Data Science und Business Analytics

Prof. Dr. Christian Bockermann

24. Juni 2026

Inhaltsverzeichnis

Einleitung

Wer heute ein BWL-Studium aufnimmt, wählt ein Fach, das sich gerade grundlegend wandelt. Nicht weil sich die Kernfragen der Betriebswirtschaft verändert hätten – wie werden knappe Ressourcen effizient eingesetzt, wie entstehen nachhaltige Wettbewerbsvorteile, wie werden gute Entscheidungen unter Unsicherheit getroffen? – sondern weil sich die Werkzeuge, mit denen diese Fragen beantwortet werden, in einem Tempo entwickeln, das vor wenigen Jahren noch kaum vorstellbar war. Im Mittelpunkt dieses Wandels stehen Daten.

Daten sind heute allgegenwärtig. Jede Bestellung in einem Online-Shop, jede Produktionsunterbrechung in einer Fertigungsanlage, jeder Klick auf einer Website hinterlässt digitale Spuren. Unternehmen, die diese Spuren systematisch lesen und verstehen, gewinnen Einblicke, die früher unsichtbar blieben: Welche Kundensegmente sind besonders profitabel? Wo im Lieferprozess entstehen Engpässe? Welche Marketingmaßnahme hat tatsächlich Wirkung gezeigt – und welche nur zufällig mit einem Umsatzanstieg zusammengefallen? Die Fähigkeit, solche Fragen datengestützt zu beantworten, ist keine Spezialität mehr, die man Statistikerinnen und Data Scientists überlassen kann. Sie ist zur Kernkompetenz moderner Unternehmensführung geworden.

Dieses Skript entstand aus der Überzeugung, dass *Data Literacy* – die Fähigkeit, mit Daten souverän umzugehen, sie zu interpretieren und kritisch einzuordnen – für Absolventinnen und Absolventen eines BWL-Studiums heute ebenso unverzichtbar ist wie das Lesen einer Bilanz oder das Verstehen von Finanzierungsstrukturen. Dabei geht es nicht darum, Algorithmen zu programmieren oder statistische Verfahren herzuleiten. Es geht darum, zu verstehen, *was* Daten aussagen können – und was nicht. Es geht darum, die richtigen Fragen zu stellen, Ergebnisse im

betriebswirtschaftlichen Kontext einzuordnen und auf dieser Grundlage tragfähige Entscheidungen zu treffen.

Diese Anforderung hat eine neue Dringlichkeit erhalten, seitdem generative Künstliche-Intelligenz-Systeme wie große Sprachmodelle in den Arbeitsalltag von Unternehmen eingezogen sind. Aufgaben, die früher spezialisiertes technisches Wissen erforderten – das Schreiben von Datenbankabfragen, die Erstellung von Analyseauswertungen, das Erzeugen von Visualisierungen – lassen sich heute in Teilen durch die Eingabe einer Textanfrage erledigen. Ein Modell generiert Code, erklärt Konzepte, fasst Dokumente zusammen. Die technische *Einstiegshürde* für viele Analyseaufgaben ist damit deutlich gesunken.

Das klingt nach einer Entlastung – und das ist es in gewisser Weise auch. Wer heute im Controlling eine Auswertung benötigt, muss dafür keine wochenlange Projektanfrage an die IT stellen. Wer verstehen möchte, warum ein Modell eine bestimmte Empfehlung ausspricht, kann sich die Erklärung auf Knopfdruck generieren lassen. Doch genau hier liegt ein Missverständnis, das sich in vielen Unternehmen hartnäckig hält: Die Reduzierung technischer Hürden verändert nicht, was Daten und KI-Systeme *nicht* können. Sie erzeugen Wahrscheinliches, kein Wahres. Sie reproduzieren Muster aus der Vergangenheit, keine Garantien für die Zukunft. Und sie beantworten die Fragen, die gestellt werden – nicht die Fragen, die gestellt werden *sollten*.

Die Einordnung von KI-Ergebnissen in den konkreten betriebswirtschaftlichen Kontext bleibt deshalb genuiner menschlicher Beitrag. Ein Klassifikationsmodell, das das Retourenrisiko einer Bestellung einschätzt, liefert eine Wahrscheinlichkeit. Ob diese Wahrscheinlichkeit eine präventive Maßnahme rechtfertigt, welche Kosten ein Fehlalarm erzeugt und ob das Modell bestimmte Kundengruppen systematisch benachteiligt – das sind Fragen, die kein Modell für sich selbst beantworten kann. Es sind unternehmerische Urteile, die betriebswirtschaftliches Verständnis, Kontextwissen und Verantwortungsbewusstsein erfordern.

Dabei ist eine weitere Dimension in den letzten Jahren sichtbar geworden, die lange unterschätzt wurde: der *Ressourcenverbrauch* moderner KI-Systeme. Das Training und der Betrieb großer Sprachmodelle ist mit erheblichem Aufwand an

Rechenkapazität, Energie und damit auch Kosten verbunden. Dieser Aufwand fällt nicht nur beim Anbieter an – er schlägt sich direkt im Unternehmensbudget nieder, sobald Mitarbeiterinnen und Mitarbeiter KI-Dienste in ihre tägliche Arbeit integrieren.

Ein jüngst bekanntgewordenes Beispiel verdeutlicht das Ausmaß dieser Herausforderung eindrücklich: Ein Unternehmen berichtete, dass unkontrollierte und ziellose KI-Nutzung durch Mitarbeitende zu Tokenverbrauchskosten in dreistelliger Millionenhöhe geführt hatte – nicht durch einen einzelnen Fehler, sondern durch die systematische Abwesenheit einer klaren Strategie. Ohne definierte Anwendungsfälle, ohne Kosten-Nutzen-Abwägung, ohne Verantwortlichkeiten wird KI-Nutzung zur unkontrollierten Betriebsausgabe. Was als Investition in Effizienz und Innovation gedacht war, verkehrt sich zu einem schlecht steuerbaren Kostentreiber.

Dies ist keine Warnung gegen den Einsatz von KI – im Gegenteil. Es ist ein Argument dafür, KI dort einzusetzen, wo ihr Einsatz klar definierte betriebswirtschaftliche Ziele verfolgt, wo Nutzen messbar und Aufwand kalkulierbar ist. Genau das aber setzt voraus, dass diejenigen, die KI-Projekte verantworten, eine Vorstellung davon haben, was diese Systeme tun, was sie kosten und wie ihre Ergebnisse zu bewerten sind. Es setzt, kurz gesagt, genau die Kompetenzen voraus, die dieses Skript zu vermitteln versucht.

Das vorliegende Skript ist entlang eines konkreten Unternehmens organisiert: der fiktiven *Brick-Flow AG*, die bestückte Baukasten-Paletten produziert, über einen eigenen Online-Shop vertreibt und Lagerung sowie Versand an einen externen Logistik-Dienstleister auslagert. Dieses Szenario ist bewusst einfach gehalten – es geht nicht um die Besonderheiten einer bestimmten Branche, sondern um Muster, die sich in nahezu jedem Unternehmen wiederfinden: Daten entstehen entlang von Prozessen; sie müssen integriert, bereinigt und modelliert werden, bevor sie entscheidungsrelevant sind; und die Verbindung zwischen einer Analyse und einer konkreten Handlung ist das eigentlich schwierige Problem.

Die sieben Kapitel bauen aufeinander auf. Sie beginnen mit der strategischen Einordnung von Daten als Ressource und Wertschöpfungsquelle, behandeln die technischen und organisatorischen Grundlagen des Datenmanagements, führen in die kritische Interpretation statistischer Analyseergebnisse ein und erarbeiten das

konzeptionelle Verständnis von Machine Learning und Data Science. Datenvisualisierung, Business Intelligence und die strukturierte Entwicklung datengetriebener Use Cases schließen den methodischen Bogen. Den Abschluss bildet eine Auseinandersetzung mit Datenschutz, Ethik und Fairness – Themen, die im KI-Zeitalter nicht mehr am Rand stehen, sondern ins Zentrum unternehmerischer Verantwortung gerückt sind.

Dieses Dokument richtet sich an Sie als BWL-Studierende ohne vertieften technischen Hintergrund. Es werden keine Programmierkenntnisse vorausgesetzt, keine mathematischen Formeln hergeleitet, die über das Notwendige hinausgehen. Was es voraussetzt, ist die Bereitschaft, sich auf ein Denken einzulassen, das systematisch, neugierig und kritisch zugleich ist. Die Fähigkeit, eine Zahl zu hinterfragen, eine Modellaussage einzuordnen und den Unterschied zwischen Korrelation und Ursache zu erkennen, ist keine technische Fertigkeit. Sie ist eine intellektuelle Haltung. Und sie ist, darin liegt die eigentliche Überzeugung hinter diesem Skript, erlernbar.

Bochum, 2025 Prof. Dr. Christian Bockermann

Kapitel 1

Daten als strategischer Erfolgsfaktor in der digitalen Wirtschaft

Die zunehmende Digitalisierung hat dazu geführt, dass Daten in nahezu allen Unternehmensbereichen verfügbar sind. Prozesse, Kundeninteraktionen und betriebliche Abläufe hinterlassen heute digitale Spuren, die systematisch erfasst und ausgewertet werden können. Dieser Fokus auf Daten bezieht sich nicht nur auf viele Unternehmensbereiche – es gibt auch kaum eine Branche, die von datenorientierten Strategien nicht profitiert oder gar ohne Daten nicht mehr funktioniert (vgl. Abbildung 1.1).

Damit verändern sich nicht nur operative Abläufe, sondern auch die Grundlagen unternehmerischer Entscheidungen. Während in der Vergangenheit Entscheidungen häufig auf Erfahrung und Intuition basierten, stehen heute umfangreiche Daten als zusätzliche Informationsquelle zur Verfügung. Dies eröffnet neue Möglichkeiten, erhöht jedoch gleichzeitig die Anforderungen an Unternehmen, diese Daten sinnvoll zu nutzen.

Daten werden damit zu einem zentralen Bestandteil moderner Unternehmensführung. Dies hat als Konsequenz jedoch auch Auswirkungen auf die Kompetenzen bzw. Anforderungen an Manager und Entscheider: Der Umgang mit Daten und die Bewertung von Daten sind *Future Skills*, die nicht mehr nur im Bereich Technik, sondern auch im Bereich Management/Unternehmensführung essentiell sind.

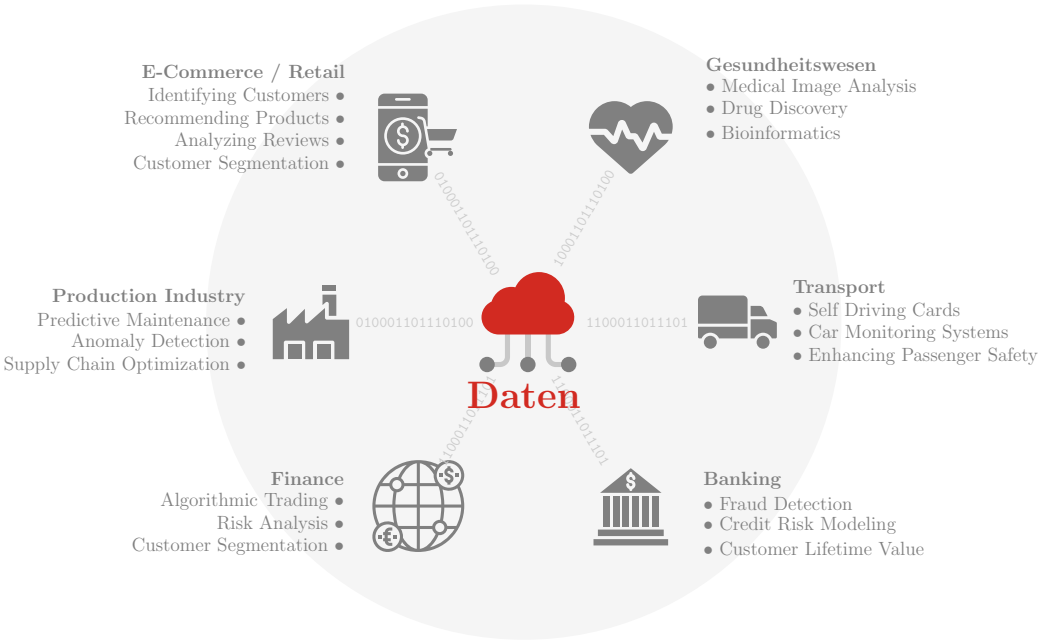


Abbildung 1.1: Daten sind zur Grundlage von Entscheidungen und Geschäftsmodellen in nahezu allen Branchen geworden.

1.1 Strategische Bedeutung von Daten

Aus Managementsicht stellt sich zunächst die Frage, warum Daten überhaupt eine strategische Ressource darstellen. Im Unterschied zu klassischen Produktionsfaktoren wie Arbeit oder Kapital zeichnen sich Daten dadurch aus, dass sie beliebig oft genutzt und kombiniert werden können. Ihr Wert entsteht nicht durch Besitz allein, sondern durch ihre Analyse und Anwendung in konkreten Entscheidungssituationen.

Unternehmen nutzen Daten beispielsweise, um:

- Prozesse effizienter zu gestalten,
- Kundenbedürfnisse besser zu verstehen,
- fundiertere Entscheidungen zu treffen.

Der strategische Nutzen liegt somit vor allem darin, dass Daten helfen, Unsicherheiten zu reduzieren und Handlungsoptionen besser zu bewerten. Gleichzeitig können Unternehmen durch systematische Datennutzung Wettbewerbsvorteile aufbauen, etwa durch schnellere Reaktionen auf Marktveränderungen oder durch individuell zugeschnittene Angebote.

Daten sind kein Selbstzweck: Ihr Wert entsteht erst, wenn sie systematisch nutzbar gemacht und angewendet werden – sei es zur Verbesserung eigener Entscheidungen und Prozesse oder als eigenständiges datenbasiertes Angebot, dessen Nutzen sich beim Käufer entfaltet.

1.2 Datengetriebene Geschäftsmodelle

Neben der Verbesserung bestehender Prozesse ermöglichen Daten auch neue Formen der Wertschöpfung. In datengetriebenen Geschäftsmodellen stehen Daten nicht nur unterstützend im Hintergrund, sondern sind ein zentraler Bestandteil des Angebots.

Typische Ausprägungen sind:

- Plattformen, die unterschiedliche Nutzergruppen zusammenbringen und deren Interaktionen durch Daten steuern,
- datenbasierte Dienstleistungen, etwa personalisierte Angebote oder Prognosesysteme,
- Geschäftsmodelle, die auf der Analyse und Nutzung großer Datenmengen basieren.

Ein wesentlicher Punkt ist dabei, dass diese Geschäftsmodelle häufig auf bereits vorhandenen Daten aufbauen. Unternehmen entwickeln also neue Leistungen aus bestehenden Informationen, die ursprünglich in operativen Prozessen entstanden sind. Damit wird deutlich, dass Daten nicht isoliert betrachtet werden können, sondern immer im Zusammenhang mit Geschäftsmodell, Prozessen und Kundennutzen stehen.

Beispiel Brick-Flow AG: Die in diesem Lehrbrief durchgängig betrachtete Brick-Flow AG – ein produzierendes Unternehmen, das bestückte Paletten über einen eigenen Online-Shop vertreibt und am Ende dieses Kapitels ausführlich vorgestellt wird – erzeugt im laufenden Betrieb umfangreiche Daten. Über die Bestellungen im Online-Shop fallen Transaktionsdaten an; über Kundenkonten und ein Kundenbindungsprogramm werden zusätzlich erweiterte Kundendaten erfasst (z. B. demografische Daten, Produktpräferenzen). Durch die Analyse dieser Daten können beispielsweise

- Marketingkampagnen gezielter gesteuert,
- Absatzprognosen erstellt und
- Logistikprozesse verbessert werden (etwa durch genauere Personalplanung).

Gleichzeitig stellen die Abverkaufsdaten für die Baustein-Hersteller und Lieferanten einen erheblichen Wert dar, weil sie Einblicke in Trends, Zielgruppen und Preisgestaltung geben. Für die Brick-Flow AG bietet sich der Verkauf dieser – konsequent *anonymisierten* – Daten als zusätzliche Wertschöpfung in ihrem Geschäftsmodell an. Welche Anforderungen eine solche Anonymisierung erfüllen muss, behandelt das spätere Kapitel zu Datenschutz und Ethik.

> Datengetriebene Geschäftsmodelle bauen häufig auf Daten auf, die ohnehin in operativen Prozessen entstehen – der Wert liegt darin, diese Daten gezielt in neue Leistungen zu überführen.

1.3 Einfluss moderner Analysetechnologien

Die strategische Bedeutung von Daten wird maßgeblich durch Fortschritte in der Datenanalyse verstärkt. Verfahren der Künstlichen Intelligenz und moderne Analysetools ermöglichen es, große Datenmengen effizient auszuwerten und komplexe Zusammenhänge zu erkennen.

Für Unternehmen ergeben sich daraus zwei zentrale Einsatzbereiche. Zum einen dienen Analysen der Unterstützung von Entscheidungen, etwa durch Prognosen oder Mustererkennung. Zum anderen ermöglichen sie die Automatisierung von Prozessen, beispielsweise in der Produktion oder in der Logistik.

Der entscheidende Wettbewerbsvorteil entsteht jedoch nicht durch die Technologie allein, sondern durch deren gezielte Anwendung. Unternehmen müssen in der Lage sein, Daten, Analyseverfahren und organisatorische Prozesse sinnvoll miteinander zu verbinden. Vor diesem Hintergrund lässt sich das Feld *Data Science* als Schnittmenge von Methoden aus der Informatik und der Mathematik/Statistik *mit einem konkreten Anwendungsgebiet* definieren (vgl. Abbildung 1.2).

Dies verdeutlicht insbesondere, dass Technologie erst dann *strategisch relevant* wird, wenn sie in konkrete betriebliche Entscheidungen und Abläufe integriert wird.

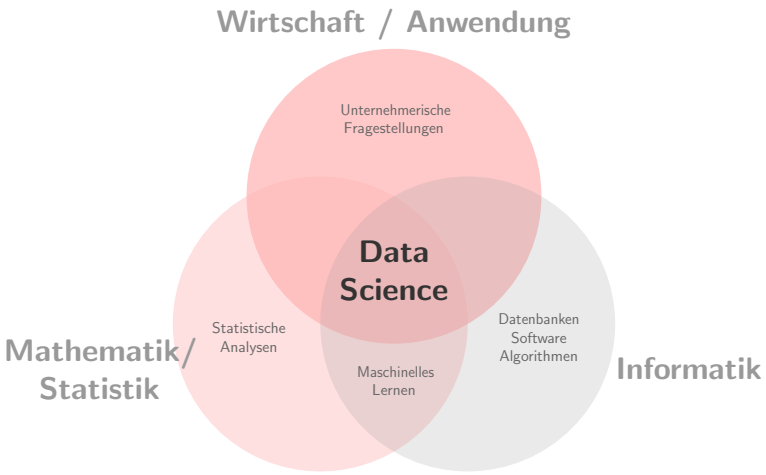


Abbildung 1.2: Data Science als Einsatz von Methoden aus der Mathematik/Statistik und Informatik im Kontext einer konkreten Anwendung.

Weiterführende Literatur

- *Competing on Analytics: The New Science of Winning* [5]: Standardwerk zur strategischen Nutzung von Analytics im Unternehmen.
- *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You* [19]: Grundlage zum Verständnis digitaler Plattformen und datengetriebener Ökosysteme.
- *Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things* [16]: Praxisorientierter Einstieg in Datenstrategie und datenbasierte Wertschöpfung.

1.4 Use Case: Die Brick-Flow AG

Als durchgängiges Beispiel zur Veranschaulichung dient in diesem Lehrbrief die *Brick-Flow AG*. Dabei handelt es sich um ein fiktives, produzierendes Unternehmen, dass aus Rohstoffen in einer Produktionslinie bestückte Paletten fertigt. Die fertigen Paletten werden vom Unternehmen über einen eigenen Online-Shop vertrieben, der Versand/Lagerung erfolgt über einen externen Logistik-Dienstleister. Abbildung 1.3 zeigt den Aufbau der Brick-Flow AG mit Online-Shop sowie der Produktionslinie.

Organisationsstruktur

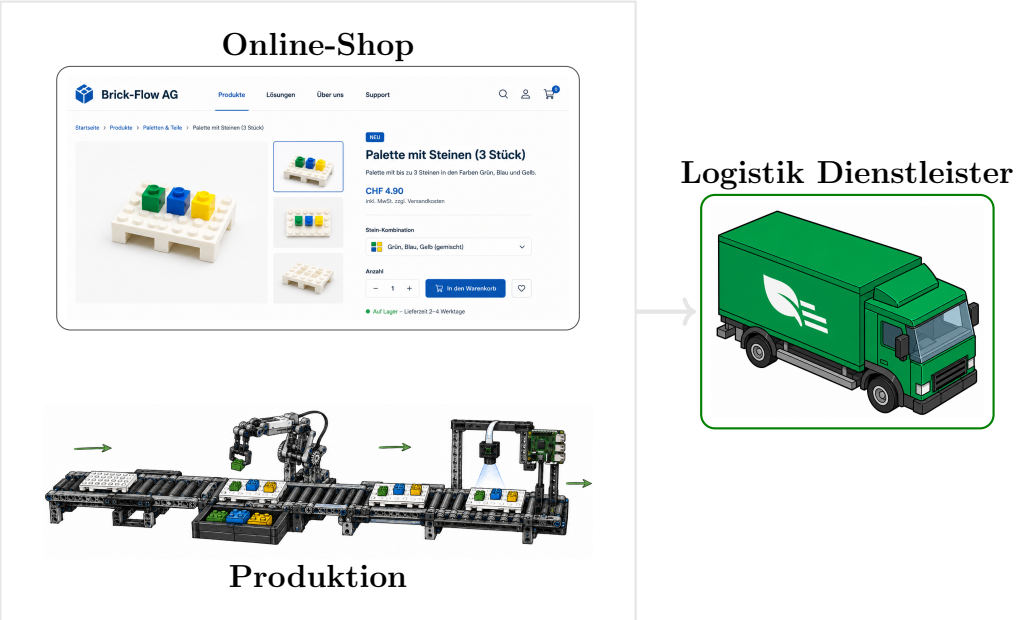


Abbildung 1.3: Produktionslinie und Online-Shop der Brick-Flow AG, sowie externer Logistik-Dienstleister.

Innerhalb der Brick-Flow AG gibt es verschiedene Abteilungen, wie z. B. dem Vertrieb/Online-Shop, Marketing, Produktion und Controlling, die sich um das operative Geschäft der AG kümmern. Der Logistik-Dienstleister übernimmt sowohl die Lagerhaltung der fertigen Produkte als auch deren Versand nach Bestellung an die Endkunden. Die Abbildung 1.4 zeigt die Organisation der Brick-Flow AG mit ihren verschiedenen Abteilungen.

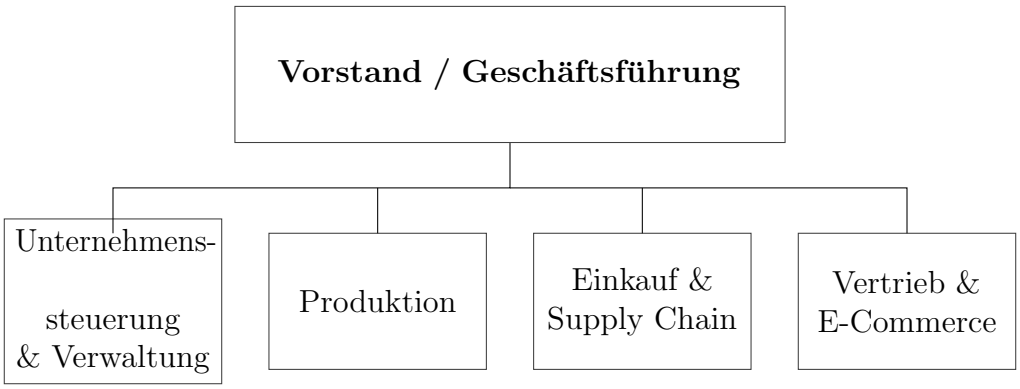


Abbildung 1.4: Die Abteilungen der Brick-Flow AG.

Kapitel 2

Datenmanagement entlang von Geschäftsprozessen

Nachdem im vorherigen Kapitel die strategische Bedeutung von Daten eingeordnet wurde, stellt sich nun die Frage, *wie Daten im Unternehmen konkret entstehen, verarbeitet und genutzt werden.* In der Praxis zeigt sich, dass Daten selten isoliert existieren. Sie sind vielmehr eng mit den operativen Abläufen eines Unternehmens verknüpft und entstehen entlang von Geschäftsprozessen.

Für das Management bedeutet dies: Wer Daten sinnvoll nutzen möchte, muss zunächst verstehen, *wo diese Daten entstehen, wie sie fließen und in welchen Systemen sie verarbeitet werden.* Dieses Kapitel zielt daher darauf ab, die Verbindung zwischen Prozessen, Daten und IT-Systemen transparent zu machen und daraus Anforderungen an ein wirksames Datenmanagement abzuleiten.

2.1 Daten und Datenflüsse in Geschäftsprozessen

In Unternehmen entstehen Daten nicht zufällig, sondern als Nebenprodukt oder Ergebnis operativer Tätigkeiten. Jeder Geschäftsprozess – von der Kundenbestellung über die Produktion bis hin zur Auslieferung – erzeugt und verarbeitet Daten. Diese Daten spiegeln den Zustand und die Entwicklung der Prozesse wider.

Das Thema Prozessmodellierung – etwa mit BPMN – ist Bestandteil des Kurses *Grundlagen der Wirtschaftsinformatik* [14]. Darin werden Geschäftsprozesse als

Abfolge von Aktivitäten definiert. Jede dieser Aktivitäten ist potenziell mit Daten verknüpft: Es werden Daten erzeugt, verändert oder weitergegeben. Damit lassen sich Prozesse nicht nur als Abfolge von Aufgaben, sondern auch als *Abfolge von Datenflüssen* verstehen.

Ein einfaches Beispiel aus der Logistik verdeutlicht dies: Eine Bestellung löst einen Prozess aus, in dessen Verlauf Daten zu Kunden, Produkten, Lagerbeständen und Lieferzeiten verarbeitet werden. Ohne diese Daten wäre der Prozess weder steuerbar noch auswertbar.

Die Betrachtung von Datenflüssen erweitert die klassische Prozesssicht um eine wichtige Dimension. Während Prozesse beschreiben, *was* passiert, geben Datenflüsse Aufschluss darüber, *welche* Informationen dabei entstehen und genutzt werden. Für die Analyse von Datenflüssen sind insbesondere folgende Fragen relevant:

- An welchen Stellen im Prozess entstehen Daten?
- Um welche Arten von Daten handelt es sich? (z. B. Uhrzeit vs. Foto/Bild)
- Welche Daten werden zwischen Aktivitäten weitergegeben?
- Wo werden Daten gespeichert und später wieder verwendet?

Die Beantwortung dieser Fragen schafft Transparenz über die Datenbasis eines Unternehmens. In der Praxis zeigt sich häufig, dass Daten mehrfach erfasst, unvollständig weitergegeben oder in unterschiedlichen Systemen unterschiedlich interpretiert werden. Ein mangelndes Verständnis von Datenflüssen führt daher oft zu Medienbrüchen zwischen Systemen, redundanter Datenerfassung und inkonsistenten Informationen.

Für das Management ist es daher wichtig, Prozesse nicht nur funktional, sondern auch datenorientiert zu analysieren.

2.2 Von operativen Systemen zum Data Warehouse

Die meisten unternehmensrelevanten Daten entstehen in operativen IT-Systemen. Dazu zählen beispielsweise Online-Shops, ERP-Systeme, CRM-Systeme oder Produktionssteuerungssysteme. Diese Systeme unterstützen die täglichen Geschäftsprozesse und erfassen dabei kontinuierlich Daten.

Ein zentrales Merkmal operativer Systeme ist, dass sie auf die Durchführung von Transaktionen ausgelegt sind. Sie erfassen Daten möglichst schnell und effizient, etwa bei der Eingabe einer Bestellung oder der Buchung eines Wareneingangs. Für analytische Zwecke sind diese Daten jedoch nicht immer unmittelbar geeignet, da sie häufig stark fragmentiert vorliegen, auf einzelne Prozesse fokussiert sind und nicht für übergreifende Auswertungen strukturiert wurden. Dennoch bilden sie die Grundlage für nahezu alle weiterführenden Analysen.

Bevor Daten in Datenbanken oder größere Systeme fließen, begegnen sie uns in der Praxis häufig in einer einfacheren Form: als *Dateien*. Tabellenkalkulationen im Excel-Format, kommagetrennte Textdateien (CSV) oder moderne Spaltenformate wie Parquet sind in vielen Unternehmen alltägliche Austausch- und Speicherformate – zwischen Abteilungen, gegenüber externen Partnern oder als Export aus Fachsystemen. Controlling-Planzahlen, Lieferantenlisten oder Berichte liegen typischerweise in solchen Dateien vor.

Für einfache, punktuelle Analysen sind dateibasierte Formate durchaus praktisch. Mit wachsender Datenmenge und steigender Komplexität stoßen sie jedoch an strukturelle Grenzen: Mehrere Versionen derselben Datei kursieren parallel ohne klare Aktualität, gleichzeitiger Zugriff durch mehrere Personen führt zu Konflikten, und Fehler bei Datentypen oder Vollständigkeit bleiben häufig unbemerkt. Für unternehmensweite Auswertungen, die Daten aus verschiedenen Bereichen zusammenführen müssen, reichen Dateien als alleiniges Speicherformat nicht aus. *Strukturierte Datenbanksysteme* beheben diese Schwächen durch klare Schemata, kontrollierte Zugriffsmechanismen und Integritätssicherung. Das Data Warehouse nimmt dabei eine besondere Rolle ein: Es integriert *heterogene Quellen* – operative Datenbanken, Fachanwendungen und dateibasierte Daten – zu einer konsistenten Analysebasis (vgl. Abbildung 2.1).

2.2.1 Datenbanken und SQL aus Anwendungssicht

Datenbanken bilden das technische Rückgrat der Datenspeicherung in Unternehmen. Sie sorgen dafür, dass Daten strukturiert, konsistent und langfristig verfügbar sind. Aus organisatorischer Sicht ist dabei weniger die technische Funktionsweise entscheidend, sondern die Frage, *wie Daten in Datenbanken organisiert und genutzt werden.*

SQL als Abfragesprache ermöglicht es, gezielt auf Daten zuzugreifen und diese für Auswertungen bereitzustellen. Für Fachbereiche bedeutet dies nicht, komplexe Abfragen selbst zu entwickeln, sondern zu verstehen, welche Daten grundsätzlich verfügbar sind, wie sie logisch miteinander verknüpft sind und welche Einschränkungen bei der Nutzung bestehen. Dieses Verständnis ist wichtig, um Anforderungen an Analysen formulieren und mit IT-Abteilungen oder Data-Teams effektiv zusammenarbeiten zu können.

2.2.2 OLTP und OLAP: zwei Welten der Datenverarbeitung

Moderne *Data Warehouse*-Architekturen bestehen aus verschiedenen Ebenen, in denen Daten gespeichert, transformiert und ausgewertet werden. In [14] werden zwei wesentliche Arten von Informationssystemen genannt: OLTP – *Online Transactional Processing* – und OLAP – *Online Analytical Processing*. OLTP-Datenbanken werden für feingranulare Transaktionsdaten verwendet, die durch die Geschäftsprozesse häufig in einem beschränkten Kontext betrachtet werden. Die Analyse von Daten erfordert hingegen in der Regel eine starke Aggregation, die im Data-Warehouse-Verbund durch OLAP-Systeme in sogenannten *Cubes* bereitgestellt wird. Abbildung 2.1 zeigt ein typisches Data Warehouse mit verschiedenen angeschlossenen Systemen. Bei diesen Systemen kann es sich um Datenbanken, Informationssysteme oder in einigen Fällen auch Dateien handeln (z. B. Controlling-Planzahlen aus Excel-Dateien). In sogenannten ETL-Prozessen (manchmal auch ETL-Strecken genannt) werden die Daten aus den angeschlossenen Systemen in das Data Warehouse integriert und dabei aufbereitet und aggregiert.

Cubes und Data Marts. OLAP-Systeme organisieren Daten in sogenannten *Cubes* (mehrdimensionale Würfel). Statt einzelner Transaktionszeilen enthält ein Cube vorberechnete *Kennzahlen* – etwa Umsatz oder Bestellmenge – die entlang mehrerer *Dimensionen* gleichzeitig auswertbar sind. Typische Dimensionen sind Zeit (Jahr, Quartal, Monat), Produkt (Kategorie, Variante) oder Region. Diese Vorstrukturierung erlaubt sehr schnelle Abfragen, weil Aggregationen nicht erst zur Laufzeit über Millionen von Einzelzeilen berechnet werden müssen. Abbildung 2.2 zeigt einen Cube der z.B. Absatzzahlen nach Produkt, Region und Quartal berechnet.

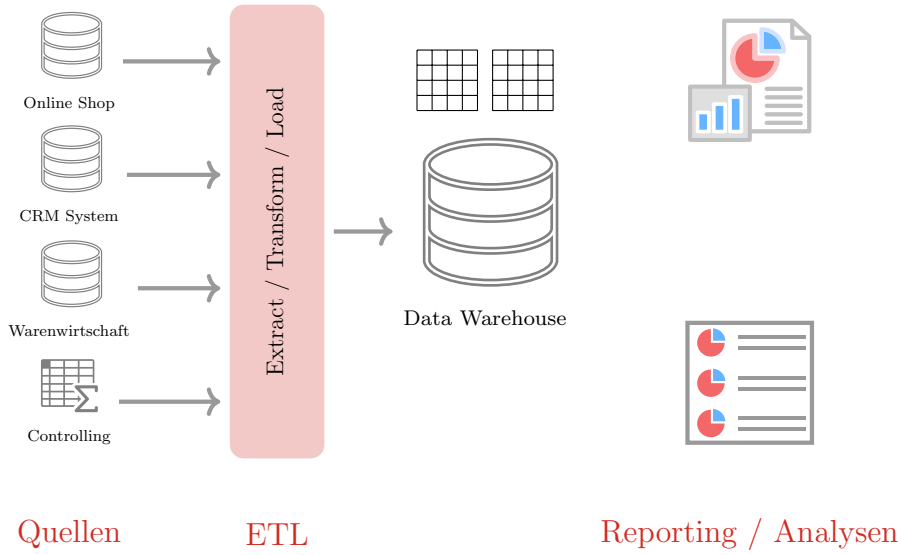


Abbildung 2.1: Beispiel einer modernen Data Warehouse Architektur.

Eng verwandt ist das Konzept des *Data Mart*: ein thematisch eingegrenzter Ausschnitt des Data Warehouse, der auf einen bestimmten Fachbereich zugeschnitten ist. Ein Vertriebs-Data-Mart stellt beispielsweise nur umsatz- und produktbezogene Cubes bereit, ein Controlling-Data-Mart Budget- und Kostendaten. Der Unterschied zum vollständigen Data Warehouse liegt im Fokus: Während das Data Warehouse alle Unternehmensdaten integriert, arbeiten Fachbereiche im Data Mart ausschließlich mit den für sie relevanten Daten und Cubes.

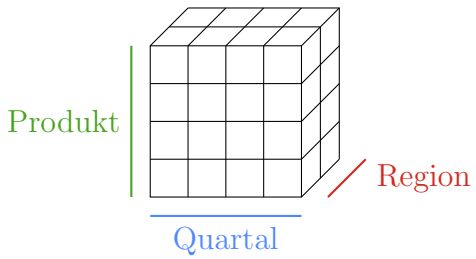


Abbildung 2.2: Ein OLAP-Cube entlang der Dimensionen *Produkt*, *Quartal* und *Region*. Jeder Würfel in diesem Cube enthält die Bestelldaten für eine bestimmte Produktkategorie in einem bestimmten Quartal in einer bestimmten Region.

Beispiel – OLTP, OLAP und Cube bei der Brick-Flow AG. Im laufenden Tagesgeschäft greifen Kundinnen und Kunden auf die *OLTP-Datenbank* des Online-Shops zu – etwa um eine Bestellung aufzugeben oder die eigene Bestellhistorie einzusehen. Dabei werden gezielt wenige Datensätze gelesen oder geschrieben; Geschwindigkeit und Datenkonsistenz im Einzelzugriff stehen im Vordergrund.

Die Controlling-Abteilung stellt hingegen aggregierte Fragen: Welche Produktkategorie haben im dritten Quartal den höchsten Umsatz erzielt? In welcher Region hat welche Produktkategorie sich am erfolgreichsten entwickelt? Für solche Auswertungen ist ein *OLAP-Cube* zuständig (siehe Abbildung 2.2), der die Bestelldaten der Brick-Flow AG entlang dreier Dimensionen vorstrukturiert:

- **Zeit:** Jahr → Quartal → Monat
- **Produkt:** Kategorie (Standard-Palette, Themen-Palette) → Farbvariante
- **Kanal:** Online-Shop, Logistikpartner

Als Kennzahlen enthält der Cube je Schnittpunkt dieser Dimensionen den kumulierten Umsatz sowie die durchschnittliche Lieferzeit. Die Controlling-Abteilung kann durch den Cube *navigieren*: ein einzelnes Quartal herausschneiden (*Slice*), von der Kategorieebene auf einzelne Farben wechseln (*Drill-down*) oder Monatswerte zu Jahreswerten zusammenfassen (*Roll-up*) – ohne jedes Mal die gesamte Bestelldatenbank neu abfragen zu müssen. Da derartige Analysen auf dem OLTP-System zu erheblichen Laufzeiten und damit zu einer Verlangsamung des laufenden Web-Shops führen würden, wird das OLAP-System parallel betrieben und z. B. nächtlich mit den neuen Bestelldaten aktualisiert (vgl. Abschnitt 2.3).

2.3 Von Batch zu Echtzeit: ETL-Prozesse und Streaming

ETL-Prozesse (Extract, Transform, Load) bilden das Herzstück klassischer Data-Warehouse-Architekturen. Sie laufen typischerweise im *Batch-Betrieb*: In festgelegten Intervallen – häufig nächtlich – werden Daten aus den Quellsystemen extrahiert, aufbereitet und in das Data Warehouse geladen. Dieses Vorgehen ist robust und für viele analytische Fragestellungen ausreichend: Monatsberichte, Vertriebsauswertungen oder strategische Kennzahlen benötigen keine sekündliche Aktualität.

Das Batch-Prinzip hat jedoch einen strukturellen Nachteil: Die Daten im Data Warehouse sind immer nur so aktuell wie der letzte Ladezyklus. Für zeitkritische Anwendungen – Echtzeit-Bestandsüberwachung, Betrugserkennung oder Live-Dashboards im Kundenservice – ist diese Latenz nicht akzeptabel.

Streaming-Architekturen. Für die Echtzeit- oder Near-Time-Verarbeitung ist eine grundlegend andere Architektur erforderlich. Anstatt Daten periodisch zu sammeln, werden sie *ereignisbasiert* und kontinuierlich übermittelt: Jedes relevante Ereignis – eine Bestellung, ein Klick, ein Sensorwert – wird unmittelbar nach seinem Entstehen als kleines Datenpaket weitergeleitet.

Technisch werden solche Architekturen häufig durch sogenannte *Message Broker* realisiert. Apache Kafka ist eines der bekanntesten Beispiele: Quellsysteme (sogenannte *Producer*) senden Ereignisse in Echtzeit an einen zentralen Broker; Analyse- und Auswertungssysteme (sogenannte *Consumer*) abonnieren diese Ereignisströme und aktualisieren sich inkrementell. Statt einmal pro Nacht einen kompletten Datensatz zu übertragen, spiegelt das Auswertungssystem den aktuellen Zustand mit einer Verzögerung von Sekunden oder weniger wider.

Ein wesentlicher Vorteil dieser Architektur ist die *Entkopplung* von Quell- und Zielsystemen: Der Broker fungiert als Puffer – Quellsysteme müssen nicht wissen, welche Konsumenten ihre Daten beziehen, und Analysesysteme können unabhängig ergänzt oder ausgetauscht werden.

Voraussetzungen und Grenzen. Streaming-Architekturen haben eine entscheidende Voraussetzung: Alle Quellsysteme, deren Daten in Echtzeit auswertbar sein sollen, müssen ereignisbasiert angebunden werden können. In der Praxis ist das – insbesondere bei älteren Legacy-Systemen – eine erhebliche Herausforderung. Hinzu kommt, dass Streaming-Infrastrukturen deutlich komplexer zu betreiben und zu überwachen sind als klassische Batch-Strecken, mit entsprechend höherem technischen Aufwand und höheren Infrastrukturkosten. Für viele Unternehmen ist daher ein *hybrider Ansatz* sinnvoll: zeitkritische Datenströme werden ereignisbasiert verarbeitet, weniger zeitkritische Daten weiterhin per Batch integriert.

ETL-Batch-Prozesse sind robust und für strategische Analysen in der Regel ausreichend – operative Echtzeit-Anforderungen erfordern grundlegend andere Architekturen mit Streaming und Message Brokern wie Apache Kafka. Der Wechsel setzt jedoch voraus, dass alle relevanten Quellsysteme ereignisbasiert angebunden werden können.

2.4 Datenqualität als Voraussetzung für Analysen

Die Qualität von Daten ist ein entscheidender Faktor für den Erfolg datengetriebener Entscheidungen. Schlechte Datenqualität führt unmittelbar zu fehlerhaften Analysen und damit zu potenziell falschen Entscheidungen. Wesentliche Dimensionen der Datenqualität sind:

- **Vollständigkeit**: Sind alle relevanten Daten vorhanden?
- **Konsistenz**: Stimmen Daten in verschiedenen Systemen überein?
- **Aktualität**: Sind die Daten auf dem neuesten Stand?

In der Praxis entstehen Qualitätsprobleme häufig durch unklare Prozesse, fehlende Standards oder manuelle Eingaben. Besonders kritisch ist dabei, dass Probleme der Datenqualität oft erst sichtbar werden, wenn Analysen durchgeführt werden.

Für das Management bedeutet dies, dass Datenqualität nicht als technisches Detail betrachtet werden darf, sondern als *zentrale Voraussetzung für verlässliche Entscheidungsgrundlagen*. Gerade bei komplexen Datenarchitekturen (vgl. Abbildung 2.1) ist eine sorgfältige und kontinuierliche Validierung der Daten erforderlich. Dies ist häufig mit großem Aufwand verbunden – insbesondere bei komplexen, historisch gewachsenen Strukturen – und wird in vielen Fällen vernachlässigt, weil daraus im operativen Geschäft kein direkt sichtbarer Nutzen entsteht. Langfristig erzeugt eine fehlende Validierung aber eine unklare Daten- und Informationslage.

Datenqualität ist kein technisches Detail, sondern die Grundlage jeder verlässlichen Analyse – Fehler in den Daten setzen sich häufig unmittelbar in falschen Entscheidungen fort.

2.5 Data Governance und organisatorische Verantwortung

Ein wirksames Datenmanagement erfordert klare Regeln und Verantwortlichkeiten. Unter dem Begriff Data Governance werden alle Maßnahmen zusammengefasst, die den Umgang mit Daten im Unternehmen strukturieren. Dazu gehören insbesondere:

- Festlegung von Verantwortlichkeiten für Daten (z. B. Data Owner)
- Definition von Standards für Datenerfassung und -nutzung
- Regelungen für Zugriff und Datenschutz

Ein zentrales Problem in vielen Unternehmen ist, dass Daten zwar vorhanden sind, aber niemand eindeutig für deren Qualität oder Pflege verantwortlich ist. Dies führt langfristig zu Inkonsistenzen und Vertrauensverlust. Data-Governance-Strategien definieren dazu verschiedene Rollen, die unterschiedliche Aufgaben im Umgang mit Daten übernehmen. Dabei ist wichtig zu verstehen, dass diese Rollen nicht zwingend einzelnen Personen entsprechen müssen, sondern auch organisatorisch gebündelt sein können. Zu den zentralen Rollen zählen insbesondere:

- **Data Owner**:
  Verantwortlich für die fachliche Definition, Qualität und Nutzung von Daten. Der Data Owner entscheidet, welche Daten wie verwendet werden dürfen und stellt sicher, dass diese den Anforderungen des Fachbereichs entsprechen.
- **Data Steward**:
  Unterstützt den Data Owner operativ und sorgt für die Umsetzung von Standards, etwa bei Datenpflege, Strukturierung oder Qualitätssicherung.
- **Data Analyst / Data Scientist**:
  Analysiert Daten und entwickelt Modelle zur Entscheidungsunterstützung. Diese Rolle ist stark anwendungsorientiert und verbindet Daten mit konkreten Fragestellungen.
- **Fachbereich / Management**:
  Nutzt Daten als Grundlage für Entscheidungen und definiert Anforderungen an Analysen und Kennzahlen.

Diese Rollen verdeutlichen, dass Datenmanagement keine isolierte IT-Aufgabe ist, sondern eine gemeinsame Verantwortung von Fachbereichen und technischen Einheiten.

Für Fachbereiche besteht dabei die Herausforderung, Anforderungen an Daten klar zu formulieren, ohne selbst tief in technische Details einzusteigen. Typische Anforderungen umfassen die Verfügbarkeit relevanter Daten, verständliche und konsistente Datenstrukturen sowie eine transparente Herkunft und Bedeutung der Daten. Mit einer zentralen *Data-Governance-Strategie* stellen Unternehmen sicher, dass Daten nicht nur vorhanden, sondern auch verlässlich und nutzbar sind.

Weiterführende Literatur

- *Fundamentals of Data Engineering: Plan and Build Robust Data Systems* [21]: Überblick über moderne Datenarchitekturen und Data Engineering.
- *Datenmodellierung in Data-Warehouse-Systemen: Konzepte, Technologien und Methoden für die Modellierung entscheidungsunterstützender Daten in Unternehmen* [8]: Modellierung von Daten für die Entscheidungsunterstützung in Unternehmen.
- *Fundamentals of Business Process Management* [6]: Anschlussfähige Vertiefung zu Geschäftsprozessmanagement und BPMN.
- *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program* [15]: Umfassende Einführung in Data Governance aus organisatorischer Perspektive.

Kapitel 3

Angewandte Datenanalyse und Statistik im Business-Kontext

Datenanalysen sind heute ein fester Bestandteil betrieblicher Entscheidungsprozesse. In vielen Unternehmen stehen umfangreiche Daten und leistungsfähige Analysetools zur Verfügung. Dennoch zeigt die Praxis, dass datenbasierte Entscheidungen häufig *falsch interpretiert oder überbewertet* werden. Der Grund dafür liegt selten in fehlender Technik, sondern vielmehr in einem unzureichenden Verständnis dafür, wie Analyseergebnisse einzuordnen und zu bewerten sind.

Dieses Kapitel knüpft an Ihre vorhandenen statistischen Kenntnisse an und verschiebt den Fokus bewusst: Weg von der Berechnung einzelner Kennzahlen, hin zur *Interpretation und kritischen Reflexion von Analyseergebnissen im Managementkontext*. Ziel ist es, Daten nicht nur zu „lesen", sondern ihre Aussagekraft im Hinblick auf konkrete Entscheidungen beurteilen zu können.

3.1 Von der Geschäftsfrage zur Analyse

Der Ausgangspunkt jeder sinnvollen Datenanalyse ist nicht der Datensatz, sondern die betriebswirtschaftliche Fragestellung oder *Geschäftsfrage*. In der Praxis zeigt sich jedoch häufig ein umgekehrtes Vorgehen: Unternehmen analysieren verfügbare Daten, ohne zuvor klar zu definieren, welches Problem eigentlich gelöst werden soll.

Ein strukturierter Ansatz beginnt daher mit der Übersetzung einer Managementfrage in eine analytische Fragestellung. Beispielsweise wird aus der Frage "Warum sinkt unser Umsatz?" eine differenziertere Analysefrage wie "In welchen Kundensegmenten oder Regionen zeigt sich ein Rückgang?"

Dieser Schritt ist entscheidend, da er festlegt, welche Daten benötigt werden, welche Analyse sinnvoll ist und wie die Ergebnisse später interpretiert und genutzt werden können.

Gute Analysen sind somit immer zielgerichtet. Ohne klare Fragestellung besteht die Gefahr, zufällige Muster zu identifizieren, die für Entscheidungen wenig relevant sind. Wir werden in Kapitel 6 dazu noch Methoden betrachten, die helfen, diesen Aspekt bei Analysen in den Fokus zu nehmen.

3.2 Deskriptive Analysen im Kontext verstehen

Deskriptive Analysen bilden in vielen Unternehmen die Grundlage für Berichte und Dashboards. Kennzahlen wie Durchschnittswerte, Wachstumsraten oder Verteilungen sind Ihnen bereits aus Modulen wie *Wirtschaftsstatistik* bekannt. An diese Stelle sei dazu nochmal auf die Lehrbriefe Wirtschaftsstatistik, Lerneinheit 1 und 2 verwiesen [12, 13]. Entscheidend ist jedoch nicht die Berechnung dieser Werte, sondern ihre *Interpretation im jeweiligen Kontext.*

Ein Durchschnittswert kann beispielsweise einen positiven Eindruck vermitteln, obwohl die zugrunde liegende Verteilung sehr ungleich ist. Ein anschauliches Beispiel liefern die *Bestellwerte der Brick-Flow AG*: Das Unternehmen bedient sowohl viele kleine Endkundenbestellungen (B2C) als auch wenige große Bestellungen des Fachhandels (B2B). Abbildung 3.1 zeigt die Verteilung der Bestellwerte.

Der *durchschnittliche Bestellwert* liegt bei rund 254 €. Diese Zahl allein vermittelt jedoch ein verzerrtes Bild: Der *Median* liegt mit nur 124 €bei weniger als der Hälfte. Die typische Bestellung ist also deutlich kleiner, als der Mittelwert suggeriert – denn die wenigen sehr großen B2B-Bestellungen ziehen den Durchschnitt nach oben. Wer auf Basis des Mittelwerts etwa Versandverpackungen oder Rabattgrenzen plant, würde die Mehrheit der tatsächlichen Bestellungen falsch einschätzen.

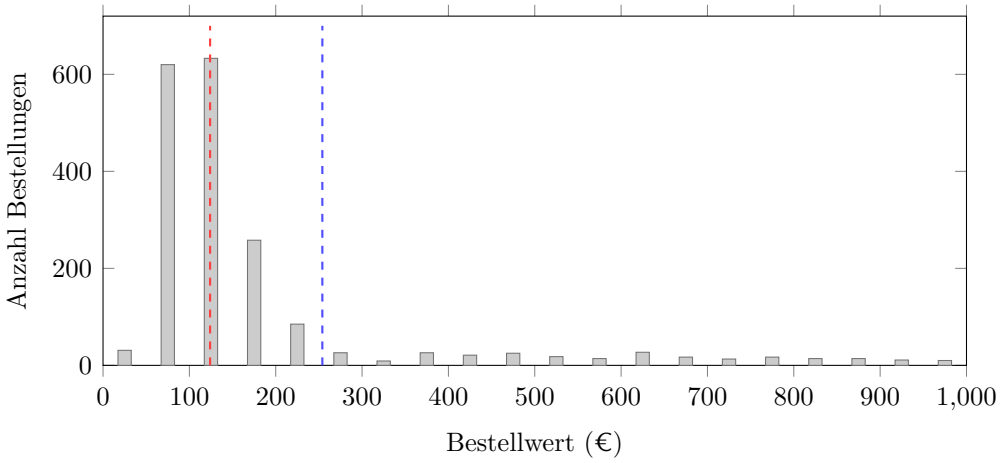


Abbildung 3.1: Verteilung der Bestellwerte der Brick-Flow AG. Die wenigen großen B2B-Bestellungen (langer Ausläufer nach rechts, bis über 3000 €) ziehen den Mittelwert deutlich über den Median.

Für die Praxis bedeutet das: Kennzahlen sollten immer im Zusammenhang betrachtet, Entwicklungen im Kontext eingeordnet werden, und Einzelwerte sind selten ausreichend für fundierte Entscheidungen.

> Ein Mittelwert beschreibt eine Verteilung nur unvollständig. Erst der Blick auf Streuung und Form – etwa über Median und Verteilungsdiagramm – zeigt, ob eine Kennzahl die Realität angemessen abbildet.

3.3 Korrelation, Kausalität und typische Fehlschlüsse

Ein besonders häufiges Problem im Umgang mit Daten ist die Verwechslung von *Korrelation und Kausalität*. Eine Korrelation beschreibt lediglich einen statistischen Zusammenhang zwischen zwei Variablen, sagt jedoch nichts darüber aus, ob eine Variable die andere verursacht.

In Kapitel 4 werden wir näher auf die Umsatzentwicklung der Brick-Flow AG eingehen. Dabei wird sichtbar, dass der Monatsumsatz der Brick-Flow AG und die durchschnittliche Lieferzeit stark zusammenhängen: In umsatzstarken Monaten steigen die Lieferzeiten. Hier ist ein *ursächlicher* Mechanismus plausibel – ein hohes Bestellaufkommen bringt den externen Versanddienstleister an seine Kapa-

zitätsgrenze. Allein aus der Korrelation folgt das aber nicht; sie ist nur ein Hinweis, dem man nachgehen muss.

Anders liegt der klassische Fall scheinbarer Ursächlichkeit: Wenn steigende Marketingausgaben mit höheren Umsätzen einhergehen, liegt die Annahme nahe, dass die Marketingmaßnahmen den Umsatzanstieg bewirken. Tatsächlich können jedoch *weitere Faktoren* (etwa das saisonale Weihnachtsgeschäft) sowohl die Marketingausgaben als auch den Umsatz gleichzeitig nach oben treiben. Der beobachtete Zusammenhang wäre dann teilweise eine Scheinkorrelation.

Daraus folgt: Ein Zusammenhang ist kein Beweis für Ursache und Wirkung, Entscheidungen sollten nicht allein auf Korrelationen beruhen, und zur Absicherung sind zusätzliche Analysen oder Experimente nötig.

Neben dieser Verwechslung treten weitere typische Denkfehler auf, die meist nicht durch mangelndes Fachwissen, sondern durch vereinfachte Interpretation entstehen:

- Überinterpretation einzelner Kennzahlen ohne Kontext
- Verwechslung kurzfristiger Effekte mit langfristigen Trends
- Ignorieren von Unsicherheiten
- selektive Wahrnehmung von Ergebnissen (man sieht, was man erwartet)

Diese Fehler verdeutlichen, dass Daten zwar objektiv erscheinen, ihre Interpretation jedoch immer subjektiv geprägt ist.

3.4 Von der Korrelation zur Ursache: Hypothesen und A/B-Tests

Wenn eine Korrelation allein keine Ursache belegt – wie lässt sich ein Wirkzusammenhang dann überhaupt belastbar nachweisen? Das wirksamste Mittel ist das *kontrollierte Experiment*. Statt nur vorhandene Daten zu beobachten, wird gezielt *eine* Größe verändert und ihre Wirkung gemessen.

Der Ausgangspunkt ist eine klar formulierte *Hypothese*, etwa: „Eine übersichtlichere Produktseite erhöht die Kaufabschlussquote.Üm sie zu prüfen, werden zwei *Vergleichsgruppen* gebildet:

- eine *Kontrollgruppe*, die die bisherige Version sieht,

- eine *Testgruppe*, die die veränderte Version sieht.

Entscheidend ist, dass die Zuordnung *zufällig* erfolgt. Dadurch unterscheiden sich die Gruppen im Mittel nur in der getesteten Änderung – alle anderen Einflüsse (Saison, Kundentyp, Marktlage) verteilen sich gleichmäßig. Ein anschließend gemessener Unterschied lässt sich dann *ursächlich* auf die Änderung zurückführen. Genau dieses Vorgehen wird als *A/B-Test* bezeichnet.

Beispiel Brick-Flow AG: Der Online-Shop möchte wissen, ob ein neuer Checkout-Prozess mehr Bestellungen erzeugt. Die Besucher werden zufällig auf die alte (A) und die neue (B) Variante aufgeteilt; verglichen wird die Conversion Rate beider Gruppen. Liegt sie in Gruppe B verlässlich höher, ist die Änderung ursächlich wirksam – und nicht nur zufällig mit höheren Umsätzen korreliert.

Auch ein A/B-Test bleibt jedoch eine Stichprobe: Kleine Unterschiede können zufällig entstehen. Bevor man eine Änderung dauerhaft einführt, muss daher beurteilt werden, ob der gemessene Effekt groß genug und ausreichend abgesichert ist – ein Übergang zur Frage der Unsicherheit.

3.5 Unsicherheit, Prognosen und Entscheidungen

Datenanalysen sind grundsätzlich mit *Unsicherheit* behaftet. Das gilt besonders für Prognosen, mit denen Unternehmen zukünftige Entwicklungen planen. Ein prognostizierter Wert ist keine sichere Vorhersage, sondern eine *Schätzung auf Basis von Annahmen und historischen Daten.*

Für die Interpretation bedeutet das, dass Prognosen als *Bandbreiten* statt als exakte Werte verstanden werden sollten, dass die zugrunde liegenden Annahmen transparent sein müssen und dass Unsicherheit ein integraler Bestandteil der Analyse ist – kein Makel, der zu verbergen wäre.

Genau deshalb entfalten Datenanalysen ihren Wert erst dann, wenn sie *in Entscheidungen einfließen, ohne diese zu ersetzen.* Eine fundierte Nutzung von Analyseergebnissen erfordert die Einordnung in den Kontext, die Berücksichtigung von Unsicherheiten und die kritische Reflexion möglicher Fehlinterpretationen. Entscheidungen sollten daher stets auf einer Kombination aus *Analyse, Erfahrung und fachlichem Verständnis* beruhen.

Weiterführende Literatur

- *Naked Statistics: Stripping the Dread from the Data* [24]: Verständliche Auffrischung statistischer Grundideen mit Alltags- und Wirtschaftsnähe.
- *How to Lie with Statistics* [9]: Klassiker zur kritischen Bewertung statistischer Aussagen und Fehlinterpretationen.
- *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python* [2]: Praxisorientierte Brücke zwischen Statistik und Data Science.

Kapitel 4

Datenvisualisierung und Business Intelligence

Datenanalysen entfalten ihren Wert erst dann, wenn ihre Ergebnisse verstanden und in Entscheidungen überführt werden. In der Praxis zeigt sich jedoch, dass genau an dieser Stelle häufig Probleme entstehen: Analysen sind zwar vorhanden, werden aber *nicht richtig interpretiert oder nicht wirksam kommuniziert.*

Datenvisualisierung und Business Intelligence (BI) setzen genau hier an. Sie dienen dazu, Daten so aufzubereiten, dass Entscheidungsträger komplexe Zusammenhänge schnell erfassen und fundierte Entscheidungen treffen können. Dieses Kapitel vermittelt daher nicht nur Grundlagen der Visualisierung, sondern zeigt vor allem, wie Daten *zielgerichtet für Managemententscheidungen eingesetzt werden.*

4.1 Datenvisualisierung für Managemententscheidungen

Im Managementkontext müssen Entscheidungen oft unter Zeitdruck und auf Basis unvollständiger Informationen getroffen werden. Gut aufbereitete Daten können dabei helfen, komplexe Sachverhalte zu strukturieren und relevante Muster sichtbar zu machen.

Visualisierungen sind dabei mehr als nur grafische Darstellungen. Sie übersetzen Daten in eine Form, die *schnell erfasst und intuitiv verstanden werden kann.* Während eine Tabelle mit monatlichen Umsatzzahlen schwer zu überblicken ist, macht ein Liniendiagramm Trends und Muster sofort sichtbar. Abbildung 4.1 zeigt dies am Beispiel der Brick-Flow AG: Die monatlichen Umsätze schwanken saisonal stark

– mit einem deutlichen Höhepunkt im Weihnachtsgeschäft –, während die eingezeichnete Trendgerade den langfristigen Wachstumspfad über drei Geschäftsjahre offenlegt. Beide Aussagen – die Saisonalität *und* der Wachstumstrend – sind aus der reinen Zahlenkolonne kaum, aus der Grafik dagegen unmittelbar ablesbar.

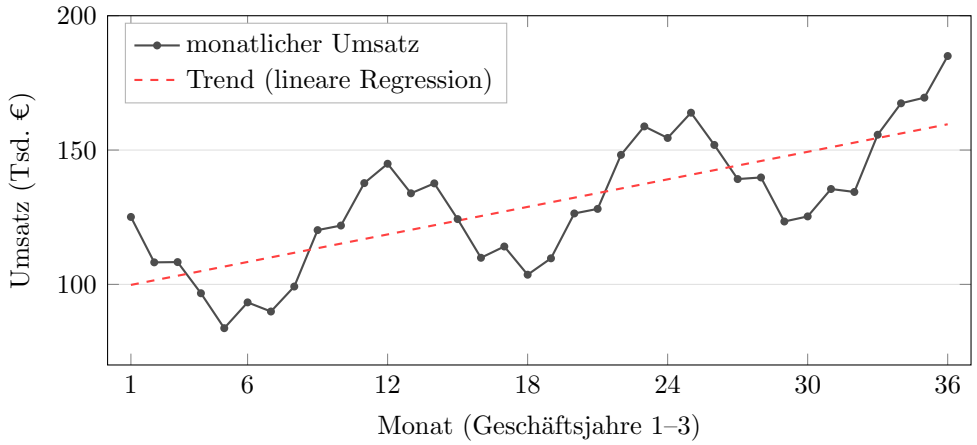


Abbildung 4.1: Monatliche Umsatzentwicklung der Brick-Flow AG über drei Geschäftsjahre mit saisonalem Verlauf und linearer Trendgerade.

Datenvisualisierung ist somit ein entscheidendes Bindeglied zwischen Analyse und Entscheidung. Damit eine Visualisierung diesen Zweck erfüllt, muss sie klar, verständlich und zielgerichtet gestaltet sein. Wichtig sind dabei die *Reduktion auf das Wesentliche*, eine klare Struktur, die *Vermeidung von Verzerrungen* (etwa durch manipulierte Achsen) sowie die Orientierung an der Zielgruppe. Ebenso wichtig ist die Wahl des passenden Diagrammtyps: Liniendiagramme eignen sich für zeitliche Entwicklungen wie in Abbildung 4.1, Balkendiagramme für Vergleiche zwischen Kategorien, Kreisdiagramme – mit Vorsicht – für Anteile. Entscheidend ist nicht die Vielfalt verfügbarer Diagramme, sondern die passende Darstellung für die jeweilige Fragestellung.

Über die rein technische Gestaltung hinaus spielt schließlich die Kommunikation eine zentrale Rolle. Beim *Storytelling mit Daten* werden Erkenntnisse nicht isoliert präsentiert, sondern in eine nachvollziehbare Argumentation eingebettet, die eine klare Fragestellung verfolgt und zentrale Aussagen hervorhebt.

Eine gute Visualisierung reduziert Komplexität – eine schlechte Visualisierung erzeugt zusätzliche.

4.2 Diagrammtypen für unterschiedliche Fragestellungen

Die Wahl des passenden Diagrammtyps ist keine ästhetische, sondern eine analytische Entscheidung. Jeder Diagrammtyp beantwortet einen bestimmten Fragetyp – und verliert seine Aussagekraft, sobald er für den falschen Zweck eingesetzt wird. Tabelle 4.1 gibt einen Überblick über die im Managementalltag relevantesten Formen.

Tabelle 4.1: Ausgewählte Diagrammtypen und ihre Fragestellungen.

Fragestellung	**Diagrammtyp**	**Typische Anwendung**
Zeitlicher Verlauf	Liniendiagramm	KPI-Entwicklung, Umsatz über Perioden
Vergleich zwischen Kategorien	Balkendiagramm	Umsatz nach Region, Produkt oder Kanal
Anteile und Zusammensetzung	Gestapeltes Balkendiagramm	Kanalaufteilung im Umsatz
Zusammenhang zweier Variablen	Streudiagramm	Korrelation Umsatz vs. Lieferzeit
Verteilung einer Variable	Histogramm	Bestellwertverteilung, Qualitätskennzahlen
Verteilungsvergleich (mehrere Gruppen)	Boxplot, Violin-Plot	Lieferzeiten nach Quartal oder Region
Muster in zwei Dimensionen	Heatmap	Bestellaufkommen nach Wochentag und Monat
Veränderung und Beitrag einzelner Faktoren	Wasserfalldiagramm	Deckungsbeitragsanalyse, GuV-Brücke

Die Diagrammtypen der ersten beiden Zeilen wurden bereits in den vorigen Abschnitten und in Kapitel 3 eingesetzt. Die folgenden Abschnitte vertiefen drei Typen, die im Managementalltag besonders informativ sind – und deren Mehrwert gegenüber einfacheren Darstellungen oft unterschätzt wird.

4.2.1 Verteilungen vergleichen: Boxplot und Violin-Plot

Histogramme zeigen, wie häufig bestimmte Werte auftreten – aber immer nur für eine Gruppe. Sobald mehrere Gruppen nebeneinander verglichen werden sollen, wird das Histogramm unübersichtlich. Hier kommen *Boxplots* und *Violin-Plots* ins Spiel.

Ein *Boxplot* fasst eine Verteilung in fünf Kennzahlen zusammen: Median (mittlere Linie), unteres und oberes Quartil (Kastenränder, gemeinsam der *Interquartilsabstand* IQR), die Whisker (typischerweise bis zum 1,5-fachen IQR) sowie Ausreißer als einzelne Punkte. Er ist kompakt und erlaubt den direkten Vergleich vieler Gruppen – auch dann noch, wenn ein Violin-Plot zu schmal und damit unleserlich würde.

Beispiel Brick-Flow AG: Abbildung 4.2 zeigt die Lieferzeitverteilung für alle zwölf Monate des Geschäftsjahres. Bereits auf einen Blick ist das saisonale Muster erkennbar: Über den Sommer sinken Median und Streubreite deutlich, im Herbst steigen beide wieder an. Die gepunktete Linie verbindet die monatlichen Mediane und macht den U-förmigen Verlauf noch expliziter.

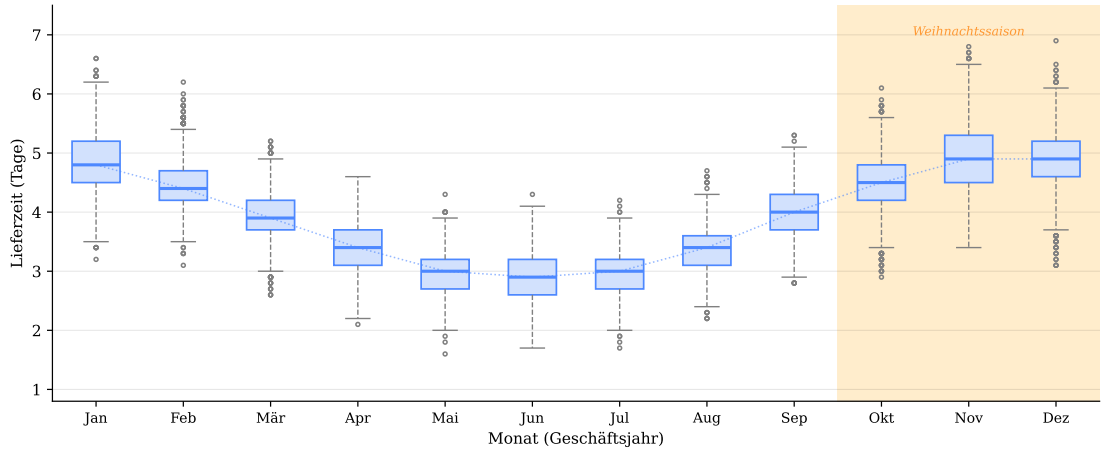


Abbildung 4.2: Monatliche Lieferzeitverteilung der Brick-Flow AG als Boxplot. Jede Box zeigt Median, IQR und Whisker; einzelne Punkte sind Ausreißer. Die gepunktete Linie verbindet die monatlichen Mediane. Die orange Hinterlegung markiert die Weihnachtssaison.

Der Boxplot eignet sich für diesen Überblick über zwölf Monate gut, weil er auch bei vielen Gruppen platzsparend bleibt. Er verzichtet jedoch auf eine Information: *Wo* genau innerhalb jeder Box liegen die meisten Bestellungen? Dieses Detail liefert der *Violin-Plot*.

Die Breite der "Geigenkurve" an jeder Stelle der y-Achse entspricht der lokalen Datendichte – wo viele Bestellungen ähnliche Lieferzeiten aufweisen, ist die Kurve breit; wo wenige Werte liegen, wird sie schmal. Im Inneren des Violin-Plots ist typischerweise ein Boxplot eingebettet, der beide Informationsebenen kombiniert. Abbildung 4.3 zeigt dieselben Lieferzeitdaten, nun auf Quartalsbasis verdichtet.

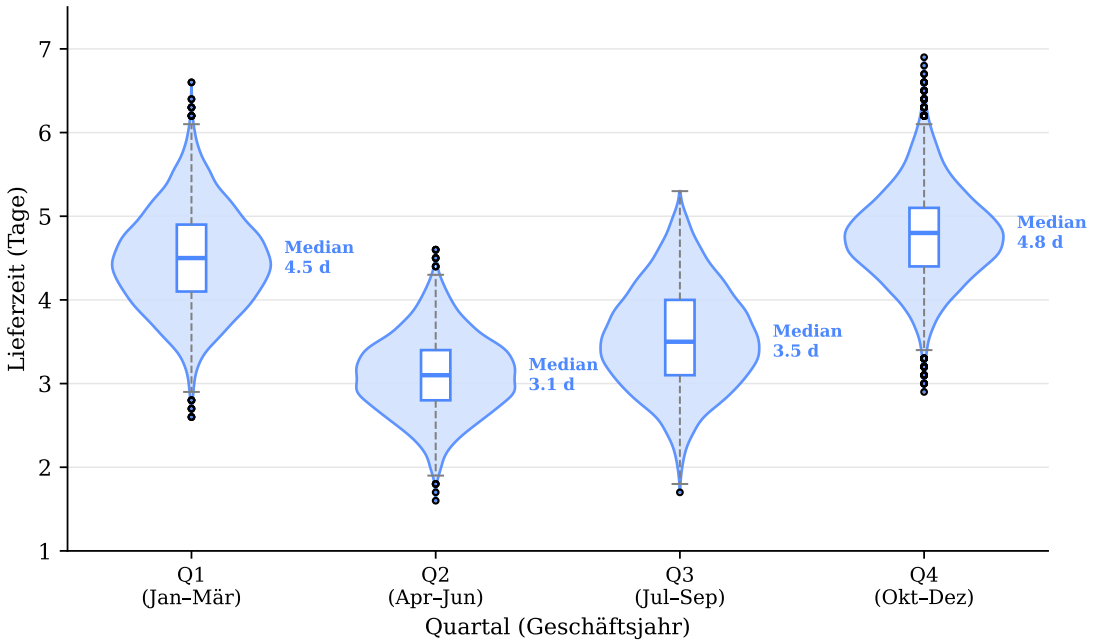


Abbildung 4.3: Lieferzeitverteilung der Brick-Flow AG nach Quartal als Violin-Plot. Die Breite der Kurve zeigt die Datendichte; der eingebettete Boxplot markiert Median und Quartile. Q2 ist schmal und niedrig (schnelle, gleichmäßige Lieferung in der Nebensaison); Q4 ist breit und nach oben verschoben (hohes Bestellvolumen, Kapazitätsengpass beim Versanddienstleister).

Beide Abbildungen zeigen denselben Sachverhalt – aber auf unterschiedlichen Aggregationsstufen und mit unterschiedlichem Informationsgehalt. Der Boxplot erlaubt den Monatsvergleich auf einen Blick; der Violin-Plot offenbart zusätzlich, ob die

Lieferzeiten innerhalb eines Quartals eher gleichmäßig verteilt oder an bestimmten Werten konzentriert sind. Für das Management ergibt sich eine klare Implikation: Nicht der Mittelwert allein, sondern die *Streuung* ist das eigentliche Steuerungsproblem. Ein Vertrag mit dem Logistikdienstleister, der nur Durchschnittswerte reguliert, würde die Ausreißer in Q4 und November systematisch übersehen.

4.2.2 Muster in zwei Dimensionen: Heatmaps

Eine *Heatmap* überträgt numerische Werte auf eine Farbskala und ordnet sie in einer Matrix an – typischerweise eine Zeitdimension auf der x-Achse und eine Kategorie auf der y-Achse. Dadurch werden zweidimensionale Muster sichtbar, die weder ein Liniendiagramm noch ein Balkendiagramm zeigen könnte: Beide Dimensionen und ihre Wechselwirkung werden gleichzeitig erkennbar.

Beispiel Brick-Flow AG: Abbildung 4.4 zeigt das mittlere tägliche Bestellaufkommen nach Wochentag (Zeilen) und Monat (Spalten). Je dunkler eine Zelle, desto mehr Bestellungen gehen an diesem Wochentag im jeweiligen Monat im Durchschnitt ein.

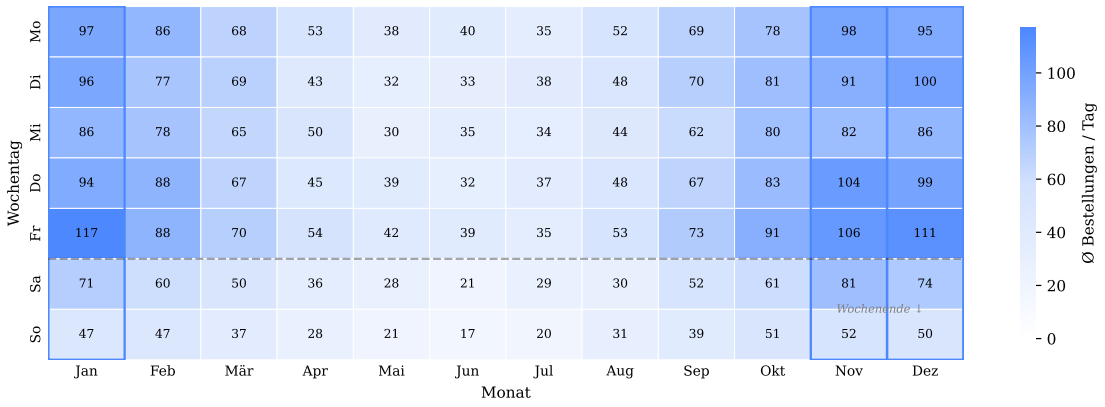


Abbildung 4.4: Mittleres tägliches Bestellaufkommen der Brick-Flow AG nach Wochentag und Monat (Ø Bestellungen/Tag). Saisoneffekt aus den synthetischen Betriebsdaten; Wochentag-Gewichtung illustrativ (B2C+B2B-Muster). Blau umrandete Spalten markieren die Hochsaisonmonate Januar, November und Dezember.

Die Heatmap macht zwei Effekte gleichzeitig sichtbar: Den *Saisoneffekt* entlang der Spalten – Januar, November und Dezember sind durchgehend dunkelblau, die

Sommermonate Mai bis Juli dagegen hell – und den *Wochentag-Effekt* entlang der Zeilen, mit dem Freitag als stärkstem Bestelltag und dem Sonntag als schwächstem. Der stärkste Einzelwert (Fr/Jan: 117 Bestellungen) und der schwächste (So/Jun: 17 Bestellungen) liegen damit fast um Faktor 7 auseinander – ein Unterschied, der in einer aggregierten Monatskennzahl vollständig verschwindet.

Für die Personalplanung des Logistikdienstleisters liefert das eine konkrete Steuerungsgröße: Nicht nur die Hochsaison als Ganzes, sondern besonders die Freitage im November und Dezember sind die kritischen Spitzentage. Heatmaps finden auch als geografische Variante Anwendung – Umsatz oder Retourenquote nach Postleitzahlregionen lassen sich in BI-Tools wie Power BI oder Apache Superset als interaktive Karte visualisieren und decken regionale Muster auf, die aggregierte Kennzahlen verbergen.

4.2.3 Zusammensetzungen und Veränderungen: Das Wasserfalldiagramm

Das *Wasserfalldiagramm* ist im Controlling und in der Finanzanalyse weit verbreitet, in Visualisierungsübersichten aber oft vergessen. Es eignet sich immer dann, wenn ein Gesamtwert aus mehreren Teilbeiträgen – positiven wie negativen – aufgebaut oder erklärt werden soll.

Jeder Balken repräsentiert einen Beitrag zum Gesamtergebnis: Positive Werte heben den laufenden Stand an, negative Werte senken ihn. Gestrichelte Verbindungslinien zeigen, wo der nächste Balken ansetzt. Am Ende steht ein Ergebnisbalken, der direkt auf der Nulllinie beginnt. Klassische Einsatzfelder sind Deckungsbeitragsrechnungen, Jahresabschlussbrücken (Umsatz des Vorjahres → Umsatz des laufenden Jahres, aufgeschlüsselt nach Wachstumstreibern) und Kostenanalysen.

Beispiel Brick-Flow AG: Abbildung 4.5 zeigt eine vereinfachte Deckungsbeitragsrechnung für das dritte Geschäftsjahr. Vom Umsatz in Höhe von 1.800 Tsd. € werden schrittweise die variablen Kosten abgezogen: Wareneinsatz (–900), Logistik und Versand (–270), Retouren und Reklamationen (–90) sowie Marketing und Vertrieb (–135). Der verbleibende Deckungsbeitrag von 405 Tsd. € entspricht einer Marge von rund 22 %.

Der Vorteil gegenüber einer Tabelle liegt darin, dass sowohl die *Größenverhältnisse* der einzelnen Kostenpositionen als auch ihr *Beitrag zur Gesamtmarge* auf einen

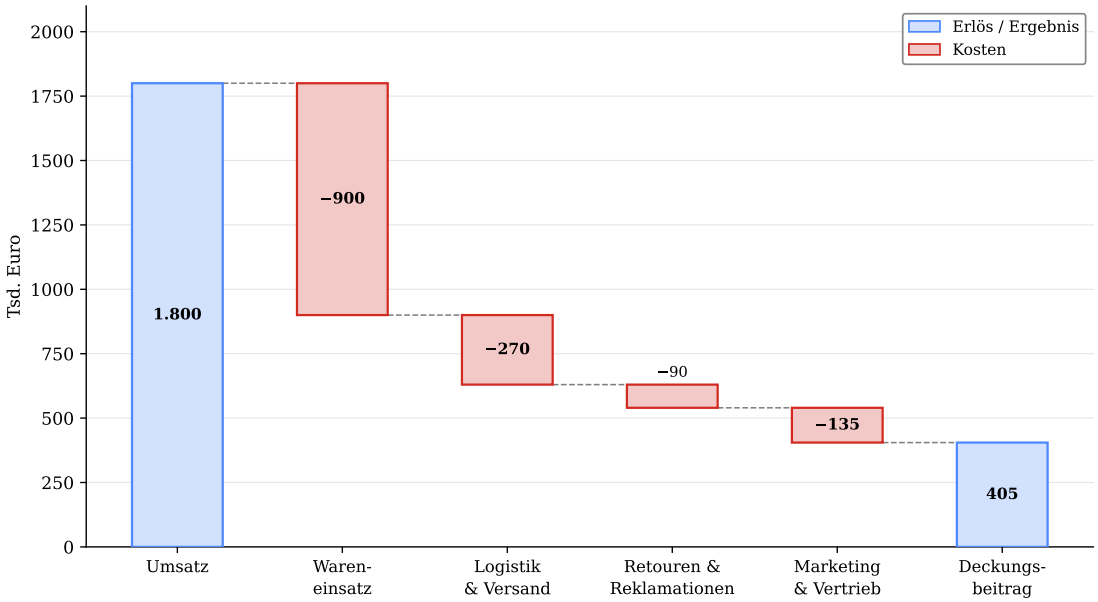


Abbildung 4.5: Deckungsbeitragsrechnung der Brick-Flow AG (Geschäftsjahr 3, illustrativ, in Tsd. €). Blaue Balken stehen für Erlöse und das Ergebnis, rote Balken für Kostenpositionen.

Blick erfassbar sind. Im Beispiel wird sofort deutlich, dass der Wareneinsatz mit Abstand den größten Kostenblock darstellt – und damit der wirksamste Hebel für eine Margensteigerung wäre.

Boxplot und Violin-Plot machen Verteilungen und Streuung sichtbar, die Mittelwertbalken verbergen. Heatmaps decken zweidimensionale Zeitraum- und Kategoriemuster auf. Das Wasserfalldiagramm erklärt, wie ein Ergebnis aus einzelnen Beiträgen entsteht. Jeder dieser Typen beantwortet eine Frage, die mit einem einfachen Balken- oder Liniendiagramm unbeantwortet bliebe.

4.3 Steuerung mit Dashboards und Kennzahlen

Die vorigen Abschnitte haben gezeigt, wie einzelne Diagrammtypen für spezifische Fragestellungen ausgewählt werden. In der betrieblichen Praxis werden diese Visualisierungen selten isoliert betrachtet – sie sind Teil eines strukturierten Informationssystems: des *Dashboards*. Ein Dashboard bündelt mehrere Charts und

Kennzahlen zu einer kohärenten Steuerungssicht, die auf eine bestimmte Zielgruppe und einen konkreten Entscheidungskontext zugeschnitten ist. Die zentrale Designfrage lautet dabei nicht "Welche Daten können wir zeigen?", sondern "Welche Entscheidungen soll dieses Dashboard ermöglichen?"

4.3.1 Dashboards als Steuerungsinstrumente

In vielen Unternehmen werden Daten heute über sogenannte Dashboards bereitgestellt. Diese bündeln relevante Kennzahlen und Visualisierungen und ermöglichen einen schnellen Überblick über zentrale Entwicklungen.

Je nach Steuerungsebene unterscheiden sich Dashboards erheblich in Zielgruppe, Aktualisierungsfrequenz und *Leitfrage*. Tabelle 4.2 zeigt die drei klassischen Ebenen.

Ebene	**Zielgruppe**	**Aktualisierung**	**Leitfrage**	**Typische KPIs**
Operativ	Schichtleitung, Teamleads	Stündlich bis täglich	Läuft der Betrieb heute reibungslos?	Offene Bestellungen, aktuelle Lieferverzögerungen
Taktisch	Bereichsleitungen	Wöchentlich bis monatlich	Erreichen wir unsere Periodenziele?	Umsatz vs. Plan, Retourenquote, Deckungsbeitrag
Strategisch	Geschäftsführung, Vorstand	Monatlich bis quartalsweise	Entwickeln wir uns in die richtige Richtung?	Umsatzwachstum, Marktanteil, Kundenzufriedenheit

Tabelle 4.2: Dashboard-Ebenen nach Steuerungshorizont und Zielgruppe.

Ein operatives Echtzeit-Dashboard nützt einer Geschäftsführerin wenig bei der strategischen Quartalsplanung; ein strategisches Monatsdashboard hilft einer Schichtleitung nicht, kurzfristig Kapazitäten zu disponieren. Die Konsequenz ist, dass Dashboards für ihre Zielgruppe gebaut werden müssen – nicht für alle auf einmal.

Ein häufiger Fehler in der Praxis ist, möglichst viele Kennzahlen auf einer Seite zu bündeln, mit dem Ergebnis, dass das Dashboard keine Frage mehr klar beantwortet.

4.3.2 KPI-Systeme verstehen und kritisch hinterfragen

Key Performance Indicators (KPIs) sind die Grundbausteine jedes Dashboards. Sie verdichten komplexe Sachverhalte zu messbaren Größen und ermöglichen eine systematische Bewertung von Leistung und Entwicklung. Ein gutes KPI-System zeichnet sich dadurch aus, dass es strategische Ziele widerspiegelt, klar definiert und verständlich ist sowie regelmäßig überprüft und angepasst wird.

Gleichzeitig bergen KPIs ein inhärentes Risiko: Sobald eine Kennzahl zum zentralen Steuerungsinstrument wird, entstehen Anreize, genau diese Kennzahl zu optimieren – manchmal auf Kosten des eigentlichen Ziels. Dieses Phänomen ist unter dem Begriff *Goodhart'sches Gesetz* bekannt: *Wird eine Kennzahl selbst zum Ziel, hört sie auf, eine gute Kennzahl zu sein.*

Beispiel Brick-Flow AG: Angenommen, der Logistikdienstleister wird vertraglich auf eine durchschnittliche Lieferzeit von 3,5 Tagen verpflichtet. Wie die Abbildungen 4.2 und 4.3 zeigen, ist die Lieferzeitverteilung in Q4 jedoch stark rechtsschief: Der Mittelwert kann die Vertragsbedingung formal erfüllen, während gleichzeitig ein erheblicher Teil der Bestellungen deutlich längere Lieferzeiten aufweist – für Kundschaft spürbar, in der Kennzahl unsichtbar. Ein wirksamerer Parameter wäre das 90. Perzentil: "90 % aller Bestellungen werden innerhalb von 5 Tagen geliefert." Ähnliche Fehlanreize entstehen, wenn der Kundendienst ausschließlich an der Retourenquote gemessen wird: Dann besteht der Anreiz, Retouren bürokratisch zu erschweren – was die Quote senkt, aber die Kundenzufriedenheit schädigt.

Für das Management folgt daraus, dass KPIs nie isoliert betrachtet werden sollten. Einzelkennzahlen brauchen *Gegenkennzahlen*, die verhindern, dass die Optimierung einer Größe auf Kosten einer anderen geht. Umsatz als alleinige Vertriebskennzahl kann Rabattexzesse begünstigen; erst in Kombination mit dem Deckungsbeitrag entsteht ein ausgewogenes Bild.

4.4 Business-Intelligence-Tools

Die bisher betrachteten Konzepte – insbesondere Kennzahlen, Dashboards und Visualisierungen – werden in der Praxis durch sogenannte Business-Intelligence-(BI)-Tools umgesetzt. Diese Werkzeuge bilden die technologische Grundlage dafür, Daten aus unterschiedlichen Quellen zu integrieren, aufzubereiten und in Form von Berichten oder Dashboards bereitzustellen.

Grundlegend lassen sich BI-Tools als Systeme verstehen, die drei zentrale Funktionen miteinander verbinden: Datenintegration, Datenaufbereitung und Datenvisualisierung. Für das Management bedeutet dies, dass BI-Tools eine Brücke schlagen zwischen operativen Daten und entscheidungsrelevanten Informationen. Sie ermöglichen es, Kennzahlen regelmäßig und konsistent zu berechnen und in einer Form darzustellen, die eine schnelle Interpretation erlaubt.

4.4.1 KPIs, Data Lineage und Nachvollziehbarkeit

BI-Tools übernehmen bei der Unternehmenssteuerung eine entscheidende Rolle: Sie sorgen dafür, dass Kennzahlen *systematisch berechnet, visualisiert und verfügbar gemacht werden*. Dabei ist insbesondere ein Aspekt von großer Bedeutung: die *Nachvollziehbarkeit von Kennzahlen*. In der Praxis reicht es nicht aus, einen KPI anzuzeigen – vielmehr muss klar sein, wie dieser KPI zustande kommt. Genau hier kommt das Konzept der *Data Lineage* ins Spiel.

Unter Data Lineage versteht man die Fähigkeit, den Weg eines Datenelements über alle Verarbeitungsschritte hinweg nachzuvollziehen – von der ursprünglichen Quelle bis zur finalen Kennzahl im Dashboard. Dies umfasst die Herkunft der Daten (z. B. aus einem ERP-System), die Transformationen und Berechnungen sowie die Aggregation zu einer KPI. Für das Management bedeutet dies: *Nur wenn die Entstehung einer Kennzahl nachvollziehbar ist, kann sie als verlässliche Entscheidungsgrundlage dienen.* Abbildung 4.6 zeigt den Verlauf der Daten von verschiedenen Quellen bis beispielsweise zur Anzeige im Dashboard. Durch die zentrale Definition von KPIs und einer Data Lineage in den Systemen, lässt sich die Berechnung transparent zurückverfolgen.

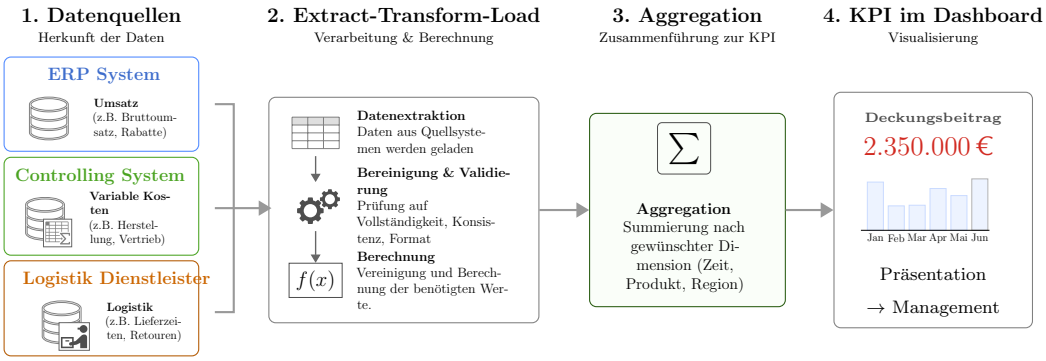


Abbildung 4.6: Data Lineage einer zentral definierten KPI ermöglicht die Rückverfolgung aller Quellen, auf denen die KPI basiert.

Ein typisches Beispiel aus der Praxis ist die Kennzahl "Deckungsbeitrag" im Controlling. Diese KPI wird häufig zentral definiert, um eine einheitliche Steuerung zu gewährleisten. Im BI-Tool wird beispielsweise festgelegt, dass die Umsatzdaten aus dem ERP-System stammen, die variablen Kosten aus einem Controlling-System übernommen werden und sich der Deckungsbeitrag als Differenz dieser Werte ergibt (vgl. Abbildung 4.7). Aus der hier gewählten farblichen Darstellung lässt sich direkt Ableiten aus welchen Systemen die Komponenten zur Berechnung des Deckungsbeitrages kommen.

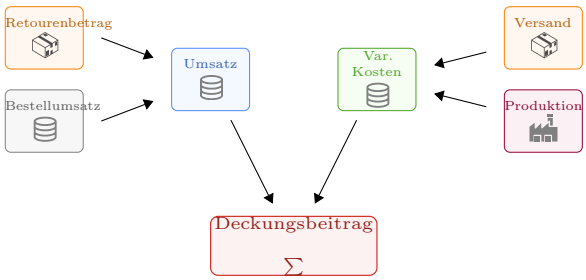


Abbildung 4.7: Definition der KPI *Deckungsbeitrag* in Abhängigkeit von anderen KPIs bzw. Basis-Kennzahlen aus unterschiedlichen Systemen.

Durch eine implementierte Data Lineage kann ein Controller von einem aggregierten Wert im Dashboard schrittweise bis zu einzelnen Transaktionen navigieren und prüfen, aus welchen Systemen die Daten stammen und wie sich die Kennzahl im Detail zusammensetzt. Dies hat zwei zentrale Vorteile:

- **Vertrauen in die Daten** wird gestärkt, da die Berechnung transparent ist,

- **Fehler können schneller identifiziert werden**, etwa bei falschen Zuordnungen oder unvollständigen Daten.

BI-Tools unterstützen diese Transparenz häufig durch Funktionen wie Drill-Down (vom KPI zu Detaildaten), Dokumentation von Berechnungslogiken und Visualisierung von Datenflüssen. Ohne Data Lineage besteht die Gefahr, dass Kennzahlen zwar genutzt, aber nicht verstanden werden – ein häufig unterschätztes Risiko in datengetriebenen Organisationen.

Beispiel Brick-Flow AG: Das Controlling erzeugt täglich einen Bericht, der akkumulierte Umsatzzahlen, Umsätze des letzten Werktags und die zugehörigen Deckungsbeiträge liefert. Durch einen Absturz eines Datenbanksystems in der Nacht ist den Verantwortlichen im IT Bereich klar, dass die Kosten nicht aktualisiert wurden. Durch Data Lineage ist es jetzt möglich herauszufinden, welche KPIs darauf basieren und welche Berichte/Dashboards daher vorrübergehend keine richtigen Daten liefern.

In einem anderen Fall wird bei der Brick-Flow AG im Controlling die KPI *Umsatz* verwendet. Bei einem Marketing Meeting wird überlegt, ob es sich dabei um den Umsatz nach oder vor Bereinigung von Retouren handelt. Durch Data Lineage lässt sich hier die genaue Definition direkt im System dokumentieren und abrufen.

4.4.2 Self-Service BI und Governance

Ein wichtiger Trend in den letzten Jahren ist die Entwicklung hin zu sogenannter Self-Service Business Intelligence. Darunter versteht man Ansätze, bei denen Fachbereiche eigenständig auf Daten zugreifen und Analysen durchführen können. Dies ermöglicht schnellere Analysen, größere Flexibilität und eine stärkere Einbindung der Fachbereiche.

Gleichzeitig entsteht jedoch ein Spannungsfeld: Wenn viele Nutzer eigene Kennzahlen definieren, besteht die Gefahr, dass unterschiedliche Versionen derselben KPI entstehen. Gerade vor diesem Hintergrund gewinnt Data Lineage zusätzlich an Bedeutung – sie hilft sicherzustellen, dass zentrale Kennzahlen einheitlich definiert bleiben, Abweichungen erkennbar sind und Analysen nachvollziehbar bleiben.

Für das Management ergibt sich daraus eine klare Implikation: BI-Tools und Self-Service BI sind nicht nur technische Lösungen, sondern zentrale Elemente der Unternehmenssteuerung. Ihr Nutzen hängt maßgeblich davon ab, ob es gelingt, Transparenz, Konsistenz und Flexibilität in Einklang zu bringen. Self-Service BI erfordert daher immer eine Kombination aus Flexibilität und klaren Governance-Strukturen.

> Kennzahlen sind nur dann verlässlich, wenn ihre Entstehung nachvollziehbar ist – Data Lineage macht diese Transparenz sichtbar und schafft Vertrauen in datenbasierte Entscheidungen.

4.5 Praxisbeispiel: BI-Dashboard bei der Brick-Flow AG

Die bisher besprochenen Konzepte – Kennzahlen, Dashboards, Data Lineage und Self-Service BI – lassen sich anhand eines konkreten Anwendungsfalls veranschaulichen. Die fiktive *Brick-Flow AG* betreibt einen Online-Shop für Baukastensysteme und beliefert neben Endkunden auch Fachhandelpartner im B2B-Bereich. Das Controlling der Brick-Flow AG steht vor einer typischen Herausforderung: Umsatz- und Bestelldaten liegen in verschiedenen Systemen vor, Auswertungen werden bisher manuell in Excel zusammengestellt und der Vertrieb wartet wöchentlich auf Reports der IT-Abteilung.

Ziel ist es, ein zentrales BI-Dashboard einzuführen, das den Fachbereichen jederzeit aktuelle Steuerungsinformationen bereitstellt – ohne jede Auswertung als IT-Ticket einreichen zu müssen.

4.5.1 Fragestellungen und Kennzahlen

Aus Controlling- und Vertriebssicht stellt die Brick-Flow AG folgende typische Fragen an ihre Datenbasis:

- Welche Produktkategorien (z. B. Fahrzeuge, Architektur, Technik-Sets) erzielen den höchsten Deckungsbeitrag?
- In welchen Regionen liegen Händlerbestellungen unter dem geplanten Zielwert?

- Wie entwickelt sich die Retourenquote im Vergleich zwischen Online-Endkundengeschäft und B2B-Handel?
- Welche Artikel weisen eine überdurchschnittliche Stornoquote auf?

Diese Fragen lassen sich alle auf Basis operativer Transaktionsdaten beantworten – vorausgesetzt, die Daten sind integriert, bereinigt und in einem konsistenten Modell verfügbar. Für das Dashboard werden dazu zunächst die zentralen Steuerungsgrößen definiert. Tabelle 4.3 zeigt eine Auswahl:

KPI	**Berechnung**	**Steuerungsrelevanz**
Umsatz (brutto)	Summe Verkaufspreise	Gesamtentwicklung Vertrieb
Deckungsbeitrag	Umsatz − variable Kosten	Rentabilität je Kategorie/Kanal
Retourenquote	Retouren / Bestellungen	Qualität und Kundenzufriedenheit
Stornoquote	Stornos / Bestellungen	Lieferprobleme, Sortimentsfehler
Durchschnittlicher Bestellwert	Umsatz / Anzahl Bestellungen	Sortimentssteuerung
Planerfüllung Händler	Ist-Umsatz / Planumsatz	Zielerreichung im B2B-Bereich

Tabelle 4.3: Ausgewählte KPIs im Brick-Flow-Vertriebsdashboard

Jede dieser Kennzahlen muss im BI-System eindeutig definiert sein. Wird etwa die Retourenquote einmal auf Basis der Bestellpositionen und einmal auf Basis der Bestellköpfe berechnet, entstehen inkonsistente Werte – ein klassisches Problem, das durch klare Governance und Data Lineage vermieden wird.

Die bereits zu Beginn des Kapitels gezeigte Umsatzentwicklung (Abbildung 4.1) ist genau ein solches Dashboard-Element: Im BI-Tool entsteht sie automatisch aus den Transaktionsdaten und wird bei jedem neuen Geschäftsmonat fortgeschrieben. Die dort eingezeichnete Trendgerade ist ein einfaches Beispiel dafür, wie ein Dashboard über die reine Darstellung hinaus auch analytische Elemente integrieren kann – ein Übergang, der zu den prognoseorientierten Verfahren in Kapitel 5 überleitet.

4.5.2 BI-Tools im Marktüberblick

Der Markt für BI-Software ist breit und reicht von leistungsstarken kommerziellen Plattformen bis hin zu frei verfügbaren Open-Source-Lösungen. Tabelle 4.4 gibt einen Überblick über die in der Praxis am häufigsten eingesetzten Werkzeuge.

Kommerzielle Werkzeuge wie Power BI oder Tableau sind in großen Unternehmen weit verbreitet: Sie bieten tiefe Integrationen in bestehende IT-Landschaften, professionellen Herstellersupport und eine breite Nutzercommunity. Ihr Nachteil liegt in den Lizenzkosten sowie in einer gewissen Abhängigkeit vom jeweiligen Anbieter – ein Wechsel ist aufwändig, da Datenmodelle, Berichte und Nutzerschulungen an das Werkzeug gebunden sind.

Open-Source-Lösungen wie Apache Superset oder Metabase haben in den letzten Jahren stark an Verbreitung gewonnen. Sie bieten viele vergleichbare Funktionen ohne Lizenzkosten, erfordern aber in der Regel mehr technischen Aufwand beim Betrieb und kommen ohne kommerzielle Supportzusagen aus. Für viele Unternehmen ist der entscheidende Vorteil die *Unabhängigkeit*: Der Quellcode ist einsehbar, anpassbar und nicht an die Geschäftsstrategie eines einzelnen Anbieters gebunden.

Unabhängig vom gewählten Werkzeug gilt: Die zugrunde liegenden Konzepte – KPI-Definition, Datenmodellierung, Dashboard-Gestaltung und Data Lineage – sind auf alle Plattformen übertragbar. Wer diese Grundlagen versteht, findet sich in Power BI ebenso wie in Superset zurecht.

4.5.3 Open-Source-BI-Tool: Apache Superset

Für die technische Umsetzung des Dashboards und die Übungsaufgaben dieses Kapitels liegt der Fokus auf Apache Superset.[1] Es wurde ursprünglich bei Airbnb entwickelt und wird heute von zahlreichen Unternehmen produktiv eingesetzt.

Superset bietet eine browserbasierte Oberfläche zur Erstellung von Diagrammen, Dashboards und SQL-basierten Analysen. Fachbereiche können ohne Programmierkenntnisse eigene Auswertungen erstellen – sofern die Datenbasis entsprechend

[1] Apache Superset ist ein Open-Source-Projekt der Apache Software Foundation: `https://superset.apache.org`

Tool	Anbieter	Lizenz / Kosten	Besonderheiten
Kommerzielle Lösungen			
Microsoft Power BI	Microsoft	Kommerziell (Desktop kostenlos, Pro ca. 10 €/Monat)	Sehr weite Verbreitung, tiefe Microsoft-365-Integration; Einstieg für Excel-Nutzerinnen und -Nutzer niedrigschwellig
Tableau	Salesforce	Kommerziell (Subscription)	Branchenstandard für anspruchsvolle Datenvisualisierung; besonders starke Drag-and-Drop-Oberfläche
Qlik Sense	Qlik	Kommerziell (Subscription)	Assoziatives Datenmodell erlaubt flexible Ad-hoc-Analysen; im DACH-Raum gut etabliert
SAP Analytics Cloud	SAP	Kommerziell (Subscription)	Native Integration in SAP S/4HANA; für Unternehmen mit SAP-Systemlandschaft relevant
Open-Source-Lösungen			
Apache Superset	Apache Foundation	Open Source (Apache 2.0)	SQL-nativ, flexibel und erweiterbar; weit verbreitet in datenaffinen Unternehmen; kommerziell als Preset verfügbar
Metabase	Metabase Inc.	Open Source (BSL, teilw.)	Sehr einfache Bedienung auch ohne SQL-Kenntnisse; für Fachbereiche mit wenig technischem Hintergrund geeignet

Tabelle 4.4: Ausgewählte BI-Tools im Marktüberblick – kommerzielle und Open-Source-Lösungen.

vorbereitet wurde. Typische Funktionen, die für das Brick-Flow-Szenario relevant sind:

- **Diagramme**: Umsatz- und Deckungsbeitragsverläufe als Linien- oder Balkendiagramm
- **Karten**: regionale Auswertung der Händlerplanerfüllung
- **Filter**: interaktive Auswahl nach Zeitraum, Produktkategorie oder Vertriebskanal
- **Drill-Down**: vom aggregierten Monatsumsatz bis zur Einzelbestellung

Tabelle 4.5 zeigt einen kurzen Vergleich von Superset mit *Metabase*, einem weiteren verbreiteten Open-Source-BI-Werkzeug:

	Apache Superset	**Metabase**
Zielgruppe	Analyst:innen, technische Fachbereiche	Einsteiger:innen, nicht-technische Nutzer:innen
Bedienung	Flexibel, aber Einarbeitungsaufwand	Sehr einfache Oberfläche
SQL-Unterstützung	Vollständig, SQL-Editor integriert	Einfache Abfragen ohne SQL möglich
Datenbankanbindung	Viele Datenbanken unterstützt	Ebenfalls breit, etwas eingeschränkter
Selbst-Hosting	Ja (Docker empfohlen)	Ja (auch als Cloud-Version verfügbar)
Lizenz	Apache 2.0 (vollständig Open Source)	Business-Source-Lizenz (teilweise)

Tabelle 4.5: Vergleich ausgewählter Open-Source-BI-Werkzeuge

Für die Brick-Flow AG bietet sich Superset an, wenn ein Analyst:innenteam mit SQL-Kenntnissen vorhanden ist und Flexibilität bei der Datenmodellierung benötigt wird. Metabase ist die bessere Wahl, wenn die Fachbereiche weitgehend selbstständig und ohne technische Vorkenntnisse arbeiten sollen.

4.5.4 Einordnung: Was ein BI-Tool leisten kann – und was nicht

Das Brick-Flow-Beispiel verdeutlicht, dass ein BI-Tool allein noch kein BI schafft. Die eigentliche Arbeit liegt in der Datenintegration, der konsistenten KPI-Definition und der Sicherstellung von Datenqualität – Aspekte, die in Kapitel 2 und 3 behandelt wurden.

Ein BI-Tool übernimmt die Visualisierung und die benutzerfreundliche Bereitstellung. Es setzt aber voraus, dass:

- Daten aus den Quellsystemen (Shop, Warenwirtschaft, CRM) zusammengeführt wurden,
- Kennzahlen einheitlich und nachvollziehbar definiert sind,
- und Zugriffsrechte und Governance geregelt sind.

Ein BI-Dashboard ist nur so gut wie die Datenbasis, auf der es aufbaut. Ob kommerziell (Power BI, Tableau) oder Open Source (Apache Superset, Metabase) – die eigentliche Wertschöpfung entsteht durch saubere Daten, klare KPI-Definitionen und organisatorische Rahmenbedingungen, nicht durch das Werkzeug allein.

Weiterführende Literatur

- *Storytelling with Data: A Data Visualization Guide for Business Professionals* [11]: Praxisnahes Standardwerk zur adressatengerechten Kommunikation mit Daten.
- *The Visual Display of Quantitative Information* [23]: Grundlegendes Werk zu Prinzipien quantitativer Visualisierung.
- *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring* [7]: Vertiefung zu Dashboard Design und entscheidungsorientierter Informationsdarstellung.

Kapitel 5

Data Science und Machine Learning im Business

In den vorherigen Kapiteln wurde deutlich, wie Daten in Unternehmen entstehen, verarbeitet und für Entscheidungen genutzt werden. Aufbauend darauf stellt sich nun die Frage, wie moderne Analyseverfahren diese Möglichkeiten erweitern. Insbesondere Data Science und Machine Learning eröffnen neue Wege, Daten nicht nur zu beschreiben, sondern aktiv für Prognosen, Mustererkennung und Entscheidungsunterstützung einzusetzen.

In der Praxis sind diese Begriffe häufig mit hohen Erwartungen verbunden. Gleichzeitig besteht oft Unsicherheit darüber, wann ihr Einsatz sinnvoll ist und welche Voraussetzungen erfüllt sein müssen. Ziel dieses Kapitels ist es daher, ein *konzeptionelles Verständnis aus Managementperspektive* zu vermitteln – von der begrifflichen Einordnung über die Grundlogik des maschinellen Lernens bis hin zu Anwendungsfeldern, Risiken und organisatorischen Voraussetzungen.

5.1 Von Business Analytics zu Data Science

Business Analytics und Data Science werden häufig synonym verwendet, verfolgen jedoch unterschiedliche Schwerpunkte. Business Analytics beschreibt die systematische Nutzung von Daten zur Unterstützung betrieblicher Entscheidungen. Im Fokus stehen dabei meist strukturierte Daten, klar definierte Fragestellungen und etablierte Methoden wie Berichte, Kennzahlen oder Prognosen.

Data Science geht darüber hinaus und ist stärker explorativ ausgerichtet. Hier liegt der Fokus auf der *Entdeckung neuer Muster und Zusammenhänge*, häufig auch in großen oder unstrukturierten Datenmengen. Während Business Analytics bestehende Fragen beantwortet, zielt Data Science darauf ab, *neue Erkenntnisse zu generieren und Hypothesen zu entwickeln.*

Für das Management ist diese Unterscheidung relevant, da sie unterschiedliche Arbeitsweisen erfordert. Business Analytics ist oft fest in bestehende Prozesse integriert, während Data Science eher projektbasiert und experimentell organisiert ist. Beide Ansätze ergänzen sich und bilden gemeinsam die Grundlage datengetriebener Entscheidungen.

Ein zentrales Werkzeug der Data Science ist *Machine Learning*: Verfahren, die aus historischen Daten lernen und dieses Wissen nutzen, um neue Daten zu bewerten oder Vorhersagen zu treffen. Wichtig ist zu verstehen, dass Machine Learning keine „intelligente Entscheidungïm menschlichen Sinne trifft – die Modelle basieren auf statistischen Mustern und sind daher stark von der Qualität und Struktur der zugrunde liegenden Daten abhängig.

5.2 Grundlagen des maschinellen Lernens

Um Machine-Learning-Verfahren einordnen und bewerten zu können, sind zwei grundlegende Konzepte hilfreich: das *Lernparadigma* beschreibt, wie ein Modell trainiert wird; die *Lernaufgabe* beschreibt, welche Art von Ausgabe es erzeugen soll. Beide Konzepte zusammen ermöglichen es, eine betriebliche Fragestellung strukturiert einzuordnen – ohne die technischen Details der Algorithmen kennen zu müssen.

5.2.1 Lernparadigmen: Supervised und Unsupervised Learning

Ein grundlegendes Strukturmerkmal von Machine-Learning-Verfahren ist die Frage, ob beim Training des Modells bekannte Ergebnisse vorliegen oder nicht.

Beim *überwachten Lernen* (Supervised Learning) werden Modelle anhand von Daten trainiert, zu denen die korrekten Ergebnisse bereits bekannt sind. Das Modell lernt,

Eingaben auf bekannte Ausgaben abzubilden, und kann dieses Wissen anschließend auf neue, unbekannte Daten anwenden.

Beispiel Brick-Flow AG: Das Unternehmen möchte vorhersagen, ob ein Kunde seinen Warenkorb abschließen oder abbrechen wird. Aus historischen Bestelldaten ist bekannt, welche Sitzungen zu einem Kauf geführt haben und welche nicht. Ein Modell kann daraus lernen, welche Merkmale – etwa Verweildauer, angesehene Kategorien, Tageszeit – einen Abschluss begünstigen.

Beim *unüberwachten Lernen* (Unsupervised Learning) liegen keine vorgegebenen Ergebnisse vor. Das Modell analysiert die Datenstruktur eigenständig und identifiziert Muster, Gruppen oder Ausreißer.

Beispiel Brick-Flow AG: Das Marketingteam möchte verstehen, welche Kundentypen es gibt – ohne vorab festzulegen, wie viele Gruppen existieren oder wie sie sich unterscheiden. Ein Clusteringverfahren kann aus Kaufverhalten, Bestellhäufigkeit und Produktpräferenzen eigenständig Kundensegmente ermitteln.

	Supervised Learning	**Unsupervised Learning**
Trainingsdaten	mit bekannten Ergebnissen (Labels)	ohne vorgegebene Ergebnisse
Ziel	Vorhersagen für neue Daten treffen	Struktur und Muster entdecken
Typische Aufgaben	Prognose, Klassifikation	Segmentierung, Anomalieerkennung
Voraussetzung	historische Daten mit Zielvariable	ausreichend Daten, Interpretierfähigkeit
Brick-Flow-Beispiel	Kaufabbruch vorhersagen	Kundensegmente entdecken

Tabelle 5.1: Gegenüberstellung von Supervised und Unsupervised Learning

Für die Praxis ist diese Unterscheidung relevant, weil sie bestimmt, welche Daten benötigt werden und welche Fragen überhaupt gestellt werden können. Supervised Learning setzt voraus, dass vergangene Ergebnisse erfasst und verfügbar sind – eine Anforderung, die nicht immer erfüllt ist.

5.2.2 Lernaufgaben: Was soll das Modell ausgeben?

Eine Lernaufgabe beschreibt, welche Art von Ausgabe ein Modell erzeugen soll. Wer eine Fragestellung als Lernaufgabe formulieren kann, ist in der Lage, das passende Verfahren zu wählen, Erwartungen an die Ergebnisse zu formulieren und diese kritisch zu bewerten – ohne die mathematischen Details der Algorithmen zu kennen.

Auch, wenn wir in diesem Lehrbrief nicht auf die genauen mathematischen Details von maschinellen Lernenverfahren eingehen wollen, ist eine grundlegende mathematische Formulierung der Probleme, die mit Lernverfahren gelöst werden hilfreich. Die Algorithmen gehen grundsätzlich von einer Darstellung der realen Welt in Form von messbaren Daten aus. Diese Daten beschreiben Merkmale von Vorgängen, Ereignissen, Kunden oder anderen Sachverhalten im Geschäftskontext:

reale Welt $\longrightarrow$

Datum	Bestellungen	Umsatz	Lieferzeit
2024-01-01	79	14567.6	4.62
2024-01-02	93	17149.2	5.07
2024-01-03	81	14936.4	4.61
2024-01-04	88	16227.2	4.9
2024-01-05	108	19915.2	5.55

Abbildung 5.1: Anzahl und Umsatz von Bestellungen der Brick-Flow AG aggregiert nach Tag.

In Abbildung 5.1 ist dies am Beispiel von aggregierten Kennzahlen (z.B. aus einem BI Tool, vgl. Kapitel 4) dargestellt. Mit Hilfe eines mathematischen Modells soll nun eine Kennzahl für die Steuerung des Geschäftes prognostiziert werden. Ein solches Modell kann als mathematische Funktionen dargestellt werden, die anhand von *Eingaben* $\mathbf{x}$ eine *Vorhersage* $\hat{y}$ berechnen:

$$f(\mathbf{x}) = \hat{y}$$

Die Eingabe $\mathbf{x}$ enthält dabei die Merkmale, auf deren Grundlage die Berechnung der Vorhersage erfolgt. In Abbildung 5.2 ist das Vorgehen nochmal grafisch dargestellt: Über einen Algorithmus wird auf historischen Daten (=Trainingsdaten) eine optimale Funktion f gesucht (=Lernen), die dann auf neuen Daten für die Vorhersage benutzt werden kann.

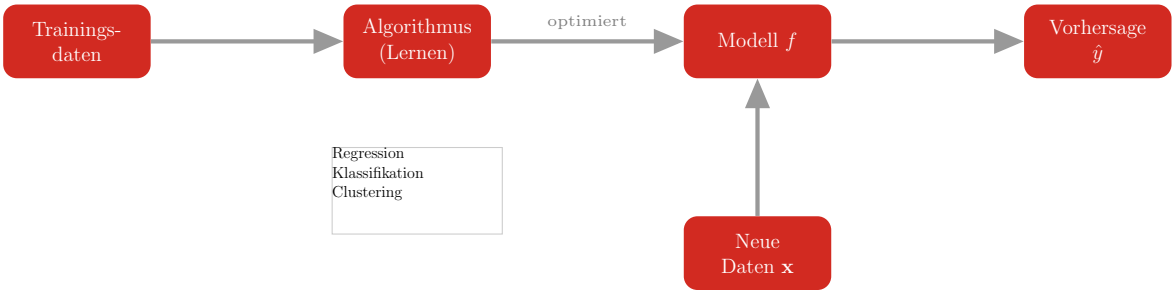


Abbildung 5.2: Schema: Maschinelles Lernen *optimiert* eine Funktion f als Vorhersagemodell für neue Daten.

Regression: Wie viel?

Regressionsmodelle erzeugen als Ausgabe einen *numerischen Wert*. Sie eignen sich für Fragestellungen, bei denen eine Mengen- oder Wertgröße vorhergesagt werden soll – etwa: Wie hoch wird der Umsatz im nächsten Quartal sein? Wie lange dauert die Lieferung eines Auftrags? Bezogen auf die mathematische Formulierung eines Vorhersage-Modells ist die Regressionsaufgabe also die Suche einer Funktion

$$f : X \to Y$$

die für z.B. die Merkmale von Wochentag, Monat, Feiertag und anderen den Umsatz für einen Tag vorhersagt. Dafür werden Daten X benötigt, die die Merkmale enthalten und eine Spalte Y, die für die entsprechenden Merkmale den Umsatz enthält. Wenn wir nun Funktion f finden, die anhand der Merkmalsdaten X immer den richtigen Umsatz Y vorhersagt, haben wir ein Prognosemodell.

Beispiel Brick-Flow AG: Für den Versanddienstleister ist eine zentrale Frage: Wie lange dauert die Lieferung – gemessen von der Bestellung bis zur Haustür – an einem Tag mit besonders hohem Bestellaufkommen? Abbildung 5.3 zeigt die Tagesdaten des Geschäftsjahres: Jeder Punkt steht für einen Tag, mit der Anzahl eingegangener Bestellungen auf der x-Achse und der durchschnittlichen Lieferzeit auf der y-Achse. Mit Hilfe der sogenannten *linearen Regression* lässt sich auf historischen Daten nun die Regressionsfunktion $\hat{f}$ berechnen:

$$\hat{f}(x) = 1.79 + 0.035 \cdot x$$

wobei x in diesem Fall die Anzahl der Bestellungen beschreibt. In komplexeren Auswertungen kann dies auch über mehrdimensionale Eingaben $\mathbf{x} = (x_1, \ldots, x_k)$ erfolgen, wobei dann beispielsweise x_1 die Anzahl der Bestellungen an einem Tag enthält, x_2 die Anzahl der Besucher der Webseite an diesem Tag, usw.

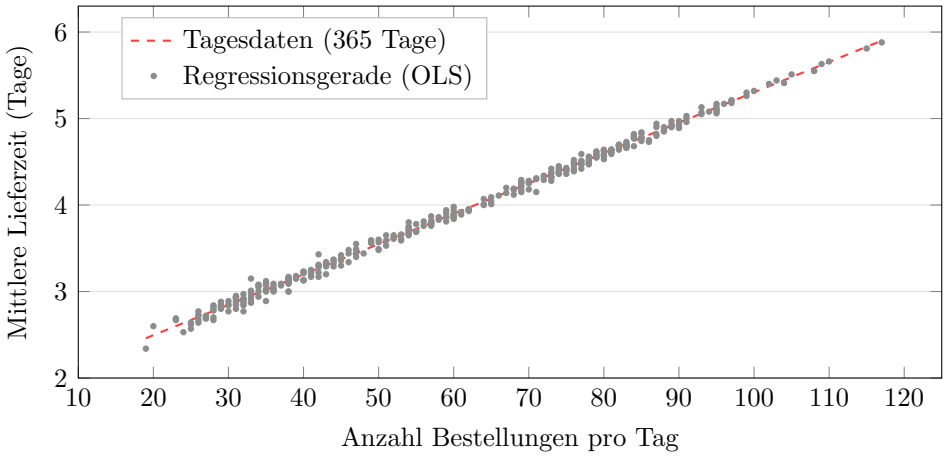


Abbildung 5.3: Zusammenhang zwischen täglichem Bestellaufkommen und mittlerer Lieferzeit der Brick-Flow AG (Geschäftsjahr 2024, $n = 365$ Tage). OLS-Regressionsgerade: $\hat{y} = 1{,}79 + 0{,}035 \cdot x$.

Die eingepasste Regressionsgerade macht den Zusammenhang sichtbar: Pro zusätzliche Bestellung steigt die Lieferzeit im Mittel um rund 0,035 Tage – bei 100 Bestellungen also etwa 3,5 Tage Mehraufwand gegenüber einem ruhigen Tag mit 30 Bestellungen.

Klassifikation: Welche Kategorie?

Klassifikationsmodelle erzeugen als Ausgabe eine *Kategorie* oder Klasse. Sie kommen zum Einsatz, wenn eine Entscheidung zwischen zwei oder mehr vordefinierten Gruppen getroffen werden soll – etwa: Wird dieser Kunde kündigen oder bleiben? Handelt es sich bei dieser Transaktion um einen Betrugsversuch?

Mathematisch unterscheidet sich die Klassifikation von der Regression dadurch, dass die Zielmenge Y keine kontinuierlichen Zahlenwerte, sondern eine endliche Menge von *Klassen* (auch Labels genannt) enthält:

$$f : X \to \{c_1, c_2, \ldots, c_k\}$$

Das Modell f ordnet also jeder Eingabe $\mathbf{x} \in X$ genau eine der k vordefinierten Klassen zu. Die Klassen müssen vor dem Training bekannt und in den Trainingsdaten als Labels vorhanden sein – damit ist Klassifikation stets ein Supervised-Learning-Verfahren.

Ein wichtiger Sonderfall ist die *binäre Klassifikation* mit genau zwei Klassen ($k = 2$), etwa $Y = \{0, 1\}$ für „kein Risiko" vs. „hohes Risiko". Sie ist besonders häufig in der Praxis und bildet die Grundlage vieler Entscheidungsautomatisierungen.

Beispiel Brick-Flow AG: Der Kundendienst möchte Bestellungen schon beim Eingang auf erhöhtes Retourenrisiko prüfen. Das Klassifikationsmodell bildet jede Bestellung – beschrieben durch Merkmale wie Produktkategorie, Bestellwert, Kundenhistorie und Versandregion – auf eine der drei Risikoklassen ab:

$$f(\mathbf{x}) \in \{\text{niedriges Risiko},\ \text{mittleres Risiko},\ \text{hohes Risiko}\}$$

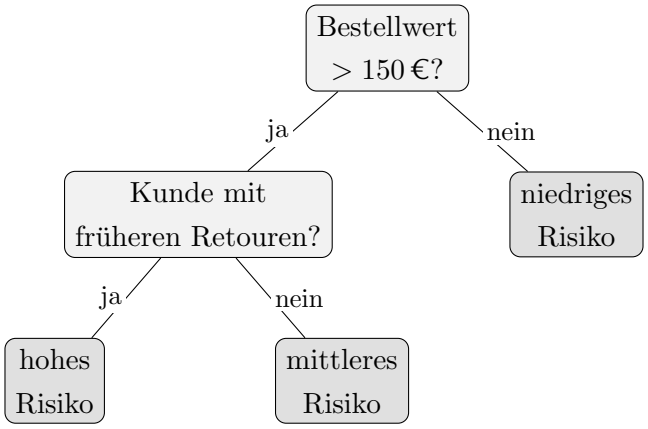
Die Eingabe $\mathbf{x}$ enthält dabei die Merkmale der Bestellung; das Modell f wurde auf historischen Bestellungen trainiert, für die das tatsächliche Retouren-Ergebnis bekannt ist.

Vertiefung: Wie ein Klassifikationsmodell lernt. *Der folgende Abschnitt wirft einen genaueren Blick darauf, was beim „Lernenëines Modells technisch geschieht. Er ist als Vertiefung gedacht und kann beim ersten Lesen übersprungen werden – er hilft jedoch zu verstehen, wie Overfitting – ein zentrales Risiko beim maschinellen Lernen – entsteht und sich erkennen lässt.*

Ein Klassifikationsmodell lässt sich formal als *Funktion* verstehen, die jeder Eingabe genau eine Klasse zuordnet:

$$f:\ \text{Eingabemerkmale} \longrightarrow \text{Klasse}$$

Im Brick-Flow-Beispiel bildet f eine Bestellung – beschrieben durch Merkmale wie Bestellwert, Kundenhistorie und Versandregion – auf eine Risikoklasse ab. Ein *Entscheidungsbaum* ist eine besonders anschauliche Form einer solchen Funktion: Jeder Pfad von der Wurzel zu einem Blatt entspricht einer Regel, die einen Teil der Eingaben einer Klasse zuweist.



„Lernen" bedeutet nun, aus einer ganzen *Familie* $\mathcal{F}$ möglicher Funktionen – hier: allen denkbaren Bäumen – diejenige auszuwählen, die zu den Trainingsdaten am besten passt. Als Maß dient der *Trainingsfehler*: der Anteil der Trainingsbeispiele, die das Modell falsch klassifiziert. Das Training ist damit im Kern ein *mathematisches Optimierungsproblem* – gesucht ist die Funktion f, die den Trainingsfehler minimiert:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left[f(x_i) \neq y_i \right]$$

Dabei sind x_i die Merkmale und y_i die bekannte korrekte Klasse des i-ten Trainingsbeispiels; der Ausdruck in Klammern zählt jede Fehlklassifikation als 1, jede korrekte Zuordnung als 0.

Genau hier liegt jedoch eine Falle. Wird *ausschließlich* der Trainingsfehler minimiert, kann das Modell die Trainingsdaten *auswendig lernen*: Ein hinreichend tiefer Entscheidungsbaum bildet für nahezu jedes einzelne Trainingsbeispiel eine eigene Regel und senkt den Trainingsfehler so bis auf null. Auf neuen, ungesehenen Daten versagt ein solches Modell jedoch, weil es zufällige Eigenheiten der Trainingsdaten statt verallgemeinerbarer Muster gelernt hat. Die Minimierung des Trainingsfehlers ist also *nicht das eigentliche Ziel* – entscheidend ist der Fehler auf neuen Daten.

Trainings- und Testdaten: den Generalisierungsfehler schätzen. Damit stellt sich eine praktische Frage: Wie lässt sich der Fehler auf "neuen Daten" überhaupt beurteilen, *bevor* das Modell im Echtbetrieb eingesetzt wird? Die Antwort besteht darin, die verfügbaren Daten von vornherein in zwei Teile zu zerlegen (vgl. Abbildung 5.4):

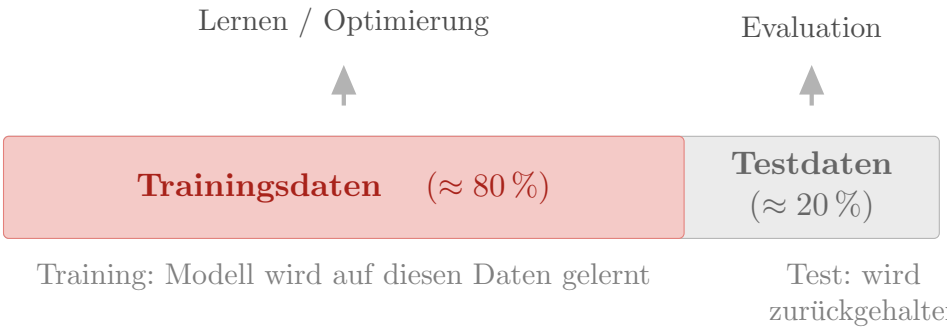


Abbildung 5.4: Aufteilung in Trainings- und Testdaten für die Evaluierung eines Vorhersagemodells.

- Die *Trainingsdaten* dienen dazu, das Modell zu lernen – auf ihnen wird der Trainingsfehler minimiert.
- Die *Testdaten* werden bewusst zurückgehalten und beim Training *nicht* verwendet. Erst nach dem Training wird das fertige Modell auf ihnen geprüft.

Der Fehler, den das Modell auf den zurückgehaltenen Testdaten macht, heißt *Testfehler* oder *Generalisierungsfehler*. Er ist eine Schätzung dafür, wie gut das Modell auf Daten abschneidet, die es noch nie gesehen hat – und damit der eigentlich aussagekräftige Maßstab. Im Brick-Flow-Beispiel würde man etwa die Bestellungen eines vergangenen Zeitraums als Trainingsdaten verwenden und das Modell anschließend an einem späteren, separaten Zeitraum testen.

Mit dieser Unterscheidung wird *Overfitting messbar*: Es liegt genau dann vor, wenn der Trainingsfehler niedrig, der Testfehler aber deutlich höher ist. Das Modell hat dann die Trainingsdaten gut „gelernt", aber nicht das verallgemeinerbare Muster. Die Lücke zwischen Trainings- und Testfehler ist somit das zentrale Diagnosewerkzeug, um die Verlässlichkeit eines Modells einzuschätzen.

Ein niedriger Trainingsfehler allein sagt nichts über die Qualität eines Modells aus. Erst der Testfehler auf zurückgehaltenen Daten zeigt, ob ein Modell generalisiert – eine große Lücke zwischen beiden ist das klarste Anzeichen für Overfitting.

Datenbias: Wenn die Trainingsdaten verzerrt sind. Neben dem Overfitting gibt es eine zweite grundlegende Ursache für Modellversagen, die nicht im Modell selbst liegt, sondern in den Daten: *Datenbias*. Während Overfitting entsteht, weil ein Modell *zu viel* aus den Trainingsdaten lernt, tritt Bias auf, weil die Trainingsdaten die relevante Realität *verzerrt* abbilden. Das Modell reproduziert diese Verzerrung – und gibt sie in seinen Vorhersagen weiter.

Beispiel Brick-Flow AG: Das Retourenmodell wurde auf Bestellungen der vergangenen drei Jahre trainiert. In diesem Zeitraum bestellten Kunden aus städtischen Regionen besonders häufig Paletten mit hohem Premiumfarbenanteil – und retournierten sie seltener. Sind Kunden aus ländlichen Regionen im Trainingsdatensatz unterrepräsentiert, lernt das Modell möglicherweise, dass ein hoher Premiumfarbenanteil kein Risikofaktor ist. Der Testfehler *insgesamt* erscheint akzeptabel – doch für bestimmte Kundengruppen liegt das Modell systematisch falsch.

Datenbias ist deshalb schwerer zu erkennen als Overfitting: Er zeigt sich nicht im globalen Testfehler, sondern erst in einer differenzierten Auswertung nach Subgruppen – etwa nach Region, Produktkategorie oder Kundentyp. Anders als Overfitting lässt sich Datenbias auch nicht allein durch mehr Daten beheben; entscheidend ist, *welche* Daten gesammelt werden. Datenbias ist damit nicht nur ein technisches, sondern auch ein organisatorisches und ethisches Problem: Wer entscheidet, welche Daten erfasst werden? Wessen Perspektive ist in den Trainingsdaten vertreten? Diese Fragen werden in Kapitel 7 vertieft.

Fehlertypen und ihre betriebswirtschaftlichen Kosten. Wie gut ein Klassifikationsmodell ist, lässt sich nicht an einer einzigen Zahl ablesen. Gerade bei binären Entscheidungen ist die *Art* der Fehler entscheidend. Betrachtet man die Frage „hohes Retourenrisiko: ja oder nein?", lassen sich vier Fälle unterscheiden:

Die beiden Fehlertypen haben unterschiedliche betriebswirtschaftliche Konsequenzen:

- Ein *False Negative* (übersehene Retoure) bedeutet, dass keine vorbeugende Maßnahme ergriffen wird – die Retoure samt Rücksendekosten tritt unbemerkt ein.

	Modell sagt: Risiko	**Modell sagt: kein Risiko**
Tatsächlich Retoure	richtig erkannt (True Positive)	übersehen (False Negative)
Tatsächlich keine Retoure	Fehlalarm (False Positive)	richtig erkannt (True Negative)

Tabelle 5.2: Fehlertypen bei der Klassifikation des Retourenrisikos (Konfusionsmatrix)

- Ein *False Positive* (Fehlalarm) löst eine unnötige Maßnahme aus, etwa einen Kundenkontakt oder einen Rabatt, obwohl die Bestellung unproblematisch gewesen wäre.

Welcher Fehler schwerer wiegt, ist keine technische, sondern eine *betriebswirtschaftliche Frage*. Sind die Kosten einer übersehenen Retoure hoch, wird man das Modell so einstellen, dass es im Zweifel eher Alarm schlägt – und umgekehrt. Diese Abwägung lässt sich über die Entscheidungsschwelle des Modells steuern und ist eine typische Schnittstelle zwischen Fachbereich und Data Science.

Clustering: Welche Gruppe?

Clustering-Verfahren sind dem unüberwachten Lernen zuzuordnen. Sie erzeugen als Ausgabe *Gruppen* (Cluster) von Datenpunkten, die sich untereinander ähneln. Anders als bei der Klassifikation sind diese Gruppen nicht vorab definiert – sie entstehen aus der Struktur der Daten selbst.

Mathematisch lässt sich Clustering ebenfalls als Funktion schreiben, die jedem Datenpunkt eine Gruppe zuweist:

$$f : X \to \{1, 2, \ldots, k\}$$

Auf den ersten Blick ähnelt das der Klassifikation – doch der entscheidende Unterschied liegt in der Natur der Ausgabe: Bei der Klassifikation sind die Klassen *vorab benannt und mit Bedeutung belegt* (z. B. „hohes Risiko"). Beim Clustering sind

die Gruppenindizes $1, \ldots, k$ zunächst bedeutungslose Ziffern. *Welche Merkmale eine Gruppe charakterisieren, ergibt sich erst durch die Analyse der zugewiesenen Datenpunkte* – und damit durch menschliche Interpretation, nicht durch das Modell.

Ein weiterer Unterschied: Die Anzahl der Gruppen k ist beim Clustering häufig kein vorgegebener Parameter, sondern muss vom Anwender gewählt oder durch das Verfahren selbst bestimmt werden. Und weil kein Label-Datensatz vorliegt, gibt es auch keinen definierten „Fehler", der minimiert werden könnte – das Modell optimiert stattdessen eine interne Güte, etwa wie kompakt die Gruppen sind und wie weit sie voneinander entfernt liegen.

Beispiel Brick-Flow AG: Das Marketingteam möchte gezielte Kampagnen entwickeln. Als Merkmale werden für jeden Kunden zwei Werte berechnet: die *Blau-Affinität* (Anteil der Käufe in der Kategorie "Blau" am Gesamtumsatz) und die *Diskont-Affinität* (Anteil der Käufe mit Rabattcode). Abbildung 5.5 zeigt die Kundschaft der Brick-Flow AG in diesem zweidimensionalen Merkmalsraum. Jeder Punkt repräsentiert eine Kundin oder einen Kunden; die Farbe gibt das vom Modell zugewiesene Cluster an.

Die Ziffern 1, 2, 3 aus der Modellausgabe erhalten erst durch Betrachtung der zugehörigen Datenpunkte ihre Bedeutung – das Team interpretiert die blaue Wolke (oben rechts) als *Kategorie-Liebhaber*, die rote Wolke (oben links) als *Schnäppchenjäger* und die graue Wolke (Mitte) als *Gelegenheitskäufer*. Diese Interpretation ist nicht Teil des Modells, sondern des betriebswirtschaftlichen Urteils.

Generative Modellierung: Was kommt als Nächstes?

Eine vierte Lernaufgabe hat in den letzten Jahren erheblich an Bedeutung gewonnen: die *generative Modellierung*. Das Grundprinzip ist dabei zunächst unscheinbar: Das Modell lernt, gegeben einem Kontext das wahrscheinlichste nächste Element vorherzusagen – ein Wort, ein Token, ein Zeichen.

Dieses Prinzip liegt Sprachmodellen wie *ChatGPT* oder *Claude* zugrunde. Beide Systeme wurden darauf trainiert, auf Basis enormer Textmengen – Bücher, Webseiten, wissenschaftliche Artikel, Code – vorherzusagen, wie ein Text sinnvoll weitergeführt wird. Die Lernaufgabe lautet also schlicht: *Vervollständige den folgenden Text.*

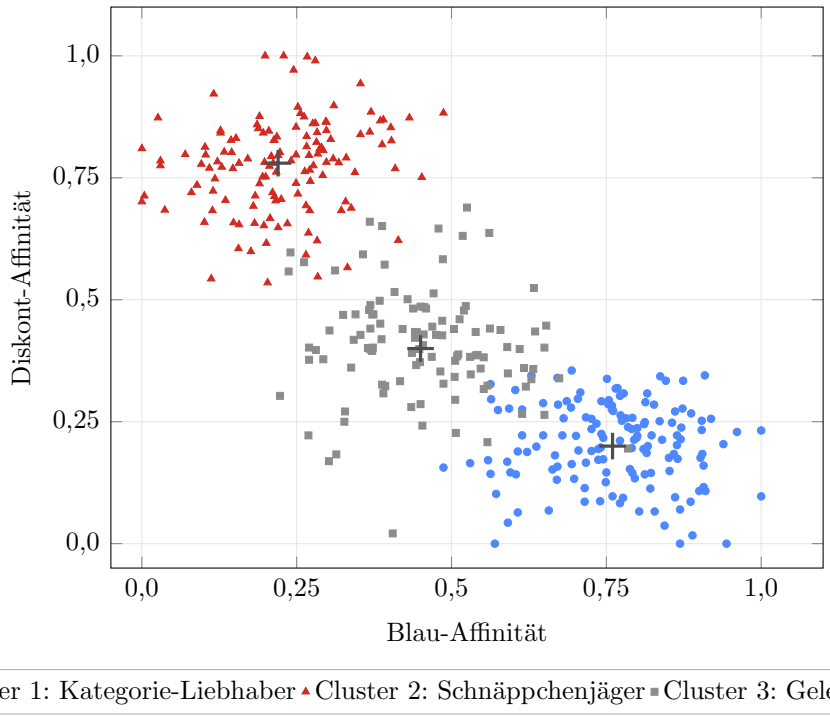


Abbildung 5.5: Kundensegmentierung der Brick-Flow AG anhand von Blau-Affinität und Diskont-Affinität ($n = 380$ Kunden, $k = 3$ Cluster). Die Clusterzentren sind als + markiert.

Aus dieser vergleichsweise einfachen Aufgabe entstehen Fähigkeiten, die beim ersten Blick überraschen: Das Modell kann Fragen beantworten, Texte zusammenfassen, Argumente entwickeln oder Code schreiben. Diese Fähigkeiten wurden nicht explizit trainiert – sie ergeben sich daraus, dass ein Modell, das Sprache gut vorhersagen kann, zwangsläufig etwas über die Struktur von Texten, Argumenten und Wissen gelernt haben muss.

Beispiel: Ein Nutzer gibt in ChatGPT ein: *„Erkläre mir, was ein Deckungsbeitrag ist."* Das Modell vervollständigt diesen Text auf Basis von Millionen ähnlicher Erklärungen in seinen Trainingsdaten – nicht weil es den Begriff „versteht", sondern weil es gelernt hat, wie solche Erklärungen typischerweise aussehen.

Damit wird auch ein zentrales Merkmal – und Risiko – dieser Systeme sichtbar: *Das Modell kann nur wiedergeben, was in seinen Trainingsdaten repräsentiert ist.* Ereignisse nach dem Trainingsabschluss sind ihm unbekannt, und da das Modell keine Ausgabe *verweigert*, sondern stets etwas Wahrscheinliches produziert, entstehen sogenannte *Halluzinationen* – sachlich falsche Aussagen, die sprachlich überzeugend klingen.

Lernaufgaben im Überblick

Die folgende Tabelle zeigt die Formulierung der wichtigsten Lernaufgaben des maschinellen Lernens, sowie deren Zielsetzung und Einordnung.

Lernaufgabe	**Ausgabe**	**Typische Frage**	**Lernparadigma**
Regression	Zahl	Wie viel? Wann? Wie lange?	Supervised
Klassifikation	Kategorie	Welche Klasse? Ja oder Nein?	Supervised
Clustering	Gruppe	Welches Segment?	Unsupervised
Gen. Modellierung	Text, Bild ...	Was kommt als Nächstes?	Supervised (auf Sequenzen)

Tabelle 5.3: Überblick über grundlegende Lernaufgaben im Machine Learning

5.3 Data-Science-Projekte in der Praxis

Die Grundlagen des maschinellen Lernens – Lernparadigmen, Lernaufgaben, Overfitting und Datenbias – beschreiben, was Modelle leisten können und wo ihre strukturellen Grenzen liegen. In der betrieblichen Wirklichkeit stellt sich jedoch die weiterführende Frage: Wie wird aus einer methodischen Möglichkeit ein nutzbares Ergebnis? Dieser Abschnitt betrachtet Data Science aus der Projektperspektive – von der strukturierten Vorgehensweise über typische Anwendungsfelder bis hin zu den organisatorischen Voraussetzungen für einen nachhaltigen Einsatz.

5.3.1 Der Data-Science-Projektzyklus: CRISP-DM

Die Entwicklung eines ML-Modells folgt keinem einmaligen linearen Ablauf, sondern einem *iterativen Prozess*. Für das Management ist es wichtig, diesen Prozess zu kennen – nicht um ihn selbst durchzuführen, sondern um realistische Erwartungen zu haben, Projekte begleiten zu können und an den richtigen Stellen Entscheidungen zu treffen.

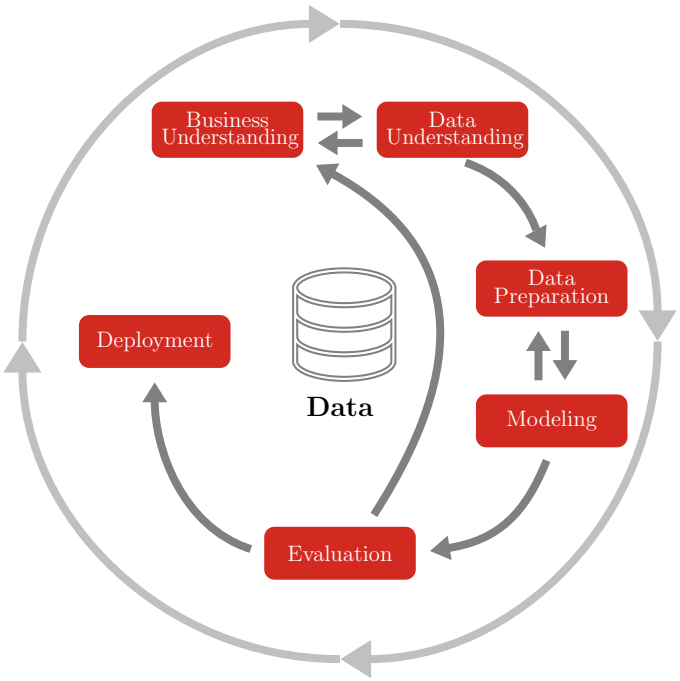


Abbildung 5.6: Der CRISP-DM Prozess für eine standardisierte Vorgehensweise bei Data Mining/Data Analytics Projekten.

Ein weit verbreitetes Referenzmodell ist *CRISP-DM* (Cross-Industry Standard Process for Data Mining). CRSIP-DM ist ein Vorgehensmodell, dass in Kooperation von verschiedenen Business Analytics, Data Mining Anbieter und Unternehmensberatungen (u.a. Teradata, ISL/SPSS, Cap Gemini) mit Anwenderunternehmen (z.B Daimler AG) im Rahmen eines EU Projektes entwickelt wurde. Es beschreibt sechs Phasen, die zyklisch durchlaufen werden und an einigen Stellen Rücksprünge erlaubt, um z.B. Erkenntnisse aus einer späteren Phase neu bewerten zu können. Abbildung 5.6 zeigt eine grafische Darstellung der CRISP-DM Phasen.

Jede Phase hat eine spezifische Zielrichtung:

1. **Business Understanding**: Was soll das Modell leisten? Welches Geschäftsproblem wird gelöst? Diese Phase ist die wichtigste – eine unklare Fragestellung führt unweigerlich zu einem nutzlosen Modell.
2. **Data Understanding**: Welche Daten sind verfügbar? Wie vollständig und verlässlich sind sie? Welche Merkmale könnten relevant sein?
3. **Data Preparation**: Bereinigung, Transformation und Aufbereitung der Daten – in der Praxis der aufwändigste Schritt.
4. **Modeling**: Auswahl und Training geeigneter Verfahren. In dieser Phase entstehen erste Modelle und werden verglichen.
5. **Evaluation**: Beurteilung der Modellgüte. Liefert das Modell die gewünschte Leistung? Beantwortet es die ursprüngliche Frage?
6. **Deployment**: Integration des Modells in Prozesse und Systeme sowie Betrieb im Echtbetrieb.

Beispiel Brick-Flow AG: Das Modell zur Retourenvorhersage (vgl. Abschnitt 5.2.2) durchläuft alle sechs Phasen. Nach der Evaluation stellt sich heraus, dass für Artikel einer bestimmten Kategorie kaum Trainingsdaten vorliegen – das Projekt kehrt zur Phase *Data Understanding* zurück, bevor ein überarbeitetes Modell erneut evaluiert wird.

In der Praxis scheitern Data-Science-Projekte häufig nicht an der Modellierungstechnik, sondern an organisatorischen und konzeptionellen Problemen: unklaren Fragestellungen, überschätzter Datenverfügbarkeit, fehlender Zielvariable oder mangelnder Integration der Ergebnisse in operative Entscheidungen.

Der ML-Prozess beginnt und endet mit dem Geschäftsproblem – nicht mit der Technologie. Die häufigsten Ursachen für das Scheitern von Data-Science-Projekten sind unklare Fragestellungen, mangelhafte Daten und fehlende Integration in Prozesse.

5.3.2 Anwendungsfelder im Unternehmen

Machine Learning wird heute in vielen Unternehmensbereichen eingesetzt. Besonders häufig finden sich Anwendungen dort, wo große Datenmengen vorliegen und wiederkehrende Entscheidungsstrukturen existieren. Tabelle 5.4 gibt einen Überblick über typische Einsatzbereiche und die jeweils geeigneten Lernaufgaben.

Unternehmensbereich	**Typische Fragestellung**	**Lernaufgabe**
Vertrieb & Marketing	Welche Kunden springen ab? Welches Angebot passt am besten?	Klassifikation, Clustering
Logistik & Produktion	Wie entwickelt sich die Nachfrage? Wann droht eine Störung?	Regression, Klassifikation
Finanzen & Compliance	Handelt es sich um eine verdächtige Transaktion?	Klassifikation
Personalwesen	Welche Bewerberinnen und Bewerber passen strukturell am besten?	Clustering, Klassifikation
Kundenservice	Welches Anliegen hat die Kundin? Wie hoch ist die Dringlichkeit?	Klassifikation

Tabelle 5.4: Typische Anwendungsfelder von Machine Learning im Unternehmen

Ein gemeinsames Merkmal dieser Anwendungsfälle ist, dass sie auf operativen Daten basieren und darauf abzielen, Entscheidungen zu verbessern oder Prozesse zu automatisieren. Der Mehrwert entsteht dabei nicht durch die Methode selbst, sondern durch ihre *gezielte Anwendung auf relevante Geschäftsprobleme.*

Beispiel Brick-Flow AG: Im Vertrieb identifiziert ein Klassifikationsmodell Kunden mit erhöhtem Retourenrisiko bereits beim Bestelleingang. Das Marketingteam nutzt Clustering, um differenzierte Kundensegmente für gezielte Kampagnen abzuleiten. In der Logistik hilft ein Regressionsmodell, das tägliche Bestellaufkommen zu antizipieren und die Kapazitätsplanung vorausschauend zu steuern.

5.3.3 Chancen, Grenzen und Risiken

Der Einsatz von Data Science und Machine Learning bietet Unternehmen erhebliche Potenziale: Entscheidungen können stärker datenbasiert getroffen, Prozesse automatisiert und neue Zusammenhänge entdeckt werden. Dadurch können Unternehmen schneller reagieren und Wettbewerbsvorteile aufbauen.

Gleichzeitig sind die Grenzen datengetriebener Modelle zu berücksichtigen. Die in Abschnitt 5.2.2 eingeführten Konzepte – Overfitting und Datenbias – beschreiben zwei grundlegend verschiedene Ursachen für Modellversagen: Overfitting entsteht, wenn ein Modell zu sehr auf seine Trainingsdaten angepasst ist und neue Daten nicht mehr zuverlässig bewertet; Datenbias tritt auf, wenn die Trainingsdaten die Realität verzerrt abbilden und das Modell diese Verzerrung in seinen Vorhersagen reproduziert. Beide Risiken lassen sich durch geeignete Evaluierungsverfahren sichtbar machen, erfordern aber auch ein organisatorisches Bewusstsein für ihre möglichen Konsequenzen.

Darüber hinaus stellen sich grundlegende Fragen zur *Erklärbarkeit* (Explainability): Viele leistungsfähige Modelle – insbesondere neuronale Netze und Ensemble-Verfahren – sind in ihrer internen Logik kaum nachvollziehbar. Für Entscheidungen, die rechtlich oder ethisch begründet werden müssen, ist dies ein ernsthaftes Problem. In der Europäischen Union schreibt die KI-Verordnung (AI Act) für bestimmte Hochrisikoanwendungen eine nachvollziehbare Entscheidungslogik explizit vor.

Diese Aspekte verdeutlichen, dass datengetriebene Modelle nicht als ,,Black Box" eingesetzt werden sollten. Aus Managementsicht ist entscheidend, Potenziale realistisch einzuordnen, Ergebnisse kritisch zu hinterfragen und die Verantwortung für Modellentscheidungen klar zu benennen.

Datengetriebene Modelle liefern wertvolle Unterstützung für Entscheidungen – ihr Nutzen entsteht jedoch erst durch die richtige Anwendung, kritische Interpretation und ein Bewusstsein für ihre strukturellen Grenzen. Die Fragen nach Overfitting, Datenbias und Erklärbarkeit sind keine rein technischen Themen, sondern haben unmittelbare betriebswirtschaftliche und rechtliche Relevanz.

5.3.4 Data Science als Organisationsaufgabe

Der Erfolg von Data-Science-Projekten hängt maßgeblich von der Zusammenarbeit zwischen Fachbereichen und technischen Expertinnen und Experten ab. Typische Rollen in einem Data-Science-Projekt umfassen:

- **Data Scientists** verantworten die Modellierung und die Auswahl geeigneter Verfahren.
- **Data Engineers** sorgen für die Verfügbarkeit, Qualität und technische Verarbeitung der Daten.
- **Domänenexpertinnen und -experten** aus den Fachbereichen definieren die betriebliche Fragestellung und bewerten die Ergebnisse inhaltlich.
- **Projektverantwortliche** koordinieren den CRISP-DM-Prozess und stellen die Integration in bestehende Systeme und Entscheidungsabläufe sicher.

Data Science ist damit weniger ein rein technisches Thema als vielmehr ein interdisziplinärer Ansatz zur Lösung betrieblicher Probleme. Wie datengetriebene Use Cases problem- und nicht technologiegetrieben entwickelt, strukturiert und priorisiert werden, behandelt Kapitel 6 ausführlich. Fragen zur Verantwortlichkeit, zu Datenschutz und ethischen Implikationen von Modellentscheidungen werden in Kapitel 7 (Governance, Ethik und Organisation) vertieft.

Weiterführende Literatur

- *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking* [20]: Standardwerk zur Managementperspektive auf Data Science und datengetriebene Entscheidungslogik.
- *Prediction Machines: The Simple Economics of Artificial Intelligence* [1]: Ökonomische Perspektive auf künstliche Intelligenz als Prognosetechnologie.

- *Artificial Intelligence Basics: A Non-Technical Introduction* [22]: Niedrigschwellige Einführung in zentrale KI-Konzepte und Unternehmensanwendungen.

Kapitel 6

Data-Driven Business: Von der Idee zum Use Case

In den vorherigen Kapiteln wurde gezeigt, wie Daten entstehen, analysiert, visualisiert und mit Verfahren des maschinellen Lernens ausgewertet werden. Insbesondere der Machine-Learning-Prozess aus Kapitel 5 beschreibt, wie ein Modell von der Problemdefinition bis zum Einsatz entwickelt wird. Eine Frage blieb dabei jedoch offen: *Woher kommt überhaupt die richtige Fragestellung?*

In der Praxis scheitern Analytics-Initiativen selten an der Technologie. Häufiger entstehen sie technologiegetrieben – ein Werkzeug oder eine Methode ist vorhanden und sucht nach einem Anwendungsfall. Das Ergebnis sind Projekte ohne klaren Nutzenbezug. Dieses Kapitel setzt daher früher an: Es geht um die strukturierte Entwicklung eines Use Cases, *bevor* der eigentliche ML-Prozess beginnt. Dafür betrachten wir zwei etablierte Werkzeuge, die sich gut ergänzen: Design Thinking für die Problemfindung und den Canvas-Ansatz für die Strukturierung.

6.1 Von der Geschäftsfrage zum Use Case

Ein datengetriebener Use Case beschreibt, wie ein konkretes betriebliches Problem mithilfe von Daten und Analysen gelöst werden soll – und welcher Nutzen dadurch entsteht. Entscheidend ist die Reihenfolge: Am Anfang steht nicht die Methode, sondern das Problem.

Problem-first statt Technology-first bedeutet, zuerst zu fragen, welche Entscheidung verbessert oder welcher Prozess optimiert werden soll, und erst danach, welche Daten und Verfahren dafür geeignet sind. Diese Haltung verhindert, dass aufwändige Modelle entwickelt werden, die an den eigentlichen Bedürfnissen vorbeigehen.

Damit übernimmt dieses Kapitel die Rolle des *vorgelagerten Schritts* zum Machine-Learning-Prozess (vgl. Kapitel 5): Erst wenn ein Use Case klar formuliert und als sinnvoll bewertet ist, lohnt sich der Einstieg in Datenaufbereitung und Modellierung. Die folgenden Abschnitte stellen zwei Werkzeuge vor, die diesen vorgelagerten Schritt strukturieren.

> Ein datengetriebener Use Case beginnt mit dem Problem, nicht mit der Technologie. Die zunächst formulierten Wünsche sind selten präzise genug – erst durch systematisches Hinterfragen entsteht eine Fragestellung, die sich als Lernaufgabe umsetzen lässt.

6.2 Design Thinking für datengetriebene Lösungen

Design Thinking ist ein Ansatz zur Lösungsentwicklung, der den Menschen und sein Problem konsequent in den Mittelpunkt stellt. Ursprünglich aus der Produktentwicklung stammend, eignet er sich besonders gut für die frühe Phase datengetriebener Projekte, weil er das Problemverständnis vor die Lösung stellt.

Der Prozess wird üblicherweise in fünf Phasen beschrieben, die iterativ durchlaufen werden:

1. **Verstehen & Beobachten**: Das Problem und der Kontext der betroffenen Nutzer werden gründlich untersucht. Wer hat welches Problem, und warum?
2. **Sichtweise definieren**: Aus den Beobachtungen wird eine klar formulierte Problemstellung abgeleitet.
3. **Ideen entwickeln**: Für die Problemstellung werden möglichst viele Lösungsansätze gesammelt – zunächst ohne Bewertung.
4. **Prototyp erstellen**: Eine vielversprechende Idee wird in eine einfache, schnell überprüfbare Form gebracht.
5. **Testen**: Der Prototyp wird mit den Nutzern erprobt; die Erkenntnisse fließen zurück in den Prozess.

Übertragen auf Data Science verschiebt dieser Ansatz den Fokus: *Nicht die Technologie steht im Vordergrund, sondern das zu lösende Problem.* Ein „Prototyp"kann hier auch eine einfache Analyse oder ein schnell erstelltes, grobes Modell sein, das prüft, ob ein Ansatz überhaupt tragfähig ist – lange bevor in ein aufwändiges Modell investiert wird.

Beispiel Brick-Flow AG: Der Vertrieb formuliert zunächst den Wunsch „Wir wollen künstliche Intelligenz einsetzen". Im Sinne des Design Thinking wird dieser Wunsch nicht direkt umgesetzt, sondern hinterfragt. Durch Gespräche mit dem Kundenservice (*Verstehen & Beobachten*) zeigt sich, dass viele Rücksendungen Aufwand und Kosten verursachen. Daraus wird eine klare Problemstellung (*Sichtweise*): „Wie können wir Bestellungen mit hohem Retourenrisiko frühzeitig erkennen?Ërst diese Problemformulierung führt zu einem sinnvollen, datengetriebenen Use Case.

> Design Thinking diszipliniert die Use-Case-Entwicklung: Es verlangt, das Problem wirklich zu verstehen, bevor eine Lösung skizziert wird. Im Data-Science-Kontext ersetzt das vage Wünsche durch belastbare Problemdefinitionen – und verhindert, dass Methoden nach einem passenden Problem suchen.

6.3 Canvas-Methoden zur Strukturierung

Wenn ein Problem klar formuliert ist, hilft ein *Canvas*, den Use Case strukturiert auszuarbeiten. Ein Canvas ist eine visuelle Strukturierungshilfe, die eine komplexe Fragestellung auf eine kompakte Darstellung mit vordefinierten Feldern reduziert. Ziel ist es, alle relevanten Aspekte gemeinsam sichtbar zu machen und ein geteiltes Verständnis zwischen Fachbereich und Technik zu schaffen.

Bekannt ist vor allem der *Business Model Canvas*, der ein gesamtes Geschäftsmodell strukturiert. Für datengetriebene Vorhaben eignet sich der spezialisierte *Machine Learning Canvas* besser: Er stellt gezielt die Elemente eines Analytics-Use-Cases zusammen – vom Ziel über Daten und Lernaufgabe bis zur Entscheidung und Erfolgsmessung. Abbildung 6.1 zeigt einen solchen Canvas, ausgefüllt für den Brick-Flow-Use-Case aus dem vorigen Abschnitt.

Ziel & Wertversprechen

Bestellungen mit hohem Retourenrisiko frühzeitig erkennen, um proaktiv gegenzusteuern und Rücksendekosten zu senken.

Datenquellen	ML-Aufgabe (Lernaufgabe)	Entscheidung & Aktion
Bestelldaten (Shop), Kundenhistorie, Produktstammdaten, Retourendaten der Logistik.	Binäre Klassifikation: hohes vs. niedriges Retourenrisiko je Bestellung.	Bei hohem Risiko: Hinweis / Kundenkontakt vor Versand; Prüfung des Sortiments.
Merkmale (Features)	**Geschäftsnutzen**	**Risiken & Einschränkungen**
Produktkategorie, Bestellwert, bisherige Retourenquote des Kunden, Versandregion.	Geringere Retourenquote, weniger Rücksendekosten, höhere Kundenzufriedenheit.	Datenqualität, Bias gegen Kundengruppen, Fehlalarme (False Positives), Datenschutz (DSGVO).

Erfolgsmessung & Evaluation

Retourenquote vor/nach Einsatz; Testfehler des Modells; Kosten der Maßnahmen im Verhältnis zur Einsparung.

Abbildung 6.1: Machine Learning Canvas, ausgefüllt für den Use Case „Retourenprognose" der Brick-Flow AG.

Der Wert des Canvas liegt nicht im Ausfüllen der Felder selbst, sondern in den Diskussionen, die dabei entstehen. Stellt sich etwa heraus, dass für die geplante Lernaufgabe gar keine geeigneten Daten vorliegen, wird dies früh sichtbar – bevor Aufwand in die Umsetzung fließt. Ein Canvas wird daher typischerweise iterativ genutzt und im Projektverlauf weiterentwickelt.

> Der ML Canvas schafft ein gemeinsames Verständnis zwischen Fachbereich und Data Science – nicht durch das Ausfüllen der Felder selbst, sondern durch die Diskussionen, die dabei entstehen. Bleiben wesentliche Felder unklar oder leer, ist das ein frühes Warnsignal, das den Projekteinstieg verzögern sollte.

6.4 Use Cases priorisieren und bewerten

In der Praxis konkurrieren meist mehrere Use-Case-Ideen um begrenzte Ressourcen. Bevor ein Vorhaben gestartet wird, sollte daher bewertet werden, ob sich der Aufwand lohnt. Eine einfache, aber wirkungsvolle Heuristik ist die Gegenüberstellung

von *erwartetem Nutzen* und *Umsetzungsaufwand* in einer Vier-Felder-Matrix (Abbildung 6.2).

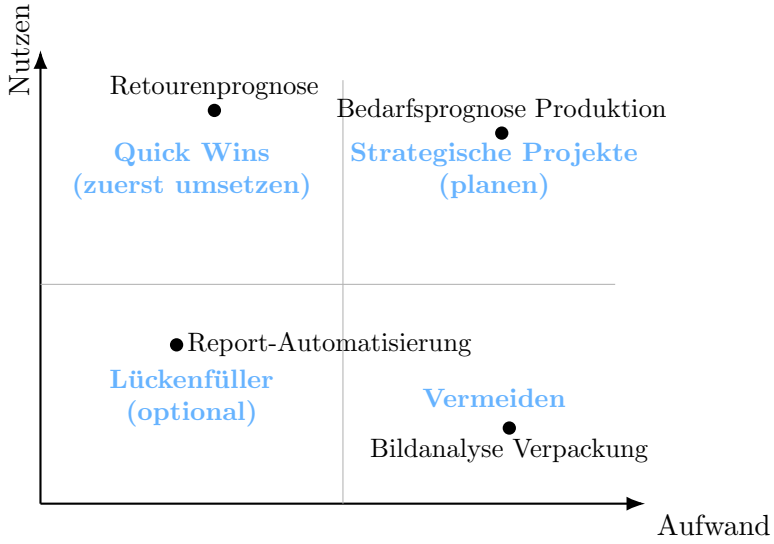


Abbildung 6.2: Aufwand-Nutzen-Matrix zur Priorisierung datengetriebener Use Cases (beispielhafte Einordnung für die Brick-Flow AG).

Use Cases mit hohem Nutzen und geringem Aufwand (*Quick Wins*) sollten zuerst umgesetzt werden – sie schaffen schnell sichtbaren Wert und Akzeptanz. Aufwändige, aber wertvolle Vorhaben (*strategische Projekte*) erfordern Planung und Ressourcen. Vorhaben mit geringem Nutzen bei hohem Aufwand sollten vermieden werden.

Bei der Aufwandsschätzung spielt vor allem die *Datenverfügbarkeit* eine zentrale Rolle: Ein Use Case, für den keine geeigneten Daten vorliegen, ist trotz hohen Nutzens nicht umsetzbar.

Datenverfügbarkeit systematisch prüfen. Bevor ein Use Case gestartet wird, empfiehlt sich eine strukturierte Prüfung der Datenlage anhand weniger Kernfragen (Tabelle 6.1). Kritisch ist vor allem das Vorhandensein der *Zielvariable*: Fehlt sie vollständig – weil das Unternehmen bislang nicht erfasst, was das Modell vorhersagen soll –, ist der Use Case unabhängig vom Nutzenpotenzial nicht umsetzbar. In diesem Fall muss zunächst die Datenbasis systematisch aufgebaut werden.

Dies erfordert häufig tiefere, strukturelle Veränderungen im Unternehmen und ist daher – vor dem Hintergrund einer Nutzenanalyse – in die Datenstrategie des Unternehmens abzugleichen.

Kriterium	**Leitfrage**	**Priorität**
Zielvariable	Ist das Ziel (z. B. Retoure: ja/nein) als Spalte in den Daten vorhanden?	kritisch
Historienreichweite	Wie viele Datenpunkte liegen vor? Sind seltene Ereignisse ausreichend vertreten?	wichtig
Datenqualität	Wie hoch ist der Anteil fehlender, inkonsistenter oder fehlerhafter Einträge?	wichtig
Datenzugang	Dürfen die nötigen Daten für diesen Zweck verwendet werden (DSGVO, IT-Sicherheit)?	kritisch
Aktualität	Sind die Daten noch repräsentativ für das heutige Geschäft?	wichtig

Tabelle 6.1: Datenreife-Checkliste zur Bewertung eines Data-Science-Use-Cases vor Projektstart

Proof of Concept: Klein starten, schnell lernen. Selbst wenn ein Use Case die Priorisierungsmatrix übersteht und die Datengrundlage prinzipiell vorhanden ist, empfiehlt sich ein schrittweises Vorgehen. Ein *Proof of Concept* (PoC) setzt den Use Case zunächst stark vereinfacht um – mit einem kleinen Datensatz, einem einfachen Modell und ohne produktionstaugliche Infrastruktur. Ziel ist nicht ein fertiges System, sondern eine schnelle Antwort auf die Frage: *Ist der Ansatz grundsätzlich tragfähig?*

Typische PoC-Fragen sind: Lässt sich mit den verfügbaren Daten überhaupt ein Modell trainieren, das besser abschneidet als eine einfache Daumenregel? Welcher Fehler ist realistisch erreichbar? Wo liegen die größten Datenlücken? Die Erkenntnisse fließen direkt in den CRISP-DM-Prozess zurück (vgl. Kapitel 5) und helfen, realistische Erwartungen für das eigentliche Projekt zu setzen – lange bevor erhebliche Ressourcen gebunden sind.

Die organisatorischen Erfolgsfaktoren – Datenqualität, Zusammenarbeit zwischen Fachbereich und Data Science sowie die Integration in bestehende Prozesse – wurden in Kapitel 5 ausführlich behandelt und gelten für jeden Use Case gleichermaßen.

Vor dem Projektstart stehen drei Fragen:

1. Lohnt es sich (Nutzen gegenüber Aufwand)?
2. Haben wir die nötigen Daten (Datenreife-Check)?
3. Und lässt sich der Ansatz mit einem einfachen Prototyp schnell validieren (Proof of Concept)?

Erst wenn alle drei Fragen bejaht werden können, ist der Projektstart sinnvoll.

6.5 Typische Fehler bei der Use-Case-Formulierung

Selbst mit Design Thinking, Canvas und Priorisierungsmatrix entstehen in der Praxis immer wieder ähnliche Fehler – fast immer nicht im technischen, sondern im konzeptionellen Teil des Projekts, also genau in der Phase, die dieses Kapitel behandelt.

Beispiel Brick-Flow AG: Das Marketingteam möchte die *Kundenzufriedenheit* vorhersagen, um bei potenziell unzufriedenen Kunden frühzeitig gegenzusteuern. Die Idee ist nachvollziehbar, der Nutzen klar. Beim Ausfüllen des ML Canvas zeigt sich jedoch: Brick-Flow AG erfasst keine Zufriedenheitsbewertungen – keine Sternebewertungen, keine NPS-Umfragen, keine Kommentare nach dem Kauf. Das Feld *Datenquellen* bleibt leer, weil keine Zielvariable existiert. Die Lernaufgabe lässt sich nicht formulieren. Das Projekt scheitert nicht an der Methode, sondern daran, dass die Grundvoraussetzung fehlt. Die Konsequenz: Zunächst muss ein systematisches Feedback-System aufgebaut werden – erst dann ist ein Vorhersagemodell denkbar.

Dieser Fall zeigt ein typisches Muster: Die Idee ist gut, aber der Canvas deckt früh auf, dass sie nicht umsetzbar ist. Tabelle 6.2 fasst die häufigsten Fehler zusammen, ergänzt um das jeweilige Frühwarnsignal im Canvas.

Fehler	**Beschreibung**	**Frühwarnsignal im Canvas**
Zielvariable fehlt	Das Ziel wird nirgends in den Daten erfasst.	Feld *Datenquellen* bleibt leer oder enthält nur Proxy-Größen.
Zu vage Fragestellung	„Wir wollen KI einsetzen" ist keine lösbare Aufgabe.	Lernaufgabe lässt sich nicht als Regression, Klassifikation oder Clustering formulieren.
Kein Erfolgskriterium	Ohne messbare KPIs ist unklar, ob das Projekt erfolgreich war.	Feld *Erfolgsmessung* enthält nur qualitative Aussagen.
Technologiegetrieben	Die Methode steht fest, das Problem wird gesucht.	Diskussion beginnt mit „Wir könnten neuronale Netze einsetzen für…"
Datenschutz zu spät	Rechtliche Prüfung erfolgt erst nach der Implementierung.	Feld *Risiken & Einschränkungen* enthält keinen Hinweis auf DSGVO oder Datenzugang.

Tabelle 6.2: Typische Fehler bei der Use-Case-Formulierung und ihre Frühwarnsignale im ML Canvas

Die häufigsten Fehler bei der Use-Case-Entwicklung entstehen nicht im Modell, sondern davor: fehlende Zielvariablen, vage Problemdefinitionen und ungeklärte Datenzugänge machen selbst methodisch einwandfreie Projekte nutzlos. Canvas und Datenreife-Check helfen, diese Fallen früh zu erkennen – bevor Aufwand in die Umsetzung fließt.

Weiterführende Literatur

- *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World* [10]: Verbindung von Daten, KI, Plattformen und Unternehmensstrategie.
- *Lean Analytics: Use Data to Build a Better Startup Faster* [4]: Praxisorientierte Perspektive auf Kennzahlen, Experimente und datengetriebene Geschäftsentscheidungen.
- *Data Driven: Creating a Data Culture* [17]: Kompakte praxisnahe Einführung in datengetriebene Organisationen und Projekte.

Kapitel 7

Datenschutz und Ethik im Umgang mit Daten

Daten sind eine zentrale Ressource moderner Unternehmen – ihr Einsatz ist jedoch nicht nur eine Frage von Technologie und Wirtschaftlichkeit, sondern auch von *Verantwortung und Gestaltung*. Mit der zunehmenden Nutzung von Daten und Analysen rücken rechtliche und ethische Fragestellungen stärker in den Fokus.

In der Praxis zeigt sich, dass datengetriebene Initiativen nicht nur an Daten oder Methoden scheitern können, sondern auch an mangelnder Transparenz, fehlendem Vertrauen oder der Missachtung rechtlicher Vorgaben. Die organisatorischen Voraussetzungen – klare Verantwortlichkeiten und Rollen im Rahmen einer *Data Governance* – wurden bereits in Kapitel 2.5 behandelt. Dieses Kapitel konzentriert sich auf drei eng verbundene Themen: den *rechtlichen Rahmen* für Datenschutz und KI-Einsatz, die *Datenethik* als normative Orientierung sowie Anforderungen an *verantwortungsvolle Entscheidungen* in der Praxis.

7.1 Rechtliche Rahmenbedingungen

Der Einsatz von Daten und KI-Systemen ist in Europa durch zwei zentrale Regelwerke reguliert: die *Datenschutz-Grundverordnung (DSGVO)* und den *EU AI Act*. Beide ergänzen sich: Die DSGVO regelt den Umgang mit personenbezogenen Daten als Grundlage; der AI Act bestimmt, unter welchen Bedingungen KI-Systeme mit diesen Daten Entscheidungen treffen oder beeinflussen dürfen. Für Unterneh-

men, die Daten analysieren oder KI-Modelle einsetzen, sind beide Rahmenwerke gleichzeitig relevant.

7.1.1 DSGVO: Grundprinzipien und Datenkategorien

Die Datenschutz-Grundverordnung gilt seit 2018 europaweit und schützt das Recht natürlicher Personen auf informationelle Selbstbestimmung. Im Kern müssen personenbezogene Daten *rechtmäßig, transparent und zweckgebunden* verarbeitet werden. Weitere Grundsätze sind Datensparsamkeit, Richtigkeit und Speicherbegrenzung.

Eine zentrale Unterscheidung betrifft den Personenbezug von Daten:

- **Personenbezogene Daten** sind alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen – nicht nur Name oder Adresse, sondern auch Kundennummern, Standortdaten oder Online-Identifikatoren.
- **Pseudonymisierte Daten** werden so verändert, dass sie ohne zusätzliche Informationen nicht mehr direkt zugeordnet werden können. Der Personenbezug bleibt grundsätzlich erhalten – sie unterliegen daher weiterhin der DSGVO.
- **Anonyme Daten** lassen keinen Personenbezug mehr zu; eine Re-Identifikation ist praktisch ausgeschlossen. Sie fallen nicht mehr unter die DSGVO, sind in der Praxis aber schwer zu erreichen.

Beispiel Brick-Flow AG: Im Online-Shop entstehen zahlreiche personenbezogene Daten – Kundenkonten, Bestellhistorien und Versandadressen. Für interne Analysen (etwa das Retouren-Modell aus Kapitel 5) können diese Daten pseudonymisiert werden, sodass das Data-Science-Team mit Kundennummern statt Klarnamen arbeitet. Möchte die Brick-Flow AG hingegen aggregierte Abverkaufsdaten an ihre Lieferanten weitergeben, müssen diese so *anonymisiert* sein, dass keine Rückschlüsse auf einzelne Kundinnen und Kunden möglich sind.

Für das Management ist entscheidend zu verstehen, dass Datenschutz nicht nur eine rechtliche Pflicht ist, sondern unmittelbare Auswirkungen auf datengetriebene Projekte hat – etwa darauf, welche Daten überhaupt für eine Analyse genutzt werden dürfen.

7.1.2 Art. 22 DSGVO: Automatisierte Einzelentscheidungen

Ein für datengetriebene Unternehmen besonders relevantes Recht findet sich in Art. 22 DSGVO: Betroffene Personen haben das Recht, *nicht ausschließlich einer automatisierten Entscheidung* unterworfen zu werden, die ihnen gegenüber rechtliche oder ähnlich bedeutsame Wirkung entfaltet. Damit adressiert die DSGVO direkt den Einsatz von Algorithmen und Modellen in unternehmensrelevanten Entscheidungen.

Typische Anwendungsfälle, die unter diesen Artikel fallen können, sind:

- automatisiertes Kredit-Scoring oder Bonitätsprüfungen
- algorithmisches Bewerber-Ranking im Recruiting
- automatisierte Preis- oder Konditionendifferenzierung
- KI-gestützte Risikobeurteilungen im Versicherungsbereich

Für Unternehmen ergibt sich daraus eine direkte Anforderung: In diesen Bereichen muss eine *menschliche Überprüfung* möglich sein, und Betroffene haben das Recht auf eine verständliche Erklärung, warum eine Entscheidung getroffen wurde. Rein automatisierte Systeme ohne Eingriffsmöglichkeit sind in diesen Kontexten rechtlich nicht zulässig. Dies hat unmittelbare Konsequenzen für die Wahl von Modellen und die Gestaltung von Entscheidungsprozessen – ein Aspekt, auf den Abschnitt 7.4 zurückkommt.

7.1.3 EU AI Act: Risikobasierte KI-Regulierung

Mit der Verordnung (EU) 2024/1689 – dem sogenannten *EU AI Act* – hat die Europäische Union den weltweit ersten umfassenden Rechtsrahmen für den Einsatz künstlicher Intelligenz geschaffen. Er gilt für alle Unternehmen, die KI-Systeme in der EU entwickeln oder einsetzen, und ist ab 2025 schrittweise anwendbar.

Kernprinzip ist ein *risikobasierter Ansatz*: KI-Systeme werden danach eingestuft, welches Risiko sie für Grundrechte, Sicherheit oder gesellschaftliche Werte darstellen.

- **Verbotene KI-Systeme**: Grundsätzlich unzulässige Anwendungen – etwa Social Scoring durch Behörden, manipulative KI-Systeme oder biometrische Echtzeit-Überwachung im öffentlichen Raum.
- **Hochrisiko-KI**: Systeme in sensiblen Bereichen wie Personalwesen (Recruiting, Leistungsbewertung, Kündigung), Kreditvergabe, Bildungszugang,

kritische Infrastrukturen oder Strafverfolgung. Für sie gelten strenge Anforderungen: Transparenz, menschliche Aufsicht, Risikobeurteilung, technische Robustheit und lückenlose Dokumentation.

- **Begrenzt riskante KI**: Systeme mit spezifischen Transparenzpflichten – z. B. Chatbots, die kenntlich machen müssen, dass man mit einem KI-System kommuniziert.
- **Minimales Risiko**: Die Mehrzahl aller KI-Anwendungen – etwa Spam-Filter, Empfehlungssysteme oder KI in Computerspielen – unterliegt keinen besonderen Anforderungen.

Beispiel Brick-Flow AG: Das Retouren-Prognosemodell aus Kapitel 5 dient der internen Prozessoptimierung und ist wahrscheinlich als KI mit minimalem Risiko einzustufen. Anders verhielte es sich, wenn die Brick-Flow AG dasselbe Modell nutzte, um Kundinnen und Kunden je nach prognostiziertem Retourenverhalten unterschiedliche Lieferbedingungen oder Servicelevels anzubieten – dann könnten Merkmale der Hochrisiko-Kategorie greifen, verbunden mit erheblichem Compliance-Aufwand.

DSGVO und EU AI Act greifen ineinander: Die DSGVO schützt personenbezogene Daten als Grundlage; der AI Act regelt, wie KI-Systeme mit diesen Daten Entscheidungen treffen oder beeinflussen dürfen. Wer KI-Systeme in der EU einsetzt, muss beide Rahmenbedingungen gleichzeitig einhalten.

7.1.4 Datenschutz und Datensicherheit

Vom Datenschutz zu unterscheiden ist die *Datensicherheit*. Während der Datenschutz regelt, *ob* und *wozu* personenbezogene Daten genutzt werden dürfen, sorgt die Datensicherheit für den *technischen* Schutz vor unbefugtem Zugriff, Verlust oder Manipulation. Beide Aspekte ergänzen sich: Eine rechtmäßige Datennutzung ist wertlos, wenn die Daten technisch nicht ausreichend geschützt sind.

Für diese Aufgaben existieren spezialisierte Rollen, etwa der *Datenschutzbeauftragte* oder der IT-Sicherheitsverantwortliche (CISO). Sie ergänzen die in Kapitel 2.5 beschriebenen Data-Governance-Rollen um die Schutzperspektive.

7.2 Ethische Dimensionen datengetriebener Entscheidungen

Neben rechtlichen Anforderungen gewinnen ethische Fragen zunehmend an Bedeutung. Datenanalysen beeinflussen Entscheidungen, die direkte Auswirkungen auf Menschen haben können – etwa bei der Kreditvergabe, im Recruiting oder bei der Preisgestaltung. Das Besondere: Datengetriebene Entscheidungen werden häufig als *objektiv* wahrgenommen, schließlich entscheidet das Modell, nicht der Mensch. Diese scheinbare Neutralität kann jedoch trügerisch sein. Modelle beruhen auf Annahmen, Datenstrukturen und Zielvorgaben, die von Menschen gesetzt werden – und reproduzieren damit deren Perspektiven, blinde Flecken und Präferenzen.

Grundlegende ethische Prinzipien. Für den verantwortungsvollen Umgang mit Daten und Algorithmen haben sich einige grundlegende Prinzipien als orientierend erwiesen:

- **Nichtschaden**: Daten und Modelle dürfen keine unverhältnismäßigen Nachteile für Personen oder Gruppen erzeugen.
- **Fairness**: Entscheidungen sollen nicht systematisch bestimmte Gruppen bevorzugen oder benachteiligen (vgl. Abschnitt 7.3).
- **Transparenz**: Die Grundlage einer Entscheidung muss nachvollziehbar und erklärbar sein.
- **Autonomie**: Betroffene sollen über die Nutzung ihrer Daten informiert sein und eine reale Entscheidungsmöglichkeit behalten.
- **Verhältnismäßigkeit**: Die Intensität der Datennutzung muss dem verfolgten Zweck angemessen sein.

Ethical by Design. Ethische Anforderungen sind dann am wirksamsten, wenn sie von Anfang an in die Entwicklung datengetriebener Systeme eingebettet werden – nicht erst nachträglich als Prüfschritt. Dieses Prinzip, häufig als *Ethical by Design* bezeichnet, entspricht dem aus dem Datenschutz bekannten ,,Privacy by Design": Bereits bei der Auswahl von Daten, der Definition der Zielgröße und der Wahl des Modells werden ethische Konsequenzen mitgedacht. Statt ,,Wie können wir rechtliche Probleme vermeiden?" steht die Frage: ,,Für wen und auf welche Weise wirkt diese Entscheidung?"

In der Praxis helfen dabei einige Leitfragen:

- Ist die Nutzung dieser Daten für den vorgesehenen Zweck gerechtfertigt – und von den Betroffenen so erwartet?
- Welche Auswirkungen hat das Modell auf verschiedene Personengruppen, auch solche, die nicht direkt adressiert werden?
- Wer profitiert von der Entscheidung, wer trägt das Risiko?
- Wie würde die Entscheidungslogik öffentlich wahrgenommen werden?

Datenethik ist damit ein integraler Bestandteil moderner Unternehmensführung – keine nachgelagerte Compliance-Aufgabe, sondern ein Teil der strategischen Gestaltungsverantwortung.

7.3 Bias und Fairness in datengetriebenen Modellen

Ein besonders wichtiges ethisches Thema ist der Umgang mit *Bias* – systematischen Verzerrungen. Wie bereits in Kapitel 5 angesprochen, sind Machine-Learning-Modelle nicht ,,neutral": Sie lernen aus historischen Daten und reproduzieren damit die darin enthaltenen Muster und Strukturen. Während Kapitel 5 Bias als methodisches Risiko eingeführt hat, steht hier die ethische und gesellschaftliche Dimension im Vordergrund.

7.3.1 Entstehung von Bias

Bias entsteht meist nicht bewusst, sondern an verschiedenen Stellen im Daten- und Analyseprozess:

- **Historische Verzerrungen**: Die Daten spiegeln vergangene Entscheidungen wider, die selbst bereits verzerrt waren.
- **Unvollständige Daten**: Bestimmte Gruppen sind unterrepräsentiert.
- **Variablenauswahl**: Relevante Einflussfaktoren werden nicht berücksichtigt – oder problematische Merkmale fließen unreflektiert ein.
- **Interpretation**: Ergebnisse werden einseitig bewertet.

So können Modelle bestehende Ungleichheiten nicht nur abbilden, sondern sogar verstärken.

Beispiel Recruiting: Ein Unternehmen möchte mithilfe eines Modells vorhersagen, welche Bewerberinnen und Bewerber besonders erfolgreich sein werden. Wurde in der Vergangenheit überwiegend eine bestimmte Personengruppe eingestellt, spiegelt sich diese Struktur in den Trainingsdaten wider. Das Modell bevorzugt dann Profile, die den bisherigen Einstellungen ähneln, und bewertet andere Gruppen systematisch schlechter. Das Problem liegt nicht im Modell selbst, sondern in den Daten und ihrer Nutzung.

Bezug zur Brick-Flow AG: Dieselbe Mechanik kann auch das Retouren-Modell aus Kapitel 5 betreffen. Sind in den Trainingsdaten Retouren bestimmter Kundengruppen – etwa aus einzelnen Regionen – überrepräsentiert, könnte das Modell diese Gruppen pauschal als ,,risikobehaftet" einstufen und ihnen häufiger zusätzliche Hürden auferlegen. Eine zunächst rein betriebswirtschaftliche Optimierung erhält damit eine ethische Dimension.

7.3.2 Fairness als Gestaltungsentscheidung

Vor diesem Hintergrund gewinnt das Konzept der *Fairness* an Bedeutung. Dabei geht es nicht nur um technische Lösungen, sondern um grundlegende Fragen: Welche Kriterien gelten als fair? Welche Unterschiede sind akzeptabel? Welche Ziele verfolgt das Unternehmen? Fairness ist daher keine rein mathematische Eigenschaft, sondern eine *bewusste Gestaltungsentscheidung*.

Für den praktischen Umgang mit Bias haben sich mehrere Ansatzpunkte etabliert: Bewusstsein schaffen, Daten auf Herkunft und Struktur prüfen, Modelle regelmäßig evaluieren und Annahmen transparent machen. Diese Maßnahmen erfordern eine enge Zusammenarbeit zwischen Fachbereich, Management und technischen Expertinnen und Experten.

> Datenmodelle sind nicht neutral – sie spiegeln bestehende Muster wider. Fairness muss daher aktiv gestaltet und kontinuierlich überprüft werden.

7.4 Verantwortungsvolle datengetriebene Entscheidungen

Datenschutz und Ethik münden in eine gemeinsame Anforderung: Datengetriebene Entscheidungen müssen *nachvollziehbar und verantwortbar* sein. Drei Aspekte sind dabei aus Managementsicht zentral:

- **Transparenz**: Es muss nachvollziehbar bleiben, wie eine datenbasierte Entscheidung zustande kommt. Die in Kapitel 4 behandelte Data Lineage ist hierfür eine wichtige Voraussetzung.
- **Mensch im Entscheidungsprozess**: Modelle liefern Empfehlungen, ersetzen aber nicht die Verantwortung des Managements. Gerade bei Entscheidungen mit Auswirkungen auf Menschen – und im Lichte von Art. 22 DSGVO – sollte eine menschliche Prüfung erhalten bleiben.
- **Rechenschaft**: Es muss klar sein, wer für eine datenbasierte Entscheidung verantwortlich ist – eine Frage, die unmittelbar an die Governance-Rollen aus Kapitel 2.5 anknüpft.

Erklärbarkeit: Das „Warum" hinter dem Modell. Eine besondere Herausforderung stellt die *Erklärbarkeit* von KI-Modellen dar (englisch: Explainability, kurz XAI). Leistungsstarke Modelle – insbesondere tiefe neuronale Netze – liefern häufig sehr genaue Vorhersagen, lassen aber kaum erkennen, welche Faktoren zu einem bestimmten Ergebnis geführt haben. Man spricht von *Black-Box-Modellen*. Für Bereiche mit direkten Auswirkungen auf Personen ist das rechtlich und ethisch problematisch: Art. 22 DSGVO fordert die Möglichkeit zur menschlichen Überprüfung, und der EU AI Act verlangt Transparenz für Hochrisiko-Systeme.

In der Praxis ergibt sich daraus häufig ein bewusster Trade-off: Ein erklärbares, aber weniger genaues Modell kann einem intransparenten, dafür leistungsfähigeren Modell vorzuziehen sein – abhängig vom Anwendungsfall und dem Risiko der Entscheidung. Für viele BWL-typische Anwendungen reichen interpretierbare Modelle wie Entscheidungsbäume oder logistische Regression aus; sie lassen sich zudem gegenüber Fachbereichen und Führungskräften wesentlich leichter kommunizieren.

Responsible AI: Ethik als organisatorische Aufgabe. Verantwortungsvoller KI-Einsatz ist keine Frage einzelner Mitarbeitender, sondern erfordert klare

organisatorische Strukturen. In der Praxis haben Unternehmen dafür verschiedene Ansätze entwickelt:

- **AI-Leitlinien**: Dokumentierte Grundsätze für den KI-Einsatz im Unternehmen, die festlegen, welche Ziele und Grenzen gelten.
- **Review-Prozesse**: Strukturierte Prüfung neuer KI-Projekte auf rechtliche und ethische Implikationen vor dem Rollout – analog zu bestehenden Compliance-Prüfungen.
- **Interdisziplinäre Teams**: Ethische, rechtliche und technische Expertise arbeiten von Beginn an zusammen, nicht nacheinander.
- **Monitoring und Audit**: Regelmäßige Überprüfung im laufenden Betrieb, ob Modelle weiterhin fair, genau und regelkonform arbeiten.

In der Praxis bedeutet dies, datenbasierte Analyse mit Erfahrung und situativer Einschätzung zu verbinden. Erfolgreiche Unternehmen integrieren Daten in ihre Entscheidungsprozesse, *ohne die Verantwortung des Managements zu ersetzen* – und ohne rechtliche und ethische Anforderungen als nachgelagerte Hürde zu behandeln, sondern als integralen Teil des Gestaltungsprozesses.

> Verantwortungsvoller Dateneinsatz erfordert mehr als Compliance: erklärbare Modelle, klare Rechenschaft und organisatorische Strukturen, die Ethik und Recht von Anfang an mitdenken.

Weiterführende Literatur

- *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* [18]: Kritische und anschauliche Einführung in gesellschaftliche Risiken algorithmischer Entscheidungen.
- *AI Ethics* [3]: Systematische Einführung in zentrale Fragen der KI-Ethik.
- *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program* [15]: Vertiefung zu Governance-Strukturen, Rollen und organisatorischer Verankerung.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Ajay Agrawal, Joshua Gans und Avi Goldfarb. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business Review Press, 2018.

[2] Peter Bruce, Andrew Bruce und Peter Gedeck. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. 2. Aufl. Sebastopol, CA: O'Reilly Media, 2020.

[3] Mark Coeckelbergh. *AI Ethics*. Cambridge, MA: MIT Press, 2020.

[4] Alistair Croll und Benjamin Yoskovitz. *Lean Analytics: Use Data to Build a Better Startup Faster*. Sebastopol, CA: O'Reilly Media, 2013.

[5] Thomas H. Davenport und Jeanne G. Harris. *Competing on Analytics: The New Science of Winning*. Boston, MA: Harvard Business School Press, 2007.

[6] Marlon Dumas u. a. *Fundamentals of Business Process Management*. 2. Aufl. Berlin und Heidelberg: Springer, 2018.

[7] Stephen Few. *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring*. 2. Aufl. Burlingame, CA: Analytics Press, 2013.

[8] Peter Gluchowski, Melanie Pfoh und Anja Tetzner. *Datenmodellierung in Data-Warehouse-Systemen: Konzepte, Technologien und Methoden für die Modellierung entscheidungsunterstützender Daten in Unternehmen*. 2. Aufl. Wiesbaden: Springer Fachmedien Wiesbaden, 2025. ISBN: 978-3-658-48847-5. DOI: 10.1007/978-3-658-48848-2.

[9] Darrell Huff. *How to Lie with Statistics*. New York: W. W. Norton & Company, 1954.

[10] Marco Iansiti und Karim R. Lakhani. *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World*. Boston, MA: Harvard Business Review Press, 2020.

[11] Cole Nussbaumer Knaflic. *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Hoboken, NJ: Wiley, 2015.

[12] Wolfgang Kohn und Riza Öztürk. *Wirtschaftsstatistik, Lerneinheit 1 – Eindimensionale Datenanalyse*. 2024. Aufl. Im Alten Holz 131, D-58903 Hagen: IfV NRW, 2024.

[13] Wolfgang Kohn und Riza Öztürk. *Wirtschaftsstatistik, Lerneinheit 2 – Regression, Mehrdimensionale Datenanalyse und Verteilungen*. 2024. Aufl. Im Alten Holz 131, D-58903 Hagen: IfV NRW, 2024.

[14] Jochen Küster und Peter Hartel. *Grundlagen der Wirtschaftsinformatik - Lerneinheit 1: Informationssysteme*. Hagen: Institut für Verbundstudien der Fachhochschulen Nordrhein-Westfalen - IfV NRW, 2019.

[15] John Ladley. *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program*. 2. Aufl. London: Academic Press, 2019.

[16] Bernard Marr. *Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things*. London: Kogan Page, 2017.

[17] Hilary Mason und DJ Patil. *Data Driven: Creating a Data Culture*. Sebastopol, CA: O'Reilly Media, 2015.

[18] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.

[19] Geoffrey G. Parker, Marshall W. Van Alstyne und Sangeet Paul Choudary. *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. New York: W. W. Norton & Company, 2016.

[20] Foster Provost und Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Sebastopol, CA: O'Reilly Media, 2013.

[21] Joe Reis und Matt Housley. *Fundamentals of Data Engineering: Plan and Build Robust Data Systems*. Sebastopol, CA: O'Reilly Media, 2022.

[22] Tom Taulli. *Artificial Intelligence Basics: A Non-Technical Introduction*. Berkeley, CA: Apress, 2019.

[23] Edward R. Tufte. *The Visual Display of Quantitative Information*. 2. Aufl. Cheshire, CT: Graphics Press, 2001.

[24] Charles Wheelan. *Naked Statistics: Stripping the Dread from the Data*. New York: W. W. Norton & Company, 2013.