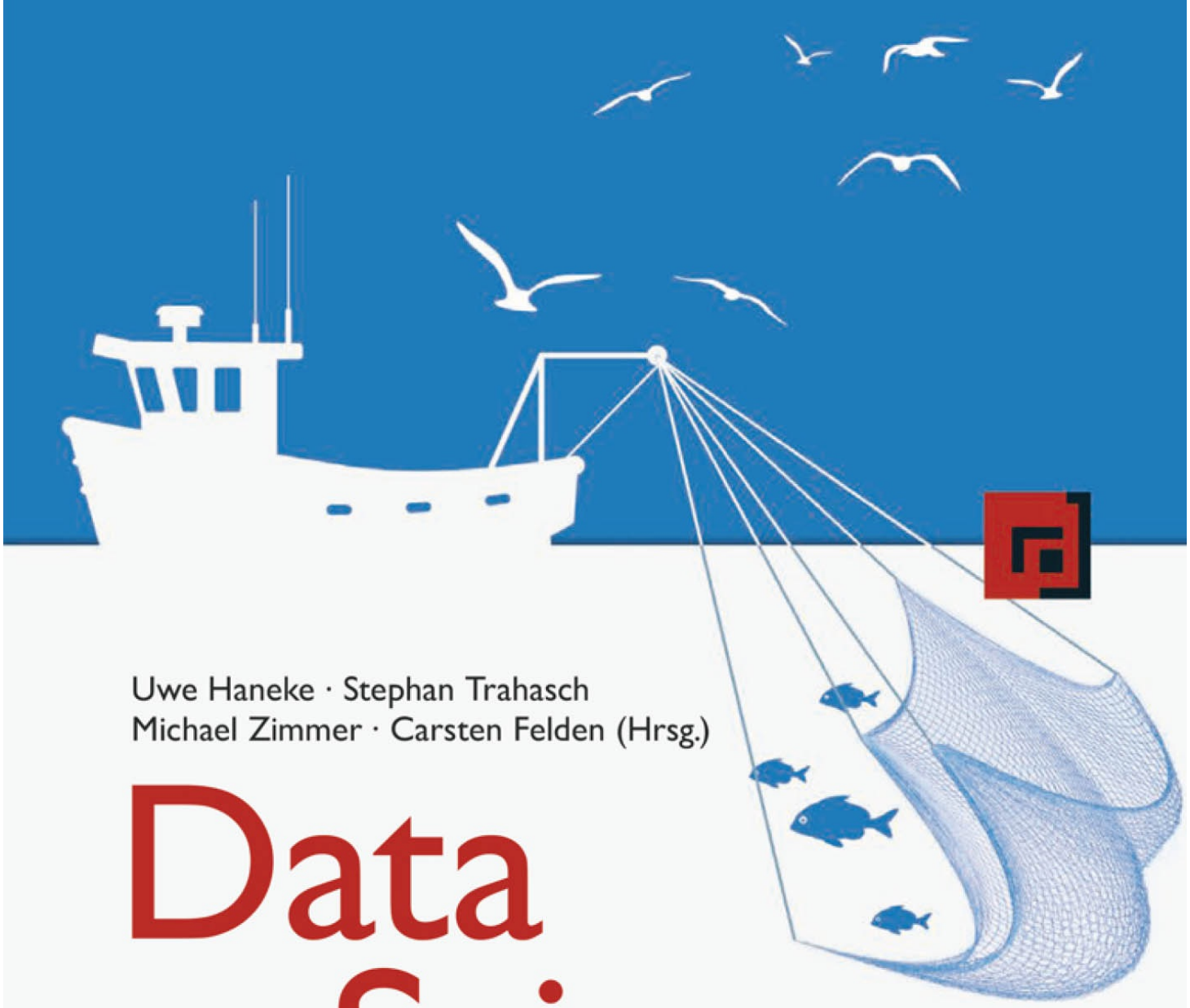

Uwe Haneke · Stephan Trahasch
Michael Zimmer · Carsten Felden (Hrsg.)

Data Science

Grundlagen, Architekturen und Anwendungen

tdwi WISSEN

dpunkt.verlag

der Kognitionswissenschaft auch Erkenntniswissenschaft. Entsprechend lassen sich zwei Ausprägungen unterscheiden: die starke KI und die schwache KI. Während die starke KI das Ziel hat, menschliche Problemlösungskreativität, Selbstbewusstsein und Emotionen abzubilden, fokussiert die schwache KI auf die Lösung konkreter Anwendungsprobleme durch Simulation von Intelligenz durch Methoden der Informatik, der Statistik und der Mathematik.

Hinsichtlich dieses hohen Maßes an Interdisziplinarität gibt es eine große Überlappung zur Data Science. Der Ursprung dieses noch recht jungen Zweigs wird zeitlich unterschiedlich verortet. Gehen Kelleher und Tierney [Kelleher & Tierney 2018] und andere häufig von Jeff Wus [Wus 1997] gehaltener Vorlesung »Statistics = Data Science?« aus, so führt Cao den Namen auf die Nennung des Begriffs im Vorwort eines 1974 publizierten Buches zu Berechnungsmethoden zurück, in dem es heißt, Data Science sei »the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences« [Cao 2017, S. 3]. Noch weiter zurück geht Donoho, der erste Ansätze bereits Mitte der 1950er-Jahre sieht [Donoho 2015, S. 1]. Bei Donoho findet sich auch die folgende Definition für Data Science:

> »This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications.«

Neben der Interdisziplinarität der Data Science rückt Donoho damit auch die Verknüpfung von wissenschaftlicher Entdeckung und Praxis in den Vordergrund. Die Data Science Association sieht ihre 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.«[4]

Entsprechend ist der Data Scientist »[…] a professional who uses scientific methods to liberate and create meaning from raw data […] The data scientist has a solid foundation in machine learning, algorithms, modeling, statistics, analytics, math and strong business acumen […].«

Damit wird deutlich, dass Machine Learning oder maschinelles Lernen eine der Methoden ist, die neben zahlreichen anderen in der Data Science zum Einsatz kommt. Maschinelles Lernen ist nach Wrobel, Joachims und Mrozik:

»[…] ein Forschungsgebiet, das sich mit der computergestützten Modellierung und Realisierung von Lernphänomenen beschäftigt« [Wrobel et al. 2013, S. 406].

Bei den eingesetzten Lernverfahren unterscheidet man das überwachte Lernen (supervised learning), das unüberwachte Lernen (unsupervised learning) sowie das Verstärkungslernen (reinforcement learning). Vielfach kommen hier neuronale Netze zum Einsatz, doch werden je nach Kontext und Fragestellung auch andere Verfahren genutzt. Die Autoren sehen Machine Learning, Data Mining und die »Knowledge Discovery in Databases« (KDD) als Teilgebiete der KI, die in den vergangenen Jahren zunehmend Eingang in praktische Anwendungen in Industrie und Wirtschaft gefunden haben. Die klassische Definition von KDD stammt von Fayyad, Piatetsky-Shapiro und Smyth:

»Knowledge Discovery in Databases describes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data« [Fayyad et al. 1996].

Data Mining ist dabei als der Teilschritt dieses Prozesses zu sehen, der sich mit der Analyse beschäftigt. Im kommerziellen Bereich verschwimmt die Unterscheidung zwischen KDD und Data Mining jedoch häufig.

Die Entwicklungen rund um Data Science fußen nicht zuletzt auf der enormen Menge an Daten, die Wissenschaftlern, Regierungen und natürlich auch den Unternehmen heute zur Verfügung stehen. Unter dem Schlagwort Big Data wird diese Entwicklung zusammengefasst. Big Data umfasst Methoden und Technologien für die hochskalierbare Integration, Speicherung und Analyse polystrukturierter Daten. Dabei bezieht man sich häufig auf die sogenannten 3Vs (Volume, Velocity und Variaty), die zum Teil durch weitere Vs, wie etwa für Value, ergänzt werden (vgl. [Cai & Zhu 2015, S. 2]). Skalierbarkeit bezieht sich insbesondere auf die in der Regel hohen Datenvolumina (Data Volume), das schnelle Anfallen der Daten und die dafür notwendige hohe Datenverarbeitungs- und analysegeschwindigkeit (Data Velocity) sowie eine breite Quellen- und Datenvielfalt (Data Variety) (vgl. [Dittmar 2016, S. 56 f.]).

1.3 Vorgehen in Data-Science-Projekten

Bei Data-Science-Projekten hat sich ein iteratives, agiles Vorgehen bewährt, das sich in der Regel an dem Vorgehensmodell CRoss-Industry Standard Process for Data Mining, kurz CRISP-DM, orientiert (siehe Abb. 1–3).

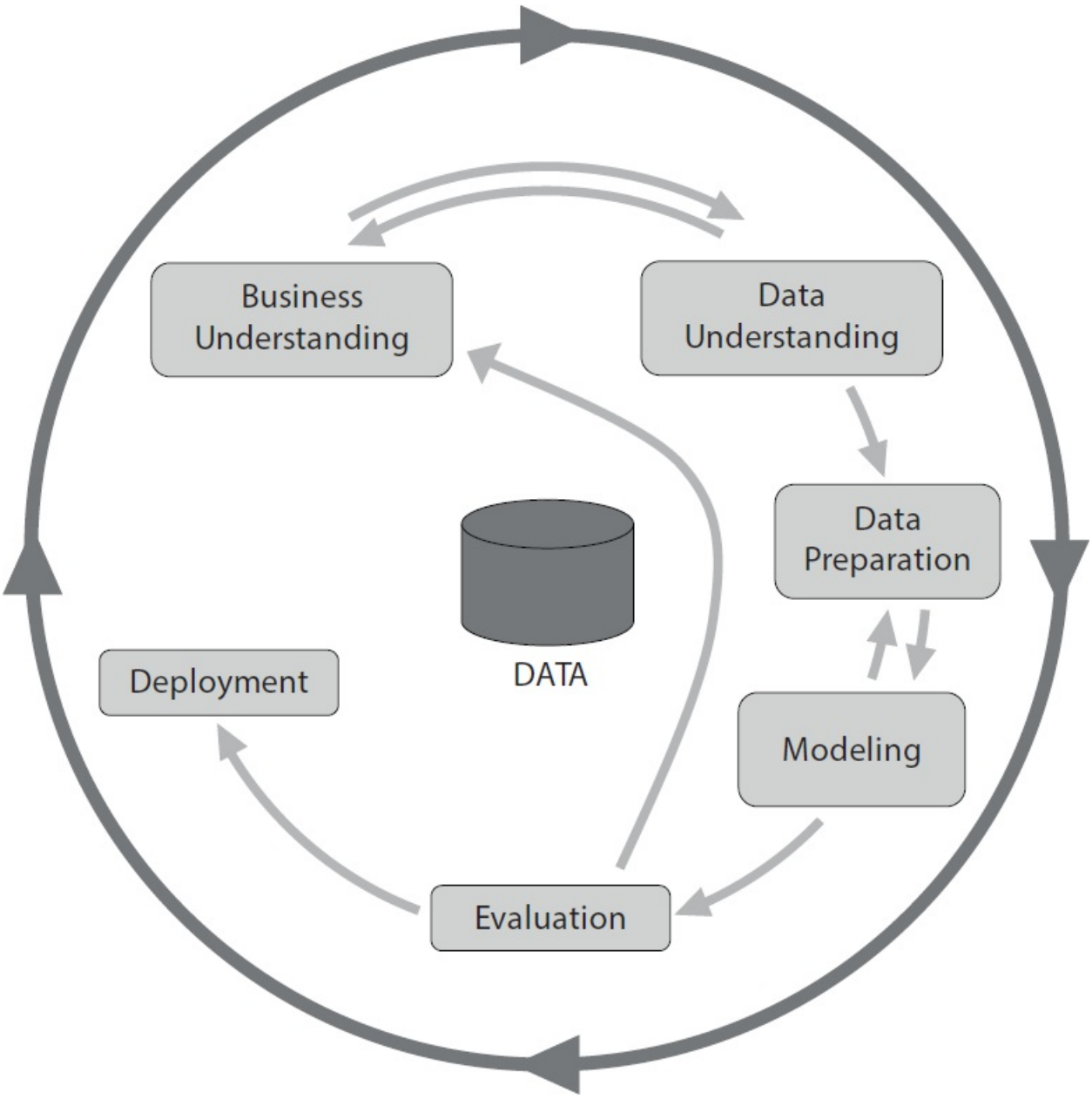

Abb. 1–3 *CRISP-DM mit sechs Phasen*

CRISP-DM besteht aus sechs Phasen, die als zyklischer Prozess zu verstehen sind. Das Business Understanding (fachliches Verständnis) umfasst die Bestimmung der Geschäftsziele, die Beurteilung der aktuellen Situation sowie die konkreten fachlichen Ziele des Data-Science-Projekts und – verbunden damit – die Planung der weiteren Aktivitäten. Im Data Understanding (Verständnis der Daten) werden die Daten und Datenquellen identifiziert, die zur Beantwortung der analytischen Fragestellung notwendig sind. Dieser Schritt enthält auch eine erste Datenerfassung, Datenbeschreibung und die Überprüfung der Datenqualität. Sind die Datenquellen identifiziert und die Daten zusammengestellt, erfolgt eine explorative Datenanalyse, um erste erkennbare Muster zu

sichten. Neben der visuellen Analyse und den deskriptiven statistischen Verfahren können auch BI-typische Datenaufbereitungen und -navigationen hilfreich sein, um erste Erkenntnisse über den vorliegenden Datenbestand zu gewinnen. Grundsätzlich folgen solche Analysen einem Prozess, um einen zielorientierten und nachvollziehbaren Ablauf der jeweiligen Datenanalyse zu ermöglichen. Bereits die Business Intelligence liefert hier einen allgemeinen Ablauf, der mit der Datenextraktion, der Transformation und dem Laden in das Data Warehouse beginnt und im weiteren Vorgehen vorab definierte Auswertungen mit einem entsprechenden Analysewerkzeug ermöglicht.

Im Rahmen der Data Preparation (Datenvorbereitung) sind die Daten so aufzubereiten, dass diese im nächsten Schritt für das Training der Modelle verwendet werden können. Modeling (Modellierung) benennt die Parametrisierung und das eigentliche Lernen eines Modells mithilfe von Data-Mining-Algorithmen zur Lösung der Aufgabenstellung. Diese können Regressionsanalyse, Assoziationsanalyse, Klassifikations- oder Clusteranalysen sein. Die Evaluierung erfolgt einerseits bezogen auf die Ergebnisqualität des gelernten Modells und andererseits gegen das Ziel der fachlichen Aufgabenstellung sowie der betriebswirtschaftlichen Bewertung. Die Gewinnung des Geschäftsverständnisses ist ein iteratives Prozedere, in dem die Ergebnisse durch unterschiedliche Algorithmen und Visualisierungen ausgewertet werden, um ein tieferes Verständnis über die erzielten Ergebnisse zu erhalten. Das abschließende Deployment ist die Übertragung der Ergebnisse in die organisationalen Operationen, seien es Vorhersagen zu Marketingaktivitäten oder zu Wartungszyklen der Maschinen in der Fertigung. Zu einem Deployment gehört allerdings auch, dass diese Modelle auf Veränderungen der Betriebsbedingungen zu überwachen sind, da sich Bedingungslagen und Strukturen ändern können, sodass die Gültigkeit von Ergebnissen nicht mehr vorliegt und ein neues Verfahren zu initiieren ist.

Neben CRISP-DM gibt es alternative Ansätze wie beispielsweise der KDD-Prozess nach Fayyad oder SEMMA. Der fayyadsche Ansatz kennzeichnet sich durch die expliziten Phasen Datenauswahl, Datentransformation, Data Mining und die darauffolgende Interpretation (vgl. Abb. 1–4). Implizit wird dabei auch davon ausgegangen, dass Schritte iterativ ausgeführt werden.

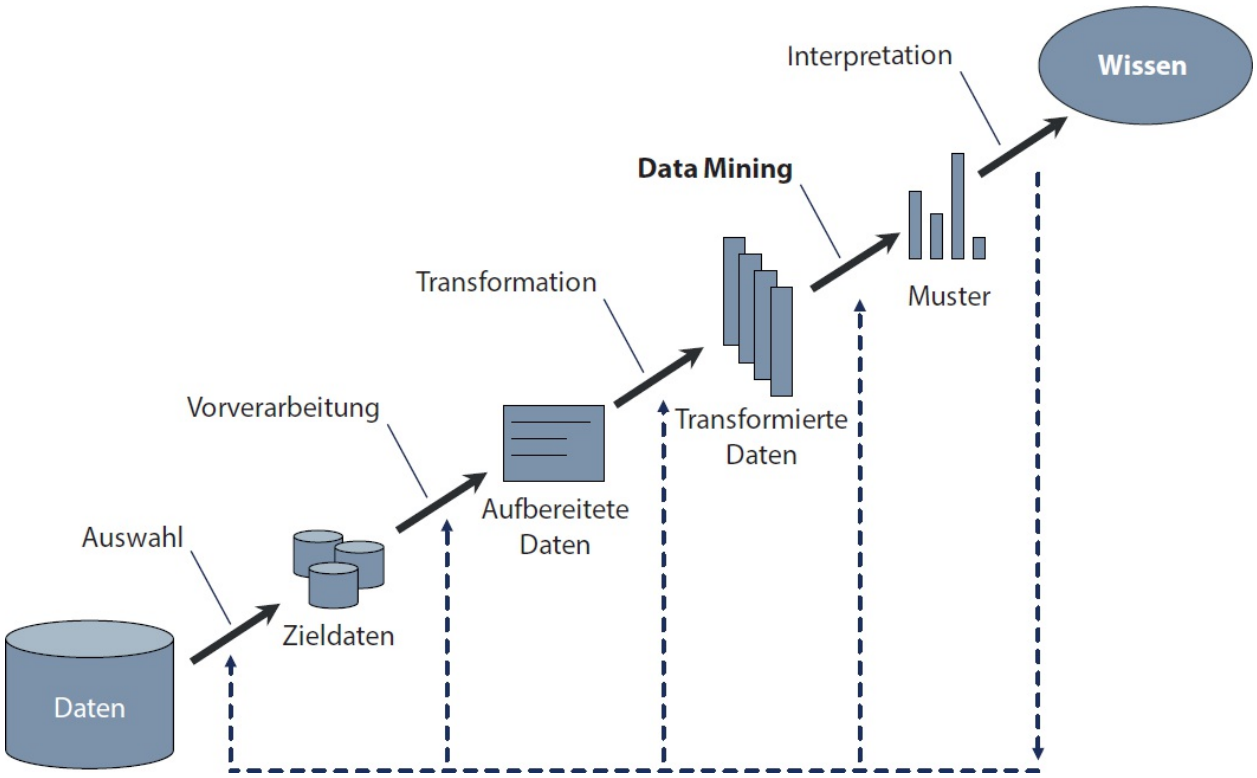

Abb. 1–4 *Überblick über den KDD-Prozess (nach [Fayyad et al. 1996])*

SEMMA, ein früher herstellernaher Ansatz, geht auch phasenorientiert vor, wobei hier von Datenauswahl (Sampling), Datenverständnis (Explore), Modifikation, Algorithmusanwendung (Model) und Ergebnisevaluation (Assess) gesprochen wird.

Die Vorgehensweise ist in fast jedem Data-Science-Projekt iterativ und die Phasen werden mehrmals durchlaufen. Dies bedingt, dass die Nachvollziehbarkeit der einzelnen Schritte wie Datenauswahl, Transformationen etc. und auch das Training in den verschiedenen Phasen ein wesentlicher Punkt ist, der von Projektbeginn an berücksichtigt werden muss. Nur wenn die Nachvollziehbarkeit der Analyse sichergestellt ist, sind eine fundierte Bewertung der Ergebnisse und die Reproduktion der Analyse in der Produktivumgebung und damit das Deployment möglich.

1.4 Struktur des Buches

Das vorliegende Werk ist in einen Grundlagenteil und einem Praxisteil mit Fallstudien gegliedert. Im Grundlagenteil werden verschiedene Aspekte von Data Science erläutert und im zweiten Teil des Buches werden die Grundlagen anhand von konkreten Fallstudien aus Data-Science-Projekten mit deren spezifischen praktischen Problemstellungen und Lösungsansätzen dargestellt. Die Projektberichte nehmen Bezug auf die Grundlagen des ersten Teils, sind in sich jedoch geschlossen und können in einer frei wählbaren Reihenfolge gelesen werden.

In Kapitel 2 diskutiert Uwe Haneke, ob Analytics wirklich das neue BI ist und welche