Kessler, René; Gómez, Jorge Marx

Implikationen von Machine Learning auf das Datenmanagement in Unternehmen

HMD Praxis der Wirtschaftsinformatik

Provided in Cooperation with:
Springer Nature

Suggested Citation: Kessler, René; Gómez, Jorge Marx (2020) : Implikationen von Machine Learning auf das Datenmanagement in Unternehmen, HMD Praxis der Wirtschaftsinformatik, ISSN 2198-2775, Springer Fachmedien Wiesbaden, Wiesbaden, Vol. 57, Iss. 1, pp. 89-105, https://doi.org/10.1365/s40702-020-00585-z

HMD (2020) 57:89–105
https://doi.org/10.1365/s40702-020-00585-z

SCHWERPUNKT

Implikationen von Machine Learning auf das Datenmanagement in Unternehmen

René Kessler · Jorge Marx Gómez

Eingegangen: 12. Oktober 2019 / Angenommen: 18. Januar 2020 / Online publiziert: 4. Februar 2020

Zusammenfassung Machine Learning ist ein Forschungsfeld mit großen Potenzialen und weitreichenden Anwendungspotenzialen. Big Data kann dabei als Enabler angesehen werden, da große und qualitativ hochwertige Daten stets die Grundlage für erfolgreiche Machine Learning-Algorithmen und -Modelle darstellen. Aktuell gibt es noch keinen voll etablierten Standardprozess für den Machine Learning-Life Cycle, wie es im Data Mining mit dem CRISP-DM beispielsweise der Fall ist, was zur Folge hat, dass gerade die Operationalisierung von Machine Learning-Modellen Unternehmen vor große Herausforderungen stellen kann. In diesem Beitrag werden anhand der Sicht auf die Beschaffenheit der Daten, die verschiedenen Rollen in Machine Learning-Teams und den Lebenszyklus von Machine Learning-Modellen Implikationen für das Datenmanagement in Unternehmen herausgearbeitet.

Schlüsselwörter Datenmanagement · Life Cycle-Modell · Künstliche Intelligenz · Machine Learning · Big Data

Implications of Machine Learning on Data Management in Companies

Abstract Machine Learning is a trend research area with great potential and far-reaching application potentials. Big Data is an enabler, as large and high-quality data are always the basis for successful machine learning algorithms and models. There is currently no fully established standard process for the machine learning life cycle, as is the case in data mining with the CRISP-DM-Process, which means that the

R. Kessler (✉) · J. M. Gómez
Department of Computing Science, Business Informatics (Very Large Business Applications), University of Oldenburg, Ammerländer Heerstr. 114–118, 26129 Oldenburg, Deutschland
E-Mail: rene.kessler@uni-oldenburg.de

J. M. Gómez
E-Mail: jorge.marx.gomez@uni-oldenburg.de

operationalization of machine learning models in particular can present companies with major challenges. In this article, the implications for data management in companies are worked out on the basis of the view of the nature of the data, the various roles in machine learning teams and the life cycle of machine learning models.

Keywords Data Management · Life cycle model · Artificial Intelligence · Machine Learning · Big Data

1 Zwischen Hype und Realität: Big Data und Machine Learning

Kaum ein IT-Trend ist aktuell so prägend wie *Big Data*. Begründet werden kann das Aufkommen durch die zunehmende Vernetzung, sowohl im Alltag (z. B. Social Media) als auch in der Wirtschaft (z. B. Industrie 4.0). So steigt das global verfügbare Datenvolumen immer stärker an und wird von aktuell ca. 33 Zettabyte auf bis zu 175 Zettabyte im Jahr 2025 anwachsen (Reinsel et al. 2018). Auswirkungen sind nicht nur in der Forschung spürbar, sondern auch der Impact auf Unternehmen ist klar erkennbar und kann an der Relevanz des Suchbegriffs *Big Data* deutlich gemacht werden[1]. Verschärft wird dieser Trend insbesondere auch dadurch, dass durch die am Markt verlangte Individualisierung und Kundenorientierung ein hoher Bedarf an Echtzeitinformationen besteht, um die Customer Experience zu erhöhen (Reinsel et al. 2018). Wird dieser Kundenwunsch nicht erfüllt, sind für Unternehmen Nachteile und der Verlust von Kunden zu erwarten (Reinsel et al. 2018).

Eine weitere durch Big Data geprägte Veränderung betrifft die Datenhaltung. Während 2010 noch ca. 90 % aller Daten in unternehmensinternen Rechenzentren gespeichert wurden, geht der Trend mehr und mehr in Richtung Cloud-Nutzung. Schon heute werden 40–50 % der Daten in der Cloud gespeichert, Tendenz weiter steigend (Reinsel et al. 2018). Zu unterscheiden ist dabei aber zwischen Cloudstorage der von Unternehmen zur herkömmlichen Speicherung und Verwaltung der internen Daten genutzt wird und Open Data, also öffentlich zugänglichen, geteilten daten.

Aus *Big Data* resultieren für Unternehmen Chancen, aber auch Herausforderungen. Von der Verfügbarkeit von großen Datenmengen allein kann nicht profitiert werden. Um unternehmerisch zu profitieren, müssen diese Daten verarbeitet und analysiert werden. Entscheidungen werden in Unternehmen häufig auf Basis von anekdotischer Evidenz getroffen, also basierend auf Gefühlen oder Erfahrungen („das haben wir schon immer so gemacht“). Besser, objektiver und von höherer Güte sind allerdings Entscheidungen, die auf empirischer Evidenz beruhen (z. B. quantifizierbare Kennzahlen) (Anderson 2015; BARC 2017; Hoeken und Hustinx 2009; Beisswenger 2016; Brynjolfsson und McElheran 2016; McElheran und Brynjolfsson 2017). Der zukünftige Weg von Unternehmen sollte es daher sein, die großen verfügbaren Datenmengen zu verarbeiten und analysieren, um die gewonnen Erkenntnisse in die Entscheidungsprozesse einzubeziehen, bis hin zur vollständigen Etablierung von datengetriebenen Entscheidungsprozessen.

[1] https://trends.google.com/trends/explore?data=all&q=Big%20Data.

In der Vergangenheit wurde bei der Extraktion von Information und der Generierung von Erkenntnissen vor allem auf die Anwendung von klassischen analytischen Verfahren gesetzt. Eine weitere und aktuell sehr prägende Disziplin der Datenwissenschaft, die bei der Entwicklung solcher Data Products genutzt werden kann, ist die künstliche Intelligenz (KI) oder im speziellen Machine Learning und Deep Learning (Polyzotis et al. 2018). Die Bedeutung dieser Disziplin und die damit verknüpften Erwartungen können anhand verschiedener Indikatoren deutlich gemacht werden:

- **Technologischer Fortschritt** (McKinsey & Company 2017): KI ist nicht neu, sondern hat ihren Ursprung in den 1950er Jahren. Schon immer wurden große Erwartungen an die Anwendung der KI gestellt, diese konnten aber in der Vergangenheit selten erfüllt werden, weshalb das Interesse aus Forschung und Praxis zyklisch verlaufen ist und von mehreren KI-Wintern, also Zeiträumen in denen kaum Interesse an diesem Thema bestand, gesprochen werden kann. Mittlerweile ist die Entwicklung von Hardware und Datenbanken aber so weit fortgeschritten, dass die notwendige Leistung für die Entwicklung und Anwendung von KI-Algorithmen zu einem bezahlbaren Preis vorhanden ist. Ebenso wurden Fortschritte und neue Erkenntnisse im Bereich der Algorithmik erzielt. Durch das gesteigerte Interesse in den letzten Jahren haben sich in Kooperation von Forschung und Praxis zudem Standards etabliert und Frameworks zur schnelleren Implementierung und Umsetzung von KI-Verfahren sind entstanden.
- **Branchen- und anwendungsübergreifende Potenziale** (Batra et al. 2018): Die Anwendungsmöglichkeiten von KI sind sehr breit gefächert und decken sämtliche Zweige der Wirtschaft ab, wie z. B. Mobilität, Energiewirtschaft, Finanzdienstleistungen/Versicherungswesen, Gesundheitswesen/Medizin, Produktion, Konsumelektronik, Landwirtschaft, Logistik, Marketing, Recht oder Sicherheit (Polyzotis et al. 2018; Hecker et al. 2017). Überall wo also Daten vorliegen und verarbeitet werden sollen, kann potenziell der Einsatz von KI sinnvoll sein. Die Forschung kann hiervon profitieren, da durch die vielen Einsatzgebiete die Diskussion angeregt werden kann und so neue Erkenntnisse generiert oder bereits etablierte Verfahren in neue Domänen übertragen werden können.
- **Hohe Erwartungen in der Praxis** (Holst 2018): Die Investitionen in die Disziplin KI seitens der Unternehmen sind in den letzten Jahren massiv angestiegen. Während im Jahr 2013 ca. 4,5 Mrd US $ investiert wurden, ist dieser Betrag im Jahr 2017 schon bei 39,2 Mrd US $ angelangt. Wird nun beachtet, dass der KI-Hype sich ab 2017 weiter beschleunigt hat (medialer Hype, bspw. durch Googles AlphaGo und zahlreiche politische KI-Intitiativen), kann davon ausgegangen werden, dass die Investitionen für die Folgejahre noch höher liegen und auch weiter ansteigen werden. Bestätigt wird dieser Eindruck auch durch die hohe Anzahl an Unternehmen, die sich auf die Entwicklung von KI-Anwendungen spezialisiert haben (4925 UN im Jahr 2018). Auffällig ist auch, dass insbesondere große Tech-Unternehmen wie Amazon, IBM, Microsoft oder Google die Entwicklungen massiv vorantreiben und große Forschungsabteilungen für KI eingerichtet haben. Deutlich wird dies insbesondere bei der Betrachtung der von diesen großen Tech-Unternehmen angemeldeten Patenten und akquirierten Startups.

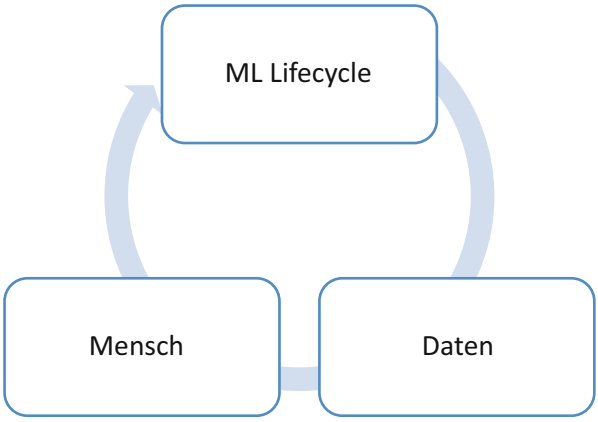

Abb. 1 Herausforderungen für das unternehmensweite Datenmanagement

Eine zentrale Herausforderung für Unternehmen ist es daher in der heutigen Zeit, wie die Entwicklungen im Technologieumfeld aufgegriffen und zum eigenen Vorteil genutzt werden können, mit dem Ziel die Marktposition zu halten oder sogar zu stärken (Bundesministerium für Wirtschaft und Energie 2019). Dabei resultieren verschiedene Anforderungen an Unternehmen. Natürlich muss zunächst Know How im Unternehmen aufgebaut werden, wenn künstliche Intelligenz adaptiert werden soll. Unzureichend ist es allerdings, wenn der Fokus ausschließlich auf die Erarbeitung und Implementierung von einzelnen KI-Use Cases gelegt wird. Neben den fachlichen Anforderungen resultieren aus diesem Adaptions- bzw. Transformationsprozess hin zu einer neuen technologischen Ausrichtung organisatorische Anforderungen, wie z. B. Schulungen zum Know-How-Aufbau oder Maßnahmen zur Akzeptanzsteigerung. Dieser Veränderungsprozess dauert länger, je weniger agil ein Unternehmen agieren kann (Hannan und Freeman 1977).

Die Grundlage eines jeden Machine Learning-Algorithmus stellt eine ausreichende und qualitativ hochwertige Datenbasis dar. Modelle hoher Güte, also zuverlässige, robuste und nachvollziehbare Modelle, benötigen nicht nur passende und dem Anwendungsfall entsprechend ausgewählte Algorithmen, sondern eben auch qualitativ hochwertige und in großer Menge vorliegende Daten, da aus diesen gelernt wird (Polyzotis et al. 2018). Daher spielt das unternehmensweite Datenmanagement innerhalb von ML-Projekten und -Strategien eine elementare Rolle. Herausforderungen und Implikationen die aus der Anwendung von KI für das unternehmensweite Datenmanagement resultieren, werden im weiteren Verlauf dieses Beitrags aus dem Machine Learning-Life Cycle, der Zusammensetzung von ML-Teams und den zugrundeliegenden Daten abgeleitet (siehe Abb. 1) und erläutert.

2 Modelle des Machine Learning-Life Cycles

2.1 Bestehende Ansätze

Der Lebenszyklus eines Machine Learning-Modells erstreckt sich von der Konzeption über die Implementierung bis hin zum Deployment des Modells und dessen Wartung im Betrieb. Phasenmodelle oder Life Cycle Modelle sind aktueller Gegenstand der Forschung und Praxis. Während es im klassischen Data Mining mit dem CRISP-DM und seinen Phasen *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment* ein voll etabliertes Standard-Life Cycle Modell gibt (Shearer 2000), ist dies im Bereich des Machine Learnings anders. Aktuell sind in der Literatur und in der Wirtschaft verschiedene Methodologien zu finden. Zwar haben die meisten einen starken Bezug zum CRISP-DM-Modell, da es sich auch bei Machine Learning um datengetriebene Analysen handelt, allerdings unterscheiden diese sich in einigen für den Machine Learning-Bereich relevanten Phasen oder ergänzen weitere. Herkömmliche Algorithmen können als deterministisch angesehen werden, da deren Zustände zu jeder Zeit im Programmcode nachvollzogen werden können. Im Bereich des Machine Learnings ist dies anders. Zwar können auch hier Kennzahlen oder Metriken herangezogen werden, allerdings ist der interne Zustand eines Modells und dessen Layer nicht aussagekräftig. Dies hat zur Folge, dass bestehende Vorgehensmodelle wie das CRISP-DM Modell und insbesondere die Ausgestaltung der einzelnen Phasen nicht abgestimmt sind auf die Anforderungen, die in der Domäne des Machine Learnings bestehen. Wie in Abschn. 1 beschrieben, wird die Entwicklung im Forschungsfeld KI besonders durch große Tech-Unternehmen geprägt und vorangetrieben (Holst 2018). Im Fokus der Recherche zu den bestehenden Lifecycle-Modellen standen also vor allem diese Unternehmen und deren Veröffentlichungen.

IBM stellte im Jahr 2017 einen Prozess vor, der sich stark am CRISP-DM Modell orientiert (Allen et al. 2017). Ergänzt wurden dabei die Phasen *Plan, Build ground truth, Monitor* sowie *Capture Feedback*. Neu hinzugekommen ist zudem ein iterativer Feedbackloop bei der Modellerstellung. Intensiv betrachtet wurde in diesem Ansatz die Vorverarbeitung von Daten und das Training der Modelle. Unklar bleibt hierbei die Schaffung der Datengrundlage und die Nutzung von verschiedenen Datenquellen. Auch das zugrundeliegende Datenmanagement wird dabei nicht näher untersucht. Ziel des Ansatzes war es zudem, DevOps-Ansätze auf AI-Anwendungen zu übertragen, weshalb besonderes Augenmerk auf den Betrieb und das Monitoring von Modellen gelegt wurde.

Microsoft veröffentlichte im Jahr 2017 einen Ansatz für das Life Cycle Management von Data Science-Anwendungen, der aber auch „intelligente Applikationen“ umfasst (Ericson et al. 2017). Dieser Lebenszyklus besteht dabei aus den fünf Phasen *Business Understanding, Data Acquisition and understanding, Modeling, Deployment* und *Customer Experience*. Auffällig ist hierbei, die Datensammlung und -vorverarbeitung im Gegensatz zu anderen Modellen sehr ausführlich beschrieben wird. Die Optimierung von bereits deployten Modellen wird allerdings nur am Rande untersucht.

Einen ähnlichen Ansatz wählte das Unternehmen Oracle im Jahr 2019 (Talagala 2019). Der vorgestellte ML-Lifecycle orientiert sich ebenfalls an den Phasen des CRISP-DM-Prozesses. Die wesentliche Veränderung besteht darin, dass die Modeling-Phase in die Phasen *Develop Models, Train Models* und *Test Models* aufgesplittet wurde. Ebenso wurde eine Phase für das Monitoring und Optimieren von Modellen eingeführt. Die genaue Ausgestaltung dieser Phasen mit Aktivitäten bleibt dagegen offen.

Forscher von Apple beschreiben in ihrer Veröffentlichung aus dem Jahr 2019 einen ML-Workflow in dem das Datenmanagement fokussiert wird (Agrawal et al. 2019). Die Ausgangsbasis, die Rohdaten, werden in einer ersten Phase exploriert und annotiert, um diese als gepflegte Daten in einem Data Store zu verwalten. Im Bezug auf die Datenstruktur wird dabei ein Konstrukt aus *Curated Data, Annotations, Splits* und *Packages* eingeführt. Basierend auf dieser vorverarbeiteten Datenbasis kann dann im Schritt *Feature Engineering* ein für das ML-Modell sinnvolles Featureset ermittelt und festgelegt werden. In den Schritten *Experimentation und Evaluation* können dann erste ML-Modelle auf diesem festgelegten Featureset trainiert und evaluiert werden. Je nach Ausgang der Evaluation werden dann neue Features hinzugefügt oder sogar weitreichendere Änderungen der Daten und deren Annotation vorgenommen.

Ein weiterer Ansatz für den Data Lifecycle in produktiven ML-Umgebungen wird in Polyzotis et al. (2018) dargestellt. Dabei wird ein Zyklus mit den drei Phasen *Data Understanding, Data Validation & Cleaning* und *Data Preparation* beschrieben. Jede Phase beinhaltet dabei weitere Subphasen. In der Phase *Data Understanding* soll durch einen *Sanity Check* und ersten deskriptiven Analysen der Daten der Zustand der vorliegenden Daten ermittelt werden. Insbesondere Fehler, unzureichende Qualität und die Aussagekraft der Datenquelle sollen festgestellt werden. In der Phase *Data Validation & Cleaning* werden die zuvor identifizierten Fehler oder Optimierungspotenziale in den Daten umgesetzt. Beispielsweise könnten Daten transformiert oder einzelne Datenfelder vereinheitlicht werden. Die dritte und letzte Phase des Beitrags behandelt die *Data Preparation* mit den Subphasen *Feature Engineering* und *Data Enrichement*. Im Kern geht es hierbei um die Zusammenstellung der optimalen Features, die in das Training der Modelle eingehen können. Zudem wird in dieser Phase untersucht, ob weitere Datenquellen einbezogen werden können, um die bestehende Datenbasis zu erweitern.

Bei Betrachtung dieser Ansätze wird klar, dass kein Ansatz den Ansprüchen eines ganzheitlichen Machine Learning-Lifecycles gerecht wird, da in keinem Ansatz das gesamte Spektrum von der Erschließung von Datenquellen bis zur Optimierung der bereits in den Betrieb überführten Modellen in voller Tiefe abgedeckt wird. Nötig war es daher, basierend auf den gewonnenen Erkenntnissen und aus zahlreichen Projekten identifizierten Anforderungen einen Life Cycle-Ansatz zu konzipieren, der sich ebenfalls am etablierten CRISP-DM-Modell orientiert und anhand dessen weitere Implikationen für das Datenmanagement in Unternehmen verdeutlicht werden können.

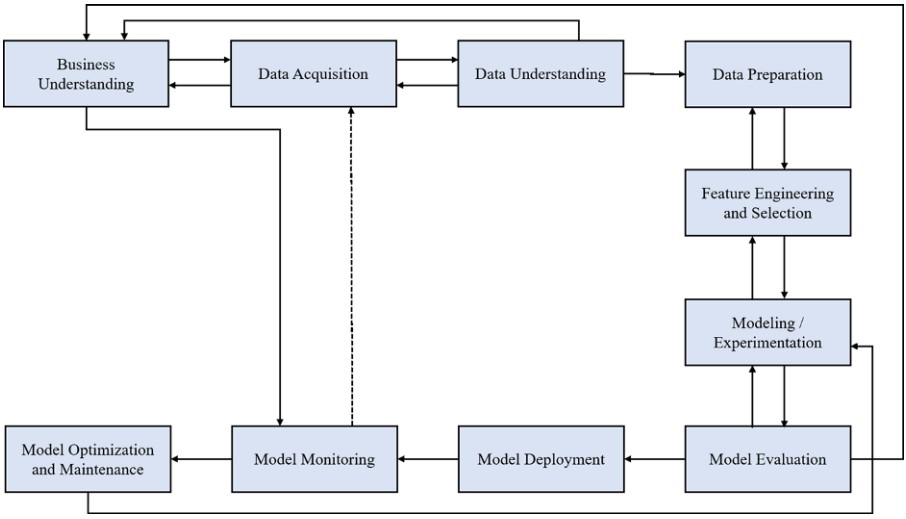

Abb. 2 Machine Learning-Lifecycle

2.2 Modell eines ganzheitlichen Machine Learning-Life Cycles (Abb. 2)

Der hier vorgestellte Machine Learning-Lifecycle basiert auf den zuvor beschriebenen, veröffentlichten Ansätzen, kombiniert und erweitert diese. Zur Qualitätssicherung wurden mehrere Plausibilisierungsgespräche mit Data Science-Experten aus Forschung und Praxis durchgeführt. Dieser neue Vorschlag eines ganzheitlichen Machine Learning-Lifecycles gliedert sich in zehn verschiedene Phasen. Die erste Phase, das *Business Understanding*, betrachtet insbesondere den Aufbau von Domänenwissen, die Zusammenstellung des ML-Teams (siehe Abschn. 3), die Spezifikation der zu unterstützenden Entscheidung sowie die Erstellung der Zieldefinition. Von Bedeutung ist hierbei auch der Einbezug der späteren Verwendung des zu entwickelnden Modells. Die Auswahl von verwendbaren Datenquellen und notwendige Datenvorverarbeitungsschritte hängen zwangsläufig davon ab, in welchem Kontext das fertige Modell verwendet werden soll und welche Eingabedaten dort erwartet werden. Gleichzeitig kann damit die Phase des *Model Deployments* unterstützt werden. Besondere Wichtigkeit bekommt die Phase *Data Acquisition* zugewiesen. Die Qualität eines ML-Modells hängt sehr stark von der Qualität der zum Training verwendeten Daten ab. Basierend auf den zuvor definierten Anforderungen aus dem *Business Understanding* müssen in dieser Phase Datenquellen (intern/extern) ausgewählt werden, welche die nötigen Daten bereitstellen. Aufgaben die hierbei anfallen sind die Prüfung von Eigenschaften der Datenquellen (bspw. Aktualität der Daten oder Zugriffsmöglichkeiten), aber auch die Beschaffenheit der Daten, die Herkunft oder auch die Rechtsgrundlage (Darf die Datenquelle für das Vorhaben überhaupt genutzt werden?).

Sobald die Datenquellen für die Erstellung eines ML-Modells ausgewählt wurden, kann mit dem *Data Understanding* begonnen werden, also dem Aufbau eines Verständnisses über die vorliegenden Daten. Hierbei kommen klassische Verfahren

der Exploration zum Tragen, wie bspw. erste händische Analysen oder die Kontrolle des Aufbaus der Daten. Jegliche Eigenschaften die hierbei erkannt werden, müssen im Datenmanagement gepflegt werden, um einer erneuten Nutzung dieser Datenquelle entsprechend vorzusorgen. Anzumerken ist hierbei, dass diese Phase oftmals in mehreren Loops mit den vorherigen Phasen *Business Understanding* und *Data Acquisition* durchlaufen wird. Erst wenn das Verständnis über die Datengrundlage aufgebaut wurde und die Auswahl der Datenquellen festgelegt ist, kann mit der *Data Preparation* begonnen werden. In dieser Phase werden die ausgewählten Datenquellen zusammengeführt und von Fehlern bereinigt. Gleichzeitig werden die Daten transformiert und validiert. Hierbei wird deutlich, dass all diese Schritte Aufgaben des Datenmanagements sind. Sinnvoll ist es daher, wenn diese Aufgaben zentralisiert erledigt werden und nicht in einzelnen Datensilos. Dadurch kann die Wiederverwendbarkeit dieser Daten und Verarbeitungsschritten ermöglicht werden. Nach der ersten Validierung der Daten kann mit der Ergänzung von Informationen begonnen werden, dem Labeling oder der Annotation der Datensätze. All diese Schritte müssen nachvollziehbar über eine Versionskontrolle erfasst werden, sodass im Nachhinein zu jedem Zeitpunkt nachvollzogen werden kann, welche Verarbeitungsschritte durchlaufen wurden. Zusätzlich kann über die Versionskontrolle die Reproduzierbarkeit der Ergebnisse unterstützt werden.

Sobald die Datensätze aufgebaut, vorverarbeitet und versionisiert sind, kann mit der Phase *Feature Engineering & Selection* fortgefahren werden. Ziel dieser Phase ist basierend auf den zur Verfügung stehenden Daten genau die Datenattribute auszuwählen, die das größte „Signal“ für ein spezifisches ML-Modell bedeuten. Wählt man zu viele Features aus, wird die Datenkomplexität und dadurch die Trainingsdauer massiv erhöht. Wählt man dagegen eine zu geringe Menge oder die falschen Features aus, werden vom Modell eventuell keine oder nur ungenaue Muster gelernt.

Die ausgewählten Features können dann in die Phase des *Modelings* überführt werden. Der Aufbau von Machine Learning-Modellen erfolgt stets in einem experimentellen Prozess. Zwar haben sich für viele Anwendungsfälle Best Practices ergeben, allerdings ist die optimale Parametrierung eines Modells schlichtweg nur über exploratives Testen der Modellkonfiguration möglich. Es wird demnach in der gegebenen Zeit solange Parameter Tuning betrieben, bis ein zufriedenstellendes Modell resultiert. Die Prüfung dessen findet im Schritt *Model Evaluation* statt, in dem Metriken definiert werden, die die Performance des Modells messen.

Sobald ein Modell trainiert und erfolgreich evaluiert wurde, kann dieses Modell das *Deployment* durchlaufen, also bspw. in eine bestehende Anwendung integriert werden. Es wird in den laufenden Betrieb überführt. Während des Betriebs muss das Modell zwangsläufig das *Model Monitoring* durchlaufen. Es muss also überwacht werden, wie sich die Performance des Modells über einen längeren produktiven Einsatz entwickelt. Die erfolgten Predictions sollten dabei erfasst und gespeichert werden. Die Nutzereingaben können währenddessen genutzt werden, um die bestehenden Datensätze zu erweitern und ggf. ein Retraining bzw. eine *Optimierung* des Modells durchzuführen. Die Optimierung und Wartung von bestehenden Modellen stellt dabei den finalen Schritt des Lifecycles dar. Im laufenden Betrieb kommt es vor, dass sich Eingabedaten verändern oder ungültig werden. In diesem Fall spricht man von einem *Drift of Concept* (Widmer und Kubat 1996). Jedes Modell wurde für

einen bestimmten Anwendungs- und Datenkontext entwickelt. Ändert sich ein Kontext, dann ist die Güte des Modells ggf. nicht mehr gegeben. Ein wichtiger Schritt des Monitorings des Modells ist daher die Erkennung von Drifts (Zliobaite 2010). Für das Datenmanagement resultiert daraus, dass klar bestimmbar sein muss, welche ML-Daten in welcher Version in welchem Modell genutzt wurden. Nur dann kann die Fehlerquelle oder Ursache eines Drifts bestimmt werden und geeignete Schritte zur Optimierung der Modelle eingeleitet werden.

3 Rollen in Machine Learning-Projekten

Forschung und Praxis sind sich einig, dass der Erfolg von ML-Projekten insbesondere von den beteiligten Personen abhängt. Wird nun im Falle eines ML-Projektes nicht nur der Aufbau und die Implementierung von ML-Algorithmen betrachtet, sondern der gesamte ML-Lifecycle (siehe Abschn. 2), wird deutlich, dass verschiedenartige Rollen und Skillprofile notwendig sind, um die anfallenden Aufgaben zu bewältigen. Überall dort, wo verschiedene Mitglieder eines Teams (oder auch mehrere Teams) auf der gleichen Datengrundlage arbeiten, rückt das zielgerichtete Datenmanagement in den Vordergrund. Zunächst müssen aber mögliche Akteure eines ML-Teams thematisiert werden.

Die Zusammensetzung von ML-Teams beschäftigt aktuell viele Unternehmen, was an Publikationen zu diesem Thema erkenntlich wird. Unterschiede sind je nach Vorschlag zwar vorhanden, im Kern spiegeln sich aber viele Gemeinsamkeiten wieder. Für unumgänglich wird es in Projekten gehalten, die Rolle des Data Engineers einzuführen (Knopp 2018; Patil 2011; Dhungana 2019; IBM Analytics 2016; Agrawal et al. 2019). Also eine Person in das ML-Team zu integrieren, die große Kompetenzen im Umgang mit Datenquellen, in der Implementierung von ETL-Strecken sowie der Kontrolle der Datenqualität aufweist. Jedes Modell hängt maßgeblich von der Datengrundlage ab, weshalb es die Hauptaufgabe des Data Engineers ist, diese Datengrundlage zu schaffen und in geeigneter Form bereitzustellen. Der Fokus der Rolle Data Engineer liegt also im Bezug auf den zuvor deklarierten ML-Lifecycle in den Teilschritten *Data Acquisition, Data Understanding, Data Preparation* und unterstützend in den Schritten *Model Monitoring*. Alle Schritte des Data Engineers sollten optimalerweise in Abstimmung mit einem Data Scientisten/Machine Learning-Engineer erfolgen (Knopp 2018; Patil 2011; IBM Analytics 2016; Polyzotis et al. 2018; Agrawal et al. 2019). Diese Rolle des Teams ist für die Auswahl, Konzeption und Implementierung von geeigneten ML-Ansätzen zuständig. Die zugrundeliegende Datenbasis wird exploriert, analysiert und in geeignete Feature Sets überführt, die dann für den Aufbau und das Training von Modellen genutzt werden können. Darüber hinaus wird von dieser Rolle die Vermittlung zwischen Fachbereich und ML-Team übernommen, weshalb neben den Schritten *Data Understanding, Feature Engineering & Selection, Modeling/Experimentation, Model Evaluation und Model Optimization/Retraining* auch die Schritte *Business Understanding* und unterstützend *Data Acquisition (Auswahl der Datenquellen)* zum Aufgabengebiet gehören (Knopp 2018; Patil 2011; IBM Analytics 2016).

Neben den bisher beschriebenen eher datenzentrierten Rollen, sind auch klassische softwarelastige Rollenprofile von Nöten. Um Nutzen aus einem ML-Modell ziehen zu können, muss dieses entsprechend operationalisiert werden, bspw. durch die Einbindung in oder Verbindung mit einer Applikation. Das *Model Deployment* wird üblicherweise von einem Softwareentwickler übernommen (Knopp 2018; Patil 2011; Dhungana 2019; Polyzotis et al. 2018; Agrawal et al. 2019). Dieser setzt die Anforderungen um, die aus den Arbeiten des Data Engineers und Data Scientisten/ML Engineers resultieren. Die Implementierung und Auswahl geeigneter Softwaretools für die Unterstützung des ML-end-to-end-Lifecycles ist ebenso eine Teilaufgabe dieser Rolle. Wünschenswert ist es darüber hinaus, dass Kompetenzen aus dem Bereich DevOps vorhanden sind, um die Kontrolle und Zustand der bereits deployten Modelle überwachen zu können und ggf. mit dem zuständigen ML-Engineer eine Optimierung anzustoßen. Grundsätzlich ist der Software Engineer also dafür zuständig, die gewonnen Erkenntnisse und Arbeiten des Data Scientisten/ML Engineers und Data Engineers zu operationalisieren und zu automatisieren. Insbesondere diese letztgenannten Aufgaben setzen die Reproduzierbarkeit der Ergebnisse voraus, wofür wiederum vollständige Transparenz innerhalb des Entwicklungsprozesses notwendig ist. Der Software Engineer deckt daher den gesamten Lifecycle in unterstützender Funktion ab mit Schwerpunktsetzung auf das *Model Deployment*.

In der Literatur werden zudem Rollen beschrieben, die keinen technischen Bezug haben, so sind zusätzlich Business Analysten und Unterstützer zur Implementierung von ML-Teams notwendig (Knopp 2018). Zum einen um die Businessprobleme klar und einheitlich für das Team zu definieren, zum anderen aber auch um die Etablierung von ML überhaupt zu ermöglichen und Unternehmensstrukturen nachhaltig darauf ausrichten zu können. Bei besonders komplexen digitalen Ethikfragestellungen kann es sinnvoll sein, zusätzliche Ethikbeauftragte zu einem Projekt hinzuzuziehen (Dhungana 2019). Dies ist insbesondere immer dann der Fall, wenn durch die am Projekt beteiligten Mitarbeiter keine objektive oder unbefangene Bewertung der Ethikfragestellung möglich ist oder die Komplexität der Fragestellung zu hoch ist. Auch Dinge wie Datenschutz, Gesetze oder Compliance und Auditing müssen in der Umsetzung von ML-Projekten betrachtet werden, weshalb der Einsatz von Legal-Beauftragten sinnvoll sein kann (Agrawal et al. 2019). Transparenz und Versionierung stellen eine Grundvoraussetzung dar.

Aus dieser Vielfalt an verschiedenen Aufgaben resultieren verschiedene Anforderungen an das Datenmanagement. Die Verwaltung, Verarbeitung und Nutzung von Daten steht im ML-Lifecycle im Mittelpunkt. Problematisch ist dabei, dass je nach Rolle ein anderes Verständnis und auch andere Herangehensweisen vorherrschen können. Verstärkt wird diese Problematik, wenn verschiedene Teams innerhalb eines Unternehmens parallel zueinander arbeiten. Schon bei der Initialisierung eines Projektes sollte also darauf geachtet werden, dass zentralisiert und transparent hinterlegt wird, welche Datenquellen genutzt werden, um Redundanzen zu vermeiden. Gleichzeitig müssen schon hier Eigenschaften der Datenquellen gepflegt werden, wie beispielsweise die Herkunft, Nutzungsvereinbarungen, die Herkunft der Datenquelle oder auch den Zweck der Nutzung. In allen weiteren Schritten erscheint es sinnvoll, wenn stets Bezug zur ursprünglichen Datenquelle genommen wird, um jederzeit sicherstellen zu können, woher die aktuell verwendeten Daten stammen. Dies

bedeutet, dass nicht nur zwischen verschiedenen Rollen, wie dem Data Engineer und Scientisten, sondern auch zwischen Teams und Abteilungen (bspw. Juristische Abteilungen oder der Fachbereich) transparente Verweise auf die Nutzung von Daten gegeben werden können. Dies ist zwingend notwendig, da nach Teilschritten der *Data Preparation* Daten derart transformiert worden sein können, dass nicht mehr erkennbar ist, aus welchen Datenquelle(n) sie bezogen wurden. Ein weiterer Vorteil der zentralisierten, versionierten und nachvollziehbaren Datenhaltung ist, dass die sehr zeit- und aufwandsintensive Datensammlung und -vorbereitung nur einmal durchgeführt werden muss und für weitere Anwendungsfälle oder Projekte wiederverwendet werden kann (Agrawal et al. 2019; Press 2016). In der Praxis ist es oft der Fall, dass viele Schritte mangels Kommunikation zwischen den Akteuren mehrfach durchgeführt werden (Agrawal et al. 2019).

4 Beschaffenheit der zugrundeliegenden Daten

Wie beschrieben, werden für hochwertige Machine Learning-Modelle aussagekräftige, große Datenmengen benötigt, um bspw. neuronale Netze damit zu trainieren. Basierend auf diesen Daten können dann Muster gelernt werden, die auf gleichartige, aber neue Inputdaten übertragen werden können. Big Data kann daher als Enabler für Machine Learning angesehen werden. Um interne oder externe Datenquellen für Machine Learning nutzbar zu machen, ist es notwendig die Daten in eine Form zu transformieren, die für Machine Learning-Ansätze geeignet ist. Dabei kann von ML-Data gesprochen werden (Agrawal et al. 2019). ML-Data besteht dabei neben den Rohdaten stets aus Versionierungsinformationen und wenn vorhanden aus Annotationen (Label oder sonstige Informationen) sowie vorgenommenen Data Preparation-Schritten (z. B. Datentransformationen). Die Art und Beschaffenheit von ML-Data führt zu neuen Implikationen für das Datenmanagement in Unternehmen, welche aus der 7V-Definition von Big Data abgeleitet werden können (Khan et al. 2014). Diese Definition stellt eine Erweiterung des 3V-Modell und des darauf aufbauenden 5V-Modells dar (Gartner 2019; Ishwarappa und Anuradha 2015).

- **Volume:** Für jeden ML-Algorithmus werden große Datenmengen benötigt. Je nach Anwendungsfall kann diese Größe der Dateien sehr stark ansteigen, insbesondere wenn unstrukturierte Daten (z. B. hoch aufgelöste Bilder oder sogar Videos) genutzt werden sollen. Hinzu kommt, dass in den meisten Fällen zwar eine Komprimierung der Daten auf ausgewählte Features stattfindet, um die Trainingsdauer der Algorithmen zu optimieren, die Daten aber zunächst roh vorliegen sollten, damit die volle Bandbreite der Informationen aus diesen Daten grundsätzlich genutzt werden kann. Erzeugte Datasets, also vorverarbeitete Daten inklusive Informationen wie Annotationen oder Label, können dennoch eine nicht zu vernachlässigende Größe einnehmen und wachsen zudem im Laufe der Zeit immer weiter an durch das Hinzufügen von neuen Daten oder Daten aus dem laufenden Betrieb und gepflegten Informationen zu diesen Daten (Agrawal et al. 2019). In Unternehmen müssen daher genügend Ressourcen vorgehalten werden, um dieses laufende Wachstum gewährleisten zu können. Es müssen Datenmodelle imple-

mentiert werden, die den Aufbau und die Verwaltung von Datasets für ML unterstützen. Statt verarbeitete Rohdaten in Datasets redundant zu speichern, sollte auf die Trennung von Rohdaten und zugehörigen Transformationen und Informationen Wert gelegt werden. Erste Ansätze sind bereits vorhanden und in der Literatur ersichtlich (Agrawal et al. 2019; Miao et al. 2016).

- **Velocity:** Rohdaten werden in der Regel batchweise oder periodisch aus einer Datenquelle extrahiert und für ML-Projekte verfügbar gemacht. Einmal gesammelte Rohdaten verändern sich in den meisten Fällen gar nicht und wenn, dann nur sehr langsam. Anders ist dies bei Annotationen, Transformationen oder der Einteilung von Subsets von Rohdaten zu Datasets, da es hierbei im Projektverlauf zu hochfrequenten Änderungen kommen kann (Agrawal et al. 2019). Beschleunigt wird die Generierung von Daten, wenn ein ML-Modell die Deploymentphase durchlaufen hat und sich im produktiven Einsatz befindet. Findet dann ein Logging der Nutzereingaben statt, können diese Daten genutzt werden, um Datasets zu erweitern und Modelle ggf. basierend auf diesen erweiterten Datasets zu optimieren.
- **Variety:** Mit Machine Learning können vielfältige Daten verarbeitet werden. Dies reicht von numerischen Werten, über Text und Sprache, Bilder bis hin zu Videos. Daraus resultieren in der Speicherung dieser Daten verschiedene Datentypen und -quellen. Hinzu kommt, dass verschiedene Datentypen zwangsläufig auch zu verschiedenen Datenannotationen führen. Denkbar sind dabei unter anderem zum einen binäre Werte bei einer Klassifikation, aber auch Bounding Boxes bei der Annotation von Bildern sowie weitere Möglichkeiten der Annotation von Daten (Agrawal et al. 2019). Bei der Erstellung eines Datenmodells zur Verwaltung von ML-Datasets müssen daher die verschiedenen Datentypen und Annotationenstypen beachtet werden.
- **Veracity** und **Validity:** Die Glaubwürdigkeit von Daten und Sicherstellung derer Qualität unterliegt gerade im Kontext von Machine Learning vielen Wechselwirkungen. Die Annotation von Datensätzen erfolgt nicht selten über Crowdsourcing. Dies bedeutet, dass ein Datensatz veröffentlicht wird und ein Kreis von Personen (meist Personen mit Interesse an den gelabelten Daten) diese mit Annotationen versieht. Der Inhalt dieser Annotationen ist dabei nicht zu überwachen. Besonders bei öffentlichen Datenquellen kann dies problematisch sein. Gleichzeitig kann aber auch davon ausgegangen werden, dass durch die Community grobe Verstöße behoben werden. Dennoch besteht ein Restrisiko. In Unternehmen ist dieses Risiko nicht so hoch wie bei öffentlichen Datenquellen, dennoch kann davon ausgegangen werden, dass verschiedene Personen Daten nicht komplett gleichartig labeln und ggf. ein Qualitätsunterschied der Annotationen messbar ist (bspw. bei der Sorgfalt einer manuell erstellten Bounding Box in einem Bild). Durch Datamanagement-Maßnahmen können diese Probleme nicht beseitigt werden, allerdings kann durch eine vorgegebene Struktur beim Annotationsvorgang, z. B. durch ein Labeltool, das manuelle Labeln vereinfacht oder unterstützt werden. So kann durch eine vorgegebene Struktur die Qualität der angehängten Informationen erhöht werden. Zur Nachverfolgbarkeit im Fehlerfall sollte hierzu zudem eine Versionierung der Daten eingeführt werden.

Einen Sonderfall stellt das Labeln von Daten über implementierte ML-Modelle oder -Services dar. Hierbei werden Daten automatisiert über ein zuvor trainiertes Modell annotiert. Zu beachten ist dabei, dass die Annotation immer nur so gut sein kann, wie es das Modell zulässt. Bei der Nutzung eines solchen Modells muss daher genau evaluiert werden, ob die Güte, also z. B. die Genauigkeit, ausreichend ist für das Ziel der Annotation. Ungenau annotierte Daten führen letztlich zu einem ungenauen Modell, da die Daten die Grundlage eines jeden Lernvorgangs darstellen.

- **Volatility** und **Value:** Die langfristige Gültigkeit von Datensätzen ist eng mit deren unternehmerischen Mehrwert verbunden. Zwar sind Daten grundsätzlich langfristig gültig, allerdings muss hinterfragt werden, wie lange die Daten sinnvoll eingesetzt werden können, um eine Entscheidung zweckdienlich unterstützen zu können. Wird bei einem deployten ML-Modell festgestellt, dass die Vorhersagewerte ungenau werden, kann diese Reduzierung des unternehmerischen Mehrwerts darin begründet sein, dass die zugrundeliegende Datenbasis nicht mehr die Realität wiederspiegelt. Zu unterscheiden ist dabei zudem zwischen einer Veränderung der Gültigkeit von Rohdaten und einer nötigen Veränderung der Annotationen. Denkbar ist, dass die Rohdaten, wie z. B. Bilder, weiterhin gültig sind, aber andere Annotationen benötigt werden. Gerade bei Prognosen mit zeitlichem Horizont muss zudem betrachtet werden, Daten welchen Zeitraums in die Prognosemodelle miteinfließen sollen, da sonst Abweichungen zu befürchten sind. Die Gültigkeit von verwendeten Daten kann nur über die Performance der darauf trainierten Modelle evaluiert werden. Daher muss zwangsläufig ein Monitoring dieser Modelle durchgeführt werden. Gleichzeitig muss dann auch gepflegt werden, welche Daten in welcher Version in welches Modell eingeflossen sind und umgekehrt, welches Modell welche Datensätze zum Training benutzt hat. Es bietet sich zudem bei der Implementierung und dem Training von ML-Modellen an, nach einem Active Learning-Ansatz vorzugehen und genau die Daten für das Training auszuwählen, die für einen ML-Algorithmus ein hohes Signal darstellen (Settles 2010). Dadurch kann die benötigte Datenmenge und damit auch die Dauer des Trainings erheblich reduziert werden.

5 Fazit und Zusammenfassung

In diesem Beitrag wurden in den vorangegangenen Abschnitten Implikationen von Machine Learning auf das Datenmanagement in Unternehmen herausgearbeitet. Dazu wurde der Machine Learning-Lifecycle, die Zusammensetzung von Machine Learning-Teams sowie die für Machine Learning benötigten Daten analysiert. Die während der Analyse gewonnenen Erkenntnisse werden im Folgenden aggregiert und zu acht Hauptimplikationen zusammengefasst:

1. **Zentralisierte Verwaltung der Daten:** Machine Learning-Projekte werden meistens von ausgebildeten Teams ausgeführt. Um Redundanzen und Datensilos zu vermeiden, sollten auch im Machine Learning Daten zentralisiert gespeichert und verwaltet werden. Wenn Daten für Machine Learning-Experimente nur lokal ge-

halten und verarbeitet werden, ist dies für weitere Beteiligte nicht transparent, nachvollziehbar und verwendbar.

2. **Datenquellen:** Je nach Zielsetzung können verschiedenartige Daten für das Training von Modellen notwendig sein. Diese Daten können in ihrer Struktur und ihrem Typ unterschiedlich sein. Eine weitere Besonderheit ist, dass im Machine Learning-Bereich häufig externe Datenquellen hinzugezogen werden, um beispielsweise Baseline-Modelle zu erstellen. Für die Konzeption und Implementierung müssen somit interne und externe Datenquellen an das Datenmanagement angebunden werden und die Eigenschaften dieser Datenquellen, wie z. B. die Herkunft, Rechte, Aktualität oder Data Ownership-Informationen, entsprechend gepflegt werden.
3. **Geeignete Datenmodelle:** Neben den Rohdaten werden im Machine Learning-Bereich oftmals zusätzliche Informationen zu den Daten gepflegt. Solche Annotationen oder Einteilungen von Rohdaten in Datasets über Splitting müssen in einem zugrundeliegenden Datenmodell berücksichtigt werden. Darüber hinaus muss ein Datenmodell unabhängig von dem Datentyp und der Art der Annotation strukturiert aufgebaut sein. Auch die Versionierung und Änderungshistorie eines Datensatzes inkl. dessen Verwendung in Modellen spielt dabei eine Rolle.
4. **Unterstützung der Datenvorverarbeitung:** Wie schon zuvor erwähnt, ist die Datenvorverarbeitung ein elementarer Schritt in der Entwicklung von Machine Learning-Modellen. Gleichzeitig ist die Datenvorbereitung allerdings auch sehr zeit- und ressourcenaufwendig. Um redundante Vorverarbeitungsschritte zu vermeiden, ist die nahtlose Dokumentation von Vorverarbeitungsschritten unumgänglich. Gleichzeitig kann durch die Automatisierung von Analysen, die typischerweise in der Datenexploration verwendet werden (z. B. Null-Werte prüfen oder Statistiken über Datensätze) Zeit und Aufwand eingespart werden.
5. **Überwachung der Datenqualität:** Die Datenqualität ist ein entscheidendes Merkmal bei der Implementierung von Modellen. Eine schlechte Datenqualität führt dazu, dass ein Modell entweder überhaupt keine Signale erhält und damit nicht lernt oder falsche bzw. nicht zweckdienliche Muster lernt. Problematisch ist dabei, dass die Datenqualität stets im Kontext der Anwendung evaluiert werden muss. So ist es denkbar, dass sich die Qualität der zugrundeliegenden Trainingsdaten erst mit der Zeit verschlechtert und nicht mehr ausreichend ist, wenn sich beispielsweise die Eingabedaten der in Betrieb genommenen Modelle verändern. Die Qualität der Daten muss also über den ganzen Lebenszyklus eines Modells hinweg überwacht werden.
6. **Versionierung:** Die Folgen von Datentransformationen oder die Qualität von Datenannotationen und -splits kann im Vorfeld eines Modelltrainings nicht vollumfänglich abgeschätzt werden. Fehlerfälle werden daher oft erst im Nachhinein identifiziert. Wichtig ist zur Optimierung eines Modells allerdings, dass nachvollziehbar ist, welche Aktivitäten unternommen wurden bei der Entwicklung des Modells. Ist diese Versionsverfolgung nicht möglich, kann auch keine Reproduzierbarkeit von Ergebnissen gewährleistet werden.
7. **Nutzungsverweise:** Die größte Fehlerquelle eines Machine Learning-Modells ist die zugrundeliegende Datenbasis, die für das Modelltraining genutzt wurde. Um einen Fehler beheben zu können ist es notwendig zu wissen, welche Daten für

das betroffene Modell genutzt wurden. Gleichzeitig kann es sinnvoll sein, aus Sicht der Datenquellenpflege zu wissen, welche Daten für welches Modell genutzt wurden. Dies ist insbesondere dann von Vorteil, wenn sich Nutzungsrechte oder sonstige Eigenschaften der Daten verändern und dies Folgen für die einsetzende Unternehmung haben kann.

8. **Transparenz und Nachvollziehbarkeit:** Ohne Transparenz und Nachvollziehbarkeit ist kein Life Cycle Management der Machine Learning-Modelle möglich. Wenn der Betrieb nicht transparent für alle Beteiligten überwacht wird, dann ist Kontrollverlust die Folge und der Erfolg der implementierten Modelle eine Blackbox. Nur wenn jederzeit erklärbar ist, welche Modelle mit welchen Schritten entwickelt und mit welchen Daten welcher Version trainiert wurden, können Fehler identifiziert und optimiert sowie die Gesundheit der Systemlandschaft gewährleistet werden.

Die Anwendung von KI kann enorme Potenziale bieten. Zunächst müssen Anwenderunternehmen aber immer in Vorleistung gehen und Kompetenzen in diesem Bereich aufbauen. Im Mittelpunkt stehen dabei immer die Daten eines Unternehmens und deren Relevanz sowie Bedeutung im Kontext KI (Geretshuber und Reese 2019). In diesem Beitrag wurden acht Herausforderungen und Implikationen herausgearbeitet, die von Unternehmen bewältigt werden müssen, um schließlich von den Vorteilen der angewandten KI profitieren zu können. Die resultierenden Vorteile erstrecken sich von der Verbesserung der Güte von getroffenen Entscheidungen über die Automatisierung von Prozessen bis hin zur Erweiterung des Produkt- oder Dienstleistungsangebot und der potenziellen Erschließung von neuen Geschäftsfeldern (Abdelkafi et al. 2019).

Funding Open Access funding provided by Projekt DEAL.

Literatur

Abdelkafi N, Döbel I, Drzewiecki J, Meironke A, Niekler A, Ries S (2019) Künstliche Intelligenz (KI) im Unternehmenskontext – Literaturanalyse und Thesenpapier. Fraunhofer-Zentrum für Internationales Management und Wissensökonomie (IMW), Leipzig (https://www.imw.fraunhofer.de/content/dam/moez/de/documents/Working_Paper/190830_214_KI_in_Unternehmen_final_FM_%C3%B6ffentlich.pdf)

Agrawal P, Arya R, Bindal A, Bhatia S, Gagneja A, Godlewski J, Low Y, Muss T, Paliwal M, Raman S, Shah V, Shen B, Sugden L, Zhao K, Wu M-C (2019) Data platform for machine learning. In: Proceeding of the 2019 International Conference on Management of Data Sigmod 19. ACM, New York, NY, USA, S 1803–1816 https://doi.org/10.1145/3299869.3314050

Allen J, Freed A, Chandrasekaran S (2017) Adapt DevOps to cognitive and artificial intelligence systems. https://developer.ibm.com/articles/cc-devops-artificial-intelligence-cognitive/. Zugegriffen: 6. Okt. 2019

Analytics IBM (2016) Data science is a team sport. Do you have the skills to be a team player? https://www.ibm.com/downloads/cas/NZRDABJV. Zugegriffen: 6. Okt. 2019

Anderson C (2015) Creating a data-driven organization. O'Reilly Media

BARC – Business Application Research Center (2017) Data-driven decision-making: 14 recommendation on how to benefit. http://barc-research.com/. Zugegriffen: 5. Okt. 2019

Batra G, Queirolo A, Santhanam N (2018) Artificial intelligence: the time to act is now. https://www.mckinsey.com/industries/advanced-electronics/our-insights/artificial-intelligence-the-time-to-act-is-now. Zugegriffen: 5. Okt. 2019

Beisswenger A (2016) Anatomie strategischer Entscheidungen. Springer, Wiesbaden

Brynjolfsson E, McElheran K (2016) Data in action: data-driven decision making in U.S. manufacturing https://doi.org/10.2139/ssrn.2722502

Bundesministerium für Wirtschaft und Energie (2019) Künstliche Intelligenz. https://www.mittelstand-digital.de/MD/Redaktion/DE/Dossiers/A-Z/kuenstliche-intelligenz.html. Zugegriffen: 5. Okt. 2019

Dhungana S (2019) On building effective data science teams. https://www.kdnuggets.com/2019/03/building-effective-data-science-teams.html. Zugegriffen: 6. Okt. 2019

Ericson G, Rohm W, Jackson S, Sharkey K, Hu J, Mabee D, Martens J (2017) The team data science process lifecycle. https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle. Zugegriffen: 6. Okt. 2019

Gartner (2019) Big data. https://www.gartner.com/it-glossary/big-data. Zugegriffen: 6. Okt. 2019

Geretshuber D, Reese H (2019) Künstliche Intelligenz in Unternehmen – Eine Befragung von 500 Entscheidern deutscher Unternehmen zum Status quo mit Bewertungen und Handlungsoptionen von PwC. https://www.pwc.de/de/digitale-transformation/kuenstliche-intelligenz/studie-kuenstliche-intelligenz-in-unternehmen.pdf. Zugegriffen: 05.10.2019

Hannan MT, Freeman J (1977) The population ecology of organizations. Am J Sociol 82(5):929–964. https://doi.org/10.1086/226424

Hecker D, Döbel I, Petersen U, Rauschert A, Schmitz V, Voss A (2017) Zukunftsmarkt Künstliche Intelligenz: Potenziale und Anwendungen. https://www.iais.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/KI-Potenzialanalyse_2017.pdf. Zugegriffen: 6. Okt. 2019

Hoeken H, Hustinx L (2009) When is statistical evidence superior to anecdotal evidence in supporting probability claims? The role of argument type. Hum Commun Res 35(4):491–510. https://doi.org/10.1111/j.1468-2958.2009.01360.x

Holst A (2018) Artificial intelligence (AI), Statista study. https://spaces.statista.com/study_id59297_artificial-intelligence-ai%20(1).pdf. Zugegriffen: 04.10.2019

Ishwarappa, Anuradha A (2015) A brief introduction on big data 5 vs characteristics and Hadoop technology. Procedia Comput Sci 48:319–324. https://doi.org/10.1016/j.procs.2015.04.188

Khan M, Uddin M, Gupta N (2014) Seven V's of big data—understanding big data to extract value. http://www.asee.org/documents/zones/zone1/2014/Professional/PDFs/113.pdf. Zugegriffen: 6. Okt. 2019 (Proceedings of 2014 Zone 1 Conference oft he American Society for Engineering Education (ASEE Zone 1))

Knopp E (2018) Building your AI team: the roles your enterprise needs. https://www.ibm.com/blogs/systems/building-your-ai-team-the-roles-your-enterprise-needs/. Zugegriffen: 6. Okt. 2019

McElheran K, Brynjolfsson E (2017) The rise of data-driven decision-making is real but uneven. IEEE Eng Manag Rev 45(4):103–105. https://doi.org/10.1109/emr.2017.8233302

McKinsey & Company (2017) Ask the AI experts: Whats driving todays progress in AI? Interview. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/ask-the-ai-experts-whats-driving-todays-progress-in-ai. Zugegriffen: 5. Okt. 2019

Miao H, Chavan A, Deshpande A (2016) ProvDB: a system for lifecycle management of collaborative analysis workflows. https://arxiv.org/abs/1610.04963. Zugegriffen: 05.10.2019

Patil D (2011) Building data science teams, O'Reilly Media. USA. http://www.datascienceassn.org/sites/default/files/Building%20Data%20Science%20Teams.pdf. Zugegriffen: 6. Okt. 2019

Polyzotis N, Roy S, Whang S, Zhinkevich M (2018) Data lifecycle challenges in production machine learning: a survey. SIGMOD Record 47(2):17–18. https://doi.org/10.1145/3299887.3299891

Press G (2016) Cleaning big data: most time-consuming, least enjoyable data science task, surveys says. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#59eecf536f63. Zugegriffen: 05.10.2019

Reinsel D, Gantz J, Rydning J (2018) The digitization of the world: from edge to core. Whitepaper. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Zugegriffen: 5. Okt. 2019

Settles B (2010) Active learning literature survey. Computer sciences technical report 1648. University of Wisconsin-Madison. http://burrsettles.com/pub/settles.activelearning.pdf. Zugegriffen: 6. Okt. 2019

Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. J Data Warehous 5:13–22

Talagala N (2019) 7 artificial intelligence trends and how they work with operational machine learning. https://blogs.oracle.com/ai/7-artificial-intelligence-trends-and-how-they-work-with-operational-machine-learning. Zugegriffen: 6. Okt. 2019

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn 23(1):69–101. https://doi.org/10.1007/bf00116900

Zliobaite I (2010) Learning under concept drift: an overview. https://arxiv.org/abs/1010.4784. Zugegriffen: 05.10.2019