Updateprotokoll) • Software-Status (Zustand, Konfiguration) • Warnmeldung (über Mensch/ Protokoll) • Warnmeldung (von der Software) • Fehlermeldungen • Standstillmeldung • Laufzeit • Betriebsart • Zeit und Ort	• Aktordaten für einen einzelnen Akteur (selektiv gesteuert) • Stellglieddaten für einen einzelnen Akteur (kontinuierlich gesteuert) • Aktordaten für ein komplexes Gesamtsystem • Sensordaten (z.B. Temperatur, Feuchte, Druck, Nähe, Füllstand, Beschleunigung) • Schwingungsdaten (Sensor) • Energieverbrauch • Störungszeiten und Stillstandszeiten • Produktionsquantität • Gutmenge • Ausschussmenge • Arbeitsbelastung	• Bestellung und Auftrag • Aggregierte Benutzeraktivitäten (z. B. Nutzung von Funktionen)	• Benutzeraktivitätsprotokoll/ Protokoll • Nutzungsprozess/ Interaktionspfad • Aktivitätsdaten über Benutzerschnittstellen • persönliche Mitarbeiterdaten • Benutzeranmeldung	• Serviceberichte (automatisiert) • Reparaturprotokoll • Wartungsprotokoll • Garantiefall • Kundenbeschwerden • Kundenrezensionen/ Bewertungen • Anregungen von Kunden

Bild A-1: Betriebsdatenübersicht

A2 Taxonomien der Betriebsdateneigenschaften

A2.1 Taxonomie der allgemeinen Betriebsdateneigenschaften

	Dimension	Ausprägungen				Indikator (beispielhaft)
Datenmerkmale (allgemein)	**Datensatz-gruppe (Typ)**	Tabel-larische Daten (strukturiert)	Relationale Daten (z. B. Tabelle, Datenmatrix)			Feste Anzahl an Daten-feldern, keine Verknüpfung zwischen Einträgen
			Graphenbasierte Daten			Verbindungen, Objektbeziehungen
			Geordnete Daten (zeitliche oder räumliche Ordnung)	Sequentielle Transaktionsdaten		Itemsets + Zeitangabe (keine bestimmte Frequenz)
				Sequenzdaten		Geordnet ohne Zeitstempel
				Zeit-reihen-daten	Signal	Messbare physikalische Parameter
					Kein Signal	Messungen über die Zeit im Allgemeinen
				Räumliche Daten		Positionen, Bereiche
		Text	Unstrukturiert			Fehlendes Format (keine Trennungen der Informationen)
			Halbstrukturiert			Tags, Metadaten
		Bild				Metadaten
		Video				Metadaten
	Dimen-sionalität	Niedrigdimensional				Deutlich weniger Features als Beobachtungen
		Hochdimensional				Mehr Features als Beobachtungen, > 100 Features

	Dimension	Ausprägungen	Indikator (beispielhaft)
Datenmerkmale (allgemein)	**Verteilung**	Spärlich	Viele Lücken in den Daten, z. B. werden nur Veränderungen sichtbar
		Dicht	Kontinuierlich aufgezeichnete Werte
	Komplexität	(Auto-)korreliert	Die Datenpunkte sind stark von ihren Nachbarn abhängig
		Unkorreliert	Die Datenpunkte sind nicht von ihren Nachbarn abhängig
	Echtzeit-verhalten	Echtzeit/ Live	Lieferung der Daten unmittelbar nach deren Erstellung
		Statisch (Batch)	Die daten werden nur einmal pro Tag übertragen/ die daten müssen vom Gerät abgegriffen werden
	Volumen (pro Tag)	Gering	Mehrere hundert Datenzeilen
		Normal	Mehrere tausend Datenzeilen
		Groß	Mehrere hunderttausend Datenzeilen

Bild A-2: Taxonomie der allgemeinen Betriebsdatenmerkmale

A2.2 Taxonomie der individuellen Betriebsdateneigenschaften

	Dimension	Ausprägungen	Bewertung	Indikator (beispielhaft)
Datenmerkmale (individuell)	Daten-qualitäts-probleme	Systematische Fehler (Datenfehler z. B. Miskalibrierung verursacht)	Zu vernachlässigen	Konstante systematische Fehler (Lieferantenspezifikation oder Domänenwissen), wo nur Verhältnisse interessant sind
			Zu berücksichtigen	Systematische Fehler, die korrigiert werden können
			Dominant	untolerierbare Sensorfehler
		Zufälliges Rauschen (Rauschverteilung folgt keinem bekanntem Modell)	Zu vernachlässigen	Zufällige Fehler, die sehr selten sind
			Zu berücksichtigen	Zufällige Fehler
			Dominant	Zufällige Fehler, die die Daten dominieren
		Ausreißer	Zu vernachlässigen	Keine oder isolierte Ausreißer
			Zu berücksichtigen	Ausreißer können klar identifiziert werden
			Dominant	Ausreißer dominieren
		Inkonsistenz (Frequenzen, Einheit, Wertebereich)	Zu vernachlässigen	Frequenzen und Einheiten sind einheitlich
			Zu berücksichtigen	Variablen haben signifikant verschiedene Wertebereiche, Frequenzen variieren
			Dominant	Informationen zu Einheiten und Wertebereichen fehlen
		Fehlende Werte	Zu vernachlässigen	Keine oder nur gelegentlich fehlende Daten
			Zu berücksichtigen	Fehlende Daten sind in der Minderheit
			Dominant	Fehlende Daten überwiegen
		Duplikate	Ja	Duplikate können klar identifiziert werden
			Nein	Keine Duplikate
	Variablentyp	Kategorisch / Qualitativ		Daten haben keine natürliche Ordnung, sind Namen (nominal) oder Daten haben eine natürliche Ordnung (ordinal)
		Numerisch		Daten haben natürliche Ordnung mit quantisierbaren Abständen
		Spezial / Hybridform		Daten haben zwei Werte (1/0); Zeitangaben (z. B. Jahr, Monat, Sekunde)

Bild A-3: Taxonomie der individuellen Betriebsdatenmerkmale

A3 Python-Tool-Übersicht

	Datenbereinigung					Datentransformation	
	Fehlende We	Ausreißer	Noise	Textbereinigu	Systematische	Skalierung un	Transformati
Scikit-Learn	Ja	Ja	Ja	Ja	Ja	Ja	Ja
XGBoost	Nein	Nein	Nein	Nein	Nein	Nein	Nein
LightGBM	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Catboost	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Annoy	Nein	Nein	Nein	Nein	Nein	Nein	Nein
H20ai	Ja	Ja	Ja	Ja	Ja	Ja	Ja
StatsModels	Nein	Nein	Nein	Nein	Ja	Nein	Nein
Pattern	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Prophet	Nein	Nein	Nein	Nein	Nein	Nein	Nein
TPOT	Nein	Nein	Nein	Nein	Nein	Nein	Nein
auto-sklearn	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Numpy	Nein	Nein	Nein	Nein	Nein	Ja	Ja
Pandas	Ja	Ja	Ja	Ja	Ja	Ja	Ja
Matplotlib	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Feature Engir	Ja	Ja	Ja	Ja	Ja	Ja	Ja
SciPy	Nein	Ja	Ja	Nein	Nein	Nein	Nein
Dtreeviz	Nein	Nein	Nein	Nein	Nein	Nein	Nein
category_enc	Nein	Nein	Nein	Nein	Nein	Nein	Nein
tslearn	Nein	Nein	Nein	Nein	Nein	Ja	Ja
sktime	Ja	Ja	Ja	Nein	Ja	Ja	Ja
TensorFlow	Ja	Ja	Ja	Nein	Nein	Ja	Ja
Keras	Nein	Nein	Nein	Nein	Nein	Ja	Ja
mljar	Nein	Nein	Nein	Nein	Nein	Nein	Nein
layzpredict	Nein	Nein	Nein	Nein	Nein	Nein	Nein
nltk	Nein	Nein	Nein	Ja	Nein	Nein	Nein
TextBlob	Nein	Nein	Nein	Ja	Nein	Nein	Nein
PM4PY	Ja	Ja	Nein	Nein	Nein	Nein	Ja

Bild A-4: Datenbereinigung und Datentransformation der Python-Tools

	Feature Engineering		
	FE Zeitreihen/	FE Text	Feature Selek
Scikit-Learn	Ja	Ja	Ja
XGBoost	Nein	Nein	Nein
LightGBM	Nein	Nein	Nein
Catboost	Nein	Nein	Nein
Annoy	Nein	Nein	Nein
H20ai	Ja	Ja	Ja
StatsModels	Nein	Nein	Nein
Pattern	Nein	Nein	Nein
Prophet	Nein	Nein	Nein
TPOT	Nein	Nein	Nein
auto-sklearn	Ja	Ja	Nein
Numpy	Nein	Nein	Nein
Pandas	Ja	Ja	Ja
Matplotlib	Nein	Nein	Nein
Feature Engir	Ja	Ja	Ja
SciPy	Ja	Ja	Ja
Dtreeviz	Nein	Nein	Nein
category_enc	Nein	Nein	Nein
tslearn	Ja	Nein	Nein
sktime	Ja	Nein	Ja
TensorFlow	Ja	Ja	Ja
Keras	Nein	Nein	Nein
mljar	Ja	Ja	Ja
layzpredict	Nein	Nein	Nein
nltk	Nein	Ja	Nein
TextBlob	Nein	Ja	Nein
PM4PY	Nein	Nein	Nein

Bild A-5: Feature Engineering der Python-Tools

	Modellierung						
	Beschreibung	Clustering	Klassifikation	Regression	Abhängigkeits	Assoziationsa	Text Mining
Scikit-Learn	Ja	Ja	Ja	Ja	Nein	Nein	Ja
XGBoost	Ja	Ja	JA	Ja	Ja	Ja	Ja
LightGBM	Nein	Nein	Ja	Ja	Nein	Nein	Nein
Catboost	Ja	Nein	Ja	Ja	Ja	Ja	Ja
Annoy	Nein	Nein	Nein	Nein	Nein	Nein	Nein
H20ai	Ja	Ja	Ja	Ja	Ja	Ja	Ja
StatsModels	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Pattern	Ja	Ja	Ja	Ja	Ja	Ja	Ja
Prophet	Nein	Nein	Nein	Nein	Nein	Nein	Nein
TPOT	Nein	Nein	Nein	Nein	Nein	Nein	Nein
auto-sklearn	Nein	Nein	Ja	Ja	Nein	Nein	Nein
Numpy	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Pandas	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Matplotlib	Nein	Nein	Nein	Nein	Nein	Nein	Nein
Feature Engin	Ja	Ja	Nein	Nein	Nein	Ja	Ja
SciPy	Nein	Ja	Nein	Nein	Ja	Nein	Nein
Dtreeviz	Nein	Nein	Nein	Nein	Nein	Nein	Nein
category_enc	Nein	Nein	Nein	Nein	Nein	Nein	Nein
tslearn	Nein	Ja	Ja	Ja	Nein	Nein	Nein
sktime	Nein	Ja	Ja	Ja	Nein	Nein	Nein
TensorFlow	Nein	Ja	Ja	Ja	Ja	Nein	Ja
Keras	Nein	Ja	Ja	Ja	Nein	Nein	Ja
mljar	Ja	Ja	Ja	Ja	Nein	Nein	Nein
layzpredict	Nein	Nein	Ja	Ja	Nein	Nein	Nein
nltk	Nein	Ja	Ja	Nein	Nein	Nein	Ja
TextBlob	Nein	Nein	Ja	Nein	Nein	Nein	Ja
PM4PY	Ja	Nein	Nein	Nein	Ja	Nein	Nein

Bild A-6: Modellierung der Python-Tools

A4 Wissensbasis

Link zur Wissensbasis (Google Drive)

A5 Interviewanalyse

Tabelle A-1: Strukturierte Auswertung der Nutzerinterviews mit Zitaten

Thema	Subthema	Interviewte	Exemplarische Zitate
Nutzen	Gute Übersicht (über mögliche Anwendungen und Verfahren)	1,3	„Für Data Science sehr geeignet, weil man damit schnell einen Überblick bekommt, was für so ein Projekt notwendig ist.“
	Richtige Fragestellungen	3	„Nötig für so ein Projekt ist nämlich zu wissen, welche Fragen man sich stellen muss, welche auch nicht…“
	Dokumentation des Prozesses auch für Seniorexperten	2	„Grundsätzlich halte ich das für ein super Tool, das Senior Data Scientists hilft, sich an die Prozesse zu halten.“
	Keine Vorkenntnisse über Methoden erforderlich	1, (2,5)	„Ich finde es total hilfreich, um zu einem Verfahren zu kommen. Und da werden gar nicht Kenntnisse darüber vorausgesetzt, dass man die Modelle kennt. Von daher finde ich das sehr hilfreich.“
	Nachvollziehbarkeit des Prozesses	5	„Ich kann jeden Schritt nachvollziehen.“
	"Learning on the job"	5	„Ich finde das Tool sehr hilfreich, da ich als Neuling direkt on the job die wichtigen Data-Science-Schritte erklärt bekomme.“
Befähigung	Einschränkung der Modelle/des Lösungsraums	1, 5	„Dann muss ich mich immerhin nur noch mit dem einen Modell beschäftigen und brauche auch nicht mehr das ganze Wissen über alle Verfahren“
	Dokumentation & Kommunikation	2	„Der größte Mehrwert liegt meiner Meinung nach in Dokumentation und Kommunikation. Also im Sinne nach außen zu den versch. Stakeholdern, aber auch nach innen.“
	Überblick über Techniken/Lösungsraum	1,2, 3	„Es zeigt ganz gut den Lösungsraum dafür auf, was ich mit Data Science in dem Fall tun kann, ohne, dass ich da jetzt ein 5-Mann-Team drauf jage.“
	Schnellere Umsetzung	3, 5	Mehrwert liegt meiner Meinung darin, dass man sich nicht mehr ewig im Internet nach passenden Techniken zuerst umsehen muss, darüber hinaus ist direkt der Domänenbezug gegeben.“; „Es verkürzt die Time-to-Analytics.“
	Sicherstellung über Beantwortung wichtiger Fragen	4	„Der Mehrwert wird durch den systematischen Ansatz sichergestellt, dass da die wesentlichen Fragen beantwortet werden und auch bei Leute mit einer Kompetenz von 2 nichts liegen bleibt.“
	Systematischer Ansatz & Struktur	2, 4, 5	„Die Fragen leiten systematisch durch den Prozess.“
	Verständnis über Großteil der Logik	1	„Es erklärt schon sehr viel am Rande,…., einen Großteil der Logik bekommt man damit schon mit und erklärt.“
	Einfachheit (z. B. Erklärungen)	2, 5	„Die Wissensbasis liefert schöne Erklärungen, die einfach zu verstehen sind, auch einfach als Fachtexte, die man sonst so dazu findet.“
	Verknüpfung/Visualisierung der Kausalkette	1, 3, 5	„Das Verständnis des ganzen Prozesses wird unterstützt, da man zielgerichtet durch den Prozess geführt wird und am Ende sieht man die Abhängigkeiten, also die Kausalkette, die zu dem Ergebnis geführt hat.“

	Erleichterung des Mappings	3	„Wenn ich am Ende herausfinde, was die Fragestellung im Business ist, aber ohne Data-Science-Wissen, dann wird das Mapping am Ende erleichtert.“
	Nachvollziehbare Fragen	4	„Die Fragen, die gestellt werden, sind alle super nachvollziehbar und lassen mich das auch gut in das Gesamtvorgehen einordnen.“
	Einordnung in Gesamtvorgehen möglich	4	s.o.
	Werkzeug zur gemeinsamen Bearbeitung durch Templates	5	„Die Templates in Form der Canvases bieten zusätzlich die Möglichkeit mit anderen tiefer ins Detail zu gehen und relevante Informationen gemeinsam zu erarbeiten.“
Stärken	Fokussiertes Lernen durch Einschränkung der relevanten Verfahren	1	„Die Stärke ist, dass ich wirklich nicht den Überblick über die Modelle oder Methoden haben muss, sondern mir dann nur noch die paar spezifischen Methoden ansehen muss, weil sonst ist das Feld ja schon sehr sehr groß.“
	Arbeitsersparnis	1, 2	„Und da kann ich dann einfach nur die ausgewählten Methoden konkret einsteigen und mir die angucken, das ist schon eine riesige Arbeitsersparnis.“
	Learning on the job	1	„Man lernt nach und nach neues kennen, was wofür genutzt wird, aber in kleinen Dosen, und man kann trotzdem schon damit arbeiten; man wird schnell befähigt.“
	Geführter Prozess	3	Eine Stärke ist auf jeden Fall, dass es ein geführter Prozess ist.
	Einfache Sprache (richtiger Detailgrad für Nicht-Experten)	3, 5	„Es ist sehr einfach gehalten, also es ist jetzt kein wissenschaftlicher Text.“
	Intuitive und nachvollziehbare Anwendung	4	„Die Anwendung des Tools ist grundsätzlich sehr intuitiv, sehr nachvollziehbar, gut kontextualisiert.“
	Visual Tools für Nicht-Experten	4	„Und grundsätzlich das Arbeiten mit den Visual-Tools kommen einem als Nicht-Experten im Data-Analytics-Bereich schon sehr entgegen.“
	Transparenz	4	„Du bekommst die Dinge vor allem vorgeschlagen und das macht auf jeden Fall sehr transparent, was man da dann als Ergebnis raushat und das ist sehr übersichtlich.“
Schwächen	Einstieg in die Datenwelt fehlt	1	„Gut wäre erstmal zu wissen, welche Daten denn so in der Produktplanung typisch sind, das finde ich gar nicht so einfach.“
	Noch transparentere Zusammenfassung, keine absolute Explainability	2, 3	„Man sieht die Ketten aus der Wissensbasis nicht im Frontend und man kann wahrscheinlich so nicht ganz nachvollziehen, warum Technik A statt B vorgeschlagen wird.“
	Abhängigkeit von (Stand) d. Knowledge Base	3	„Es ist stark davon abhängig, wie rigoros du den Lösungsraum erfasst hast, also wie gut die Knowledge Base ist.“
	Medienbruch	4	„Durch den Medienwechsel zum Conceptboard sind die Bezüge ja nicht direkt im Tool abgebildet, also diese Zusammenhänge sind dadurch nicht mehr ganz nachvollziehbar.“
	Vorschlagsansichten teilweise zu schnell	4	„Beim Durchführen durch die Pipeline kam mir der Sprung in das geeignete Modell etwas plötzlich.“
	Tool bildet Unschärfen des DS-Prozesses nicht ab	5	„Das Tool bildet die Unschärfen des Prozesses nicht ab, wo oft nicht so klare Wenn-

			Dann-Regeln existieren und langjährige Erfahrung oft nötig ist.“
Vertrauen	Logisch, nicht zufällig	1	„Es ist ein sehr logisches Verfahren und hat nicht so viel mit Zufallsgenerierung zu tun.“
	Befähigung der Nutzer (gute Angaben zu machen)	1	„Es ist letztendlich davon abhängig wie gut ein Nutzer seine Sachen angibt, aber da führst du echt gut durch.“
	Fundierte Wissensbasis	1, 4	„In dem Vergleich sehr vertrauenswürdig, beispielsweise durch die Hinweise auf die wissenschaftlich fundierte Wissensbasis.“
	Expertenwissen	2, 5	„Das Tool basiert ja auf Experten, denen ich mehr Vertrauen schenke als Chatbots wie ChatGPT, die ja auch viel halluzinieren und falsche Angaben machen…“
	Entscheidungen werden transparent	5	„Dazu macht das Tool die Entscheidungen recht transparent. Das ermöglicht, dass ich das besser nachvollziehen kann.“