

HANSER



Leseprobe

zu

Datenkompetenz

von Michael Lang (Hrsg.)

Print-ISBN: 978-3-446-47585-4

E-Book-ISBN: 978-3-446-47743-8

Weitere Informationen und Bestellungen unter

<https://www.hanser-kundencenter.de/fachbuch/artikel/9783446475854>

sowie im Buchhandel

© Carl Hanser Verlag, München

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

Vorwort	V
1 Datenkompetenz – Grundlagen	1
<i>Robert Butscher</i>	
1.1 Der Begriff Datenkompetenz (Data Literacy)	1
1.2 Definitionen	3
1.3 Vorgehensmodelle	15
1.4 Berufsfelder	25
1.5 Die wichtigsten Punkte in Kürze	32
2 Datenmodellierung	39
<i>Andreas Gadatsch und Benedikt Haag</i>	
2.1 Modelle und Datenmodelle	40
2.2 Wissenspyramide: Daten, Informationen und Wissen	41
2.3 Kategorien von Daten	43
2.4 Fehler in Daten	45
2.5 Zweck und Nutzen von Datenmodellen	48
2.6 Entwurf von Datenbanken	49
2.7 Einführung in das Entity-Relationship-Modell (ERM)	51
2.8 Erweiterungen des Entity-Relationship-Modells	56
2.9 Alternativen zur Chen-Notation	60
2.10 Die wichtigsten Punkte in Kürze	61

3	Daten sammeln, aufbereiten und speichern	63
	<i>Beate Navarro Bullock</i>	
3.1	Von der Quelle zum aufbereiteten Datensatz	63
3.2	Daten sammeln	66
3.3	Daten aufbereiten	73
3.4	Daten speichern	78
3.5	Die wichtigsten Punkte in Kürze	83
4	Datenanalyse – Einführung, deskriptive und diagnostische Analyse	85
	<i>Oliver Schwarz</i>	
4.1	Übersicht zu den Analyseformen	85
4.2	Analyseformen und Analysemethoden	87
4.3	Deskriptive Analyse	88
4.4	Diagnostische Methoden	95
4.5	Die wichtigsten Punkte in Kürze	114
5	Datenanalyse – prädiktive und präskriptive Analyse	117
	<i>Oliver Schwarz</i>	
5.1	Maschinelles Lernen – eine Übersicht	117
5.2	Klassifikation und Regression	120
5.3	Trainings- und Testdaten	122
5.4	Lineare Regressionsanalyse	122
5.5	Logistische Regression	129
5.6	Klassifikationsbäume	136
5.7	Präskriptive Analyse – ein Beispiel	142
5.8	Die wichtigsten Punkte in Kürze	144
6	Datenvisualisierung – die relevanten Daten vor Augen	147
	<i>Roland Zimmermann</i>	
6.1	Können wir unseren Augen trauen?	147
6.2	Analytische Aufgaben in visuelle Abfragen übersetzen	151
6.3	Drei-Stufen-Modell für effiziente visuelle Suchen	154
6.4	VS1 – quasi-unbewusste Wahrnehmung maximieren	157

6.5	VS2 – Mustererkennung optimieren, Gestaltungsoptionen frei halten . . .	164
6.6	VS3 – Wahrnehmung durch Planung antizipieren	169
6.7	Vorgehensmodell zur Datenvisualisierung	179
6.8	Die wichtigsten Punkte in Kürze	181

7 Data Governance 183

Kristin Weber und Christiana Klingenberg

7.1	Data Governance: Einführung	183
7.2	Empfehlungen für Data Governance	185
7.3	Das qualitätsorientierte Data Governance Framework	186
7.4	Handlungsfeld der strategischen Ebene	189
7.5	Handlungsfelder der organisatorischen Ebene	191
7.6	Handlungsfelder auf Ebene der Informationssysteme	197
7.7	Relevanz von Datenqualität über alle Ebenen des Frameworks	203
7.8	Die wichtigsten Punkte in Kürze	204

8 Datenqualität 207

Christiana Klingenberg und Kristin Weber

8.1	Probleme mit Datenqualität	207
8.2	Begriff Datenqualität – fit for use	209
8.3	Dimensionen der Datenqualität	211
8.4	Datenqualitätsregeln	213
8.5	Messen der Datenqualität	215
8.6	Bewerten der Datenqualität	217
8.7	Herausforderung Datenqualität bei der Auswertung von Daten	219
8.8	Herausforderung Datenqualität in überbetrieblichen Prozessen	220
8.9	Kosten schlechter Datenqualität	222
8.10	Die wichtigsten Punkte in Kürze	224

9 Datenschutz und Datensicherheit 227

Stefan Karg

9.1	Grundlagen und Begriffe	227
9.2	Informationssicherheit	229
9.3	Datenschutz	234

9.4	Methoden	239
9.5	Der Konvergenzbereich: TOM	249
9.6	Herausforderungen in der Praxis	255
9.7	Die wichtigsten Punkte in Kürze	257
10	Big Data und Big Data Analytics	261
	<i>Oliver Hummel</i>	
10.1	Big Data, worum geht es?	261
10.2	Big Data Analytics	264
10.3	Speicherung großer Datenmengen	270
10.4	Verarbeitung großer Datenmengen	284
10.5	Big-Data-Referenzarchitekturen	289
10.6	Resilienz in Big-Data-Systemen	291
10.7	Probabilistische Datenstrukturen in Big-Data-Systemen	294
10.8	Die wichtigsten Punkte in Kürze	296
11	Datenkompetenz: Warum es ohne Soft Skills nicht geht	299
	<i>Benedikt Haag und Andreas Gadatsch</i>	
11.1	Die Unterscheidung zwischen Soft und Hard Skill	299
11.2	Soft-Skill-Kategorien	301
11.3	Die Bedeutung von Soft Skills in der Arbeitswelt	304
11.4	Soft Skills für Datenkompetenz	306
11.5	Messung von Soft Skills	308
11.6	Ansätze zur Entwicklung von Soft Skills	309
11.7	Die wichtigsten Punkte in Kürze	311
11.8	Literatur	312
	Der Herausgeber	313
	Die Autor:innen	315
	Index	319

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 sachgerecht 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.

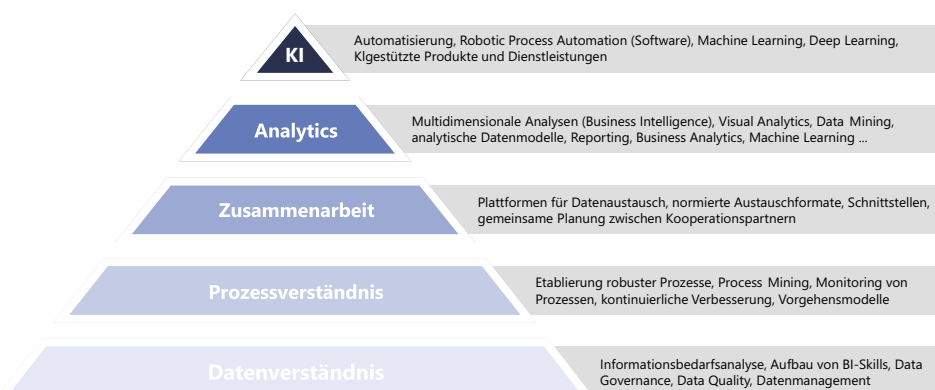


Bild 1.2 Stufen auf dem Weg hin zu einer datengetriebenen Unternehmenskultur

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 bis 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 dazugehörigen 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.

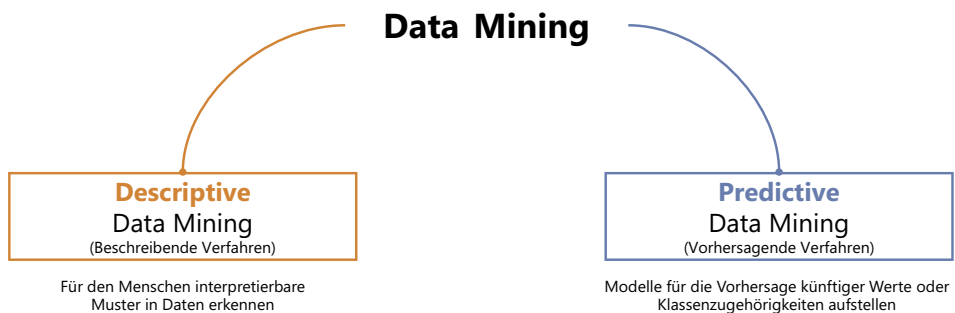


Bild 1.3 Kategorien von Data Mining

- **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.

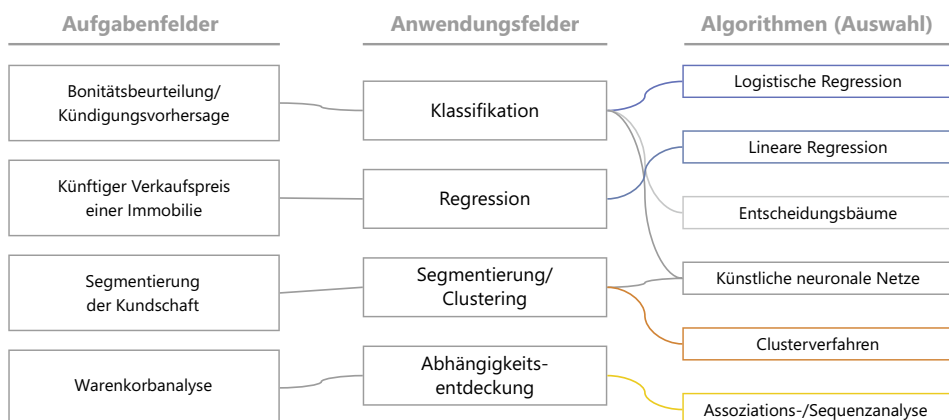


Bild 1.4 Aufgabenfelder, Anwendungsfelder und Algorithmen von Data Mining (Auswahl)

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 **computer-gestützten 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.

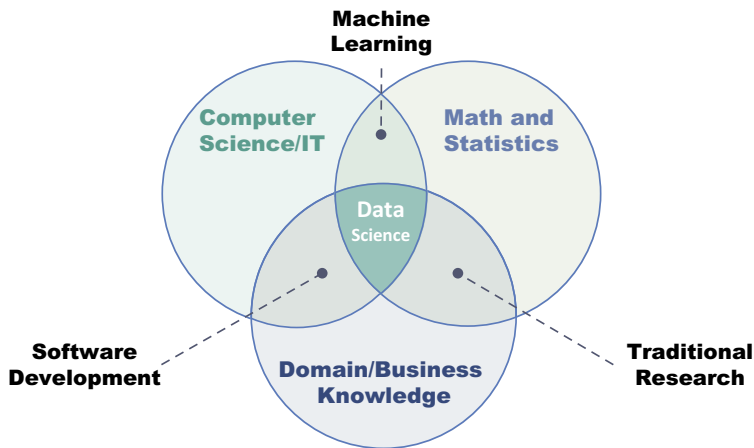


Bild 1.5 Data Science mit Berührungspunkten zu anderen Disziplinen (Luellen 2018)

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.“

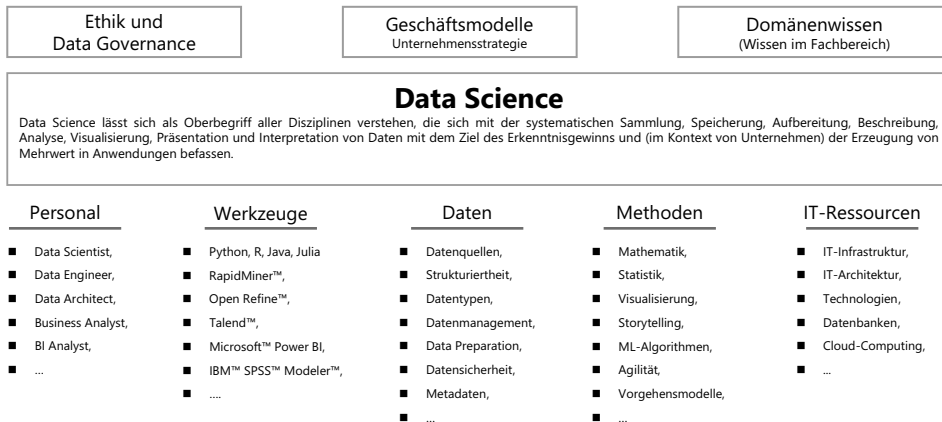


Bild 1.6 Facetten rund um Data Science

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 [...].“

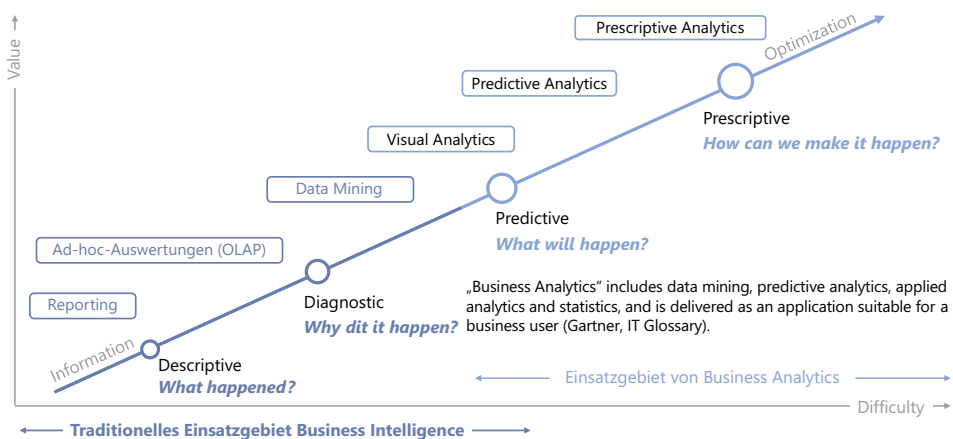


Bild 1.7 Facetten von Business Analytics (in Anlehnung an Seiter 2019)

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).

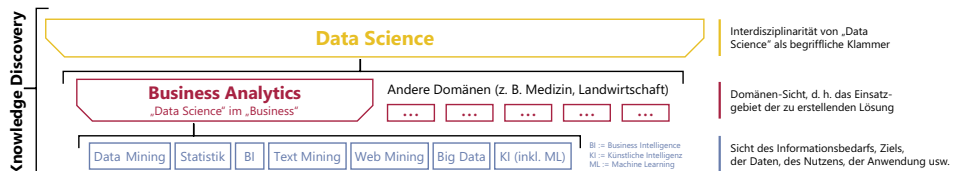


Bild 1.8 Zusammenfassung und Einordnung von Business Analytics

■ 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.

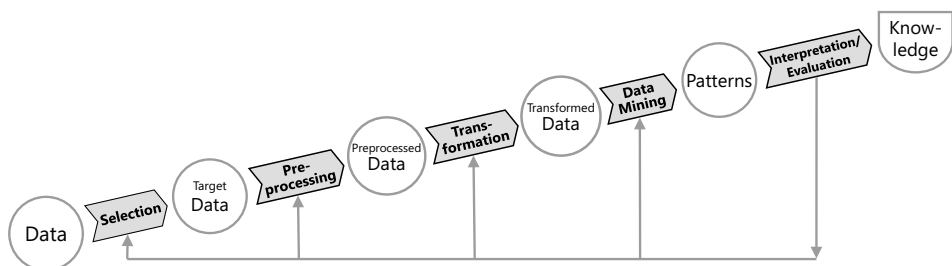


Bild 1.9 Knowledge Discovery in Databases (KDD) (Fayyad/Piatetsky-Shapiro/Smyth 1996)

Index

A

Abfrage, visuelle 151, 170
ACID 271
Alltagskompetenzen 304
Analyse 85
–, deskriptive 86, 88
–, diagnostische 86
–, prädiktive 86, 117
–, präskriptive 86, 117, 142
Analyseformen 87
Anonymisierung 243, 244
Apache Avro 71
Apache Kafka 72
ASUM-DM 22
Aufmerksamkeit lenken 150, 173, 177

B

BASE 272
Batch-Verarbeitung 71, 287
Berechtigungskonzept 250, 251
Berufsfelder 25
Bewegung 163
BI Analyst 29
Big Data
– Referenzarchitekturen 289
– Resilienz 291
– Speicherung 270
– Systemen 294
– Verarbeitung 284
– Verarbeitungsparadigmen 286
Big Data Analytics 10, 264
Business Analyst 30

Business Analytics 13
Business Intelligence (BI) 3

C

CAP-Theorem 272
CART-Algorithmus 136, 140
Chartjunk 167
Clusterverfahren
–, hierarchische 105
–, zentroide 109
Confusion Matrix 134
Controlling 191
CRISP-DM 17
– Business Understanding 18
– Data Preparation 19
– Data Understanding 18
– Deployment 21
– Evaluation 20
– Modeling 20
CSV 69

D

Darstellung, unverzerrte 163
Data Analyst 29
Data Architect 29
Data at rest 251
Data Engineer 28, 65
Data Governance
– Einführung 183
– Empfehlungen 185
– Handlungsfelder 186, 187, 188, 203

- Leitlinien 189
- Rollenmodell 192
- Data Governance Framework
 - , qualitätsorientiertes 186
- Data-Ink-Ratio 167
- Data in transit 251
- Data Lake 81
- Data Lakehouse 82
- Data Literacy 1
- Data Mining 7
- Data Pipeline 64
- Data Science 11
- Data Scientist 25
- Data Steward 31
- Data Warehouse 80
- Dateiformat 68
- Dateispeicher 79
- Daten
 - aufbereiten 63, 73
 - Ausreißer 75
 - bereinigen 73, 77
 - besondere Arten 235
 - Durchsatz 73
 - , externe 67
 - Extremwert 76
 - Fehler 45
 - integrieren 76, 77
 - Kategorien 43
 - Konvertierung 75
 - , personenbezogene 222, 228, 234, 242
 - Quellsysteme 67
 - sammeln 66
 - Schema 77
 - Schutzbedarf 239
 - , semistrukturierte 69, 265
 - Skalierbarkeit 73
 - speichern 78
 - , strukturierte 69, 265
 - transformieren 76, 77
 - , unstrukturierte 69, 265
 - zur Verfügung stellen 78
- Datenbank
 - , dokumentenbasierte 280
- Entwurf 49
- Graf- 283
- Key-Value- 279
- NoSQL- 274, 278
- , relationale 271, 276
- Datenevaluation 307
- Datenkompetenz 1
 - Definition 1
 - Soft Skills 299, 306
- Datenmanagement
 - Abteilung 195
 - Controlling 191
 - Organisation 192
 - Prozesse 195
- Datenmodelle 40
 - Zweck/Nutzen 48
- Datenmodellierung 39
- Datenorganisation 80
- Datenprodukt 78
- Datenqualität 207
 - Anforderungen 192
 - bewerten 217
 - Datenqualitätsdimensionen“ 189
 - Dimensionen 211
 - , hohe 210
 - Kosten 222
 - messen 191, 203, 213, 215
 - , niedrige 211, 223
 - Probleme 207
 - Quality Gates 191, 195, 203
 - Schwellenwert 216
- Datenqualitätsdimension
 - Glaubwürdigkeit 214
 - Gültigkeit 212
 - Korrektheit 212, 214
 - Vollständigkeit 211, 212, 214
- Datenqualitätsregeln 213
- Datenschutz 227, 234
 - besondere Arten pbDaten 235, 237
 - Grundsätze 236
 - Persönlichkeitsrecht 234
 - privacy by default 238
 - privacy by design 238
 - Sicherheit der Verarbeitung 236
 - Verantwortlicher 237

Datenschutzbeauftragter (DSB) 237, 255
– Aufgaben 237
– Einbindung 237, 256
Datensicherheit 228
Denken, kritisches 307
Denormalisierung 275
Dominanz, visuelle 165
Drei-Stufen-Modell (visuelle Suche) 154
Duplikat 74, 197, 215, 223

E

Entity-Relationship-Modell (ERM)
– Alternativen zur Chen-Notation 60
– Einführung 51
– Erweiterungen 56
Eventual Consistency 271

F

Flächen schätzen 148

G

Graphdatenbanken 283
Grundformen, visuelle 159, 161

H

Hadoop 285
Hard Skills 299
Hashing 248

I

Informationssicherheit 228, 229
– Beauftragter (ISB) 255
– Schutzziele 236, 249
Informationssicherheits-
managementsystem 232
ISMS 232, 233, 239
– ISO 27001 233, 234, 254
– IT-Grundschutz 233, 251

J

JSON 70

K

Kanäle, visuelle 158
Kappa-Architektur 290
Kernkompetenzen 302
Key-Value-Datenbank 279
Klassifikation 120
Klassifikationsbaum
– Interpretation 142
– Modellerstellung/-auswahl 138
Knowledge Discovery in Databases (KDD)
15
Kommunikation 307
Kompetenzen
– Alltags- 304
– Kern- 302
–, kognitive 302
–, körperliche/motorische 303
– Selbstmanagement- 303
–, soziale/kommunikative 303

L

Lambda-Architektur 289
Lernen
–, maschinelles 117
–, überwachtes 118
–, unüberwachtes 119
Lernfähigkeit 308

M

Merkmale
–, metrisch skalierte 89
–, nominalskalierte 88
–, ordinalskalierte 88
Microbatch 72
Micro-Batch-Verarbeitung 288
Modelle 40
Mustererkennung 148, 156, 164

N

Normalisierung 76
 Normen 232
 NoSQL-Datenbank 274, 278

O

Objektspeicher 79

P

Parquet 70
 Problemlösung 307
 Pseudonymisierung 242, 244

R

Regression 120
 –, logistische 129
 Regressionsanalyse
 – Modellerstellung 122
 – Modellinterpretation 125
 – Modellprüfung 126
 – Modellschätzung 123
 Risiko 239
 – Betroffener 228, 237
 – Ermittlung 240
 – Minderung 241, 242
 – risikobasierter Ansatz 239
 – Überprüfung 254
 – Übertragung 241
 – Unternehmen 228
 – Verarbeitung 254
 – Vermeidung 241
 Risikoanalyse 233, 238, 239
 Risikobedarfsanalyse 239
 Root Mean Square Error (RMSE) 128

S

Schutzbedarfsanalyse 239
 Schutzziele 229, 249
 – Authentizität 231, 247, 250
 – Integrität 230, 236, 248, 251, 253

– Verbindlichkeit 231
 – Verfügbarkeit 231, 232, 236, 253
 – Vertraulichkeit 230, 236, 246, 247, 252
 Selbstmanagementkompetenzen 303
 Sequenzen, analytische 173
 Skalierung
 –, horizontale 262
 –, vertikale 262
 Soft Skills 299
 – Bedeutung in der Arbeitswelt 304
 – Entwicklungsansätze 309
 – für Datenkompetenz 299, 306
 – Kategorien 301
 – Messung von 308
 Speicherbegrenzung 236
 Stammdaten
 – Material 221
 – Produkt 221
 Standardisierung 76
 Stapelverarbeitung 71
 Stream-Verarbeitung 72, 288
 Structured Query Language 271
 Suchmaschine 280

T

technische und organisatorische
 Maßnahmen 236
 Testdaten 128, 134
 TOM 236, 249
 – Integrität 253
 – Protokollierung 230, 253
 – Transportkontrolle 251
 – Trennung nach Zwecken 249
 – Verfügbarkeit 253
 – Vertraulichkeit 251, 252
 – Weitergabekontrolle 253
 – Zugangskontrolle 250
 – Zugriffskontrolle 250
 – Zutrittskontrolle 250

U

Useful Field of View (UFOV) 161

V

Verschlüsselung 243, 245, 251, 252

Visualisierung 147

– Vorgehensmodell 179

–, wahrnehmungsgerechte 179

W

Wahrnehmung

–, bewusste 156

– durch Planung antizipieren 169

–, präattentive 156

–, quasi-unbewusste 157

Wide-Column Store 282

Wissenspyramide 41

Z

Zweckbindung 236