
Michael Lang (Hrsg.)
DATEN-
KOMPETENZ
Daten erfolgreich nutzen
HANSER

Lang (Hrsg.)
Datenkompetenz

Michael Lang (Hrsg.)

Datenkompetenz

Daten erfolgreich nutzen

HANSER

Print-ISBN: 978-3-446-47585-4
E-Book-ISBN: 978-3-446-47743-8

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://dnb.d-nb.de abrufbar.

www.hanser-fachbuch.de
Lektorat: Lisa Hoffmann-Bäuml
Herstellung: le-tex publishing services, Leipzig
Titelmotiv: © stock.adobe.com/Siarhei
Coverrealisation: Max Kostopoulos
Satz: Eberl & Koesel Studio, Kempten
Druck und Bindung: CPI books GmbH, Leck
Printed in Germany

Vorwort

Durch den digitalen Wandel entstehen immer mehr Daten, die für die Geschäftstätigkeit genutzt werden können. Für Unternehmen ergeben sich damit enorme Chancen und Risiken zugleich. Somit ist es für den zukünftigen Erfolg von Unternehmen entscheidend, wie gut es ihnen gelingt, relevante Daten zu sammeln, diese systematisch auszuwerten, daraus wertvolle Erkenntnisse abzuleiten und diese für die Geschäftstätigkeit zu nutzen.

Die zentrale Grundlage dafür ist, dass die Mitarbeitenden des Unternehmens die erforderlichen Kompetenzen für eine erfolgreiche Nutzung von Daten besitzen.

Doch welche Kompetenzen sind erforderlich?

Die Antwort auf diese Frage – und viele weitere praxisrelevante Impulse – erhalten Sie in diesem Buch.

Ich freue mich sehr, dass dazu zehn ausgewiesene Expertinnen und Experten an diesem Buch mitgewirkt haben und Ihnen die relevanten Datenkompetenzen vermitteln.

Ich wünsche Ihnen viel Spaß beim Lesen des Buches und viel Erfolg beim Umsetzen der dabei gewonnenen Erkenntnisse!

Ihr Herausgeber

Michael Lang

Inhalt

1 Datenkompetenz – Grundlagen

Robert Butscher

Fortschritt und Prosperität hängen zunehmend davon ab, wie es gelingt, Daten aus den unterschiedlichsten Bereichen und Quellen verfügbar und nutzbar zu machen. In Unternehmen sind Daten längst zum zentralen Produktionsfaktor geworden. Ohne die Fähigkeit, Daten aus verschiedenen Quellen systematisch zu sammeln, aufzubereiten und zu analysieren, gäbe es weder datenbasierte Wertangebote noch Erlöse. In der Folge steigt der Bedarf nach Kompetenzen und Berufen, die auf datenspezifische Belange und Anforderungen von Unternehmen ausgelegt sind.

In diesem Beitrag erfahren Sie,

- was den Begriff Datenkompetenz auszeichnet und welche besonderen Merkmale er aufweist,
- welche typischen Fähigkeiten im Kontext von Datenkompetenz gesehen werden,
- welche unterschiedlichen Begrifflichkeiten rund um Datenanalyse im Laufe der Jahre entstanden sind,
- welche typischen Vorgehensmodelle es gibt, aus Daten Wissen und Erkenntnisse zu gewinnen,
- welche typischen datenzentrierten Berufsfelder entstanden sind.

1.1 Der Begriff Datenkompetenz (Data Literacy)

Der englische Begriff **Data Literacy** wird im Deutschen meist mit **Datenkompetenz** übersetzt. Literacy selbst steht für die Lese- und Schreibkompetenz.

Beide Kompetenzen zählt das Goethe-Institut zu den Schlüsselqualifikationen (Goethe-Institut 2022): Unter Literacy fallen Kompetenzen wie Lesen und Sinnverstehen von Texten, sprachliche Abstraktionsfähigkeit oder die Fähigkeit, Texte

eigenhändig zu verfassen oder mit eigenen Worten wiederzugeben. Literacy befähigt somit den Einzelnen, Inhalte zu erzeugen, niederzuschreiben oder Information aus Texten zu ziehen. Zu Literacy zählt auch die Fähigkeit, den Sinn gesprochener wie geschriebener Sprache zu verstehen und zu kommunizieren. Das dafür nötige Kompetenzspektrum ermöglicht den sicheren Umgang mit einer Sprache.

Bezogen auf den Umgang mit Daten bezeichnet Data Literacy bzw. Datenkompetenz die Fähigkeit, mit Daten kompetent umzugehen. Ähnlich wie für den Umgang mit einer Sprache setzt Datenkompetenz ein Kompetenzbündel voraus, um Daten sachgerecht zu verarbeiten oder aus Daten nützliche wie sinnvolle Information abzuleiten. Zu Datenkompetenz gehört etwa die Kompetenz, ein fachliches Problem in ein datenanalytisches zu überführen. Dazu zählt, Daten für ein bestimmtes fachliches Problem zu erfassen, zu sammeln und adäquat zu speichern. Eine solche Datengrundlage zusammenzutragen und nur jene Datensätze zu verwenden, die im Hinblick auf das zu lösende Problem nützlich sind, ist eine weitere Fähigkeit im Kontext der Datenkompetenz. Gleiches gilt auch, wenn umfangreiche Datenmengen für Machine-Learning-Modelle oder für Anwendungen der künstlichen Intelligenz aufzubauen und entsprechend zu kuratieren sind.

In dem Zusammenhang spielen Data Governance und betriebliches Datenmanagement als weiteres Kompetenzbündel eine wichtige Rolle: So spannt Data Governance einen Ordnungsrahmen auf und umfasst rechtliche wie ethische Vorgaben, welche Daten wie und für welche Zwecke gespeichert und verarbeitet werden dürfen. Das betriebliche Datenmanagement regelt den gesamten Datenlebenszyklus, etwa wie gesammelte Daten auszuzeichnen, in welcher IT-Infrastruktur Daten zu verarbeiten oder gesetzeskonform zu löschen bzw. zu archivieren sind.

Eine weitere Fähigkeit im Rahmen der Datenkompetenz liegt darin, Daten kritisch nach ihrer Qualität und Eignung beurteilen zu können. Ferner umfasst Datenkompetenz alle Fähigkeiten, die zur Analyse von Daten erforderlich sind, etwa geeignete Data-Mining-Algorithmen ausfindig zu machen und diese sachgerecht anzuwenden. Dazu zählt auch, Daten adäquat aufzubereiten, korrekt zu visualisieren und eventuell vorhandene Muster in Daten fachgerecht zu interpretieren. Auch die Fähigkeit, Analyseergebnisse an Dritte kommunizieren zu können, fällt unter Datenkompetenz. Weiterhin gehört dazu, Daten für neue Produkte oder Dienstleistungen nutzen zu können. Daten werden hierbei zum Produktionsfaktor für Unternehmen und bilden die Grundlage für datengetriebene Geschäftsmodelle. Datenkompetenz erstreckt sich nicht nur auf individuelle Fähigkeiten einer Person, sondern auch auf die gesamte Organisation: Datengetriebene Geschäftsmodelle setzen eine datenfokussierte Unternehmenskultur voraus.

Datenkompetenz wird somit zum immateriellen Vermögensgegenstand eines Unternehmens, eng verbunden mit der Fähigkeit zur Innovation und digitalen Transformation. Bild 1.1 zeigt zusammenfassend das Kompetenzbündel rund um Datenkompetenz.

Bild 1.1 Datenkompetenz im Überblick

Das US-Marktforschungsunternehmen Gartner definiert Datenkompetenz wie folgt (Panetta 2021):

> „Gartner defines **data literacy** as the ability to read, write and communicate data in context, including an understanding of data sources and constructs, analytical methods and techniques applied, and the ability to describe the use case, application and resulting value."

1.2 Definitionen

1.2.1 Business Intelligence

Viele Analytics-Begriffe sind Buzzwords. Unternehmensberatungen oder Softwarehersteller prägen diese, selbst wenn die Ursprünge im akademischen Bereich, wie z. B. bei künstlicher Intelligenz, liegen. In der Folge lassen sich die Begrifflichkeiten selten trennscharf voneinander abgrenzen oder einheitlich definieren: Jeder Begriff muss vor seinem zeitlichen wie technischen Hintergrund interpretiert und mit anderen in Beziehung gebracht werden. Erst so treten Unterschiede und Gemeinsamkeiten zutage.

Bedingt durch den Einsatz betrieblicher Informations- und Datenbanksysteme stieg das Datenvolumen spätestens seit den 1980er-Jahren an. Ebenso der Wunsch, diese Datenbestände auszuwerten und daraus Erkenntnisse, z. B. für die Planung und Steuerung des Unternehmens, zu gewinnen. Mit dem Ziel, mittels Datenanalyse bessere Geschäftsentscheidungen zu treffen, ist Business Intelligence verbunden. Solche Verbesserungen können Umsatzsteigerungen, Kosteneinsparungen oder Effizienzgewinne sein. Hans Peter Luhn, der bei IBM angestellt war, führte 1958 den Begriff Business Intelligence ein (Luhn 1958). Howard Dresner, der später für das US-Marktforschungsunternehmen Gartner arbeitete, griff in den 1990er-Jahren den Begriff auf und machte ihn populär. Er definiert Business Intelligence wie folgt (Kottbauer 2015):

Business Intelligence (BI) ist „[…] the process of transforming data into information and, through discovery into knowledge“.

BI stellt einen technologiegetriebenen Prozess dar, der darauf abzielt, aus Daten nützliche Information für Führungskräfte und Mitarbeitende gleichermaßen zu liefern. „Klassische“ BI-Anwendungsgebiete sind Planung, Controlling, Reporting und Unternehmensführung. Die Informationen, die von BI-Systemen bereitgestellt werden, können für verschiedene Zwecke verwendet werden, wie z. B. die Überwachung von Leistungsindikatoren, die Identifikation von Märkten, die Vorhersage von Trends oder die Optimierung von Geschäftsprozessen.

Bedingt durch die Bedeutung von Daten für operative Abläufe, strategische Entscheidungen, Geschäftsmodelle und datengestützte Wertangebote spielt BI mit der damit einhergehenden datengestützten Entscheidungs- und Führungskultur heute in allen Bereichen des Unternehmens eine wichtige Rolle. Im Einklang mit ihren Befugnissen und Aufgaben müssen alle Beschäftigte des Unternehmens Zugriff auf Daten bekommen.

Vor dem Hintergrund der digitalen Transformation ist eine offene datengetriebene Unternehmenskultur zum Erfolgsfaktor geworden. Ein weiteres Erfolgskriterium bildet die (unternehmensweite) Datenkompetenz. Bild 1.2 zeigt die Stufen auf dem Weg hin zu einer datengetriebenen Unternehmenskultur mit dem Ziel, am Ende mithilfe von künstlicher Intelligenz (KI) Daten für „smarte“ Produkte und Dienstleistungen nutzen zu können. Grundlage bildet das Datenverständnis, welches eng mit Business Intelligence verknüpft ist. Ohne die aus der Domäne stammenden Daten zu verstehen, ist es nicht möglich, aus Daten Nutzen zu schöpfen.

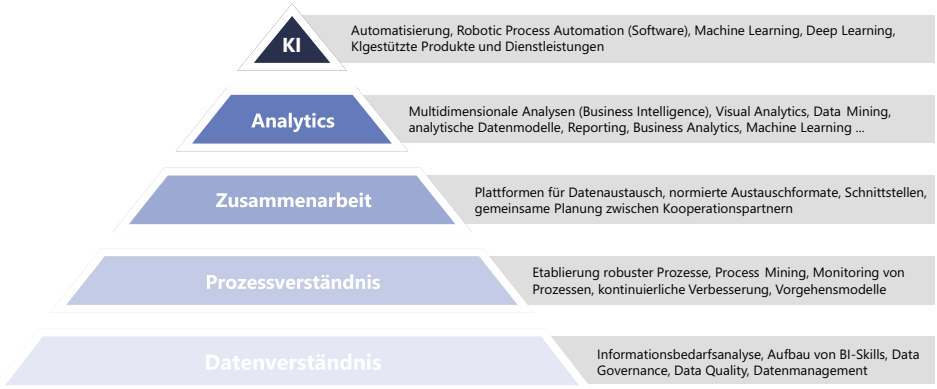

Bild 1.2 Stufen auf dem Weg hin zu einer datengetriebenen Unternehmenskultur

Ohne *Datenverständnis* einer fachgerecht aufbereiteten Datenbasis und ohne Datenkompetenz, beides gepaart mit fachlicher Expertise, bleiben BI-Werkzeuge wirkungslos. BI-Werkzeuge wenden sich zunehmend an die fachliche Nutzerschaft: Diese kennt die fachlichen Anforderungen und wird durch intuitiv benutzbare BI-Werkzeuge befähigt, eigenständig Daten zusammenzuführen, zu bereinigen, analytische Datenmodelle zu erstellen und auf deren Grundlage Daten zu analysieren. Ermittelte Ergebnisse entsprechend zu visualisieren und in Form (interaktiver oder paginierter) Berichte oder Dashboards zu kommunizieren, sind weitere Leistungsmerkmale. BI-Werkzeuge, die den Einzelnen im Umgang mit Daten befähigen, zählen zur Kategorie Self-Service-BI. Data Governance spannt einen Ordnungsrahmen für den angemessenen Umgang mit betrieblichen Daten als wichtige Wirtschaftsgüter auf (Gluchowski/Leisten/Presser 2022). Die Grundlage bilden Tätigkeiten, die regulatorische Vorschriften umsetzen oder Maßnahmen zur Datensicherheit oder Datenqualität sicherstellen.

Die Stufe Datenverständnis begünstigt auf der nächsten Stufe das **Prozessverständnis**: Der Einsatz von BI-Technologien wird nur dann effizient, wenn es gelingt, möglichst reibungslose und automatisierte Prozesse für die Datenversorgung und -analyse aufzusetzen und im Unternehmen dafür passende Vorgehensmodelle zu etablieren. Dies bedingt auch, bestehende Prozesse zu hinterfragen, zu messen, zu bewerten und entsprechend zu verbessern.

Reibungslose und nahtlose Prozesse fördern die inner- wie überbetriebliche **Zusammenarbeit**. Diese spielt eine wichtige Rolle, den stetigen Wandel durch Kooperationen und Partnerschaften mit anderen Unternehmen bewältigen zu können. Zudem ist die Kooperationsfähigkeit eine zentrale Voraussetzung, notwendiges Know-how ins Unternehmen zu bringen. Interdisziplinäre Zusammenarbeit, innerwie überbetrieblich, bildet ferner eine Voraussetzung, Analytics-Projekte erfolgreich bewältigen zu können.

Mit der Stufe **Analytics** geht ein breites Spektrum einher: angefangen von klassischen Auswertungen bin hin zu avancierten Verfahren im Zusammenhang mit Data Mining. Ein wichtiger Aspekt bei der Stufe Analytics liegt darin, eine auf Daten und Analysen aufsetzende Entscheidungskultur im Unternehmen platziert zu haben. Dies schafft zum einen das nötige Vertrauen in eine solche Vorgehensweise und zum anderen begünstigt es eine Experimentierbereitschaft im Unternehmen.

Die in der Pyramide oberste Stufe bildet **künstliche Intelligenz (KI).** Nicht nur wegen der eingesetzten Algorithmen und Verfahren, sondern vor allem wegen der vielen Voraussetzungen, die für die letzte Stufe erforderlich sind. Etwa die Fähigkeit, umfangreiche Daten zu sammeln, zu speichern und entsprechend vorzubereiten bzw. auszuzeichnen. Oder die Fähigkeit, ein tragfähiges Geschäftsmodell für smarte Wertangebote gefunden bzw. diese überhaupt erst entwickelt und funktionsfähig gemacht zu haben.

Es ergibt sich somit ein Kontinuum bis hin zur Spitze der Pyramide: Alle Stufen müssen erfolgreich durchlaufen sein und ineinanderwirken, damit ein Umfeld geschaffen ist, aus Daten das innewohnende Potenzial im Hinblick auf Automatisierung vollends auszuschöpfen. BI stellt für diese Entwicklung einen Startpunkt dar.

Auch wenn Business Intelligence zunächst unabhängig von IT-Architekturen wie Data Warehouses entstanden ist, wird BI heute in Zusammenhang mit solchen Ansätzen und insbesondere Cloud-basierten IT-Infrastrukturen gebracht. Per ETL (Extract, Transform, and Load) werden meist strukturierte Daten aus verschiedenen Datenquellen zusammengeführt und in eine Struktur überführt, auf der analytische Datenmodelle für mehrdimensionale Analysen aufbauen. Hierbei kommt dem Data Warehouse als zentralisierte analytische Datenbank die Rolle zu, die heterogenen und oftmals dezentral anfallenden Daten zu konsolidieren und in einen für die Analyse konsistenten und qualitätsgesicherten Zustand zu überführen.

Die im Zusammenhang mit Business Intelligence oft vorkommende Abkürzung OLAP (Online Analytical Processing) zielt auf die interaktive Navigation in verdichteten mehrdimensionalen Daten ab: Eine typische OLAP-Operation ist, aggregierte Daten nach ihren Einzelwerten aufzulösen, etwa wenn bei einem Summenwert die dazugehörenden Einzelwerte interessieren. In Fachbereichen (z. B. Controlling) browsen Anwenderinnen und Anwender mittels OLAP-Systemen interaktiv in multidimensionalen Daten oder stellen Datenkonstellationen anhand von (Pivot-)Tabellen oder Datenvisualisierungen nach. Mit diesem Ansatz lassen sich Auffälligkeiten in Daten interaktiv nachgehen, nachvollziehen und auch verstehen sowie an Dritte kommunizieren.

BI verwendet zur Analyse strukturierte Daten aus betrieblichen Vorsystemen, etwa Warenwirtschafts-, Buchhaltungs- oder Enterprise-Resource-Planning-Systemen. Klassische BI-Analysen sind oftmals deskriptiv und diagnostisch. Sie zielen darauf ab, nach Ursachen für bestimmte Entwicklungen oder Probleme zu suchen. Predictive Analytics und Prescriptive Analytics bleiben bei BI meist unberücksichtigt. BI umfasst Prozesse, strukturierte Daten aus verschiedenen Datenquellen zu bereinigen, aufzubereiten und zu verdichten: Sie bleiben so für den Menschen übersichtlich.

Das Zusammenspiel von Analysen, die einerseits strukturierte und andererseits unstrukturierte Daten nutzen, brächte für Unternehmen neue Erkenntnisse. Die Analyse unstrukturierter Daten in Form von Texten, etwa Rezensionen auf Online-Portalen, Blogs oder in den sozialen Medien, schüfe für Unternehmen interessante Einblicke, wie die Kundschaft Produkte des Unternehmens wahrnimmt oder beurteilt. Solche Einblicke werden im Zusammenspiel mit BI-Auswertungen interessant, die sich aus betrieblichen Informationssystemen speisen. So ließen sich z. B. Umsatzrückgänge bestimmter Produkte oder Dienstleistungen anhand der analysierten Rezensionen der Kunden weitaus besser verstehen, als nur die Daten aus internen Systemen auszuwerten und Zusammenhänge darin zu finden. Der An-

spruch, strukturierte, semistrukturierte und unstrukturierte Daten gleichermaßen analysieren zu wollen, führt zu neuen Architekturen und macht es erforderlich, klassisches BI damit zu verbinden.

Der Definitionsansatz von Dietmar Schön (2018) fasst nochmals Aspekte rund um BI abschließend zusammen:

> „Business Intelligence ist die Integration von fachlichen Managementmethoden, IT-Verfahren und analytischen Prozessen, die sowohl die Aufbereitung und Bereitstellung von Daten als auch die Aufdeckung relevanter Zusammenhänge sowie die Kommunikation der gewonnenen Erkenntnisse zur Entscheidungsunterstützung für das Management umfassen und hierzu für die Planung, die Analysen und die Prognosen leistungsfähige IT wie Data-Warehouse- und Big-Data-Technologien einsetzen."

1.2.2 Data Mining

Data Mining ist ein weiterer Analytics-Fachbegriff.

> Fayyad, Piatetsky-Shapiro und Smyth verstehen unter **Data Mining**: „Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. [...] By non-trivial, we mean that some search or inference is involved; that is, it is not a straightforward computation of predefined quantities like computing the average value of a set of numbers [...]" (Fayyad/Piatetsky-Shapiro/Smyth 1996).
>
> Nach Bissantz und Hagedorn beschreibt der „[...] Begriff Data Mining, im Folgenden übersetzt mit **Datenmustererkennung**, die Extraktion implizit vorhandenen, nicht trivialen und nützlichen Wissens aus großen, dynamischen, relativ komplex strukturierten Datenbeständen" (Bissantz/Hagedorn 1993).

Beiden Definitionsansätzen ist der Ansatz von Data Mining gemeinsam, aus Daten nützliche Muster zu extrahieren, die sich in Wissen überführen lassen.

Data Mining greift in der Regel auf strukturierte Daten zu, die in relationalen Datenbanken oder in Data Warehouses vorliegen. Sind unstrukturierte Daten wie Texte die Datengrundlage, handelt es sich um Text Mining. Data Mining zielt primär darauf ab, aus relativ komplex strukturierten Datenbeständen nützliches und nicht triviales Wissen zu gewinnen.

Data Mining lässt sich dabei in Descriptive Data Mining und Predictive Data Mining unterteilen (Bild 1.3). Die deskriptive Facette von Data Mining hat Berührungspunkte zur explorativen Datenanalyse. Explanatory Data Analysis (EDA) ist ein Teilgebiet der Statistik. Sie geht auf *John Tukey*, einem Vordenker für die computer-

gestützte Statistik, zurück. Er etablierte Explanatory Data Analysis in den 1970er-Jahren und führte auch Methoden zur grafischen Datenanalyse wie etwa das Boxplot (Box and Whisker Plot) ein. Data Mining nutzt diese Ansätze, Daten visuell zu inspizieren und darin z. B. nach Ausreißern zu suchen oder Verteilungen zu visualisieren. Die dafür typischen Ansätze bzw. Verfahren lassen sich in folgende Anwendungsfelder (Auswahl) gliedern:

- **Klassifikation**

  Bei der Klassifikation geht es darum, Objekte (z. B. Personen) anhand ihrer Merkmale automatisiert in Klassen einzuteilen. Die Klassen sind vorgegeben, die Klassenzugehörigkeit eines Objekts ist nicht bekannt. Die Herausforderung für das Klassifikationsmodell besteht darin, das jeweilige Element möglichst treffsicher einer der vorhandenen Klassen zuzuordnen. Die Klassifikation fällt in den Bereich Predictive Analytics und zählt zum sogenannten Supervised Learning (überwachtes Lernen), einem Teilbereich des Machine Learning. Das Klassifikationsmodell wird anhand einer sogenannten Trainingsmenge gebildet. In den Trainingsdaten ist pro Objekt bekannt, in welche Klasse es fällt. Das Klassifikationsmodell stellt einen Zusammenhang zwischen den ausgewählten Merkmalen (Features) und der Klassenzugehörigkeit her.

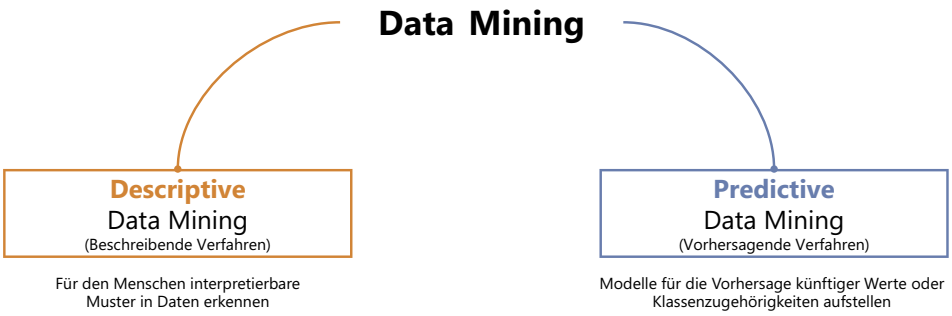

Bild 1.3 Kategorien von Data Mining

- **Segmentierung/Clustering**

  Beim Clustering geht es darum, bestehende Objekte bzw. eine Grundgesamtheit anhand ihrer Merkmale in vorab unbekannte Gruppen einzuteilen. Die Herausforderung für das Clustering-Modell liegt darin, die Gruppen so zu bilden, dass die den jeweiligen Teilgruppen zugewiesenen Objekte hinsichtlich ihrer Merkmalsausprägungen möglichst homogen sind. Zwischen den Teilgruppen sollten die Unterschiede hingegen möglichst groß sein.

- **Regression**

  Bei der Regression geht es darum, Zusammenhänge zwischen Merkmalsausprägungen von Objekten zu finden. Mit einem Regressionsmodell wird eine abhängige, stetige Merkmalsausprägung durch mehrere unabhängige erklärt. Die dafür nötige Schätzfunktion ist das sogenannte Regressionsmodell. Bei der linearen Regression entspricht das Regressionsmodell einer Funktion ersten Grades: Sind die Werte der unabhängigen Variablen bekannt, lässt sich der numerische Output der abhängigen Variable bestimmen. Auch wenn Regression und Klassifikation beide zur Kategorie Predictive Analytics zählen, unterscheiden sie sich. Während die Klassifikation eine Wahrscheinlichkeit zu einer bestimmten Klassenzugehörigkeit angibt, errechnet die Regression einen bestimmten Zahlenwert als numerischen Output.

- **Abhängigkeitsentdeckung/Assoziation**

  Ziel ist, Abhängigkeiten zwischen Merkmalen oder einzelnen Merkmalsausprägungen zu erkennen, die innerhalb des Datenbestandes bzw. einer Teilmenge des Datenbestandes vorliegen. Ein typisches Beispiel für eine Assoziationsanalyse ist die Untersuchung von Warenkörben: Finden sich etwa Produkte, die überdurchschnittlich häufig gemeinsam gekauft werden, so lassen sich daraus Empfehlungen ableiten. Wenn eine Person etwa ein Produkt gekauft hat, das überdurchschnittlich oft in der Vergangenheit mit einem anderen Produkt gemeinsam gekauft wurde, so lässt sich dieses als Empfehlung vorschlagen.

Bild 1.4 stellt Aufgabenfelder bzw. Geschäftsanwendungsfälle Analytics-Anwendungsfeldern und potenziell geeigneten Algorithmen gegenüber. Aus der Abbildung wird ersichtlich, dass es oftmals eine Reihe von Algorithmen gibt, die sich für ein Anwendungsfeld eignen.

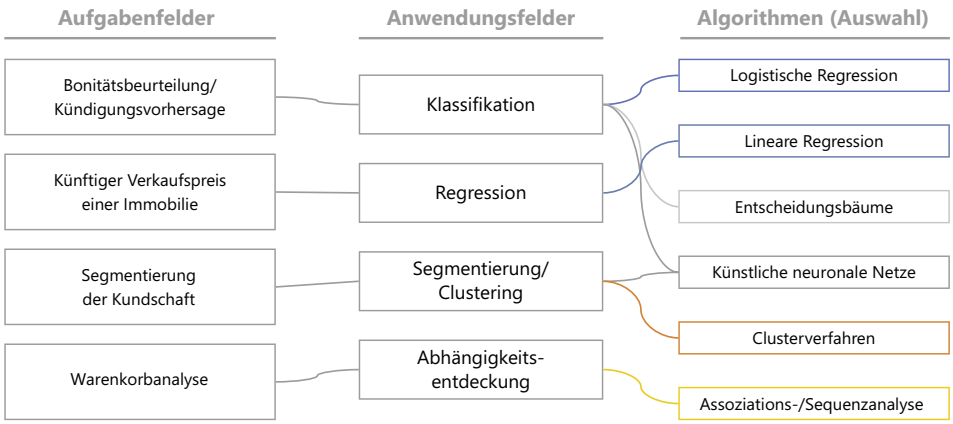

Bild 1.4 Aufgabenfelder, Anwendungsfelder und Algorithmen von Data Mining (Auswahl)

Verschiedene Vorgehensmodelle nutzen den Begriff Data Mining unterschiedlich: Bei Standardvorgehensmodellen wie z. B. Knowledge Discovery in Databases oder CRISP-DM bildet Data Mining eine eigene Phase innerhalb des Gesamtprozesses. Data Mining lässt sich aber ebenso als Oberbegriff sehen, unter den sich alle zur Wissensentdeckung erforderlichen Prozesse subsumieren lassen. Stellenweise wird die Suche nach Mustern in Big-Data-Beständen ebenso zu Data Mining gezählt, obgleich Data Mining ursprünglich strukturierte Daten nutzt, die aus (meist) relationalen Datenbanksystemen oder Data Warehouses stammen. Im Vergleich dazu greifen Big-Data-Analysen auf polystrukturierte Daten zu, die auf verteilten Dateisystemen liegen oder in NoSQL-Datenbanken gespeichert sein können. Wenn Data Mining agnostisch von der technischen Datenhaltung gesehen wird, passt der Begriff zum Verständnis von Big Data (Analytics), in sehr großen Datenbeständen nach Mustern zu suchen.

Es gibt aber hinsichtlich der Herangehensweise zwischen Data Mining und Big Data Analytics Unterschiede: Bei Data Mining sucht der Mensch computergestützt mittels (häufig statistischer) Datenanalyse- und Entdeckungsalgorithmen situativ und meist einzelfallbezogen nach neuen Mustern in einem vorliegenden Datenbestand. Das Modell ist auf die gegenwärtige Datenverteilung abgestimmt. Big Data Analytics setzt hingegen auf Methoden aus dem Umfeld von Machine Learning oder Deep Learning im Kontext künstlicher Intelligenz. Machine Learning verfolgt den Anspruch, Computer durch verschiedene Lernansätze (z. B. Supervised Learning) zu befähigen, Aufgaben zu erledigen, ohne Computer dafür im Vorfeld explizit programmiert zu haben (programming by example).

Bei Supervised Learning soll anhand von Vergangenheitsdaten computerbasiert neues Wissen generiert werden. Der Computer soll „befähigt“ werden, aus Daten eigenständig Muster zu erkennen, die sich künftig auch auf unbekannte Daten anwenden lassen. Ziel ist dabei, einen Kreislauf zu schaffen, der computerbasiertes Lernen möglich macht, ohne in späteren Stadien noch einen Menschen involvieren zu müssen.

Die Abgrenzung von Data Mining und Machine Learning ist nicht immer trennscharf. Im Zuge der computerbasierten Suche nach Mustern in Daten greift Data Mining ebenso auf Algorithmen zurück, die sich für Machine Learning nutzen lassen bzw. aus diesem Fachbereich stammen. Die Schwerpunkte beider Ansätze liegen anders: Bei Data Mining steht im Vordergrund, aus einem konkreten Datenbestand Muster zu extrahieren und einzelfallbezogen unbekannte Muster aufzudecken. Machine Learning zielt darauf ab, aus gesammelten Daten Algorithmen bzw. Modelle abzuleiten, die sich eignen, in neuen Daten nach bekannten Mustern zu fahnden. Finden sich Gesetzmäßigkeiten, kann der Computer (unter anderem) künftige Werte vorhersagen oder die Wahrscheinlichkeit für eine bestimmte Klassenzugehörigkeit bestimmen. Idealtypisch funktioniert ein kalibriertes Machine-Learning-Modell nach einer Einschwingphase zunehmend ohne menschliches

Zutun, lässt sich wiederverwenden und verbessert sich laufend selbständig. Es verfolgt somit auch den Anspruch, wiederverwendbar und möglichst generisch zu sein. Data Mining sucht nicht nach einem generischen und für alle Datenbestände anpassbaren Modell, sondern erstellt ein für die vorliegenden Daten optimales. Dies rechtfertigt auch den hohen Aufwand, den der Mensch in einem Data-Mining-Projekt hat. Idealerweise fördert Data Mining nützliche Einsichten zutage: Der Mensch muss diese interpretieren, aufbereiten, kommunizieren und danach handeln.

1.2.3 Data Science

Der Begriff Data Science wurde 1974 vom Informatiker Peter Naur in seinem Werk *Concise Survey of Computer Methods* eingeführt. Er definiert Data Science wie folgt (Naur 1974):

Data Science ist „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".

In seinem Beitrag „Future of Statistics: Data Science" schlägt Jeff Wu im Jahr 1986 vor, den Begriff Statistik durch Data Science zu ersetzen und statt Statistiker die Bezeichnung Data Scientist zu verwenden. So sollte die Bedeutung der **computergestützten Datenanalyse** zum Ausdruck gebracht werden. Ein Umstand, auf den bereits der Statistiker John Tukey in den 1960er-Jahren hinwies, wobei er die Bedeutung der aufkommenden Computer für Datenanalyse hervorhob.

Data Mining und Data Science eint wiederum das gemeinsame Ziel, aus Daten nützliche Erkenntnisse und Wissen abzuleiten. Im Vergleich zu Data Mining wird Data Science von vornherein als interdisziplinäre Datenwissenschaft definiert: Diese beschäftigt sich mit der Extraktion von Wissen und Erkenntnissen aus großen Mengen von Daten. Data Science kommt in vielen Bereichen zum Einsatz, wie etwa der Wirtschaft, der Medizin, der Wissenschaft, der Technologie und der Sozialwissenschaften. Data Science nutzt Methoden und Techniken aus verschiedenen Bereichen wie der Informatik, Statistik, Mathematik, Machine Learning und der Visualisierung, um komplexe Probleme zu lösen und Daten zu analysieren.

Diese Überschneidungen im Zusammenspiel mit dem Wissen der jeweiligen Fachdisziplin (Domain Knowledge) sind oftmals als Venn-Diagramme visualisiert (Bild 1.5). Die gemeinsame Schnittmenge aus den verschiedenen Fachdisziplinen bildet dabei Data Science. Die Definition von Data Science als Schnittmenge löst kontroverse Diskussionen aus und wirft die Frage auf, ob dies bereits ausreichend für eine eigenständige Wissenschaft sei. Auch bei anderen Analytics-Fachbegriffen

wie etwa Data Mining gibt es starke Berührungspunkte zur Mathematik/Statistik. Die bei Data Mining vorhandene Überschneidung zur IT wird aber weitaus weniger in den Vordergrund gestellt als bei Data Science. Bei Data Science werden vor allem die spezifischen Kenntnisse rund um Programmierung, Big-Data-Technologien sowie Analytics-Infrastrukturen im Generellen hervorgehoben. Data Science beleuchtet damit auch die zur Speicherung, Analyse und Verarbeitung von Daten nötigen Ressourcen.

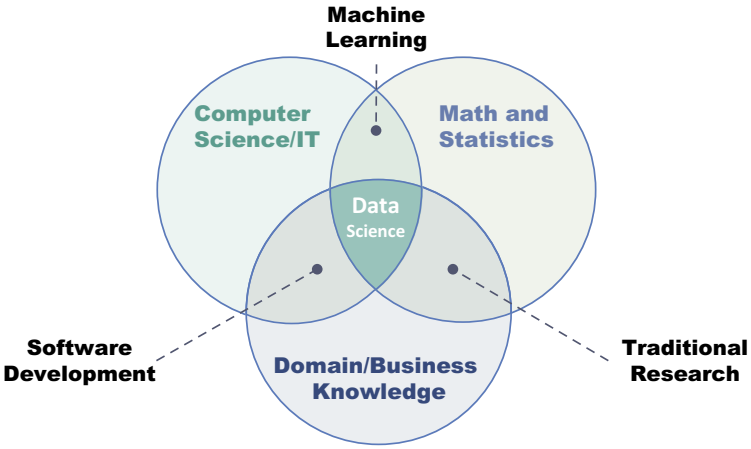

Bild 1.5 Data Science mit Berührungspunkten zu anderen Disziplinen (Luellen 2018)

Bild 1.6 zeigt den umfassenden Ansatz von Data Science, auf die verschiedensten Aspekte rund um Analytics einzugehen, damit ist Data Science deutlich breiter aufgestellt als etwa Data Mining.

Die Data Science Association definiert Data Science wie folgt (Haneke et al. 2021):

> „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."

Bild 1.6 Facetten rund um Data Science

1.2.4 Business Analytics

Für Business Analytics gibt es keine übereinstimmende Definition. *Davenport* und Harris (2007) betonen in ihrem Buch *Competing on Analytics* die strategische Bedeutung der Analytik für die Wettbewerbsfähigkeit von Unternehmen. *Provost* und Fawcett (2013) argumentieren in ihrem Werk *Data Science for Business* in eine ähnliche Richtung. Einhergehend mit Big Data stieg der Bedarf, polystrukturierte Daten zur Lösung unternehmensrelevanter Aufgaben zu nutzen bzw. darin nach Wissen zu suchen.

Vor diesem Hintergrund kam in den 2000er-Jahren der Begriff Business Analytics (BA) auf: Der Aspekt Analytics wird als Kernbereich in Business Analytics besonders hervorgehoben. Die Abgrenzung zu Business Intelligence ist nicht immer trennscharf. Oftmals werden daher BI und BA zu BI&A oder BIA zusammengezogen. Meist wird Business Analytics mit prognostischen Verfahren assoziiert, während der Fokus von Business Intelligence eher auf deskriptiven Methoden liegt. Eine solche Unterscheidung zeigt Bild 1.7. Hierbei wird Business Analytics vor allem mit Predictive Analytics und Prescriptive Analytics assoziiert. Predictive Analytics richtet den Blick auf die Zukunft und versucht mithilfe verschiedener Methoden, von Daten aus der Vergangenheit auf zukünftige Trends und Entwicklungen zu schließen. Statt Business Intelligence und Business Analytics anhand der eingesetzten Analytics-Verfahren zu trennen, verfolgen Chen, Chiang und Storey einen anderen Ansatz: Sie sehen Business Analytics und Business Intelligence als angewandtes Data Science im Geschäftskontext (Chen/Chiang/Storey 2012):

„[...] BI&A is data science in business."

Im Folgenden wird Business Analytics als Data Science im Business verstanden. Mit dem Fokus auf die betriebliche Anwendungsdomäne. Der Ansatz trägt dem Umstand Rechnung, Business Intelligence und Business Analytics mittels eingesetzter Analyseverfahren trennscharf voneinander abgrenzen zu können. Data Science berücksichtigt nicht nur die notwendigen Algorithmen und Analyseverfahren, sondern betrachtet auch die für Analytics nötigen Ressourcen. Darunter fallen unter anderem Fachleute mit dem erforderlichen Know-how, Datenmanagement, IT-Management, IT-Infrastrukturen, fachliche Aspekte der jeweiligen Anwendungsdomäne, Organisation sowie Vorgehensmodelle einschließlich Projektmanagement. Erst das reibungslose Zusammenspiel all dieser (nicht vollständig aufgelisteten) Faktoren macht es möglich, die mit Business Analytics verknüpften Erwartungen erfüllen zu können.

Gartner definiert Business Analytics wie folgt – auch hier ist die Trennung insbesondere zu BI nicht trennscharf, sodass im Folgenden Business Analytics als Data Science im Business gesehen wird.

> „Business analytics is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states. Business analytics includes data mining, predictive analytics, applied analytics and statistics, and is delivered as an application suitable for a business user [...]."

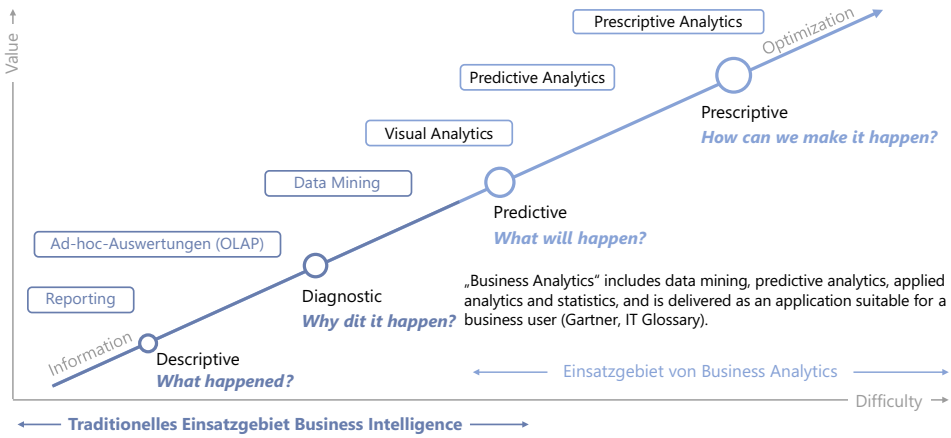

Bild 1.7 Facetten von Business Analytics (in Anlehnung an Seiter 2019)

Bild 1.8 zeigt einen Vorschlag, wie sich die verschiedenen Begrifflichkeiten einordnen ließen. Das gemeinsame Ziel aller Analytics-Aktivitäten liegt darin, aus Daten Wissen zu schöpfen. Hierfür bildet Data Science als interdisziplinäre Wissenschaft die Klammer über sämtliche Begriffe. Data Science hat Berührungspunkte zur Mathematik/Statistik, Informationstechnologie/Informatik und ebenso zu den je-

weiligen fachspezifischen Wissenschaften. Data Science in der wirtschaftswissenschaftlichen Domäne angewandt, wird zu Business Analytics. Aus Sicht des Informationsbedarfs, der Anwendungen, aber auch der Daten (Strukturiertheit, Menge usw.) ergeben sich unterschiedliche Ansätze, aus den zugrunde liegenden Daten nützliches Wissen abzuleiten oder datengetriebene Anwendungen zu erstellen. Auf unterster Ebene spielt es eine Rolle, ob die jeweiligen (Analyse-)Ergebnisse an den Menschen adressiert sind, etwa um bessere Geschäftsentscheidungen zu treffen oder ob datenbasierte Vollautomatisierung bzw. autonome, KI-basierte Anwendungen das Ziel sind. Auch wenn etwa Data Mining, BI oder KI usw. in der Darstellung getrennt sind, arbeiten sie in praxi eng zusammen und konvergieren auch technologisch, etwa im Bereich der Analytics-Architekturen (z. B. Data Lakehouse).

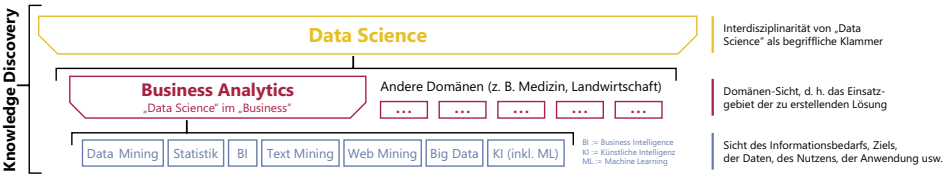

Bild 1.8 Zusammenfassung und Einordnung von Business Analytics

1.3 Vorgehensmodelle

1.3.1 Knowledge Discovery in Databases (KDD)

Ein wesentliches Ziel eines Analytics-Projekts liegt darin, aus Daten nützliche Information und Wissen zu gewinnen. Dies erfordert ein systematisches und nachvollziehbares Vorgehen. Ein Vorgehensmodell dafür stellt der KDD-Prozess (Knowledge Discovery in Databases) dar. KDD geht auf Gregory Piatetsky-Shapiro zurück, der ihn bereits im Jahr 1989 vorschlug. Nach Usama Fayyad, Gregory Piatetsky-Shapiro und Padhraic Smyth bezieht sich KDD auf den gesamten Prozess, der aus Daten nützliches Wissen gewinnt. Im KDD-Prozess bildet hingegen Data Mining einen Teilschritt in diesem Gesamtprozess und zielt darauf ab, anhand von (Data-Mining-)Algorithmen Muster in Daten zu finden (Fayyad/Piatetsky-Shapiro/Smyth 1996).

Gregory Piatetsky-Shapiro beschreibt in einem Beitrag auf LinkedIn seine Motivation („can we find such patterns in data automatically?" (Piatetsky-Shapiro 2021)). An KDD orientieren sich weitere Vorgehensmodelle, etwa CRISP-DM. KDD gilt somit als einer der ersten computergestützten Ansätze, durch systematisches sowie

standardisiertes Vorgehen aus Daten schrittweise Information und Wissen zu ziehen. KDD zielt darauf ab, durch Data Mining in Daten Muster zu finden, die sich möglichst generalisieren und in Wissen überführen lassen (Bild 1.9).

Den Ausgangspunkt für den KDD-Prozess bildet eine Grundgesamtheit an Daten in Form von Rohdaten. Ein Teil dieser Daten wird in Folgeschritten aufbereitet und in eine qualitativ gesicherte Datengrundlage überführt. Für die jeweiligen Data-Mining-Algorithmen oder Analysewerkzeuge müssen die Daten in einem bestimmten Format bzw. einer bestimmten Struktur vorliegen. Die transformierten Daten werden anschließend mittels Data-Mining-Algorithmen analysiert. Hierbei ist grundlegend, welches analytische Problem zu lösen ist. Die einzusetzenden Data-Mining-Algorithmen müssen dafür passend und geeignet sein. Am Beispiel einer Klassifikation lässt sich aufzeigen, wie der KDD-Prozess induktive und deduktive Ansätze für den Erkenntnisgewinn kombiniert. Anhand von Trainingsdaten erstellt ein Algorithmus induktiv ein Modell. Nach Interpretation/Evaluation dienen die gefundenen Regeln dazu, Aussagen über unbekannte Daten in Form von Zuordnungen machen zu können. Die Klassifikation unbekannter Daten in gegebene Klassen sind regelbasierte Aussagen: Für jede Instanz bestimmt der Algorithmus die Klassenzugehörigkeit einer Instanz. Das induktiv gewonnene, nun allgemeine Modell auf neue Daten anzuwenden, ist Deduktion.

Die von den jeweiligen Data-Mining-Algorithmen gefundenen Muster sind zu interpretieren und auch hinsichtlich ihrer Güte zu evaluieren. Die Evaluation hängt davon ab, welches analytische Problem zu lösen war. Bei einer Klassifikation etwa lassen sich die Ergebnisse anhand von Testdaten überprüfen, während dieser Ansatz bei einer Cluster-Analyse ausscheidet. Führt die Evaluation zu qualitätsgesicherten Ergebnissen, lassen sich die aus den Daten gefundenen Erkenntnisse nutzbringend verwenden. Aus Daten gewonnene Erkenntnisse führen nach Fayyad, Piatetsky-Shapiro und Smyth zu Wissen. Der KDD-Prozess liefert ein systematisches Vorgehensmodell, das sich als Ausgangspunkt für weitere Referenzmodelle eignet, die Phasen noch weiter zu verfeinern und zu konkretisieren. Der KDD-Prozess selbst ist implementierungs-, werkzeug- und anwendungsneutral.

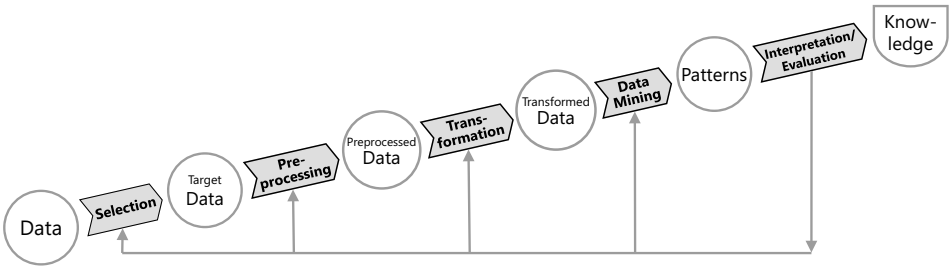

Bild 1.9 Knowledge Discovery in Databases (KDD) (Fayyad/Piatetsky-Shapiro/Smyth 1996)

1.3.2 CRISP-DM

Ein weiteres Vorgehensmodell für Analytics-Projekte stellt CRISP-DM (CRoss Industry Standard Process for Data Mining) dar. CRISP-DM ist ein branchenübergreifendes Referenzmodell, das von der EU gefördert wurde. Es entstand ab Mitte der 1990er-Jahre unter Federführung von (damals) Daimler-Benz AG, SPSS (damals ISL) und NCR Corporation, die Analytics-Lösungen im Angebot hatten oder bereits Data Mining operativ einsetzten (Alby 2022). CRISP-DM ist werkzeugneutral, implementierungsunabhängig und weitverbreitet (KDnuggets 2014).

Bild 1.10 zeigt die Phasen von CRISP-DM als Kreislaufmodell: Der äußere Kreis versinnbildlicht die Iteration des gesamten CRISP-DM-Modells. Jede Iteration kann dabei neue Fragen aufwerfen oder schrittweise das Vorgehen verbessern. Die inneren Pfeile zwischen den Phasen machen Abhängigkeiten und Rückkopplungen deutlich. Es kommt vor, dass von einer Phase auf die vorherige zurückgesprungen werden muss (Alby 2022). Neue Erkenntnisse oder Probleme können dies erforderlich machen. CRISP-DM ist als nie endender Kreislauf zu verstehen: Wenn ein Zyklus durchlaufen ist, setzt ein neuer ein – mit dem Ziel der stetigen Verbesserung bislang erreichter Ergebnisse bzw. Einsichten in Daten.

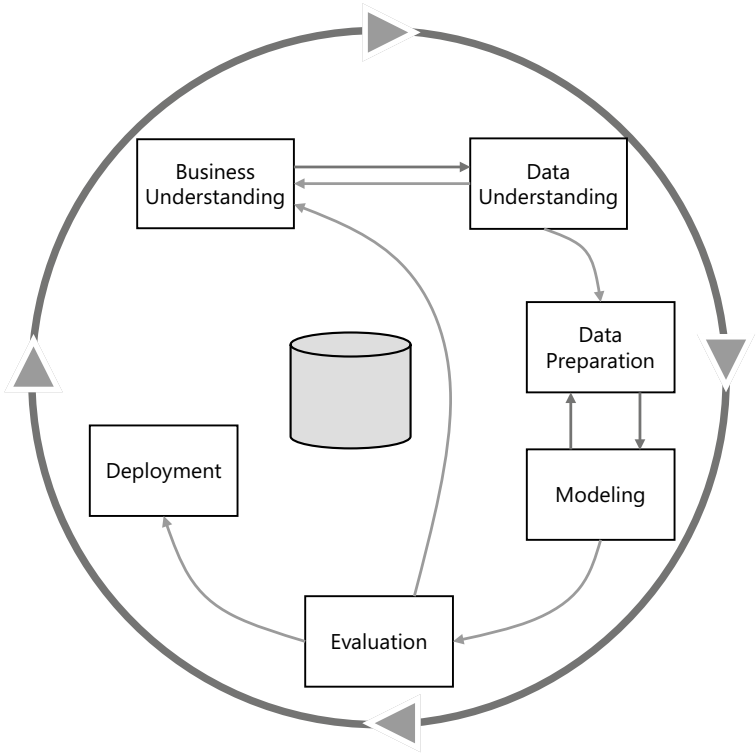

Bild 1.10 CRISP-DM als Kreislaufmodell (Alby 2022)

Zu Beginn des CRISP-DM-Modells steht die Phase Business Understanding: In einem Analytics-Projekt sollte nicht die Modellierung bzw. Analyse den Startpunkt bilden, sondern zunächst das Geschäft mit seinen Abläufen bzw. die fachliche Problemstellung im Generellen („Anwenderseite“) verstanden werden. Bei Business Understanding steht im Vordergrund, Geschäftsziele zu bestimmen und die gegenwärtige Situation zu bewerten. Darüber hinaus sind im Zuge von Business Understanding operationalisierbare Ziele sowie der Nutzen des Analytics-Projekts festzulegen. Auch potenzielle Risiken und Handlungsalternativen („Plan B“) sollte Business Understanding betrachtet haben, falls keine geeigneten Daten vorliegen oder deren mangelhafte Qualität eine tiefer gehende Analyse verhindert (Gutman et al. 2022). In der Phase Business Understanding ist die Entscheidung zu fällen, ob das beabsichtigte Analytics-Projekt starten soll oder nicht. Dafür lassen sich etwa folgende Kontrollfragen heranziehen (Herbold 2022):

- Sind die benötigten Daten vorhanden?
- Lassen rechtliche Rahmenbedingungen sowie ethische Normen und Werte das Projekt zu?
- Sind die technologischen Ressourcen vorhanden bzw. überhaupt im Zeitraum zu beschaffen?
- Steht die (rechnerische) Personalkapazität für den Projektzeitraum uneingeschränkt bereit?
- Gibt es die benötigten Fähigkeiten (Stichwort Datenkompetenz) im Unternehmen?
- Lassen die identifizierten Risiken das Projekt zu?
- Wie sehen die ökonomischen Zielvorstellungen und die Projektplanung aus?

Der Phase Business Understanding folgt Data Understanding: Hierbei stehen die Daten und Datenquellen im Mittelpunkt. Dazu zählt, die erforderlichen Daten zu sammeln, zu beschreiben, zu untersuchen und deren Datenqualität zu beurteilen. Ein Hilfsmittel dafür ist die explorative Datenanalyse (EDA), die auf den Statistiker John Tukey zurückgeht. In der explorativen Datenanalyse dienen Daten dazu, zu neuen Hypothesen und Einsichten zu kommen. Fahrmeir et al. sehen die explorative Datenanalyse „[…] konzipiert zur Suche nach Strukturen und Besonderheiten in den Daten und kann so oft zu neuen Fragestellungen oder Hypothesen in den jeweiligen Anwendungen führen“ (Fahrmeir et al. 2011). Im Rahmen von Data Understanding interessieren bisher unbekannte Strukturen, Auffälligkeiten bzw. Besonderheiten in den Daten.

EDA bildet bei CRISP-DM den „Einstieg“ in die Datenanalyse, bevor in Folgeschritten avanciertere Verfahren (z. B. Machine Learning) zum Einsatz kommen. Anhand deskriptiver Statistik und visueller Analysemethoden sind Strukturen in einem Datenbestand oder in einer gezogenen Stichprobe (= Auszug einer Grundgesamt-

heit) so prägnant wie möglich zu charakterisieren. Die explorative Datenanalyse nutzt dafür Parameter, Tabellen und insbesondere Grafiken. Welche Verfahren sich auf die Daten anwenden lassen, hängt mit dem Skalenniveau (nominal, ordinal oder metrisch) der Merkmale zusammen. Die Untersuchungen zielen darauf ab, im Datenbestand auffällige Muster zu erkennen. Darüber hinaus interessieren Zusammenhänge und Korrelationen zwischen den einzelnen Merkmalen. Ebenso eventuell vorhandene Dimensionalität in den Daten, etwa dann, wenn es einen Zeit- oder Ortsbezug gibt. Im Zuge der explorativen Datenanalyse werden auch Ausreißer oder fehlende Werte identifiziert. Anhand der Analyseergebnisse und Plots wird entschieden, welche Merkmale sich für weiterführende Analysen eignen oder nicht. Die vielseitigen Untersuchungen tragen dazu bei, die Daten zu verstehen und damit auch die Datenqualität beurteilen zu können.

Nach Data Understanding liegt ein Bild vor, welche Daten, Merkmale und Datenquellen relevant sind. Darüber hinaus wurden Strukturen und Zusammenhänge in den Daten identifiziert und auch wesentliche Merkmale für das spätere Analytics-Modell bestimmt. Data Preparation folgt auf Data Understanding und umfasst alle vorbereitenden Maßnahmen, die im Hinblick auf die Phase Modeling/Data Mining nötig sind. Darunter fallen Tätigkeiten zum Aufbau von Datenkatalogen, anhand derer Datenbestände mithilfe von Metadaten beschrieben werden. Zu Data Preparation zählen sämtliche Maßnahmen, die im Zuge der Datenbereinigung (Data Cleansing) anfallen: etwa fehlende Werte zu ersetzen, fehlerhafte Werte zu bereinigen, Werte zu konvertieren, Werte zu filtern, Werte zu (re-)codieren, unterschiedliche Größenordnungen bei Werten zu normalisieren oder doppelt vorhandene Werte zu entfernen. Tätigkeiten rund um Data Preparation sind meist zeitintensiv und wenig „glamourös“, spielen aber eine neuralgische Rolle für die späteren Ergebnisse. Ein Ziel der Phase Data Preparation liegt darin, eine qualitativ gesicherte Datengrundlage für die nachfolgenden Analyse- und Modellierungsschritte zu erstellen.

In Abhängigkeit zu den später eingesetzten Machine-Learning-Verfahren werden in der Phase „Data Preparation“ die Daten in ein für die Analysemethoden benötigtes Format bzw. Struktur überführt. Eine solche Struktur ist etwa die sogenannte Analytical Base Table (ABT), welche bei prognostischen Verfahren, z. B. einer Klassifikation, die Datengrundlage in Form einer flachen, denormalisierten Tabelle beschreibt. Zu Data Preparation zählen zudem Tätigkeiten, wenn Daten aus unterschiedlichen Quellen in eine einheitliche Struktur zu überführen sind, etwa wenn die Notwendigkeit besteht, Datenbanktabellen aus verschiedenen Datenbanken zusammenzuführen.

Die bei Business Understanding definierte Aufgabe gibt bereits vor, welches Analytics-Anwendungsfeld grundsätzlich infrage kommt. Algorithmen erzeugen kalibrierbare Machine-Learning-Modelle (ML), anhand derer sich Muster und Zusammenhänge in Daten auffinden lassen. Für ein bestimmtes Analytics-Problem

stehen meist etliche Algorithmen zur Auswahl: So gibt es beispielsweise für das Analytics-Problem der Klassifikation (Supervised Learning) eine Vielzahl infrage kommender Algorithmen: etwa logistische Regression, Naive Bayes, neuronales Netzwerk, Random Forest oder Decision Tree – um nur einige zu nennen. Dies zieht die Frage nach sich, welcher der potenziell möglichen ML-Algorithmen der am besten für den jeweiligen Datensatz geeignete ist. Bedingt durch die vielfältigen ML-Algorithmen ergibt sich ein Wechselspiel mit der Phase Data Preparation. Führt ein ML-Algorithmus nicht zu aussagekräftigen Ergebnissen, kommt ein anderer zum Zug. Dieser kann aber andere Daten, Merkmale oder Skalenniveaus voraussetzen, was wiederum Nacharbeiten bei Data Preparation bedingen kann.

Die Phase Modeling ist dann erfolgreich abgeschlossen, wenn zum Untersuchungsgegenstand und zum Datensatz passende Analysemethoden und ML-Algorithmen gefunden wurden. Darüber hinaus geeignete ML-Algorithmen und Analysemethoden auf den Datensatz angewendet und ML-Modelle erstellt wurden. Und weiterhin verschiedene Modellvarianten bzw. -reihen mit unterschiedlichen Parametern „durchgerechnet“ wurden. Die Modeling-Phase soll nützliche Ergebnisse in Form von Modellen liefern und Muster in Daten zum Vorschein bringen. Auf diesen Aspekt zielt auch die alternative Bezeichnung Data Mining für „Modeling“ ab: Data Mining sucht nach interessanten Mustern in Daten. Solche identifizierten Muster können Korrelationen, Trends, Regelmäßigkeiten, Auffälligkeiten usw. sein.

Die Phase Evaluation ist inhaltlich davon abhängig, welches Analytics-Problem bei Modeling/Data Mining zu lösen war: Je nach Analytics-Problem gibt es unterschiedliche Ansätze bei der Evaluation. Bei einer Klassifikation (Supervised Learning) wird die Datengrundlage für die Analyse in einen Trainings- und Testdatenbestand aufgeteilt. Testdaten dienen dazu, das anhand des Trainingsdatensatzes angelernte Modell mithilfe von Gütekriterien zu beurteilen. So geht es etwa bei einer Klassifikation darum, ein möglichst generalisierbares Modell generiert zu haben, das auch auf unbekannten Datensätzen eine hohe Trenngüte besitzt. So soll durch Training das Modell einen Zusammenhang zwischen den eingehenden Variablen (Features) und einem Output (Target) in Form einer Klassenzugehörigkeit finden. Durch die Evaluation lässt sich sicherstellen, wie trennscharf das Modell arbeitet. Zudem lässt sich durch eine Evaluation beurteilen, welche Nutzen die erzielten Ergebnisse haben und wie passgenau sie im Hinblick auf die Zielsetzung ausfallen.

Bei anderen Analyseproblemen (z. B. Clustering) lassen sich nur rein fachliche Beurteilungskriterien heranziehen, da es weder gelabelte Trainings- noch Testdaten gibt. Bei gelabelten Daten ist pro Datensatz bekannt, zu welcher Klasse sie zählen Die Phase „Evaluation“ zielt darauf ab, die vom ML-Algorithmus erzeugten Modelle bzw. Ergebnisse auszuwerten. Überprüft werden dabei alle Ausführungsschritte, auch aus ökonomischer Sicht. Die Ergebnisse legen offen, wie hoch der Reifegrad des Modells bzw. ob das Modell für die produktive Inbetriebnahme be-

reits genügend qualitätsgesichert ist. Die Phasen Modeling/Data Mining und Evaluation finden im Wechsel statt und werden meist mehrfach durchlaufen.

In der Phase Deployment wird das erfolgreich evaluierte ML-Modell der Nutzerschaft produktiv bereitgestellt. Die den Evaluationsprozess durchlaufenen Ergebnisse (Modelle, Analysen, Erkenntnisse usw.) werden dabei zielgruppenspezifisch aufbereitet und innerhalb der Organisation an Stakeholder kommuniziert und verteilt. Menschen oder Maschinen zählen dabei gleichermaßen zur „Nutzerschaft". Die Spannweite ist groß: Das ML-Modell kann als Visualisierung (z. B. Entscheidungsbaum) vorliegen und Grundlage für Entscheidungen durch einen Menschen bilden. Oder das ML-Modell wird in ausführbaren Programmcode überführt und ermöglicht dann autonome Entscheidungen durch eine Maschine. Das ML-Modell kann in Form von Programmcode auch als Herzstück einer „smarten" Anwendung, App oder Website fungieren. Ebenso ließe sich das ML-Modell in eine Abfragesprache übersetzen und dann direkt in einer Datenbank ausführen. Deployment markiert den Wechsel in den Verantwortungsbereich des IT-Bereichs. Produktiv gesetzte Modelle werden somit in Workflows oder Wertangebote integriert. Es erschließen sich somit Potenziale zur Verbesserung oder Automatisierung. Entscheidungen werden im Generellen vermehrt datengestützt getroffen. Deployment beendet nicht zwangsläufig ein Analytics-Projekt, sondern legt vielmehr die Grundlagen, laufend die Ergebnisse hinsichtlich Aktualität und Verbesserungspotenzial zu überwachen (Monitoring) oder Pläne zur Wartung (Maintenance) der Modelle aufzustellen. Zur Phase Deployment gehört auch, das Analytics-Projekt aus der Retroperspektive zu beurteilen (Review) und somit Verbesserungspotenziale für künftige Projekte ausfindig zu machen.

Tabelle 1.1 listet typische Kontrollfragen im Zuge der verschiedenen CRISP-DM-Phasen auf.

Tabelle 1.1 Kontrollfragen bei den einzelnen CRISP-DM-Phasen

Business Understanding („Geschäftsverständnis")	▪ Warum ist das Projekt wichtig? ▪ Wen betrifft das Problem? ▪ Was ist zu tun, wenn nicht die richtigen Daten vorliegen? ▪ Wann ist das Projekt zu Ende? ▪ Was tun wir, wenn die Ergebnisse nicht gefallen?
Data Understanding („Datenverständnis")	▪ Wer hat die Daten erhoben und wie wurden sie gewonnen? ▪ Sind die Daten repräsentativ? ▪ Wie wurden Ausreißer oder fehlende Werte behandelt? ▪ Welche Merkmale sind relevant und haben welche Rolle? ▪ Können die Daten abbilden, was zu messen beabsichtigt ist?

Tabelle 1.1 Kontrollfragen bei den einzelnen CRISP-DM-Phasen *(Fortsetzung)*

Data Preparation („Datenvorbereitung“)	▪ Wurden alle relevanten Datenquellen berücksichtigt? ▪ Wurden alle relevanten Daten bereinigt? ▪ Wurden die zur Zielsetzung passenden Daten gewählt? ▪ Wurden die Daten hin auf Anomalien usw. geprüft? ▪ Liegen die Daten in dem benötigten Format vor? ▪ Ist die Datenqualität gut genug für die Modellbildung?
Modeling („Modellierung/Analyse“)	▪ Wurden alle potenziellen ML-Algorithmen bedacht? ▪ Ist das Skalenniveau für den ML-Algorithmus geeignet? ▪ Ist ein Testentwurf generiert worden? ▪ Ist das Modell bewertet worden und ist es verständlich?
Evaluation („Überprüfung“)	▪ Wurden unterschiedliche ML-Algorithmen eingesetzt? ▪ Wurde das Modell anhand von Trainingsdaten erzeugt? ▪ Wurden die Ergebnisse anhand von Testdaten überprüft? ▪ Wurden die Evaluations-/Gütekriterien festgelegt?
Deployment („Inbetriebnahme“)	▪ Wurden Pläne für das Deployment aufgestellt? ▪ Wurde der IT-Bereich rechtzeitig eingebunden? ▪ Gibt es definierte Übergabeprozesse für ML-Modelle? ▪ Gibt es ausreichend technische Ressourcen/ Kapazitäten? ▪ Bewährt sich das Modell angesichts neuer Daten fortlaufend?

1.3.3 ASUM-DM

Das CRISP-DM-Referenzmodell wurde in den 1990er-Jahren entwickelt. Auch wenn CRISP-DM heute noch vergleichsweise weitverbreitet ist (KDnuggets 2014), gibt es neuere Ansätze, die systemische Schwächen von CRISP-DM zu beseitigen versuchen. Ein zentraler Kritikpunkt an CRISP-DM liegt darin, dass erst zu einem sehr späten Zeitpunkt, bei der Phase Deployment, IT-spezifische Belange wie etwa ein reibungsloser IT-Betrieb berücksichtigt werden.

Der CRISP-DM-Prozess zielt sehr isoliert auf Data Mining ab. Die dafür typischen Vorarbeiten wie Data Understanding oder Data Preparation sind wenig mit anderen betrieblichen Prozessen, vor allem mit denen aus dem IT-Bereich und IT-Betrieb, synchronisiert und abgestimmt. Analytics-Projekte haben heute einen weitaus stärkeren Bezug zum IT-Betrieb, als dies seinerzeit bei der Konzeption von CRISP-DM der Fall war. Ansätze wie etwa Agile Software Development haben die Softwareentwicklung geändert. Dies erfordert einen hohen Integrationsgrad bei Prozessen, Projektmanagement, Tools sowie umfangreiche Kommunikation. Eher isolierte Modelle wie CRISP-DM benötigen daher Schnittstellen zu anderen Prozessen, die aber wiederum unternehmensspezifisch sind.