O'REILLY®

Der Leitfaden für Studium und Beruf

Data Science Management

Vom ersten Konzept bis zur Governance datengetriebener Organisationen

Marcel Hebing &
Martin Manhembué

Data Science Management

Vom ersten Konzept bis zur Governance datengetriebener Organisationen

Marcel Hebing, Martin Manhembué

O'REILLY®

Marcel Hebing, Martin Manhembué

Lektorat: Alexandra Follenius
Korrektorat: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner
Umschlaggestaltung: Karen Montgomery, Michael Oréal, *www.oreal.de*

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

ISBN:
Print 978-3-96009-214-8
PDF 978-3-96010-808-5
ePub 978-3-96010-809-2

1. Auflage 2024

Wieblinger Weg 17
69123 Heidelberg

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«.
O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wisse

Inhalt

Einleitung: Ein Handbuch zum Management von Data Science

Ein typisches Szenario: Ein mittelständisches Unternehmen mit Milliardenumsatz vollzieht die digitale Transformation. Die Zukunft des Unternehmens liegt in einer effektiven und effizienten Nutzung von Daten, das ist allen Beteiligten klar. Erste Schritte werden unternommen, es wird investiert, Ziele werden gesteckt, und es vergehen einige Monate. Doch dann kommt der Prozess ins Stocken.

In den Abteilungen des Unternehmens finden sich bereits Menschen, die Datenanalysen durchführen. Dies sind Menschen mit einem Studium in Betriebswirtschaftslehre oder Wirtschaftsinformatik. Es soll wohl auch einen promovierten Physiker geben, der sehr gut in Statistik ist. Man hat vor einiger Zeit alle Mitarbeitenden mit entsprechenden Kompetenzen zusammengezogen und in der IT-Abteilung gebündelt. Die dortige Leitung weiß aber nicht so recht, welche Ziele verfolgt werden sollen, die das Unternehmen voranbringen könnten. Die Verortung in diese Abteilung scheint zwar nicht verkehrt, da man sich auf kurzem Dienstweg Zugang zu Datenbanken und anderen IT-Ressourcen verschaffen kann, aber es bleibt unklar, woran man nun konkret arbeiten soll. Es gibt viele Ideen, aber keine konkreten Projekte, die wertstiftende Ergebnisse liefern.

Das Management des Unternehmens wird langsam unruhig, hatte man doch schon vor Monaten eine Strategie verabschiedet, die das Unternehmen in eine datengetriebene Zukunft führen sollte. Nach einigen Gesprächen mit der IT-Abteilung kristallisiert sich heraus, dass sich Ziele und Mission der Datenanalystinnen und -analysten sowie der Data Scientists klar an der Strategie des Unternehmens orientieren müssen. Da die Entwicklung von Strategien zum Bereich der Geschäftsführung gehören, werden die Data Scientists organisatorisch hier verortet. Die Sprache, die Art der Kommunikation und das hierarchische Gefälle ändern sich schlagartig. Es wird klar, wohin es langfristig gehen soll. Doch leider bleibt über Wochen unklar, was konkret umgesetzt werden soll. Die Reiseziele kennen nun zwar alle, aber das Transportmittel bleibt ungewiss.

Das Unternehmen hat nun also eine Strategie, kompetente Mitarbeitende, eine technische Infrastruktur und sicherlich auch schon umfangreiche Datenschätze aus den operativen Systemen der Fachabteilungen – und doch führen die Bemühungen nicht zu den gewünschten Erfolgen. Es fehlt etwas, das die verschiedenen Komponenten zusammenhält und gleichzeitig entsprechende Prozesse in Gang bringt und antreibt.

Zum einen fehlt es an einer klaren Rolle für die Steuerung dieses Prozesses, die weder vom Topmanagement noch von der Leitung der IT-Abteilung oder einer anderen Fachabteilung wahrgenommen werden kann. Zum anderen fehlt ein Management- und Prozessmodell, um entsprechende Datenanalyseprojekte auch über längere Zeiträume hinweg planen und kalkulieren zu können – es werden zwar viele kleine Projekte angefangen, konnten bisher aber nie in größere, nachhaltige und gewinnbringende Anwendungen überführt werden.

Und es fehlt noch eine Zutat, die im Englischen oft als *Secret Sauce* bezeichnet wird: eine Kultur, die datengetriebene Entscheidungen ermöglicht und Mitarbeitende kollaborativ an Datenanalysen arbeiten lässt.

Das hier dargestellte Beispiel ist zwar fiktiv, basiert aber auf den Erfahrungen, die wir in den letzten zehn Jahren in verschiedenen Rollen als Berater, Data Scientists, Projektmanager und Professoren in der Zusammenarbeit mit Technologie-Start-ups, klassischem deutschem Mittelstand, öffentlich finanzierten Forschungsinstituten und Großkonzernen mit vielen Subunternehmen sammeln durften. Es ließen sich immer wieder zwei Hürden identifizieren, an denen viele Projekte scheitern: das Fehlen einer dezidierten Rolle für das Management von Data-Science-Projekten und unterschiedliche Vorstellungen davon, wie solche Projekte organisatorisch zu gestalten sind.

Für wen ist dieses Buch besonders geeignet?

Um Unternehmen für die oben geschilderten Herausforderungen zu wappnen, haben wir das Konzept für dieses Buch entwickelt. Hier lernen Sie und lernt ihr, was Daten sind und wie man mit ihnen umgeht, wie Datenanalysen durchgeführt werden und welche Werkzeuge hierfür heutzutage infrage kommen. Wir schauen uns den Prozess der Datenwertschöpfung von Anfang bis Ende an und analysieren, wie mit Daten ein Mehrwert für das Unternehmen generiert werden kann. Dabei nehmen wir Sie und euch mit auf eine Reise durch die Datenmodellierung und -verarbeitung und zeigen Best-Practice-Ansätze. Schließlich präsentieren wir Wege, wie man Data-Science-Projekte organisieren kann und als Unternehmen in diesem Bereich erfolgreich wird und bleibt. Zusammengefasst, bietet dieses Buch Folgendes:

- eine Einführung in das Management von Data-Science-Projekten aller Größenordnungen bis hin zur Data Science Governance von Unternehmen,
- einen umfassenden Überblick über konkrete Vorgehen in Data-Science-Projekten,
- einen Einblick in die Schritte zur Automatisierung und Operationalisierung für produktive Data-Science-Anwendungen,
- ein Schritt-für-Schritt-Vorgehen im Data-Science-Lifecycle sowie
- Techniken für den Umgang mit Daten und Stakeholdern für eine erfolgreiche Datenmodellierung.

Wir wollen mit diesem Buch Individuen und Unternehmen in die Lage versetzen, zu verstehen, was Data Science ausmacht und welcher Methoden man sich bedienen kann, um die Komplexität zu managen. Dabei steht für uns im Vordergrund, die Methoden aus dem Bereich Data Science einzuführen, aber nicht erschöpfend zu diskutieren. Für einen umfassenden Überblick über den Bereich Data Science und mögliche Anwendungsfelder empfehlen wir »Data Science für Unternehmen« von Foster Provost und Tom Fawcett (mitp 2017). Entsprechendes statistisches Grundwissen vorausgesetzt, gibt es außerdem sehr gute praktische Einführungen, beispielsweise *Datenanalyse mit Python* von Wes McKinney (O'Reilly 2023) oder *Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow* von Aurélien Géron (O'Reilly 2023). Wir wollen die Grundzüge dieser Methoden unseren Leserinnen und Lesern allerdings nahebringen, damit sie ein breites Wissen über die Arbeitsweise von Data Scientists entwickeln können. Gleichwohl wollen wir dafür werben und Verständnis dafür aufbauen, dass Data Science nicht als Monolith in Unternehmen funktioniert, sondern aktiv in bestehende Strukturen eingebettet werden muss, um zu den Zielen und dem Erfolg des Unternehmens beizutragen.

Die wertstiftende Auswertung von Daten betrifft viele Menschen in Unternehmen, da immer mehr datengetriebene Entscheidungen getroffen werden. Dasselbe gilt für die Unternehmen: Immer mehr Unternehmen analysieren ihre Daten. Das Besondere an unserem Buch ist die deutsche Sprache, die das Buch auch für Menschen in kleinen und mittelständischen Unternehmen interessant macht, die bei englischsprachigen Büchern eventuell eine zu große Sprachbarriere sehen. Ganz konkret richtet sich das Buch an:

- Entscheidungsträgerinnen und Entscheidungsträger sowie Managerinnen und Manager, die Data Science in ihrem Unternehmen einführen wollen,
- Verantwortliche für Projekte und Product Owner im Umfeld von Data Science, Big Data und Data Analytics,
- IT-Verantwortliche, die den Data-Science-Bereich ausbauen und stärken wollen,
- Data Scientists, die sich über statistische und technische Fähigkeiten hinaus fortbilden wollen,
- Studierende in den Bereichen Data Science, Statistik, Wirtschaftsinformatik, BWL, VWL, Digital Business usw. sowie an
- alle interessierten Menschen, die sich weiterbilden möchten.

Was ist Data-Science-Management?

Data Science ist eine interdisziplinäre Wissenschaft, die sich bei den Theorien und Methoden anderer Disziplinen wie Mathematik und Statistik, Computerwissenschaften bzw. Informatik sowie entsprechenden Domainwissenschaften und beim Branchenwissen (also beispielsweise der Betriebswirtschaftslehre im Kontext von Business Analytics) bedient. Das Ziel von Data Science ist es, Entscheidungsprozesse mit Daten bzw. Datenanalysen zu unterstützen.

In Abbildung E-1 ist die Interdisziplinarität visualisiert. Diese bringt es mit sich, dass sich Data-Science-Teams aus Personen mit sehr unterschiedlichen fachlichen Hintergründen zusammensetzen können und mit verschiedenen Stakeholder-Gruppen (beispielsweise anderen Fachabteilungen oder diversen Kundengruppen) zusammenarbeiten.

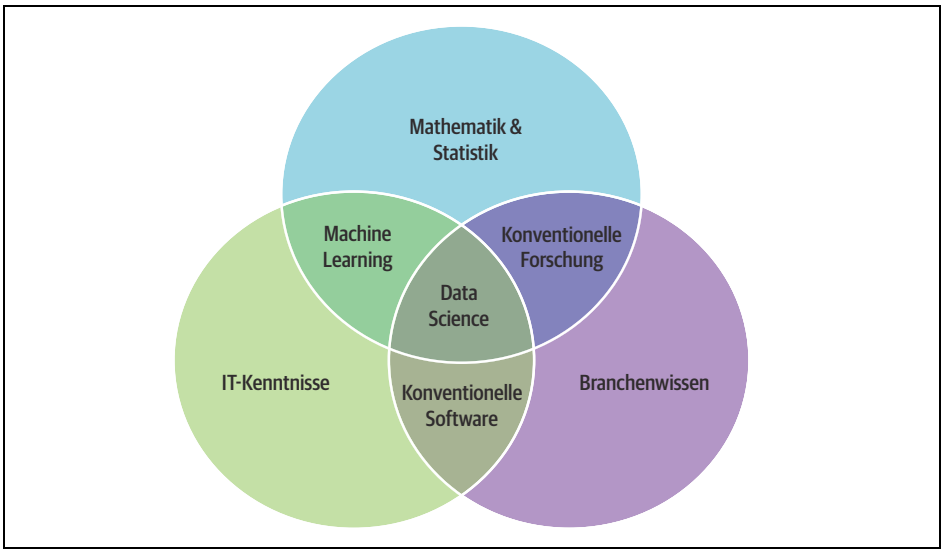

Abbildung E-1: Venn-Diagramm der Fähigkeiten und Disziplinen, die Data Science ausmachen, in Anlehnung an Drew Conway[1]

Das Venn-Diagramm nach Conway zeigt eine Schwäche: Es fehlt die soziale Komponente, die als Kitt zwischen den Disziplinen dient. Interdisziplinarität kann nur funktionieren, wenn Kommunikation stattfindet und es Strukturen gibt, die diese ordnen. Es bedarf also des Managements des Zusammenspiels zwischen Menschen mit ihren unterschiedlichen fachlichen Hintergründen und methodischen Vorgehensweisen.

Sowohl Data Science als auch Data-Science-Management sind sehr junge Bereiche, daher gibt es zahlreiche Ansätze einer Definition. In der Infobox unten finden Sie eine Definition von Data-Science-Management, die wir in diesem Buch mit Leben füllen werden.

Definition von Data-Science-Management

Data-Science-Management (DSM) umfasst Methoden und Theorien zur Organisation und Steuerung von Prozessen, Projekten und Anwendungen, in denen Wissen aus Daten extrahiert wird, um Entscheidungsprozesse zu unterstützen, Produkte zu entwickeln und Ergebnisse zu kommunizieren, die einen Mehrwert erzeugen.

1 Abgerufen unter: *http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram*

In einem typischen Fall von Data-Science-Management steht am Anfang in einem Unternehmen eine Geschäftsidee oder eine Herausforderung. Diese soll datenbasiert bearbeitet werden. Data-Science-Management hilft dabei, den Prozess der Wissensgenerierung durch Anwendung von Methoden aus dem Data-Science-Umfeld und dem klassischen sowie dem agilen Management (siehe Teil II, *Data-Science-Management*) zu strukturieren, zu initiieren, zu steuern und zum Abschluss oder zur Implementierung zu führen. Dabei gibt es einerseits Data-Science-Projekte, die im klassischen Sinne eines Projekts ein definiertes Ziel und Begrenzungen in Bezug auf den zeitlichen Umfang, die finanziellen Ressourcen und die personelle Aufstellung nach DIN 69901[2] haben. Ergebnisse dieser Projekte können beispielsweise eine Projektpräsentation, ein digitales Produkt, wie eine Software oder App, oder schlicht eine Information oder Wissen sein.

Andererseits betrachten wir in diesem Buch solche Vorhaben, die ein langfristiges Engagement zur Folge haben, wie etwa die Entwicklung und das Betreiben einer Software oder das kontinuierliche Anbieten eines datengetriebenen Service.

Data-Science-Management hat viele Gemeinsamkeiten mit dem Prozessmanagement und umfasst daher auch Aspekte wie das Coachen und Unterstützen von Teams (siehe Abschnitt »Coaching und Mentoring von Data Scientists« auf Seite 171), die strategische Ausrichtung von Produkten, Portfolios oder des gesamten Unternehmens (siehe Abschnitt »Wettbewerbsvorteile durch Data Science« auf Seite 236), die Optimierung von Prozessen, das Schaffen und Einhalten von Standards (siehe Abschnitt »Governance, Compliance und rechtliche Aspekte« auf Seite 261) bis hin zur Entwicklung und Pflege einer Organisationskultur (siehe Kapitel 23, *Digitale Kompetenzen und Data-Science-Kultur*), die auf datengetriebenen Entscheidungen basiert.

Warum brauchen Unternehmen Data-Science-Management?

»*The world's most valuable resource is no longer oil, but data*«[3] ist ein oft benutztes Zitat, das seine Bedeutung nicht eingebüßt hat. In den 20er-Jahren des 21. Jahrhunderts befinden wir uns weiterhin in einer Phase des exponentiellen Anstiegs des Datenvolumens. Allein aus diesem Grund setzen viele Firmen auf Spezialisten und Expertinnen im Umgang mit Daten.[4] Denn allein die Menge der Daten erfordert ein strukturiertes und organisiertes Vorgehen, damit diese adäquat verarbeitet und ein Mehrwert generiert werden kann.

2 Deutsches Institut für Normung e. V., *https://www.beuth.de/de/publikation/din-taschenbuch-472/325349267*

3 *The Economist:* »The world's most valuable resource is no longer oil, but data«, *https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data*

4 Den Umfang und die Variation der Methoden, Ansätze und Technologien im Umgang mit Daten betrachten wir in Teil I, *Data-Science-Grundlagen*, im Detail.

Gleichzeitig ist die Menge an Daten allein noch kein Erfolgskriterium. Denn genauso wie Öl sind Daten in Rohform erst einmal von geringem Wert. Erst durch eine Veredlung entfalten beide ihr Potenzial. Bei den Daten ist das die Gewinnung von Informationen und Wissen. Denn für Unternehmen und Individuen sind erst diese tatsächlich wertstiftend. Das liegt insbesondere an der wachsenden Bedeutung der Wissens- bzw. Informationsgesellschaft als viertem (quartärem) Wirtschaftssektor neben Rohstoffgewinnung (primär), Rohstoffverarbeitung (sekundär) und Dienstleistung (tertiär), die wir am Ende der Einleitung erläutern. Durch die Auswertung von Daten wollen die Menschen in den Unternehmen Entscheidungen, die bislang häufig durch Intuition getroffen wurden, daten- und evidenzbasiert treffen. Aufgrund des Wissensvorsprungs können sie einen Wettbewerbsvorteil nutzen und sich wirtschaftlich besser für die Zukunft aufstellen. Zugespitzt könnte man sogar sagen, dass viele Unternehmen zukünftig nur bestehen können, wenn sie datengetriebene Entscheidungen treffen.

Wenn Unternehmen und Individuen dieser Entwicklung folgen wollen, müssen sie technologisch und methodisch in der Lage sein, Daten zu verarbeiten, um daraus Informationen und Wissen zu generieren. Ein Wissenschaftsbereich, der sich insbesondere hiermit beschäftigt, ist die Data Science.

Im Jahr 2012 wurde im Harvard Business Review ein Artikel mit dem Titel »Data Scientist: The Sexiest Job of the 21st Century« veröffentlicht.[5] Darin wird das Argument von Hal Varian, Chefökonom bei Google, aufgegriffen, das er drei Jahre zuvor äußerte:

»The sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?«

Sowohl das Zitat als auch der Artikel betiteln Jobs und Berufe, die im jungen 21. Jahrhundert große Aufmerksamkeit erfahren haben. Diese wollen wir unter dem Begriff Data Science subsumieren. Wir werden im Folgenden noch darauf eingehen, welche Rollen und Aufgaben es in diesem Feld gibt. Über allem steht die Erfassung von Komplexität in Daten, die Verdichtung von Information und die wissensinduzierende Kommunikation. Der Artikel und das Zitat von Hal Varian können mindestens als ein Beschleuniger für einen bislang exponentiellen Anstieg an Data Scientists weltweit angesehen werden.[6]

Bereits vor dem Entstehen des Zitats von Hal Varian und dem Artikel gab es Menschen, die sich Data Science verschrieben haben. Eine spannende Überlegung ist an dieser Stelle, ob die folgende Phase ab etwa 2010 davon geprägt war, dass Unternehmen unter Einfluss des Phänomens der *Fear of Missing Out* (FOMO) Data Scientists einstellten, oder ob die Unternehmen händeringend nach Data Scientists suchten, die die Use Cases endlich umsetzen würden. Die Frage bleibt also: Was war zuerst da, der Hype um Data Scientists oder die Real World Problems in den Unternehmen?

5 T. H. Davenport und DJ Patil. »Data Scientist: The Sexiest Job of the 21st Century«. *Harvard Business Review*, *https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*

6 Stitch Data, *https://www.stitchdata.com/resources/the-state-of-data-science/*

Wir finden für diese Frage bislang kaum eine evidenzbasierte Antwort. Jedoch können wir festhalten, dass es anekdotische Evidenz gibt, dass Data Scientists in einigen Arbeitsumgebungen noch heute nicht optimal eingesetzt werden. Das liegt daran, dass die Unternehmen teilweise wenig vorbereitet sind, Data Scientists mit dem auszustatten, was sie benötigen, um wirksam zu sein. Zumindest gibt es einen Hinweis darauf, dass Data-Science-Projekte zu einem großen Teil scheitern: Atwal[7] berichtet, dass nur 22 % der Data-Science-Projekte hohe Einnahmen generieren. Bei Projekten mit Bezug zu Big Data sind es gar 60 bis 85 %, die gänzlich scheitern.

In diesem Spannungsfeld betrachten wir Data-Science-Management. Wenn wir auf der einen Seite einen hohen Bedarf an Menschen haben, die aus Daten Informationen und Wissen generieren sollen, und mehr Unternehmen Data Scientists anstellen, wir aber auf der anderen Seite eine Situation haben, in der die meisten Data-Science-Projekte scheitern, benötigen wir einen Rahmen, der die Herausforderungen solcher Projekte aufzeigt und Lösungen entwickelt. Mit Data-Science-Management fassen wir Werkzeuge, Methoden, Prozesse und Denkweisen zusammen, die dabei helfen sollen, Data Science plan-, steuer- und messbar zu machen.

Wie arbeitet man mit diesem Buch?

Dieses Buch ist in vier thematische Schwerpunkte unterteilt: Data-Science-Projekte, Data-Science-Management, Infrastrukturen und Architekturen für Data Science und sowie Data Science Governance. Inhaltlich fokussieren sich die Buchteile entweder auf technologisch-anwendungsbezogene oder management- und organisationsorientierte Ansätze. Abbildung E-2 gibt Ihnen einen Überblick über Ausrichtung und Inhalte der einzelnen Teile.

Die Teile I bis IV bauen aufeinander auf, wobei sich insbesondere die ersten beiden Teile den Grundlagen des Data-Science-Managements von Projekten widmen. Für Menschen, die erste Ideen in Richtung Data Science haben und starten wollen, und für die, die wenig bis keine Vorkenntnisse haben, ist dies der ideale Startpunkt in das Buch und damit in die Thematik. Erfahrene Data Scientists und Menschen aus Unternehmen, die bereits erste Data-Science-Projekte umgesetzt haben und nun einzelne Aspekte vertiefen wollen, können direkt ab Teil III, *Infrastruktur und Architektur*, einsteigen.

In Teil I, *Data-Science-Grundlagen*, geht es um die methodischen Voraussetzungen, um relevante, erfolgskritische Aspekte und um die Ressourcen für ein Data-Science-Projekt. Wir schauen uns Prozessmodelle bzw. Data-Science-Lifecycles an und führen damit eine Vorgehensweise zur Umsetzung von Data-Science-Projekten ein. Entlang des Data-Science-Lifecycle vertiefen wir in diesem Teil die Themen Designen von Projekten, Datenerhebung und -verarbeitung, Analyse und Analysemethoden,

7 H. Atawal (2020). *Practical DataOps: Delivering Agile Data Science at Scale*. Apress, *https://link.springer.com/book/10.1007/978-1-4842-5104-1*

Möglichkeiten zur Dokumentation und die zielgerichtete inhaltliche Evaluation sowie die Bemessung der Wirkung.

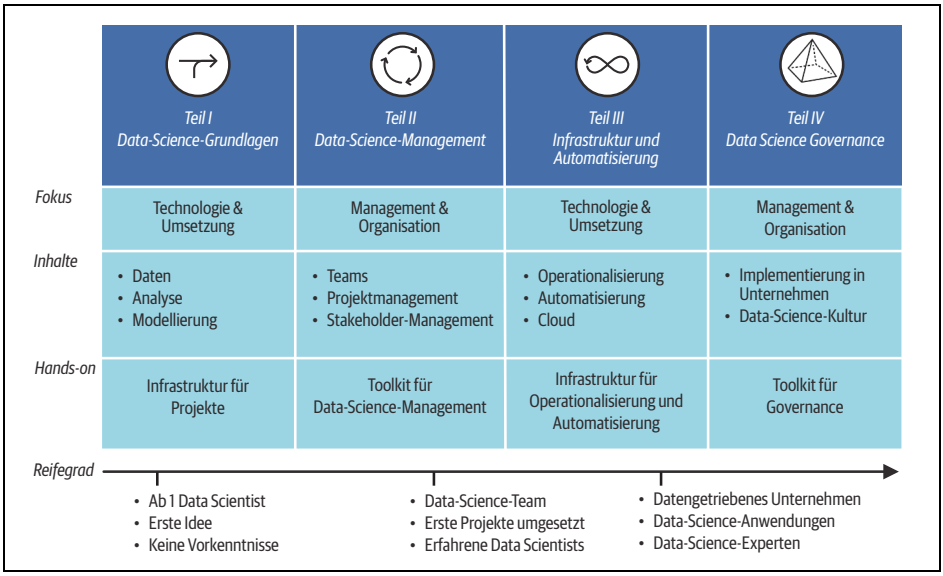

Abbildung E-2: Dieses Buch ist in vier Teile gegliedert. Je weiter man nach rechts geht, umso eher passen die Themen zu einem Unternehmen mit einem hohen Data-Science-Reifegrad.

Am Ende eines Buchteils teilen wir in einem Hands-on-Teil praktische Erfahrungen mit Tools und Methoden, die es ermöglichen sollen, die theoretischen Grundlagen möglichst schnell in die Praxis umsetzen zu können. Im Hands-on-Kapitel von Teil I stellen wir ein Analysebeispiel vor. Wir schauen uns die Entwicklung der Selbstständigen in Deutschland seit 1957 an und werden dabei insbesondere auf Fragen der Datenqualität und der Kommunikation mit Daten noch einmal anhand des Beispiels genauer eingehen.

In Teil II, *Data-Science-Management*, befassen wir uns mit den Aspekten der Organisation von Data-Science-Projekten und den Teams, die diese durchführen. Wir blicken hier insbesondere auf die Grundlagen und auf Möglichkeiten zum bestmöglichen Management. Einerseits schlagen wir Ansätze vor, die es den Data-Science-Teams ermöglichen, effizient miteinander zu arbeiten und zu kommunizieren. Dies soll eine Arbeitsatmosphäre schaffen, die den Data Scientists umfangreiche Gestaltungsmöglichkeiten bietet und zu einer wertstiftenden Umgebung führt. Andererseits beleuchten wir die Kommunikation mit den Fachbereichen und anderen Stakeholdern in Hinblick auf die Optimierung des Prozesses von einer Geschäftsidee oder einem Businessproblem hin zu einer datengetriebenen Lösung. Zudem befassen wir uns ausführlich mit der Rolle des Data-Science-Managers (Managerin und Manager) und wie diese durch modernes Leadership einen Mehrwert für die Teams und das Unternehmen erbringen können.

Im Hands-on-Kapitel von Teil II stellen wir ein Toolset aus Boards, Canvases, Checklisten und anderem vor, das sich für uns in der Praxis bewährt hat.

In Teil III, *Infrastruktur und Architektur*, widmen wir uns der Frage, wie eine nachhaltige Umgebung für die Entwicklung und den Betrieb von Data-Science-Anwendungen aussieht. Das heißt, wir verlassen teilweise die Ebene der terminierten Projekte und kommen in den Bereich der produktiven Anwendungen. Hierzu betrachten wir die technologischen Voraussetzungen sowie die agile Softwareentwicklung, um Algorithmen in einen fortlaufenden Betrieb zu bringen. Ein besonderer Fokus liegt auf dem Konzept der *Machine Learning Operations* (MLOps) für den Betrieb solcher Systeme.

Im Hands-on-Kapitel zu Teil III schauen wir uns visuelle Tools an, die bei der Konzeption und Modellierung von verschiedenen Aspekten einer Dateninfrastruktur helfen. Dies umfasst unter anderem die Modellierung von (sozialen) Prozessen, in die ein Data-Science-Projekt eingebettet ist, die Darstellung der Datenbank oder die Struktur der Software.

Teil IV, *Data Science Governance und Data-driven Culture*, behandelt – flankierend zu den technischen Aspekten des vorangegangenen Buchteils – die Voraussetzungen, die sich aus den Veränderungen in der Arbeitswelt ergeben, und die Herausforderungen bei der Implementierung von Data Science in Unternehmen. Schließlich gehen wir auf eine gelebte Data-Science-Kultur als aus unserer Sicht den höchsten Reifegrad für Unternehmen ein und betrachten die Erfolgsfaktoren vom Individuum bis zur Implementierung von Data Science im Unternehmen.

Im Hands-on-Kapitel von Teils IV betrachten wir Werkzeuge, um die Steuerung und Governance im Unternehmen methodisch zu begleiten. Hierzu zählen weitere Canvases zur Erarbeitung von digitalen Geschäftsmodellen und ein Datenstrategie-Designkit. Eine wiederverwendbare Tabelle mit einem Überblick über Schlüsselfaktoren für erfolgreiches Data Science in Unternehmen rundet das Buch schließlich ab.

Begleitend zu diesem Buch bieten die Autoren Zusatzmaterial wie Videos, Podcasts und Blogposts an: *https://datasciencemanagement.de/*

Wie alles begann oder: der Aufstieg der Digital Economy

Die Arbeitswelt und die Arbeitsbedingungen unterliegen einem ständigen Wandel. Im Zuge der industriellen Revolutionen der vergangenen drei Jahrhunderte haben sich Tätigkeiten geändert, und das soziale Umfeld der Arbeitenden wurde teils erheblich schlechter. In anderen Zeiten haben wir durch soziale Gesetzgebung und Streiks eine Verbesserung der Arbeitsverhältnisse gesehen. Vor diesem Hintergrund müssen wir auch die Entwicklungen der letzten etwa drei Jahrzehnte in den Blick nehmen und uns die Fragen stellen:

- Was hat sich verändert?
- Warum hat es sich verändert?
- Welche Auswirkungen hat das auf die Menschen und Unternehmen?

Mit diesen Fragen im Hinterkopf betrachten wir im Folgenden den Aufstieg der New Economy, wie diese das Wirtschaftssystem und die Arbeitsplätze verändert und welche Veränderungen in den nächsten Jahren denkbar sind. Mit diesem Hintergrundwissen sind wir in der Lage, besser zu verstehen, warum Data Science als Teil dieser Entwicklungen gemanagt werden muss und wie wir das umsetzen können.

Die Entwicklung der Wirtschaftssysteme weltweit unterlag bis dato mindestens drei großen industriellen Revolutionen (siehe Abbildung E-3). Häufig wird noch die vierte industrielle Revolution genannt. Allerdings ist diese bereits zuvor ausgerufen worden, man postulierte also, dass dies eine industrielle Revolution darstellen wird. Die anderen wurden erst im Nachhinein historisch betrachtet beschrieben und stellen damit einen mehr oder minder abgrenzbaren Zeitraum dar.

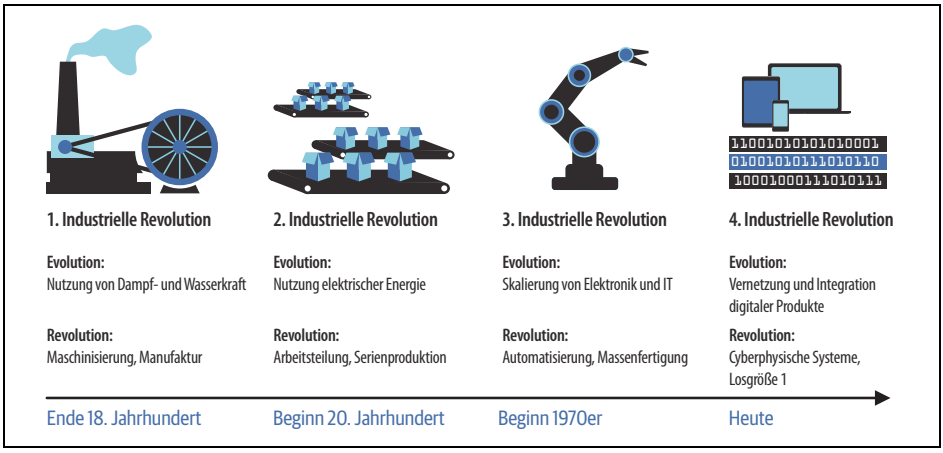

Abbildung E-3: Phasen der industriellen Revolution mit deren evolutionären und revolutionären Eigenschaften

Die erste industrielle Revolution fand ihren Anfang in Großbritannien zur Mitte des 18. Jahrhunderts. Ihr ging eine lange Zeit in Frieden voraus. Immer mehr Menschen wurden geboren, und damit war viel Arbeitskraft vorhanden, die Insellage bot einen geschlossenen Handelsraum, es gab reiche Kohlevorkommen, und die Seenähe erlaubte einen transatlantischen Handel. Ein aufkommender Erfinder- und Gründergeist führte beispielsweise zur Entwicklung der Spinning Jenny (ein Webstuhl), zu Dampfmaschine, Glühbirne, Benzinmotor und Fotografie. Von Großbritannien ausgehend breitete sich die Industrialisierung und Maschinisierung auf ganz Europa aus. Obwohl die Gesamtbevölkerung einen enormen Zuwachs an Wohlstand und eine Verbesserung der Lebensumstände erfuhr, litt die Arbeiterschaft unter 16-Stunden-Tagen[8] und widrigen Lebensumständen.

8 Geschichte der Gewerkschaften – Ausbeutung und Massenelend, *https://www.gewerkschaftsgeschichte.de/industrielle-revolution-ausbeutung-und-massenelend.html*

Die zweite industrielle Revolution, die etwa zwischen 1870 und 1880 begann, war geprägt von einer zunehmenden Verzahnung von Forschung und Industrie. Die Unternehmen betrieben selbst Forschung und erzielten dadurch Durchbrüche in chemischen und physikalischen Prozessen. Somit sind die chemische Industrie, die Elektrotechnik und der Maschinenbau prägend für diese Zeit. Telefone, Telegrafen und der Ausbau der Eisenbahn waren außerdem Treiber der Globalisierung. In den USA bildeten sich mit dem Taylorismus und dem Fordismus Produktions- bzw. Managementtechniken heraus, die eine Arbeitsteilung, das Aufteilen in Prozessschritte und eine Effizienzsteigerung mit sich brachten. Sie führten auch dazu, dass die Arbeitsbedingungen verbessert wurden und Industriearbeiterinnen und -arbeiter mehr verdienten. Autos aus der Massenproduktion, die teilweise exportiert wurden, verstärkten den Trend zur Globalisierung. Der internationale Handel ermöglichte eine Diversifizierung des Angebots und neue Absatzmärkte. Die Schattenseiten waren allerdings Imperialismus, Ressourcenausbeutung und Kolonialisierung, beispielsweise in afrikanischen Ländern.

Die in unserer Betrachtung dritte industrielle Revolution wird auch als digitale Revolution oder mikroelektronische Revolution bezeichnet, die ohne die beiden vorherigen nicht möglich gewesen wäre. Wir können ihren Beginn etwa in den 1980er-Jahren des vergangenen Jahrhunderts verorten. Noch viel stärker als die erste und zweite industrielle Revolution hat die digitale Revolution global und zeitgleich stattgefunden. Dies lag unter anderem an der Vernetzung der Akteure und des Handels und damit der schnellen Übermittlung von Informationen, was zu Innovationen und neuen Technologien führte.

Bereits vor den 1980er-Jahren wurden allerdings wichtige Schritte unternommen, die der Digitalisierung zuzuordnen sind. Dokumente und Informationen wurden digitalisiert, indem man die Informationen in Zuständen darstellte. Man hatte die Möglichkeit, eine Zahl in Einsen und Nullen abzubilden: 1 für »Strom an« und 0 für »Strom aus«. Im englischsprachigen Raum bezeichnet man dies als *Digitization*. Etwas später war es möglich, komplexe Berechnungen und ganze Prozesse als Einsen und Nullen darzustellen. Dies nennt man auch *Digitalisation*. Hiermit war es möglich, die Automatisierung von Prozessen weiter voranzutreiben.

Ein wichtiger Teil der digitalen Revolution war die Weiterentwicklung des Internets, das bis dato weitestgehend vom Militär und den Universitäten genutzt wurde. Berners-Lee und Cailliau entwickelten am Forschungszentrum CERN in Genf Hypertext-Protokolle (das HTTP), Links und Webbrowser. Auf diese Weise entwickelten sie das World Wide Web, das wir heute kennen. Nur durch diese technische Innovation, die Digitalisierung eines Prozesses, konnte die digitale Transformation stattfinden. Mit dem Web 2.0, das im Jahr 2004 erstmals in Fachartikeln erwähnt wird, also dem interaktiven und kollaborativen Internet, wurden Nutzerinnen und Nutzer weltweit Informationen verfügbar gemacht und Kommunikationswege eröffnet. Dies war der Anfang von Social Media, dem Internet of Things, Cloud-Services und damit auch der digitalen Transformation.

Neuer Wirtschaftssektor: Informationen

In der Volkswirtschaftslehre gliedert man die Wirtschaft üblicherweise in Sektoren.

- Der Primärsektor umfasst dabei die sogenannte *Urproduktion*. Das ist im Wesentlichen die Landwirtschaft.
- Der Sekundärsektor ist die Industrie und das Gewerbe. Hierzu zählen beispielsweise die Produktion von Automobilen, das Baugewerbe und die Lebensmittelverarbeitung.
- Im Tertiärsektor werden die Dienstleistungen zusammengefasst, die beispielsweise vom Staat erbracht werden. Aber auch Banken, Versicherungen, Handel und der öffentliche Verkehr gehören dazu.

Bereits 1961 hat Jean Gottmann einen weiteren Wirtschaftssektor definiert, der bis dato nicht existierte. Die Rede ist hier vom Quartärsektor bzw. dem Informationssektor. Dieser Sektor zeichnet sich durch Tätigkeiten aus dem Dienstleistungssektor bzw. dem tertiären Sektor und wohl auch durch Tätigkeiten aus dem industriellen Sektor, dem sekundären Sektor, aus, die besonders hohe intellektuelle Ansprüche und einen hohen Grad an Vorbildung und Ausbildung voraussetzen. Zudem wird hier eine große Bereitschaft vorausgesetzt, Verantwortung zu übernehmen. Der Informationssektor umfasst insbesondere die Bereiche, die mit der Datenerzeugung, Datenverarbeitung und damit auch der Generierung von Wissen beschäftigt sind. Das Geschäftsmodell in diesem Bereich ist also häufig wissens- bzw. datenbasiert. Hierzu zählen

- die Beratung, also die großen Kanzleien und Steuerberatungsunternehmen sowie natürlich die Unternehmensberatungen,
- alle IT-Dienstleister wie etwa AWS oder Microsoft,
- die Unternehmen aus der Kommunikationstechnik,
- die Hochtechnologie mit Robotik, maschinellem Lernen, Digitalisierung usw.

Im Jahr 1983 brachte die Zeitschrift Time ein Heft mit dem Titel »The New Economy«[9] heraus. Das stellt vermutlich den Startpunkt für die Beschreibung dieser neuen Art des Wirtschaftens mit Informationen dar (siehe Infobox). Die New Economy basiert auf Informationen und Dienstleistungen. Die Dienstleistungen, die dabei im Fokus stehen, bestehen eher aus immateriellen Wirtschaftsgütern, wie zum Beispiel Informationen und Wissen. Der Trend zur Nutzung von Daten, der einen ersten Höhepunkt in den 1990er-Jahren fand, erfasste die gesamte US-Wirtschaft und wurde zu einem globalen Phänomen.

Die wirtschaftliche Bedeutung wurde zum Teil durch die Etablierung und Nutzung der Computer erreicht, um die sich eine ganze Industrie aus Mikrochips- und Halbleiterherstellung aufbaute. Dieser neue Industriezweig umfasste aber auch die Produktion der Endgeräte und die Softwareentwicklung. Durch neue Sensoren und

9 C. P. A. Monday. »The New Economy«. *Time Magazine*, 1983.

Analysemethoden bot sich die Möglichkeit, Prozesse und Tätigkeiten messbar und dadurch besser steuerbar zu machen, was zu einer Effizienzsteigerung führte. Das Internet bot zudem neuen Raum für Kommunikation, Werbung und Produkte. Die Einführung der Technologiebörse NASDAQ als Alternative zum NYSE[10] war ein weiteres Puzzleteil für massive Investitionen in die Tech-Branche. Dies ging so weit, dass diese stark überbewertet wurde, in der sogenannten Dotcom-Blase Anfang der Nullerjahren einen herben Rückschlag erlitt und riesige Vermögenswerte und damit auch Vertrauen in den Markt vernichtete.

Der Aufstieg der digitalen Ökonomie war dadurch jedoch nicht gestoppt. Bis heute sehen wir, dass Tech-Giganten, insbesondere aus den USA[11], unseren Alltag prägen. Die wertvollsten Unternehmen stammen heutzutage nicht mehr nur aus der Automobilindustrie, der Rohstofferzeugung oder der Energieträgergewinnung, sondern auch aus der digitalen Ökonomie. Einige sind der Meinung, dass die digitale Ökonomie (*New Economy*) die Grundregeln der klassischen Volkswirtschaftslehre aus den Angeln hebt. Diese vermögensbasierte Ökonomie setzt darauf, dass man allem einen Wert beimessen kann, beispielsweise durch Geld oder Aktien. In der »Old Economy« mussten hingegen Waren und Dienstleistungen einen tatsächlichen (materiellen) Wert haben. Die Unternehmen, die der digitalen Ökonomie zugeordnet werden, haben teilweise über Jahre hinweg rote Zahlen geschrieben, wurden nur durch Investorengeld gehalten und trotzdem im Milliardenbereich bewertet, obwohl sie zum Großteil aus immateriellen Werten bestanden. Viele dieser Unternehmen revolutionierten jedoch unser Leben und veränderten die Art, wie wir konsumieren und kommunizieren. Herausragende Beispiele hierfür sind:

- Social Media, z. B. Facebook, WeChat oder Twitter (X)
- Onlinehandel, z. B. Amazon oder Alibaba
- Onlinebezahldienste, z. B. PayPal
- Onlinemedien, z. B. YouTube oder Netflix
- Onlinewerbung, z. B. Google
- Sharing Economy, z. B. Uber oder Airbnb
- Onlinedating, z. B. Tinder

An dem Erfolg dieser Dienste (Software) ist die Zugänglichkeit über Endgeräte (Hardware) maßgeblich beteiligt. Unternehmen wie Microsoft oder Apple gehören auch deshalb zu den wertvollsten Unternehmen der Welt, weil sie Personal Computer und Smartphones herstellen. Insbesondere bei Smartphones gibt es noch weitere global agierende Unternehmen wie Samsung oder Huawei, die den Markt prägen. Die Software dieser Geräte basiert wiederum auf den Betriebssystemen von Apple und Google.

10 NASDAQ ist ein Kursindex und eine elektronische Börse, die historisch eher Technologieunternehmen abbildet. Sie gilt als Konkurrent zur deutlich älteren NYSE (New York Stock Exchange) an der Wall Street, die die größte Wertpapierbörse der Welt ist.

11 Diese Entwicklung fand auch in anderen Ländern statt, zuvorderst China. Allerdings haben chinesische Technologieanbieter in Europa und den USA eine bedeutend kleinere Rolle als US-amerikanische Unternehmen – mit Ausnahme von TikTok.

Wie schon bei den industriellen Revolutionen zuvor hat der Aufstieg der digitalen Ökonomie in der digitalen Revolution zu vielen Veränderungen für Menschen und Unternehmen geführt. Es besteht, auch durch die voranschreitende Globalisierung, der Druck, sich zu verändern und digitaler zu werden. Das gilt sowohl für Menschen als auch für Unternehmen. Mittelständische Unternehmen handeln häufig über die deutschen Grenzen hinweg und stehen somit in globaler Konkurrenz. Deshalb ist es eine große Aufgabe, die Unternehmen in und durch die digitale Transformation zu führen und den Menschen entsprechende Fähigkeiten zu vermitteln. Dieses Buch handelt davon, wie dies in Bezug auf Data Science gelingen kann.

Danksagung

Data Science ist ein Teamsport. Jede Entdeckung, jede Innovation in diesem dynamischen Feld ist das Ergebnis von Zusammenarbeit, gegenseitiger Inspiration und dem gemeinsamen Bestreben, das Unbekannte zu erforschen und zu verstehen. Ähnlich wie im Sport, wo das Zusammenspiel verschiedener Talente und Fähigkeiten zum Erfolg führt, baut auch Data Science auf der Synergie von Fachwissen, Kreativität und technischem Know-how auf.

Als Autorenduo dieses Werks möchten wir uns deshalb ganz herzlich bei allen bedanken, die zum Gelingen dieses Buches beigetragen haben. Besonderer Dank gilt den vielen fleißigen und fachkundigen Korrekturleserinnen und -lesern und Fachgutachterinnen und Fachgutachtern, die ihre Zeit und Expertise großzügig zur Verfügung gestellt haben, um sowohl Teile als auch das gesamte Manuskript kritisch zu prüfen und zu verfeinern: Robert Bölke, Marcus Fraaß, Martin Habedank, Kevin Loncsarszky, Fabian Payer, Anne-Kristin Polster, Svenja Rohr, Sarah Stemmler, Martin Szugat, Ramon Wartala und Arif Weider.

Ein besonderer Dank gebührt unserer Lektorin Alexandra Follenius, deren tiefgreifendes Verständnis und unermüdliche Geduld das Rückgrat dieses Projekts bildeten. Ihre Fähigkeit, sowohl die großen Linien als auch die feinsten Details im Blick zu behalten, hat maßgeblich dazu beigetragen, die Qualität und Klarheit unseres Werks zu steigern.

Nicht zuletzt möchten wir unseren Familien und Freunden unseren tiefsten Dank aussprechen. Ihr habt uns durch eure Unterstützung, euer Verständnis und eure Geduld während der vielen Stunden, die wir in dieses Projekt investiert haben, beigestanden. All das ist nur durch euch möglich.

TEIL I
Data-Science-Grundlagen

Um ein Data-Science-Team effizient und effektiv leiten zu können, braucht es ein grundlegendes Verständnis davon, mit welchen Tätigkeiten und Herausforderungen ein solches Team in der täglichen Arbeit konfrontiert ist und wie es diese üblicherweise lösen wird. Und auch wenn Sie keine Teamleitung abstreben, sondern beispielsweise als Auftraggeber mit einem externen Partner zusammenarbeiten, wird Ihnen dieses Verständnis dabei helfen, das Projekt zu planen, Herausforderungen und Lösungsansätze zu bewerten, ein gemeinsames Verständnis im Team zu schaffen und, alles in allem, das Projekt zu einem erfolgreichen Abschluss zu führen.

KAPITEL 1

Eine Einführung in Data Science aus Projektsicht

In einem Data-Science-Projekt wollen wir Daten und Analysen nutzen, um einen Mehrwert für uns, unser Unternehmen oder unsere Kunden zu schaffen. Wichtig ist dabei, dass nicht alles, was mit Daten zu tun hat, automatisch Data Science ist. Die operative Nutzung von Daten, beispielsweise in der Buchhaltung, der Inventarliste oder im CRM-System, muss zunächst einmal nichts mit Data Science zu tun haben, sondern kann einfach nur der Abwicklung operativer Prozesse dienen. Data Science kommt ins Spiel, sobald wir einen zusätzlichen Mehrwert durch die Analyse dieser Daten schaffen wollen. Bei Bedarf können wir darüber hinaus zusätzliche Daten erheben, um komplexere Fragestellungen zu beantworten. Dabei stellt sich die Frage, welche Arten von Mehrwert wir mit Daten und Analysen erzeugen können. Wir gehen davon aus, dass wir Data Science in einem Unternehmen einsetzen möchten. Dann können wir grundsätzlich drei Einsatzarten unterscheiden:

Prozessoptimierung: Wir nutzen Data Science, um die Prozesse und Abläufe in unserem Unternehmen zu verbessern. Dabei kann jeder Funktionsbereich (Buchhaltung, Personalwesen, Marketing usw.) davon profitieren, wenn bessere Informationen zur Verfügung stehen. Dies kann je nach Anwendungsfall zu Kosteneinsparungen, besseren Entscheidungen oder schnelleren Prozessabläufen führen.

Datenbasierte Produkte und Geschäftsmodelle: Daneben können wir Data Science einsetzen, um unsere Produkte zu verbessern oder neue Produkte zu entwickeln. Entscheidend ist hierbei, dass die Verwendung von Data Science ein Teil des Mehrwerts wird, den wir unserer Kundschaft bieten. Manche Unternehmen entwickeln Daten und Analyseergebnisse selbst zu Produkten, andere ergänzen bestehende Produkte und machen beispielsweise eine Glühbirne »smart«.

Letztlich können auch Daten selbst ein Produkt sein, wenn die Daten einen Mehrwert für andere haben, beispielsweise die Immobilienpreise einer Region. Dies funktioniert allerdings in der Praxis nur für relativ wenige Anbieter. Die meisten setzen auf datenbasierte Produkte und Geschäftsmodelle.

Strategische Entscheidungen: Bei strategischen Entscheidungen geht es um einmalige Entscheidungen mit wichtigen Konsequenzen. Die Entscheidungen sind so schwerwiegend, dass es sich lohnt, ein Datenanalyseprojekt hierfür aufzusetzen.

Folglich ergibt sich der Mehrwert von Data Science bei der Prozessoptimierung eher durch eine Vielzahl vergleichbarer Entscheidungsprobleme, auf die entsprechend optimiert werden kann. In der Strategie hingegen geht es mehr um Einzelfallentscheidungen, bei denen die Analysen stärker in die Tiefe gehen. In der Praxis kann es dabei aber auch zu einem fließenden Übergang kommen, wie wir weiter unten im Zusammenhang mit dem Analytics Continuum sehen werden.

In der Literatur (Beispiel: Valliappa Lakshmanan. *Data Science on the Google Cloud Plattform*, O'Reilly 2022) sehen wir manchmal die Unterscheidung, dass einmalige strategische Entscheidungen als »Datenanalysen« bezeichnet werden und die Optimierung von Prozessen (mit potenziell automatisierten Analysen und Entscheidungen) als »Data Science«. Für unsere Einführung zu Data Science wollen wir den Begriff »Data Science« jedoch bewusst weiter fassen und auch einmalige Analyseprojekte einbeziehen, vor allem weil es sich hierbei eher um eine theoretische Abgrenzung handelt, die unserer Erfahrung nach nicht zur Praxis von Data-Science-Projekten und deren Management passt.

Verlauf eines Data-Science-Projekts (Prozessmodell)

In Data-Science-Projekten lassen sich gewisse wiederkehrende Abläufe identifizieren, die eigentlich immer vorkommen, sinnvollerweise in einer gewissen Reihenfolge ablaufen sollten und entsprechend als *Prozessmodell* dargestellt werden können. Das Prozessmodell, das den folgenden Darstellungen zugrunde liegt, besteht aus fünf Prozessschritten, die einerseits ein existenzieller Teil jedes Data-Science-Projekts sind, andererseits aber auch spezifische Anforderungen an das Team und dessen Kompetenzen stellen (siehe Abbildung 1-1). Das Modul wurde als Teil von Beratungsprojekten der Impact Distillery[1] entwickelt und basiert insbesondere auf dem *Generic Longitudinal Business Process Model*[2] (GLBPM) sowie dem Prozessmodell von Mischa Seiter[3].

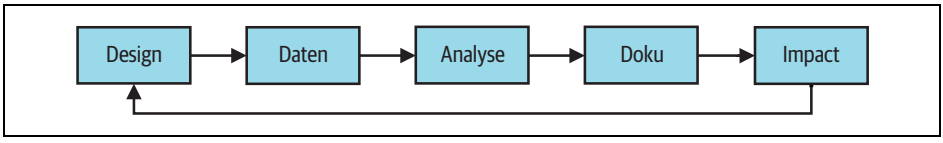

Abbildung 1-1: Prozessmodell der Impact Distillery (https://www.impactdistillery.com/de/digitale-transformation/datengetriebene-organisationskultur/)

Die fünf Schritte unseres Modells umfassen die konzeptionelle Planung (Design) des Projekts, die Arbeitsschritte, um eine belastbare Datengrundlage zu schaffen, die eigentliche Analyse der Daten, die Dokumentation der Ergebnisse und deren Umset-

1 *https://www.impactdistillery.com/*

2 I. Barkow, W. Block, J. Greenfield, A. Gregory, M. Hebing, L. Hoyle, W. Zenk-Möltgen. »Generic Longitudinal Business Process Model«. *DDI Working Paper Series – Longitudinal Best Practices*, No. 5, 2013, *https://ddialliance.org/sites/default/files/GenericLongitudinalBusinessProcessModel.pdf*

3 M. Seiter (2019). *Business Analytics: Wie Sie Daten für die Steuerung von Unternehmen nutzen*. Vahlen.