Die Dimension Zeit, die untrennbar mit dem Raum verbunden ist (Raumzeit), spielt dabei eine wichtige Rolle. Im Moment der Erfassung – wie beispielsweise bei der Überfliegung von Stuttgart – erhält jedes Datum immer auch automatisch einen Zeitstempel. Dieser veraltet, da die Raumzeit in eine Richtung verläuft, wonach die Zukunft, der wir uns entgegenbewegen, vor uns und damit alles überhaupt real möglich zu Erfassende im Hier und Jetzt oder hinter uns liegt.[26] Die Gesamtheit aller Daten über unsere reale Welt ändert sich somit fortlaufend.[27] Die Erfassung mit steigendem Digitalisierungsgrad betrifft alle unsere Lebens- bzw. Anwendungsbereiche mit der Herausforderung, die Veränderungen aktuell zu halten.[28] Das sog. Big-Data-Phänomen birgt für unsere Wissenschaften interdisziplinär Chancen und Risiken[29]: Die Chance, dass alle vorhandenen und im Zugriff stehenden digitalen Daten allumfassende sowie exakte Informationen über alle Phänomene liefern. So könnten beispielsweise Vorhersagen (Prognosemodelle) ständig überprüft, trainiert und damit verbessert werden (siehe Wetterdaten). Dem steht aber das hohe Risiko gegenüber, dass die Daten fehlerhaft sind oder nur Ausschnitte der Realität darstellen, ohne den unbekannten Teil zu kennen (schiefe Daten).

Fazit: Seit Beginn der Zivilisation versucht der Mensch, die ihn umgebende Welt mit ihren natürlichen Gesetzmäßigkeiten immer besser zu beschreiben und zu verstehen, indem er das Sammeln immer zahlreicher und präziser werdender Beobachtungen bzw. Messungen systematisiert und optimiert.

Kelleher und Tierney begründen in ihrem Buch „Data Science“ dieses Handeln damit, dass der Mensch Wissen generiert und anwendet, um die Weisheit (und damit sich selbst) weiterzuentwickeln.[30]

Der DIKW-Pyramide (Abb. 1.5) zufolge ist Wissen interpretierte Information, die als Entscheidungsgrundlage dient, auf deren Basis gezieltes Handeln ermöglicht wird. Aus dem Handeln oder auch Nicht-Handeln lernt der Mensch und zieht seine Lehren (*wisdom = acting on knowledge in an appropriate way*). Datenerfassung und Informationsgewinnung zur Ansammlung von Wissen dienen demnach dem (zukünftigen) Moment, um eine richtige bzw. immer bessere Entscheidung treffen zu können. Wir sprechen hier von einem evolutionären Prozess, denn nicht immer gelingt das.

Ein aktuelles, wenngleich trauriges Beispiel ist die Flutkatastrophe an der Ahr im Juli 2021. Eine umfassende Datengrundlage war vorhanden, die Analysen und Prognosen inkl. Warnungen waren eindeutig. Das Handeln bzw. Nicht-Handeln war im Verhältnis dazu unangemessen, was unnötig Menschenleben kostete.[31] [32] Es könnte die Frage gestellt werden, ob man sich der Notwendigkeit des Handelns im entscheidenden Moment bewusst war.

Denn im Prozess der Wissensgenerierung spielt die Kommunikation hin zum Menschen, der die Notwendigkeit zum Handeln verstehen muss, eine zentrale Rolle. Es

[26] Vgl. Hawking 2016, S. 40 ff.
[27] Vgl. Beaulieu & Leonelli 2022, S. 35 ff.
[28] Vgl. Sester 2006.
[29] Vgl. Riede 2018, S. 1 ff.
[30] Vgl. Kelleher & Tierney 2018, S. 55.
[31] Vgl. Staib 2022.
[32] Vgl. Geomarketing.de 2021.

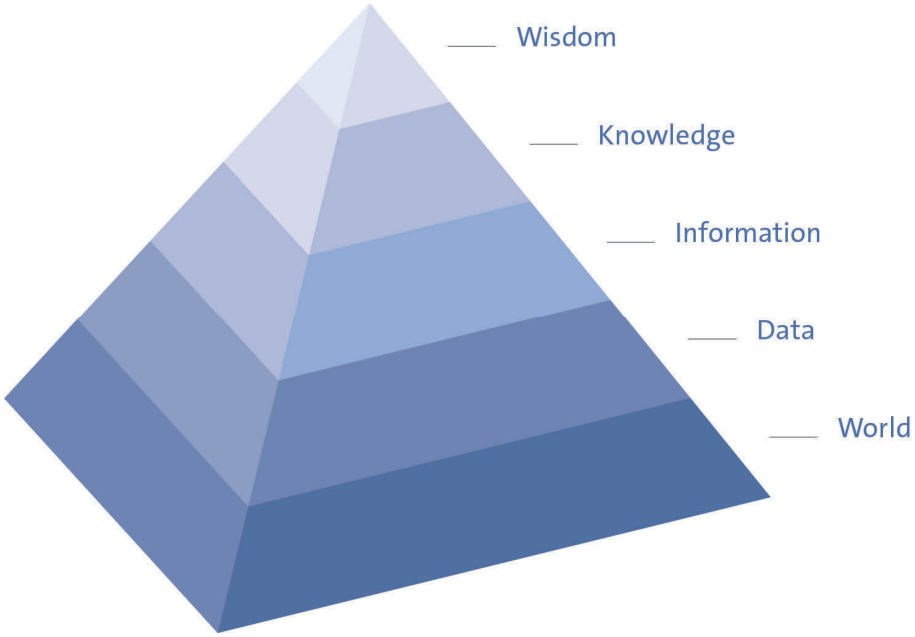

Abb. 1.5 DIKW-Pyramide[33]

handelt sich nach Colin Ware um den letzten Abschnitt im Datenvisualisierungsprozess, bei dem der Datenanalyst die grafische Informationsaufbereitung (das Mapping) manipuliert, um beim Betrachter die gewünschte „kognitive Verarbeitung" zu erzielen (*View Manipulation*).[34]

Fazit: Data Science beschäftigt sich mit der ständigen Optimierung unseres Handelns auf Basis fortlaufender Datenerfassung unserer Welt sowie von unseren evolutionären Erkenntnissen darüber. Dazu zählt auch Visualisierung als Entscheidungsbasis.

Abb. 1.6 Evolutionärer Prozess datengetriebener Handlungsoptimierung (eigene Darstellung)

1.4 Data Science – eine Definition

Es stellt sich als Herausforderung dar, Data Science in einem Satz zu definieren. Denn wir haben gesehen, dass es längst nicht ausreichend ist, dafür die Tätigkeitsfelder eines Daten-Scientisten zum Beispiel anhand des vielzitierten Venn-Diagramms von Conway heranzuziehen.

[33] Vgl. Kelleher & Tierney 2018, S. 55.
[34] Ware 2021, S. 4.

Wir sprechen vielmehr von einem über die Zeit unserer Zivilisation hinweg evolutionären Prozess datengetriebener Handlungsoptimierungen, beginnend mit der Datenerfassung, die eine höchstmögliche Interdisziplinarität bedingt. Die unterschiedlichen Anwendungsfelder führen aber auch zu unterschiedlichen Perspektiven, weshalb dieses Buch ein Kapitel enthält, in dem die jeweiligen Sichtweisen der Autor:innen auf das Feld der Data Science wiedergegeben werden (vgl. Kap. 3).

Der interdisziplinären Vielfalt zum Trotz ist eine allgemeingültige Definition für den Begriff nützlich. Im deutschen Wikipedia als „das" digitale Lexikon wird mit Stand Februar 2022 die direkte Übersetzung des Wort-Kompositums „Datenwissenschaft" ergänzt um die aus dem englischen Wikipedia stammende Erklärung: „Data Science ist ein interdisziplinäres Wissenschaftsfeld, welches wissenschaftlich fundierte Methoden, Prozesse, Algorithmen und Systeme zur Extraktion von Erkenntnissen, Mustern und Schlüssen sowohl aus strukturierten als auch unstrukturierten Daten ermöglicht"[35] *„… and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data."*

Als Referenz werden dazu lediglich zwei Online-Texte genannt, von denen beide weder die wissenschaftliche Bedeutung der Datenerfassung und -generierung (unserer Welt) noch den evolutionären Gedanken der Handlungsoptimierung dahinter andeuten.[36] Ergänzt um diesen Sachverhalt ergibt sich folgende Neudefinition:

Data Science ist ein interdisziplinäres Wissenschaftsfeld, das sich mit der exakten digitalen Erfassung, Analyse und Visualisierung vergangener, aktueller sowie zukünftiger Phänomene unserer realen Welt beschäftigt, um datengetrieben den Prozess der Wissensgenerierung als bestmögliche Entscheidungsbasis für menschliches Handeln zu optimieren.

1.5 Neue Dimensionen in Data Science

Die digitale Abbildung unserer Welt schreitet rasant voran. In Zeiten von Big Data entstehen Daten nicht mehr nur aus gezielten Beobachtungen und Messungen heraus, sondern auch durch die Digitalisierung unseres gesamten gesellschaftlichen Handelns. Wir hinterlassen fortlaufend digitale Fußabdrücke, auch als ‚Datafication' unserer Gesellschaft bezeichnet[37], was dazu genutzt werden kann, das zukünftige Handeln zu beeinflussen. Es kommt zu einer Wechselwirkung. Ein einfaches Beispiel ist die Stauvorhersage, in Echtzeit gewonnen aus dem GPS-Tracking von Autos bzw. Handys, was dazu führen kann, dass man eine andere Route wählt, die ebenfalls getrackt wird, was wiederum dann einen Einfluss auf das Vorhersagemodell hat.

Die Folge ist eine sich verstärkende Entwicklung in unserer Gesellschaft, datengetrieben zu handeln. Das beinhaltet 1.) das Vorantreiben der Digitalisierung unserer Welt (inkl. ihrer Vergangenheit) selbst sowie 2.), auf dieser Basis immer bessere

[35] Vgl. Wikipedia 2022, Artikel „Data Science".
[36] Vgl. Dhar 2013.
[37] Vgl. Beaulieu & Leonelli 2022, S. 6.

Entscheidungen zu treffen bzw. von der Maschine statt vom Menschen treffen zu lassen. Dies führt zu neuen Betrachtungsebenen bzw. Dimensionen in Data Science, die sich aus den gesellschaftlichen Fragestellungen (why, for what) ihrer Welt in Daten (what) und dem, *wie* diese erfasst werden bzw. damit umgegangen wird, ergeben.[38]

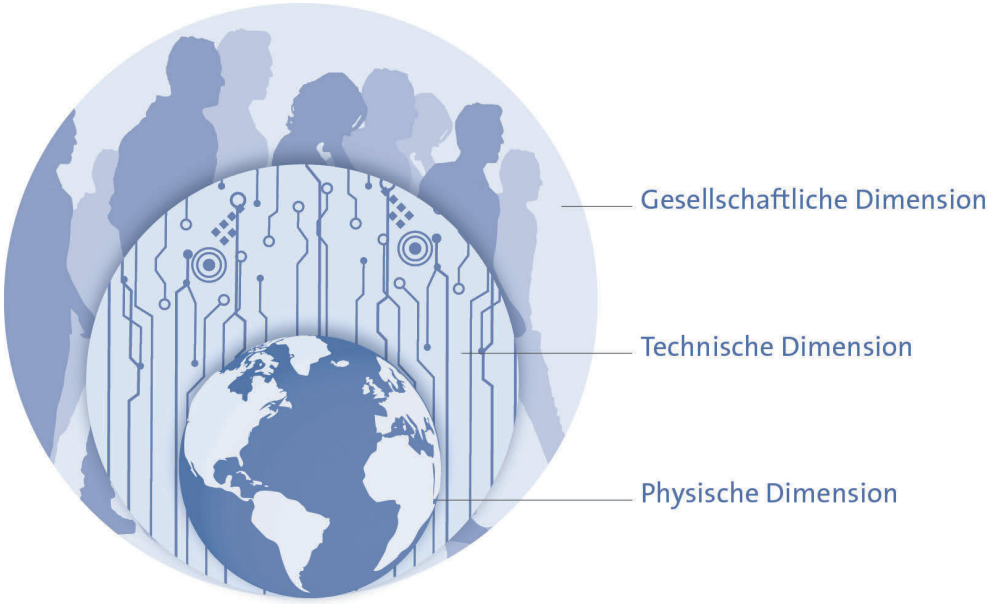

Abb. 1.7 Neue Datendimensionen in Anlehnung an „A model of the spheres of datafication“[39] (eigene Darstellung)

Die gesellschaftlichen Dimensionen (Why) behandeln in erster Linie den evolutionären Aspekt, ob entweder richtige oder falsche Informationen zu „nicht-optimalem“ Handeln führen. Kelleher und Tierney betonen in ihrem Kapitel „Privacy & Ethics“, dass die größte Herausforderung in Data Science darin bestehe, die Balance zwischen Freiheit und Privatsphäre eines Individuums sowie von Minderheiten einerseits und der Sicherheit bzw. den Interessen einer Gesellschaft zu finden.[40] Das Spannungsfeld aus den Spuren, die wir täglich hinterlassen und den Möglichkeiten, unsere Welt durch Daten zu verstehen, wirft ethische Fragestellungen auf, die unter dem neuen und wachsenden Feld der *Data Ethics* behandelt werden. So können exakte Daten, z. B. biometrische Gesichtsdaten, zu einer eindeutigen Erkennung führen, was Chancen (richtiges Handeln) und Risiken (falsches Handeln) birgt.

Eine falsche (schiefe) Informationsbasis kann aber auch aus Daten resultieren, die einen ‚Bias‘ beinhalten, der gesellschaftlich in uns verankert liegt. Beobachten beispielsweise nur Männer die Welt, werden diese Beobachtungen sehr wahrscheinlich (und unbewusst) einen genderspezifischen *Data-Bias* beinhalten. Die wachsende

[38] Vgl. Beaulieu & Leonelli 2022, S. 8.
[39] Beaulieu & Leonelli 2022.
[40] Vgl. Beaulieu & Leonelli 2022, S. 181 ff.

Bedeutung, Diversität (die menschliche Vielfalt) in unserer Gesellschaft wertzuschätzen, hat sich in unserer Datenwelt gleichermaßen in den Dimensionen Geschlecht, ethnische Herkunft und Hautfarbe, Alter, Behinderung, Religion und/oder Weltanschauung sowie sexuelle Identität widerzuspiegeln.[41] Data Science bedingt demnach nicht nur Interdisziplinarität entlang des gesamten Wissensgenerierungsprozesses – von der Datenerfassung bis hin zur datengetriebenen Entscheidung – sondern auch Diversität (*Data Diversity*).

Auf Personenebene geht damit einher, dass die Rolle der *Data Privacy & Protection* immer wichtiger wird. Welche Daten über wen erhoben, geschützt werden müssen und zu welchem Zweck ausgewertet werden dürfen, sind heute zentrale Fragestellungen in unserer Gesellschaft.[42] *Data Privacy* regelt je nach Notwendigkeit dabei nicht nur die Datenerfassung von Personen, sondern auch – wie beispielsweise im Fall der Drohnenverordnung das Überflugverbot mit Kamera von Wohngebieten – die Räume, die man digital erfassen bzw. nicht erfassen darf.[43] Eine großflächige, politisch-gesellschaftliche Debatte entbrannte dazu erstmals 2010, als Google die Straßen und Häuser abfuhr.[44] Heute komplettieren hochauflösende 3D-Satellitenbilder den Internetdienst, der seinen Anwendern den „freien" Blick sogar in den Hinterhof und -garten zu jedem Grundstück (mit und ohne Gebäude) ermöglicht. Bei Google Street View gab es noch die Möglichkeit des „Opt-out" (was zum Verschleiern der Gebäudeansicht führt). Bei den Satellitenbildern geht das nicht mehr, was vielen so gar nicht klar ist. Kein Wunder also, dass 2021 die *Data Literacy Charta* initiiert wurde. Deren Unterzeichner:innen machen deutlich, dass Datenkompetenzen für alle Menschen in einer durch Digitalisierung geprägten Welt wichtig und unverzichtbarer Bestandteil der Allgemeinbildung sind. Die *Data Literacy* umfasst die Fähigkeiten, Daten auf kritische Art und Weise zu sammeln, zu managen, zu bewerten und anzuwenden.[45]

Unter den **technischen bzw. technologischen Dimensionen (How)** lassen sich alle Methoden und Techniken zusammenfassen, die ihren Einsatz von der Datenerfassung bis hin zur Wissensgenerierung finden. Nach Leonelli und Tempini handelt es sich dabei um eine *Data Journey*, die sich vom *Data Creation Gathering* z. B. durch Satelliten oder Befragungen (Surveys) über *Data Processing, Cleaning & Exploration* bis hin zur Visualisierung und Interpretation erstreckt[46] – im Business-Bereich auch als Cross Standard Process for Data Mining (CRISP-DM) bekannt.[47]

Grundlegende Voraussetzung für eine optimale Wertschöpfung aus Daten, das Schaffen von *Data Assets*, sind professionelle Datenmanagement-Prozesse, die die dazu notwendige Dateninfrastruktur und Qualitätsmanagement bereitstellen.[48] Im Rahmen der *Data Governance* von Organisationen wird der *Asset* aus Daten zunehmend an seiner Qualität bemessen. Die DAMA (The Data Management Association)

41 Vgl. Berlin.de 2022.
42 Vgl. Bender 2022.
43 Vgl. drohnen.de 2022.
44 Vgl. Süddeutsche Zeitung 2010.
45 Vgl. Schüller et al. 2021.
46 Leonelli & Tempini 2020.
47 Kelleher & Tierney 2018, S. 57.
48 Weber & Klingenberg 2021, S. 55 ff.

definiert dazu sechs primäre Datenqualitäts-Dimensionen aus Vollständigkeit (*Completeness*), Eindeutigkeit (*Uniqueness*), Aktualität (*Timeliness*), Gültigkeit (*Validity*), Genauigkeit (*Accuracy*) und Konsistenz (*Consistency*).[49] [50] Zur Sicherung von Qualitätsstandards setzen die jeweiligen Fachdisziplinen bei der Datenerfassung verstärkt auf Normen wie DIN und ISO, wie z. B. die DIN PAS 1071 für Geodaten vom Deutschen Dachverband für Geoinformationen[51] oder die ISO 20252 der Markt-, Meinungs- und Sozialforschung.[52] Es schließt sich die in der Data Journey nachgelagerte Dimension des Data Engineering auf Basis einer effizienten IT-Infrastruktur an, bestehend aus Datenbanken, In-Memory-Technologien und Datenmodellen[53], die es zum Ziel hat, die Daten so strukturiert aufbereitet zur Verfügung zu stellen, dass eine optimale Analyse der Daten erfolgen kann. Dazu zählen auch alle Metadaten über den Datenentstehungs- und Bereitstellungsprozess bis zu diesem Zeitpunkt wie beispielsweise das Wissen um einen etwaigen *Bias* in den Daten, über den ab hier jedes maschinelle Lernen oder künstliche Intelligenz stolpern würde.[54] Gerade das Trainieren von Methoden und Algorithmen, automatisch Zusammenhänge und Muster zu erkennen und sich darin selbst zu optimieren, sogar bis hin zur völlig automatisierten Entscheidungsfindung, gewinnt zunehmend an Bedeutung.[55] Eher klassisch stellen sich da die explorativen Methoden dar, sachliche, räumliche sowie zeitliche Muster zu identifizieren bzw. zu prognostizieren. Explorativ, weil die Wahl eingesetzter Analysetools den Data Scientisten letztendlich freisteht. So können beispielsweise auftretende Fälle wie Neukunden über eine Diskriminanzanalyse, das Random-Forest-Verfahren oder eine multinomiale Regression in bestehende Kundencluster bzw. -gruppen eingeteilt werden. Dazu zählen auch visuelle Aufbereitungen von Zwischenergebnissen, die den Data Scientisten zur Überprüfung und weiteren Mustererkennung dienen. Die visuelle Aufbereitung und der Transport des Endergebnisses in jeglicher Form wie z. B. Text und Bild hin zum Empfänger stellen die letzte Dimension dar. Sie kann die alles Entscheidende sein in der Wahrnehmung über richtig oder falsch bzw., ob Handeln notwendig ist oder nicht. Umso wichtiger ist auch bei der Informationsgewinnung wie auch ihrer Kommunikation das Einhalten aller disziplinabhängiger Standards, wie beispielsweise die thematisch richtige Darstellung einer Karte.[56] [57] [58]

Die physischen Dimensionen (What) umfassen die eigentlichen Daten. Sie beschreiben unsere Welt (vgl. die erste Stufe der DIKW-Pyramide) möglichst exakt, wie sie ist, wie sie war und wie sie sein wird.[59] Der Raum zu einem bestimmten Zeitpunkt stellt dabei die Dimension dar, in der alle Gegebenheiten bzw. Phänomene

[49] Weber & Klingenberg 2021, S. 87.
[50] Beaulieu & Leonelli 2022, S. 24.
[51] Vgl. Deutscher Dachverband für Geoinformation e. V. 2022.
[52] Vgl. Wiegand 2018, S. 99.
[53] Quix 2021, S. 85 ff.
[54] Vgl. Weber & Klingenberg 2021, S. 92.
[55] Vgl. Retkowitz 2021, S. 209 f.
[56] „Passau ist Pink“: Mit dem erstmaligen Auftreten der Inzidenz über 500 wurde entgegen üblicher kartographischer Stilmittel die Farbe Pink gewählt und auch beibehalten.
[57] Geroldinger 2020.
[58] Universität Wien 2008.
[59] Stadt Zürich 2022.

auftreten. *„The real world consists of many geographies which can be represented as a number of related data layers“*, lehrt das Buch „Understanding GIS“ bereits im Jahr 1990.[60] Die reale Welt lässt sich so in beliebig viele Datendimensionen untergliedern.[61]

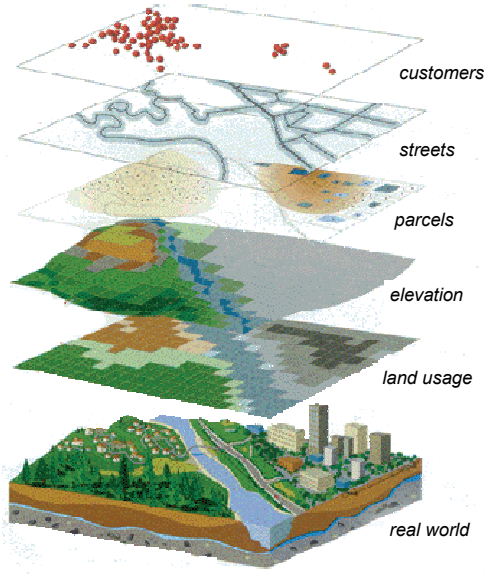

Abb. 1.8
Datenlayer in einem GIS[62]

Inhaltlich wie datenschutzrechtlich stellen der Mensch und die Daten über ihn ein besonderes Phänomen mit unterschiedlichen Eigenschaften dar. Egal ob Daten über seine Biologie, seinen Gesundheitszustand, seine Sprache, sein Verhalten, z. B. Mobilität & Konsum, oder seine Einstellungen: Alle diese sogenannten personenbezogenen Daten unterliegen beispielsweise in Europa der Datenschutzgrundverordnung (DSGVO) und dienen dem Persönlichkeitsschutz, der im Grundrecht verankert liegt. Die Dimension Mensch tritt dabei je nach Lebensbereich in unterschiedlichen Rollen auf, beispielsweise als (potenzielle) Kunden, Kündiger, Infizierte, Patienten, Gäste oder Mitglieder. Als anonymisierte Datensätze kann diese Dimension auch außerhalb des Datenschutzes behandelt werden, so z. B. Anzahl von Kunden in einem Gebiet.

Alle anderen Daten lassen sich zu objektbezogenen Daten zusammenfassen: Daten über Häuser, ihr Alter und Dachtyp, Straßen und Bäume, Daten über Autos und ihren technischen Zustand, Daten über Waschmaschinen und ihre tägliche Nutzung oder Daten über Unternehmen wie z. B. ihre Standorte und Umsätze.

Im Ergebnis entsteht so aus den n-Datenlayern unserer Welt ein vereinfachtes Vier-Dimensionen-Modell aus Wer (Mensch), Was (alle anderen Objekte) sowie Wo und

[60] Vgl. Esri 1990, S. 2.
[61] Vgl. wheelmap.org 2022.
[62] Natural Resource Management: https://www.seos-project.eu/resources/resources-c03-s02.html.