Data Science Management

Vom ersten Konzept bis zur Governance datengetriebener Organisationen

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].

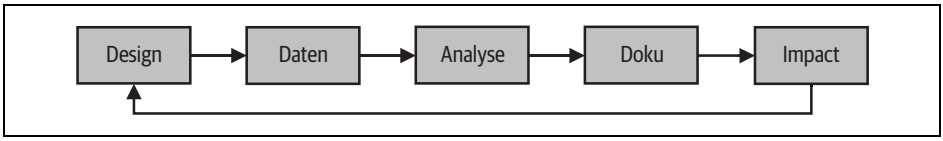

Abbildung 1-1: Prozessmodell der Impact Distillery (https://www.impactdistillery.com/de/digitale-transformation/datengetriebene-organisationskultur/)

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.

zung in praktische Maßnahmen (Impact). Außerdem setzt das Modell ein iteratives Vorgehen voraus – sobald ein solches Projekt abgeschlossen ist, stehen für gewöhnlich neue Fragestellungen im Raum, die den Ausgangspunkt für ein neues Data-Science-Projekt bilden. Die fünf Schritte wollen wir uns im Folgenden einzeln anschauen:

- **Design:** Die Designphase legt den Grundstein für das Projekt. Idealerweise starten Projekte, weil es einen praktischen Bedarf (ein Businessproblem) gibt, der aber noch zu unspezifisch ist. Ein erster Arbeitsschritt ist nun, diesen Bedarf bzw. diese Problemstellung in eine Forschungsfrage zu übersetzen, die dann im Fokus aller folgenden Arbeitsschritte stehen wird. Basierend auf der Forschungsfrage kann jetzt auch ein Zeitplan für das Projekt entwickelt und können die notwendigen Ressourcen kalkuliert werden, die insbesondere die Beschaffung von Daten, eine technische Infrastruktur und personelle Ressourcen umfasst.
- **Daten:** In der zweiten Phase (siehe Kapitel 3, *Datenbeschaffung und -aufbereitung*) geht es um den Aufbau einer entsprechenden Datenbasis für die Bearbeitung der Forschungsfrage. Wenn nicht schon entsprechende Daten verfügbar sind, müssen gegebenenfalls neue Daten erhoben werden. In jedem Fall müssen diese Daten aufbereitet, qualitätsgesichert und für die weitere Nutzung bereitgestellt werden.
- **Analyse:** Die Auswahl der entsprechenden Analysemethoden orientiert sich dann sowohl an der Forschungsfrage als auch an der Struktur der Daten und gegebenenfalls auch an bereits durchgeführten Vorstudien. Im Abschnitt »Von einfachen Analysen zur Automatisierung (Analytics Continuum)« auf Seite 32 werden Sie das Analytics Continuum kennenlernen, das uns eine Entscheidungshilfe für die Auswahl von Analysemethoden in den aufeinander aufbauenden Phasen eines Data-Science-Projekts bietet. Dabei werden wir uns sowohl Methoden der klassischen Statistik als auch neuerer Machine-Learning-Algorithmen bis hin zu neuronalen Netzen ansehen.
- **Dokumentation:** Damit die Ergebnisse der Analysen dann praktisch genutzt werden können, müssen sie dokumentiert und kommuniziert werden. Dabei geht es zum einen um eine technische Dokumentation, um Daten und Methoden später nachnutzen zu können. Und zum anderen sollen die Ergebnisse ansprechend und leicht nachvollziehbar für ein nicht technisches Publikum aufbereitet werden, beispielsweise als Report oder interaktives Dashboard (siehe den Abschnitt »Reporting« auf Seite 83). Inhaltlich sind dabei nicht nur die vorteilhaften Ergebnisse zu berichten, sondern es sollte auch explizit auf mögliche Limitationen der jeweiligen Arbeit eingegangen werden. Gleichzeitig sollten die Inhalte aber für die jeweiligen Leserinnen und Leser verständlich präsentiert und erzählt werden (siehe dazu auch den Abschnitt »Storytelling und visuelle Kommunikation mit Daten« auf Seite 85).
- **Impact:** Mit Impact meinen wir alle praktischen Maßnahmen, die einen Mehrwert für den jeweiligen Auftraggeber bringen und damit die Kosten für ein Data-Science-Projekt rechtfertigen. Dieser Mehrwert kann monetär leicht messbar (z.B. wenn eine Steigerung der Verkaufszahlen gelingt) oder auch schwerer zu

greifen sein (z. B. wenn es um eine Steigerung der Kundenzufriedenheit geht). In jedem Fall ist es sinnvoll, die entsprechenden Maßnahmen zu evaluieren, um zu überprüfen, ob sie auch die gewünschte Wirkung haben, oder um gegebenenfalls nachzusteuern.

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

Von einfachen Analysen zur Automatisierung (Analytics Continuum)

Während die vorgestellten fünf Phasen unseres Prozessmodells gut geeignet sind, um einzelne Projekte zu strukturieren, werden wir in der Praxis selten nach einem einzelnen Projekt wieder aufhören, mit Daten zu arbeiten. Vielmehr werden die fünf Phasen in aufeinander aufbauenden Iterationen immer wieder neu durchlaufen, weswegen man auch von einem *Data-Science-Lifecycle* spricht. Aus fast jedem Data-Science-Projekt wird sich eine neue Fragestellung ergeben, die wir in einer neuen Iteration bearbeiten können. Dies können sowohl die Evaluation der Maßnahmen sein als auch eine weiterführende Analyse, beispielsweise wenn wir einen spannenden Zusammenhang in unseren Daten gefunden haben und uns nun fragen, ob wir diesen vielleicht auch für Vorhersagen nutzen können. Schließlich ist es möglich, sich in späteren Iterationen bis zu einer Automatisierung der Maßnahmen vorzuarbeiten (siehe Kapitel 15, *Automatisierung und Operationalisierung im kybernetischen Regelkreis*).

Auf dieser Ebene bietet uns das Analytics Continuum[4] von Gartner eine Struktur, anhand der wir uns im Laufe der Zeit und über verschiedene Iterationen hinweg von einfachen beschreibenden Analysen bis hin zu komplexen Automatisierungsprojekten bewegen können (siehe Abbildung 1-2).

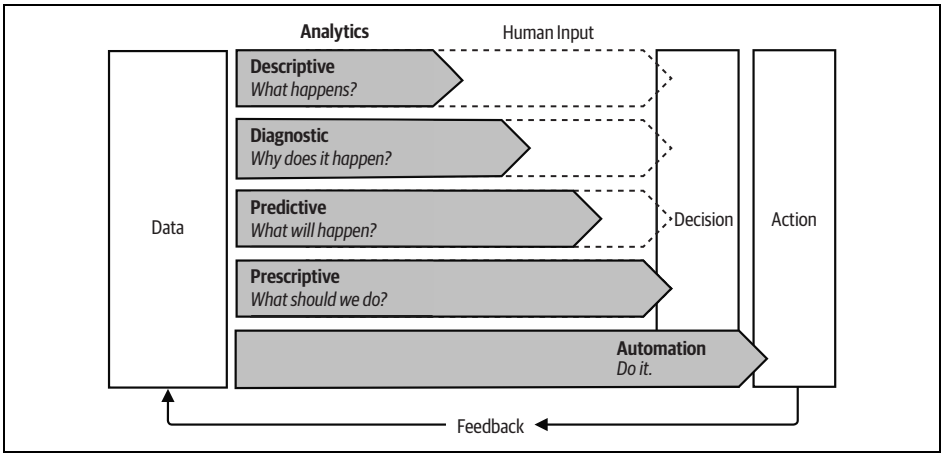

Abbildung 1-2: Analytics Continuum nach Gartner (eigene Darstellung)

4 *https://www.gartner.com/en/newsroom/press-releases/2014-10-21-gartner-says-advanced-analytics-is-a-top-business-priority*

Schauen wir uns die fünf Ebenen des Analytics Continuum im Detail an:

Beschreibende Analysen (Descriptive)
: Am Anfang eines Projekts brauchen wir einen guten Überblick über den Status quo, also das, was gerade passiert. Dabei reichen meist einfache deskriptive Analysen und Visualisierungen aus, um schnell die aktuelle Lage einschätzen zu können, gegebenenfalls auch im Vergleich mit historischen Daten. Ein einfaches Beispiel ist das Inventarverzeichnis in einem Lager. Basierend darauf können wir uns einen Überblick darüber verschaffen, welche Produkte, Produktkategorien, Marken oder Ähnliches wir aktuell vorrätig haben.

Diagnostische Analysen (Diagnostic)
: Als Nächstes werden wir uns fragen, wie es zum aktuellen Zustand gekommen ist. Wenn beispielsweise ein Produkt im Lager nicht mehr vorhanden ist, liegt das daran, dass dieses Produkt nicht mehr verfügbar ist? Oder daran, dass die Nachfrage so groß ist, dass wir in der Lieferung kaum hinterherkommen? Im Bereich der diagnostischen Analysen interessieren wir uns besonders für kausale Beziehungen. Was ist die Ursache für bestimmte Phänomene?

Vorhersagende Analysen (Predictive)
: Wenn wir die Ursachen verstanden haben, können wir versuchen, darauf aufbauend Vorhersagen zu treffen. Wenn wir merken, dass die Nachfrage nach einem bestimmten Produkt gerade sehr hoch ist, wollen wir beispielsweise wissen, wie groß die Nachfrage voraussichtlich im nächsten Monat sein wird, um entsprechende Vorkehrungen treffen zu können.

Vorschreibende Analysen (Prescriptive)
: Nachdem wir nun eine Vorstellung davon haben, wie viele Produkte im kommenden Monat nachgefragt werden könnten, stellt sich als Nächstes die Frage, wie viele wir davon nachbestellen sollten. Dies ist etwas anderes als die reine Menge der Nachfrage, denn nun müssen wir zusätzliche Faktoren miteinbeziehen: Wie lange ist die zu erwartende Lieferdauer? Wie viel Platz haben wir im Lager zur Verfügung? Wie haltbar ist das Produkt? Wir wollen nun eine Handlungsempfehlung formulieren, haben es dabei aber schnell mit einem Optimierungsproblem zu tun, wenn wir die angedeuteten Fragen miteinbeziehen. Ist beispielsweise nur begrenzt Platz im Lager, müssen wir vielleicht zwischen mehreren Produkten abwägen, die aktuell stark nachgefragt sind.

Automatisierung (Automation)
: Wenn sich unsere Vorhersagen und Handlungsempfehlungen über längere Zeit bewährt haben, werden wir in Erwägung ziehen, diese zu automatisieren. Wir können also beispielsweise in der Software der Lagerhaltung ein Programm einbauen, das automatisch nachbestellt, sobald ein Produkt knapp wird, und dabei die Ergebnisse der vorherigen Phase nutzen, um die richtigen Mengen zu kalkulieren.

Ein häufig anzutreffender konzeptioneller Fehler, den wir immer wieder in Diskussionen um den Einsatz von Data Science sehen, ist ein vorschneller Fokus auf die letzten Phasen, insbesondere auf die Automatisierung von Prozessen. Eine wesentli-

che Erkenntnis aus der langjährigen Arbeit mit dem Analytics Continuum ist, dass wir die ersten Phasen nie überspringen können. Wir werden uns immer erst mal einen Überblick über den Status quo verschaffen müssen, verstehen, wie dieser zustande gekommen ist, und erste Vorhersagen testen. Erst dann können wir uns an die Entwicklung von Empfehlungssystemen oder die Automatisierung von Prozessen machen.

Im dritten Teil des Buchs werden wir dann sehen, dass insbesondere mit zunehmender Automatisierung der Prozessabläufe (egal ob bei der Auswertung der Daten oder auch bei der Umsetzung in Maßnahmen) eine Anpassung des Prozessmodells Sinn ergeben wird. Sie werden dazu in Kapitel 15, *Automatisierung und Operationalisierung im kybernetischen Regelkreis*, den kybernetischen Regelkreis als Modell und Werkzeug zur Strukturierung von automatisierten Prozessen kennenlernen.

Welche Kompetenzen brauchen wir in einem Data-Science-Projekt?

Data Science wird gern als inter- oder transdisziplinäre Wissenschaft bezeichnet. Das bedeutet, dass Data Science ganz wesentlich auf einer Reihe anderer Disziplinen aufbaut. Conway (2010[5]) nennt dabei Programmierkenntnisse (Softwareentwicklung), Mathematik und Statistik sowie fundiertes Wissen um das jeweilige Anwendungsfeld (im Folgenden als Domain Knowledge bezeichnet) als die drei wesentlichen Fundamente für den Bereich Data Science. Wir möchten diese drei Bereiche noch um einen vierten Bereich ergänzen, der sich auf soziale Normen und Kommunikationsfähigkeit bezieht (die soziale Dimension).

Statistik und Mathematik
: Aus der Statistik übernimmt Data Science sowohl Methoden, um ein initiales Verständnis für die jeweiligen Daten zu gewinnen (deskriptive Statistik), als auch vielfältige Methoden zur Berechnung von abstrakten Modellen. Während bei der klassischen Statistik der Fokus der Modellbildung mehr auf dem Erklären von Zusammenhängen liegt, konzentriert sich die Modellbildung bei Data Science vorrangig auf die Vorhersage von Ereignissen. Beispiele für Vorhersagen können von der Wettervorhersage über die Erzeugung von Kaufempfehlungen in Onlineshops bis zur Automatisierung des Nachkaufs in einem Warenlager reichen. Ein fundiertes statistisches Grundwissen bleibt auch in Zeiten zunehmend automatisierter Analysetools unerlässlich, denn wir müssen weiterhin hinterfragen, ob die Ergebnisse verlässlich und anwendbar für unsere jeweiligen Fragestellungen sind.

Softwareentwicklung
: Neben dem Fokus auf Vorhersagen ist die Bereitstellung und Analyse der Daten im Bereich Data Science deutlich rechenintensiver als in der klassischen Statistik, was die Softwareentwicklung ins Spiel bringt. Der Begriff *Big Data* bezieht

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

sich nicht nur auf das reine Speichervolumen der Daten, sondern schließt insbesondere auch Vielfältigkeit, teilweise Korrektheit und letztlich die Geschwindigkeit der Entstehung neuer Daten mit ein – alles Faktoren, die neben der eigentlichen Analyse der Daten wachsende Anforderungen an die (automatisierte) Aufbereitung der Daten stellen. Gleichzeitig müssen viele der Methoden aus der Statistik an die neuen Gegebenheiten angepasst werden, beispielsweise weil deren Berechnung über verschiedene Teilsysteme verteilt werden muss.

Fachexpertise
: Es wird gern als Faustregel genommen, dass in einem Data-Science-Projekt nur ca. 20 % der Arbeitszeit auf die eigentliche Arbeit an den Modellen entfällt und ca. 80 % auf die Aufbereitung der Daten. Diese 80 % erfordern neben dem bereits dargestellten technischen Wissen auch ein gutes Verständnis des jeweiligen Anwendungsfalls. Von Data Scientists wird daher erwartet, dass sie entsprechendes Vorwissen im jeweiligen Fachgebiet bzw. der jeweiligen Domäne mitbringen.

Soziale Dimension
: Die Zusammenarbeit und Kommunikation mit Stakeholdern ist ein wesentlicher Teil der Arbeit in Data-Science-Teams. Es geht nicht nur darum, ein möglichst gutes Modell zu entwickeln, die Ergebnisse müssen auch angemessen präsentiert und kommuniziert werden. Darüber hinaus sehen wir in den letzten Jahren, dass sich Data Scientists zunehmend mit sozialen Aspekten der Verwendung ihrer Arbeit auseinandersetzen müssen. Insbesondere wenn es sich um personenbezogene Daten handelt, hat die Einführung der Datenschutzgrundverordnung (DSGVO) neue Maßstäbe gesetzt. Aber auch bei anderen Datenquellen sind rechtliche Aspekte nicht zu vernachlässigen, beispielsweise das Urheberrecht oder Firmengeheimnisse (siehe Kapitel 22, *Sicherheit und Datenschutz*).

Abbildung 1-3 gibt einen Überblick über die vier Bereiche und zeigt auch noch einmal zusätzliche Schnittstellen zwischen diesen auf. So können wir beispielsweise die klassische (empirische) Forschung an der Schnittstelle von Statistik und Fachwissen verorten. Klassische Unternehmensberatung findet häufig an der Schnittstelle von sozialer Dimension und fachlicher Expertise statt, insbesondere in Hinblick auf betriebliche Abläufe. Fragen der Nutzerfreundlichkeit (*Usability*), aber auch des Datenschutzes lassen sich insbesondere zwischen Programmierung und sozialer Dimension verorten. Und die Entwicklung von Machine-Learning-Algorithmen erfordert sowohl fundiertes mathematisches Wissen als auch Programmiererfahrung. Entsprechend werden wir bei einer genaueren Betrachtung dessen, was Data Science eigentlich ist, auch immer wieder Aspekte dieser verschiedenen Schnittstellen finden. Wie schon gesagt, Data Science ist eine interdisziplinäre Wissenschaft, und entsprechend gibt es viele angrenzende Bereiche, von denen wir gute Theorien und bewährte Tools übernehmen können.

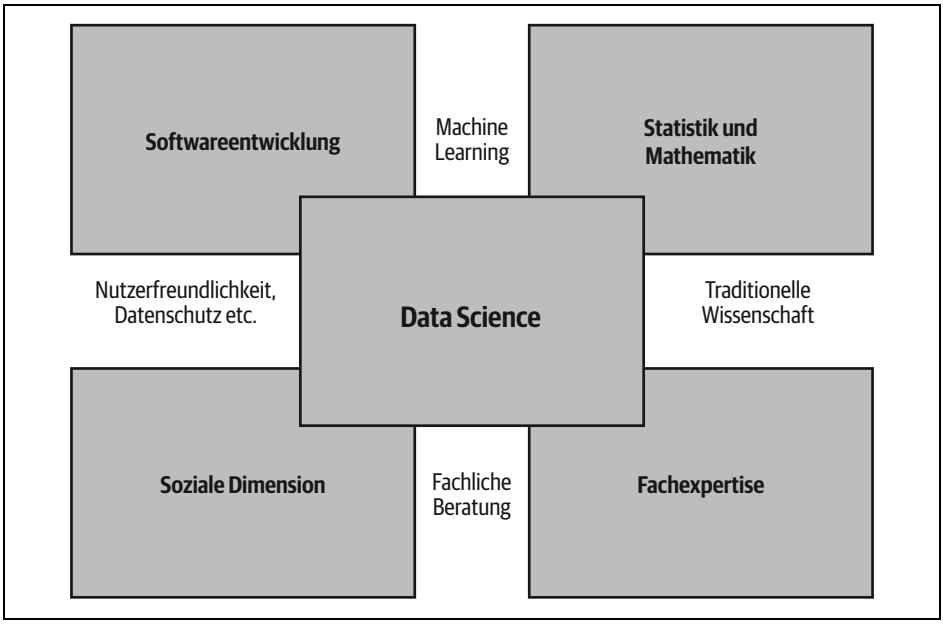

Abbildung 1-3: Data Science als interdisziplinäre Wissenschaft