Data Science mit Big Data

Techniken, Werkzeuge und Algorithmen
zur Analyse großer Datenmengen

Teil 1

Grundlagen

1
Gestatten, Data – Big Data

Ein Lehrbuch über den Umgang mit großen Datenmengen zu schreiben, ist nicht zuletzt auch eine große Herausforderung. In kaum einem anderen Technologiefeld, vielleicht mit Ausnahme der künstlichen Intelligenz, ist der technische Fortschritt so rasant und der „Zoo“ der Werkzeuge[1)] und Frameworks so umfangreich und schwer zu überblicken wie bei Big Data.

Wir haben uns daher bemüht, Ihnen, liebe Leserinnen und Leser, die Informationen und wichtigsten Hintergründe passgenau so aufzubereiten, wie wir es bei unserem Einstieg in das Themenfeld vor mittlerweile etlichen Jahren selbst gerne gehabt hätten. Damals war Big Data ein boomendes Thema, ähnlich wie die Künstliche Intelligenz es heute ist. Zahlreiche Hersteller entwickelten neuartige Werkzeuge und priesen sie als eierlegende Wollmilchsau, auch wenn diese oft noch in den Kinderschuhen steckten und die Erfahrungen im Umgang mit ihnen vielfach erst mühsam erarbeitet werden mussten. Spannend ist in diesem Kontext insbesondere, dass gerade NoSQL-Datenbanken im Vergleich mit ihren SQL-Pendants eher das absolute Gegenteil einer eierlegenden Wollmilchsau sind, da sie ihre Vorteile gewöhnlich nur für sehr eng begrenzte Anwendungsfälle ausspielen können, auch wenn die Werbung der Hersteller anderes glauben machen möchte.

Heute sind zahlreiche Big-Data-Systeme längst im Dauereinsatz und verrichten ganz unauffällig ihren Dienst in der Cloud. Die Datenwissenschaft mit Big Data ist in den letzten gut zwei Jahrzehnten so weit gereift, dass sowohl stabile Werkzeuge als auch umfangreiche Erfahrungen mit ihnen zur Verfügung stehen. Diese möchten wir Ihnen in diesem Lehrbuch, das nach unserem Stand in seiner Breite mindestens im deutschsprachigen Raum, wenn nicht sogar weltweit, einzigartig ist, kompakt und dergestalt

1 Beinahe schon legendär geworden ist Matt Turcks „MAD Landscape“, die versucht, einen Überblick über KI- und Big-Data-Tools zu geben, siehe *https://mattturck.com/landscape/mad2025.pdf*

vermitteln, dass es zum Ausprobieren und Experimentieren animiert. Schließlich waren nie so viele Werkzeuge und Daten frei verfügbar, wie das heute der Fall ist.

Primäre Zielgruppe unseres Buches sind neben Studierenden einschlägiger Studiengänge, ungefähr ab dem 4. Semester, auch all diejenigen, die sich für die berufliche Praxis einen schnellen und soliden technischen Überblick über den Umgang mit großen Datenmengen verschaffen möchten oder müssen. Sei es zur beruflichen Weiterentwicklung oder aus persönlichem Interesse. Auch Azubis im IT-Bereich dürften gegen Ende ihrer Ausbildung zahlreiche nützliche Anregungen aus diesem Buch verwenden können. Nicht zuletzt kann es natürlich auch allen Interessierten als Nachschlagewerk dienen, um spezielle Big-Data-Themen bei Bedarf schnell auffrischen zu können sowie Hinweise zu vertiefender Literatur zu finden.

1.1 Die Facetten von Big Data

Wörtlich übersetzt bedeutet „Big Data“ nichts weiter als „große Daten“ bzw. passender „große Datenmengen“[2]. Nie zuvor in ihrer Geschichte hat die Menschheit solche Massen davon produziert wie heute. Das globale Datenvolumen wächst rasant – von 26 Zettabyte (ZB) im Jahr 2017 über 104 ZB im Jahr 2022 zu geschätzten 175 ZB in 2025 [RRG18]. Diese Verdreifachung innerhalb von nur fünf Jahren setzt sich beschleunigt fort: Prognosen erwarten 284 ZB bis 2027 und 394 ZB bis 2028 [IDC24].

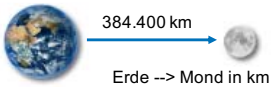
Bild 1.1 zeigt die für 2025 geschätzten 175 ZB in Bytes ausgeschrieben – was einer schier unvorstellbaren Zahl von 175 Billionen Gigabyte oder $175 * 10^{21}$ Byte entspricht.

Bild 1.1 175 Zettabyte in Bytes ausgeschrieben

175 Zettabyte sprengen jede Vorstellungskraft: Ein Stapel mit den dafür benötigten DVDs würde für rund 116 Reisen von der Erde zum Mond reichen und das Gesamtgewicht der aufeinander gestapelten DVDs entspräche etwa dem einer halben Milliarde Fahrzeuge vom Typ VW Golf (Bild 1.2). Diese Zahl übersteigt die gesamte Golf-Produktion seit Markteinführung 1974 (ca. 37 Millionen Exemplare) [Car24] um knapp das Zwölffache.

2 Teile dieses Kapitels wurden bereits in [Lan23] veröffentlicht.

Kapazität DVD (in GB)	1 ZB in # DVDs	1 DVD ≈ 1,2 mm Höhe	Erde --> Mond in km
4,7	212.765.957.447	255.319.148.936	384.400
Daten in 2025: 175 ZB	**175 ZB in # DVDs**	**Höhe in mm**	**# Reisen Erde --> Mond**
175	37.234.042.553.192	44.680.851.063.830	116



Kapazität DVD (in GB)	1 ZB in # DVDs	1 DVD ≈ 16g	Gewicht in # VW Golf 8
4,7	212.765.957.447	3.404.255.319.149	2.431.611
Daten in 2025: 175 ZB	**175 ZB in # DVDs**	**Gewicht in g**	**Gewicht in # VW Golf 8**
175	37.234.042.553.192	595.744.680.851.064	425.531.915

Bild 1.2 175 Zettabyte veranschaulicht

Ein sich selbstverstärkender Kreislauf

Diese rasant wachsenden Datenmengen sind weder Ursache noch alleinige Folge technologischer Entwicklungen, sondern Teil eines komplexen Wechselspiels. Das in Bild 1.3 dargestellte Kreislaufmodell illustriert die zirkuläre Dynamik ohne eindeutige Kausalität – keine Komponente fungiert als alleiniger Auslöser. Vielmehr entfaltet sich ein System gegenseitiger Verstärkung, bei dem jedes Element gleichzeitig Ursache und Wirkung ist.

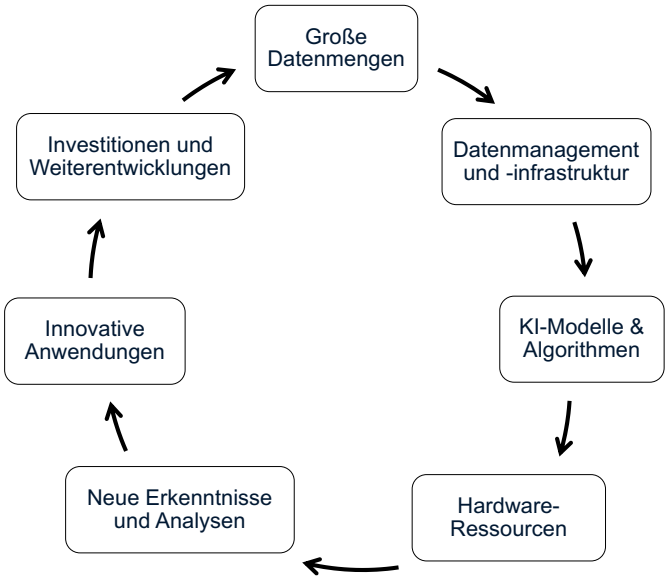


Bild 1.3 Der sich selbstverstärkende Big-Data-Kreislauf

Der Kreislauf beginnt mit der Notwendigkeit für Datenmanagement und -infrastruktur, da zu Beginn des Big-Data-Hypes (ab ungefähr 2005) herkömmliche Speichertechnolo-

gien mit den exponentiell wachsenden Datenströmen überfordert waren. Diese Herausforderung treibt Investitionen in Cloud-Infrastrukturen[3] oder eigene Hardware-Ressourcen voran, deren Entwicklung selbst bemerkenswert ist: Die Leistung von Machine-Learning-Hardware verdoppelt sich alle 2,3 Jahre, während die Performance pro Dollar jährlich um 30 Prozent steigt [Epo24]. Diese technologische Basis ermöglicht es erst, komplexe KI-Modelle und Algorithmen zu entwickeln, die aus heterogen strukturierten Datenmengen systematisch Erkenntnisse extrahieren. Die daraus generierten Analysen inspirieren datengetriebene Geschäftsmodelle oder münden in innovative Anwendungen, die wiederum neue Datenströme erzeugen und weitere Investitionen rechtfertigen.

Vielfältige Datenquellen

Bild 1.4 zeigt die diversen Ursprünge dieser Datenmengen, die den Kreislauf kontinuierlich speisen. Die Entwicklung des Internet of Things (IoT) verdeutlicht diese Dynamik beispielhaft: Die Anzahl der IoT-Geräte soll von 20,1 Milliarden 2025 auf prognostizierte 32,1 Milliarden bis 2030 gewachsen sein [IDC24]. Neue Datenquellen, vielfältige Datenformate, große Datenmengen und kontinuierliche Datenströme treiben wiederum neue Technologien voran, um die Daten speichern, wirtschaftlich verarbeiten und analysieren zu können.

Bild 1.4 Mögliche Quellen für die Herkunft von Big Data

3 In diesem Kontext ist interessant, dass bereits in den 1960er-Jahren John McCarthy das mit Cloud einhergehende Computing so vorhersagte: „*Computing may someday be organized as a public utility just as the telephone system is a public utility*“ [Gar11].

Die Evolution von Big-Data-Technologien

Diese Dynamik spiegelt sich auch in der Evolution der Technologien wider (Tabelle 1.1). Die Entwicklungen verliefen nicht linear, sondern griffen ineinander und verstärkten sich gegenseitig. Doug Laneys Definition der 3Vs (Volume, Velocity, Variety) von 2001 [Lan01] reagierte auf bereits existierende Datenphänomene und schuf gleichzeitig den konzeptionellen Rahmen für deren systematische Bewältigung. Googles MapReduce-Paper [DG04] machte ab 2004 parallele Datenverarbeitung einfach programmierbar sowie auf Standard-Hardware fehlertolerant und skalierbar. Apache Spark (2011) beschleunigte die Verarbeitung um Faktor 100 und eröffnete neue Anwendungsszenarien.

Tabelle 1.1 Ausgewählte Meilensteine im Kontext von Big Data

Jahr	Meilenstein	Erläuterung
2001	3Vs	Doug Laney definiert Big Data durch Volume, Velocity und Variety
2003	GFS	Google File System – verteiltes Dateisystem für Big Data
2004	MapReduce	Programmiermodell für parallele Datenverarbeitung
2005	Hadoop	Open-Source-Implementierung von GFS und MapReduce
2007	NoSQL-DBs	Nicht relationale Datenbanken für flexible Skalierbarkeit
2011	Apache Spark	In-Memory-Computing, bis zu 100-mal schneller als MapReduce
2013	Docker	Containertechnologie für Big-Data-Anwendungen
2014	Apache Flink	Stream-Processing mit echter Echtzeitverarbeitung
2015	Apache Kafka	Message-Broker wird Standard für Streaming-Pipelines
2016	TensorFlow 1.0	Deep-Learning-Framework für Big-Data-Pipelines
2021	Data Lakehouse	Kombination von Data Warehouse und Data Lake
2022	Generative KI	Revolution der Analyse unstrukturierter Daten
2024	Model Context Protocol (MCP)	Offener Standard für die Integration von Kontext- und Tool-Schnittstellen in generative KI-Systeme

Big Data im Kontext industrieller Revolutionen

Die in Bild 1.5 dargestellte Evolution industrieller Revolutionen verdeutlicht einen Wandel im Umgang mit Daten und Information. Während die ersten drei Revolutionen primär durch neue Energiequellen (Dampfkraft), Produktionsmethoden (Fließband) und Automatisierungstechnologien (Computer/IT) charakterisiert waren, markiert die vierte Revolution einen Paradigmenwechsel: Daten werden zum strategischen Produktionsfaktor.

Die vierte industrielle Revolution generierte Big Data als systemimmanente Konsequenz – durch Vernetzung von Fertigungsanlagen und Prozessautomatisierung. Daten

entstehen nicht mehr als Nebenprodukt, sondern dienen fortan als Basis für Geschäftsmodelle, Prozessverbesserungen und Wertangebote. Cloud Computing, verteilte Systeme und neue Analyseverfahren wurden entsprechend als notwendige Infrastruktur ausgebaut [Sei19].

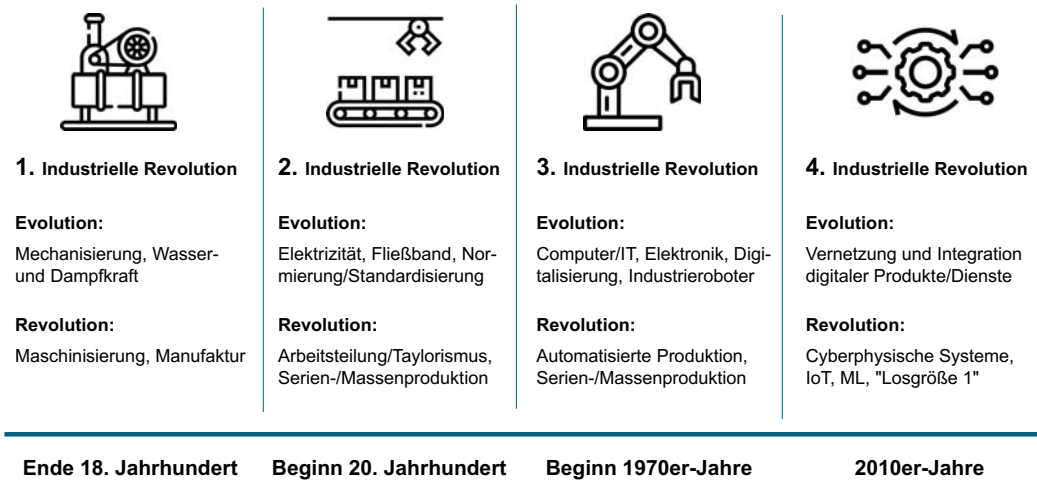


Bild 1.5 Industrielle Revolutionen im Überblick in Anlehnung an [HM24]

Der Übergang zu einer diskutierten fünften industriellen Revolution verschiebt den Fokus erneut. Die Europäische Kommission entwickelt und verfolgt mit Industrie 5.0 drei Säulen: Menschenzentriertheit, Nachhaltigkeit und Resilienz [Eur+21]. Diese Neuausrichtung interpretiert Big Data nicht mehr nur als Instrument der Prozessoptimierung, sondern auch als Mittel, gesellschaftliche Ziele zu verwirklichen, Lieferketten störungsfreier zu machen und eine nachhaltige Kreislaufwirtschaft zu etablieren [Sch+23].

KI geht schon heute Hand in Hand mit Big Data: Während Industrie-4.0-Ansätze Daten zur Prozessoptimierung nutzen, ermöglichen KI-Systeme erweiterte Mensch-Maschine-Interaktionen. Machine-Learning-Algorithmen identifizieren Muster in Datenmengen und passen Produktionsprozesse adaptiv daran an. Solche Systeme entwickeln eigenständig Lösungsansätze aus verfügbaren Daten, bleiben jedoch abhängig von Datenqualität und -umfang. Fehlende oder verzerrte Trainingsdaten führen zwangsläufig zu fehlerhaften Ergebnissen.

Die europäische Vision einer Industrie 5.0 zielt darauf ab, dass Big Data Analytics über reine Effizienzsteigerung hinaus auch gesellschaftliche Ziele verwirklichen soll [Sch+23]. Ob datengetriebene Systeme tatsächlich nachhaltiger oder resilienter werden, bleibt jedoch empirisch zu belegen. Der in Bild 1.3 dargestellte Mechanismus zwischen Datensammlung, KI-Training und Produktionssteuerung erhält eine erweiterte Funktion: Er soll technischen Fortschritt mit gesellschaftlichen Zielen systematisch verbinden.

Die Öl-Metapher und ihre Grenzen

Die Metapher von Daten als dem „neuen Öl“[4] des 21. Jahrhunderts bietet einen eingängigen Vergleich, erfasst aber nur Teilaspekte der Datenökonomie.

Die (vermeintlichen) Parallelen liegen zunächst in den Verarbeitungsprozessen: Wie Öl müssen Daten gesammelt (gefördert), gespeichert (gelagert) und aufbereitet (raffiniert) werden, um irgendeine Form von Wert zu generieren. Beide unterliegen strengen Regularien – Raffinerien folgen Sicherheitsvorschriften, Datenverarbeitung erfordert Data Governance. Diese umfasst die Katalogisierung durch Metadaten, die Einhaltung von Datenschutzrecht und die Nutzung verbindlicher Interoperabilitätsstandards. Auch die Katastrophenpotenziale ähneln sich: Datenlecks verursachen digitale „Ölpesten“, wie der Facebook-Vorfall im Jahre 2021 demonstrierte, bei dem über 500 Millionen Nutzerdatensätze kompromittiert wurden [Ham23; Del21].

Zwischen Daten und traditionellen Ressourcen wie etwa Öl existieren jedoch auch erhebliche Unterschiede: Daten vermehren sich durch ihre Nutzung, statt sich zu verbrauchen. Daten lassen sich zudem zielgerichtet erzeugen [Wer24]. Jede Analyse generiert potenziell neue Erkenntnisse, die weitere Datenerhebungen motivieren. Moderne Produkte erzeugen kontinuierlich Datenströme: Sensoren erfassen Umgebungsparameter, Nutzerinteraktionen werden protokolliert, Transaktionen gespeichert. Diese selbstverstärkende Dynamik steht im Gegensatz zur Endlichkeit fossiler Rohstoffe. Ein weiterer Unterschied betrifft die Reproduzierbarkeit. Öl verbrennt einmalig und ist dann verloren – Daten hingegen lassen sich beliebig oft kopieren und verwenden. Analysen zehren Daten nicht auf; sie sind ökonomisch betrachtet nicht rivalisierend, aber exkludierbar, etwa durch Zugangsrechte oder technische Beschränkungen.[5] Diese Eigenschaft hat jedoch eine Kehrseite: Datenlecks wirken irreversibel, da sich einmal veröffentlichte Informationen praktisch nicht mehr aus dem digitalen Raum entfernen lassen. Hier zeigt sich eine Parallele zu Ölkatastrophen: Beide hinterlassen dauerhafte Schäden in ihren jeweiligen Systemen – physisch bei Ölverschmutzungen, digital bei Datenlecks.

Einen weiteren Unterschied betont Wernicke: Im Vergleich zu physischen Rohstoffen haben Daten als nicht rivalisierende Güter keinen universellen Wert, so wie es etwa für einen Barrel Öl einen weltweit einheitlich definierten Wert pro Menge (aktueller Weltmarktpreis) gibt. Der Wert von Daten ergibt sich aus dem Kontext, in dem sich Daten nutzen lassen. Je nach Nutzer, Anwendungsfall und Erkenntnisziel können die identischen Daten entweder sehr wertvoll oder sehr wertlos sein [Wer24]. Gibt es keinen universellen Wert für Daten, so ist auch der Zusammenhang zwischen Datenmenge und Datenwert unklar [Wer24]. Während physische Rohstoffe wie Öl mit zunehmender

4 Grundlage für diese Metapher ist die Äußerung von Clive Humby aus dem Jahr 2006 [Hum06]. Michael Palmer betont den Aspekt der Raffinierung [Pal06]. Eine tiefere Auseinandersetzung mit der Metapher findet sich bei Hirsch aus dem Jahr 2014 [Hir14] oder auch bei Wernicke aus dem Jahr 2024 [Wer24].

5 So charakterisieren etwa Corrado et al. Daten als *non-rival (yet excludable)* Gut [Cor+22].

Menge in der Regel linear an Wert gewinnen – zwei Barrel sind doppelt so viel wert wie ein Barrel – gilt dies für Daten nicht. Ob zwei Terabyte an Daten mehr Wert erzeugen als eines, hängt von Faktoren wie Qualität, Exklusivität, Redundanz und Nutzbarkeit ab. Denkbare Verläufe reichen von *linear* (konstanter Grenznutzen) über *degressiv* (zunehmender, aber abnehmender Grenznutzen) bis hin zu *überproportional* im Sinne einer S-Kurve (Netzwerkeffekte, sprunghafte Verbesserungen ab einer kritischen Datenmenge). In ungünstigen Fällen kann der Wert mit wachsendem Volumen sogar *negativ* verlaufen, etwa durch überproportional steigende Kosten für Speicherung und Verarbeitung oder durch sinkende Datenqualität. Auf der anderen Seite können aber Daten auch zu sich selbst verstärkenden Kreisläufen führen: Je mehr datenbasierte Produkte gekauft werden, desto mehr entstehen weitere Daten (z. B. Nutzungsdaten), die wiederum neue Funktionalität, mehr Komfort oder höhere Zuverlässigkeit für das Produkt bedeuten können, was wiederum dem Absatz, der Wettbewerbsfähigkeit oder der Innovationskraft zugutekommt [Wer24]. Daten sind keine homogenen Güter: Ihr Wert reicht von „völlig nutzlos“ bis „sehr wertvoll“ und wird entscheidend durch den jeweiligen Anwendungskontext bestimmt.

Ironischerweise teilen beide Ressourcen auch ökologische Problematiken: Sowohl die Ölförderung und -verwendung als auch der massive Energiebedarf von Rechenzentren für Big-Data-Verarbeitung belasten die Umwelt erheblich [Bit24].

Die Öl-Metapher erfasst aber nicht die qualitative Andersartigkeit von Daten im Kontext der industriellen Transformation. Sie entsteht durch die Verbindung von Daten, hier speziell Big Data mit Künstlicher Intelligenz, worunter klassisches Machine-Learning (ML) ebenso fällt, wie auch avanciertere Ansätze wie Deep Learning: Daten werden nicht nur verarbeitet oder beinhalten Informationen, sondern ermöglichen auch lernende Systeme. Diese können sich (idealtypisch) kontinuierlich selbst verbessern und damit „intelligent“ auf sich verändernde Rahmenbedingungen reagieren oder gar zum Bestandteil datengestützter Wertangebote werden [Sei19]. Dieser selbstverstärkende Mechanismus – von der reaktiven Prozessverbesserung in Industrie 4.0 zur Vision einer menschenzentrierten Produktion in Industrie 5.0 [Eur+21] – geht weit über die Möglichkeiten physischer Rohstoffe hinaus. Die Öl-Metapher hilft beim ersten Verständnis, regt zum Nachdenken und Diskutieren an, verdeckt aber den weitreichenden Unterschied zwischen stofflichen und informationsbasierten Ressourcen.

Big Data: Von der Option zur Notwendigkeit

Bild 1.3 hat gezeigt, wie Big Data sich selbst verstärkt. Unternehmen, die Daten systematisch nutzen, setzen einen idealtypischen Kreislauf in Gang: Daten ermöglichen Analysen, diese generieren Erkenntnisse, daraus entstehen Innovationen, die zu neuen Einnahmequellen führen [Wer24]. Diese wiederum finanzieren weitere Investitionen in KI und Big Data. Was für digitale Vorreiter wie Google oder Amazon von Beginn an Bestandteil des Geschäftsmodells war, müssen traditionelle Unternehmen