

Hausarbeit

Integration der Data Science in Organisationen / Abteilungen

Vorlesung Data-Science Projektorganisation

Leon Henne

Köln, den 4. November 2023

Betreut durch Prof. Dr. Michael Schulz

Inhaltsverzeichnis

Tabellenverzeichnis	I
Abbildungsverzeichnis	II
1 Problemstellung	1
2 Grundlagen	2
3 Data Science in datengesteuerten Organisationen	4
4 Herausforderungen	7
5 Integrationsprozesse	9
5.1 CSPG Framework	9
5.2 Design Parameters	9
5.3 Experiment Evolution Model	10
6 Fazit	12
Literatur	13

Tabellenverzeichnis

Abbildungsverzeichnis

2.1	Disziplinen der Data Science	3
3.1	In der Literatur beschriebene Aspekte von datengesteuerten Organisationen .	5
3.2	Rollen und deren Zusammenarbeit in Data Science Teams	5
4.1	Herausforderungen einer datengesteuerten Organisation	7
5.1	Gestaltungsparameter einer datengesteuerten Organisation	10
5.2	Gestaltungsparameter einer datengesteuerten Organisation	11

1 Problemstellung

Viele Trends der technologischen Welt innerhalb der letzten Jahrzehnte wurden besonders durch drei bestimmende Faktoren beeinflusst. Die Entwicklung der Datenspeicherungs- und Rechenkapazitäten als zwei dieser Faktoren entstammen den Erkenntnissen und sich bestätigenden Entwicklungen von Gordon E. Moore bzw. dem Moorschen Gesetz.¹ Demnach wurde bereits sehr früh erkannt, dass integrierte Schaltungen die bekannten Telefonschaltungen ersetzen wird und Computer leistungsfähiger entsprechende Daten verarbeiten können.² Erkenntlich wird diese Entwicklung anhand der heutzutage konstanten und umfangreichen Generierung von Daten durch Anwendungen in Mobiltelefonen, Autos, IOT Geräten oder Industriemaschinen.³ Prognosen sagen dabei eine verfünffachende Entwicklung der globalen Datenmenge vorher von 33 Zettabytes in 2018 zu 175 Zettabytes in 2025.⁴ Der dritte Faktor ist die Entwicklung der Möglichkeiten zur Datenverarbeitung und Analyse anhand von zunehmend komplexeren Algorithmen.⁵

Folgend aus diesen drei Faktoren zielen immer mehr Softwareunternehmen darauf ab, sich in eine datengesteuerte Organisation zu transformieren.⁶ Höheres Interesse an dem Besitz und der Verarbeitung großer Datenmengen zeigen heutzutage jedoch Organisationen aller Bereiche wie Wirtschaft, Regierungen und Forschung.⁷ Deren Strategien weisen häufig auf, mittels Daten fokussierter Kultur und Datenanalysen möglichst umfangreiche faktenbasierte Entscheidungen zu treffen.⁸ Forschungsfeld dieser Absicht bildet die Data Science, bzw. Big Data Analytics, mit dessen Einsatz von Technologien sich Unternehmen einen Wettbewerbsvorteil erzielen.⁹ datengesteuerte Unternehmen können dadurch bis zu 26 % profitabler werden gegenüber Wettbewerbern, welche weniger bis gar keine digitalen Technologien einsetzen.¹⁰ Ein Grund dafür können bessere und schnellere Entscheidungen sein, welche aus dem voll umfänglicheren Verständnis des Kunden und höherer Transparenz des Entwicklungsprozesses resultieren.¹¹ Trotz der heute vorhandenen Fülle von erzeugten Daten bleibt die Anzahl an Unternehmen, welche erfolgreich in eine datengesteuerte Organisation transformierten eher gering.¹² Zusätzlich existiert bisher nur wenig Forschung dazu, wie solche Transformationen zu datengesteuerte Unternehmen umzusetzen sind.¹³

¹Vgl. Moore, 1998, S. 1.

²Vgl. Moore, 1998, S. 1.

³Vgl. Dalpiaz et al., 2020, S 3f.

⁴Vgl. Hupperz et al., 2021, S. 1.

⁵Vgl. Dalpiaz et al., 2020, S. 4.

⁶Vgl. Fabijan et al., 2017, S. 1.

⁷Vgl. Pratt et al., 2023, S. 1.

⁸Vgl. Dalpiaz et al., 2020, S. 18.

⁹Vgl. Dalpiaz et al., 2020, S. 3.

¹⁰Vgl. Fabijan et al., 2017, S. 1.

¹¹Vgl. Dalpiaz et al., 2020, S. 18.

¹²Vgl. Fabijan et al., 2017, S. 1.

¹³Vgl. Fabijan et al., 2017, S. 1.

2 Grundlagen

Ziel dieses Kapitels ist die Erläuterung von notwendigen Konzepten zur weiteren Betrachtung des Themas Integration der Data Science in Organisationen und Abteilungen. Dazu wird in diesem Kapitel besonders auf die grundlegende Thematik der Data Science eingegangen.

Wie in der Problemstellung identifiziert, steigt das Interesse an der Aufzeichnung und Verarbeitung von Daten innerhalb von Organisationen. Da hierfür folglich qualifiziertes Fachpersonal benötigt wird, nimmt auch die Rolle des Data Scientist an Wert für Unternehmen zu.¹ Aufgabe dieser Rolle ist es, an Datenanwendungen zu arbeiten, welche eine zeitnahe signifikante Auswirkung auf das Geschäftsmodell verursachen.² Data Science im Allgemeinen wird häufig mit einem Prozess in Verbindung gebracht, welcher durch Einsatz von Techniken des maschinellen Lernens wichtige Erkenntnisse aus Daten ableitet.³ van der Aalst definierte im Jahr 2016 das Feld der Data Science wie folgt:⁴

Data Science ist ein interdisziplinäres Feld mit dem Ziel Wert aus Daten zu generieren. Daten können dabei in strukturierter oder unstrukturierter Form, in großer oder geringer Menge, statisch oder nur in Echtzeit vorliegen. Wert entsteht durch Vorhersagen, automatisierten Entscheidungen, optimierten Modellen oder Visualisierungen, welche Erkenntnisse erzeugen. Data Science inkludiert die Extraktion, Vorverarbeitung, Erkundung, Transformation, Speicherung und den Erhalt von Daten sowie Recheninfrastruktur, Arten des Lernens, das Präsentieren von Erklärungen und die Nutzung von Erkenntnissen unter ethischen, sozialen, rechtlichen und geschäftlichen Aspekten.

Die Definition betrachtet die drei wichtigen Aspekte der Daten, des Wertbeitrags und der Aufgabenbereiche. Folgend werden die einzelnen Aspekte, ausgeschlossen des bereits im vorherigen Kapitel behandelten Wertbeitrags, detaillierter ausgeführt.

Aus der Definition lassen sich bereits einige Charakteristika von Daten hinsichtlich Form, Menge und Persistenz erkennen. In der Literatur werden besonders im Bereich von *Big Data* die 5 V's von Daten thematisiert. Übersetzt beschreiben sie die Charakteristika Volumen, Geschwindigkeit, Wahrheit, Vielfalt und Wert.⁵ Ausprägungen dieser Merkmale lassen sich wie folgt zuordnen: Volumen - Number of Bytes, Geschwindigkeit - static or real time, Richtigkeit - Grad der Vorurteilsbelastung, Vielfalt - strukturiert oder unstrukturiert, Wert - Leistungspotenzial der Daten. Von diesen Charakteristika liegt jedoch global betrachtet der Großteil der Daten in unstrukturierter Form, bspw. als Text, Bild oder Audio, vor.⁶

¹Vgl. Fabijan et al., 2017, S. 1.

²Vgl. Patil, 2011, S. 12.

³Vgl. Zhang et al., 2020, S. 1.

⁴Vgl. van der Aalst, 2016, S. 10.

⁵Vgl. Naeem et al., 2022, S. 1.

⁶Vgl. van der Aalst, 2016, S. 4.

Die Disziplin der Data Science entwickelte sich aus der Statistik und der Kombination vieler verwandter Disziplinen, welche selbst unter sich Überschneidungen aufweisen und in ihrem Einfluss auf die Data Science variieren.⁷ Die verwandten Wissenschaftsbereiche sind dabei in Abbildung 2.1 aufgezeigt.⁸

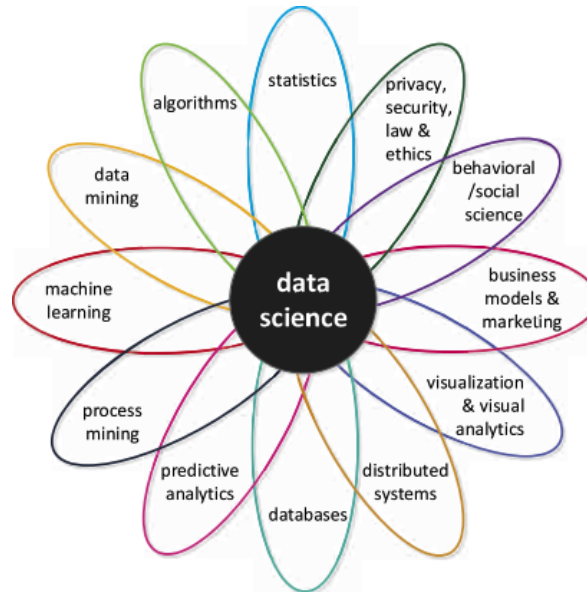


Abbildung 2.1: Disziplinen der Data Science

Anwendung der Data Science und deren Subdisziplinen findet sich in der Bearbeitung von datengetriebenen Problemstellungen der vier Kategorien: Berichterstattungen, Diagnosen, Vorhersagen und Empfehlungen.⁹ Als Arbeitsablauf dieser Bearbeitung wird in der Literatur häufig ein Prozessmodell mit drei Phasen und bis zu zehn Schritten beschrieben, welcher die Phasen der Datenvorbereitung, der Modellentwicklung und dessen Bereitstellung involviert.¹⁰ Dabei ist es sinnvoll, je nach Anwendung und Problemstellung verschiedene Phasen zu fokussieren. Beispielsweise sind zur Berichterstattung die Datenvorbereitung und Bereitstellung der Erkenntnisse zu fokussieren, da häufig kein explizites Modell zu entwickeln ist. Hingegen werden bei der Erstellung von Vorhersagen und Empfehlungen akkurate Modelle benötigt, um die Richtigkeit der Vorhersagen und Empfehlungen zu garantieren. Konkrete Aufgaben in der Data Science Rolle umfassen das Bereinigen und Vereinen von Datensätzen, die Visualisierung von Daten oder die Entwicklung umfangreicher Softwaretools zur Verarbeitung von Daten.¹¹ Das Produkt der Phasen und konkreten Aufgaben kann dabei als analytisches System in einer Organisation betrachtet werden, welches sich aus der Infrastruktur, den Modellen und insgesamt deren Betrieb zusammensetzt.¹² Neben dem bekanntesten Anwendungsbereich der Kundenintelligenz kommen derartige analytische Systeme auch im Bereich von Lieferketten- und Qualitätsmanagement sowie Risikoanalyse und Betrugsdetektion zum Einsatz.¹³

⁷Vgl. van der Aalst, 2016, S. 12.

⁸Vgl. van der Aalst, 2016, S. 12.

⁹Vgl. van der Aalst, 2016, S. 10.

¹⁰Vgl. Zhang et al., 2020, S. 1.

¹¹Vgl. Patil, 2011, S. 13.

¹²Vgl. Grossman und Siegel, 2014, S. 22.

¹³Vgl. Elgandy und Elragal, 2014, S. 221ff.

3 Data Science in datengesteuerten Organisationen

Inhalt dieses Kapitels ist die Darstellung des aktuellen Stands der Forschung zur Eingliederung von Data Science Teams in datengesteuerten Organisationen. Damit wird verfolgt, ein aktuelles Zielbild annäherungsweise zu konkretisieren, um in weiteren Kapiteln auf die Möglichkeiten zur Integration einzugehen.

Eine Herausforderung entsteht bereits dabei, dass datengesteuerte Organisationen, in der Fachsprache als *data driven Organizations* bezeichnet, einer Vielfalt an Definitionen unterliegen.¹ Fabijan et al. definierte z. B., dass datengesteuerte Organisationen Daten akquirieren, verarbeiten und Datenvorteile in einer zeitlich angebrachten Art und Weise nutzen, um Effizienzgewinne zu erzeugen, neuartige Produkte zu entwickeln und sich durch die Wettbewerbslandschaft zu navigieren. Ein gemeinsames Verständnis der Definitionen besteht in dem Prozess des Sammelns von Daten, der Gewinnung von Erkenntnissen durch Analysen und dem Treffen von Entscheidungen basierend auf den erzielten Analyseergebnissen.² Einer der wichtigsten Aspekte einer datengesteuerten Organisation ist die Manifestation einer datengesteuerten Kultur.³ Dessen Antrieb ist es, Daten nicht nur den Analytic-Abteilungen oder dem leitenden Management vorbehalten sind, sondern so weit wie rechtlich möglich, jedem Organisationsmitglied zur Verfügung gestellt werden sollte.⁴ Dieser Umgang würde es ermöglichen, alle Arten der Analyse (deskriptiv, prädiktiv, präskriptiv) auf allen Ebenen der Organisation (operativ, taktisch, strategisch) einzusetzen.⁵ In der bisherigen Praxis wird von diesen Möglichkeiten jedoch nur ein Teil angewendet.⁶ Weitere Aspekte einer datengesteuerten Organisation konnten durch die Forschung von Marius Hupperz et al. identifiziert werden. Für den untersuchten Aspekt der Data Science ist erkannt worden, dass der Wertbeitrag durch Transparenz, zielgerichtetem Marketing oder automatisierten informierten Entscheidungen zu Wettbewerbsvorteilen führen kann, jedoch keinen unmittelbaren Einfluss auf die Vermögenswerte ausübt.⁷ Zusätzlich kann Einrichtung einer Digitalisierungsabteilung zwar die Transformation zum datengesteuerten Unternehmen unterstützen, jedoch durch das alleinige Einrichten von Data Science Abteilungen keine Geschäftserkenntnisse aus Daten zu erzeugen.⁸ Alle weiteren Aspekte aus der strukturierten Literaturanalyse von datengesteuerten Organisationen sind in folgender Abbildung 3.1 dargestellt:⁹

¹Vgl. Dalpiaz et al., 2020, S. 4.

²Vgl. Dalpiaz et al., 2020, S. 4.

³Vgl. Dalpiaz et al., 2020, S. 15.

⁴Vgl. Patil, 2011, S. 6.

⁵Vgl. Dalpiaz et al., 2020, S. 4.

⁶Vgl. Dalpiaz et al., 2020, S. 4.

⁷Vgl. Marius Hupperz et al., 2021, S. 5.

⁸Vgl. Marius Hupperz et al., 2021, S. 5.

⁹Vgl. Marius Hupperz et al., 2021, S. 4.

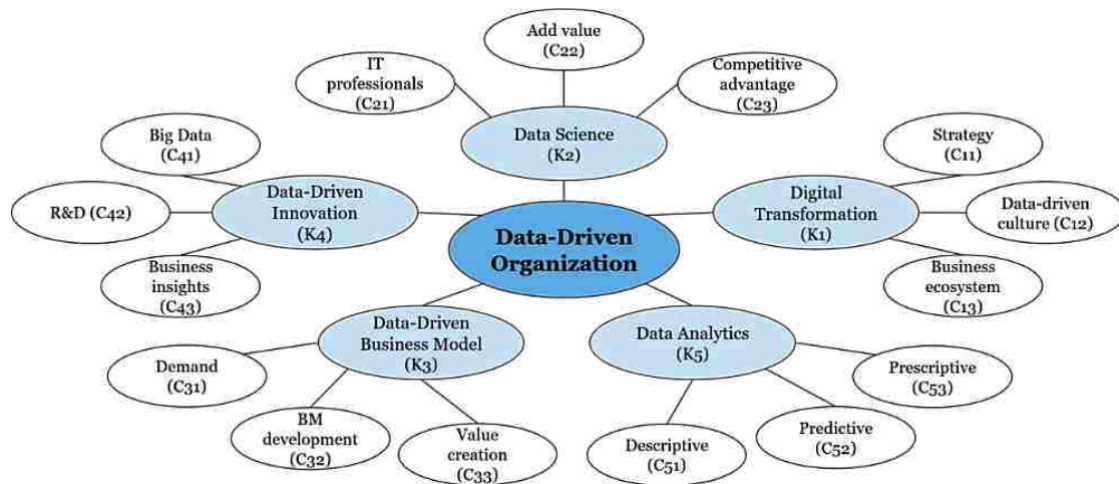


Abbildung 3.1: In der Literatur beschriebene Aspekte von datengesteuerten Organisationen

Zur praktischen Umsetzung von datengesteuerten Unternehmen führten Zhang et al. eine Online Befragung mit insgesamt 183 Teilnehmenden aus der Data Science durch. Da alle Beteiligten der Umfrage aus dem IT-Konzern IBM stammen, ist die Umfrage zwar nicht statistisch repräsentativ für alle Organisationen zeigt jedoch ein signifikantes Bild über den Aufbau und die Zusammenarbeit von Data Science Abteilungen. Mit der Umfrage sind Erkenntnisse erzielt worden, welche die fachliche, personelle und Rollen bezogene Zusammensetzung und Zusammenarbeit von Data Science Abteilungen betreffen. Ein Ergebnis der Umfrage ist es, dass Data science Abteilungen häufig mit einer Teamgröße von bis zu sechs Personen arbeiten, wobei jede Person bis zu 5 Jahre Erfahrung mit Data Science Projekten aufweisen kann.¹⁰ Dabei führen die Teammitglieder meistens mehr als eine Rolle aus, was aus der nachträglichen Abbildung 3.2 (a) hervorgeht:¹¹

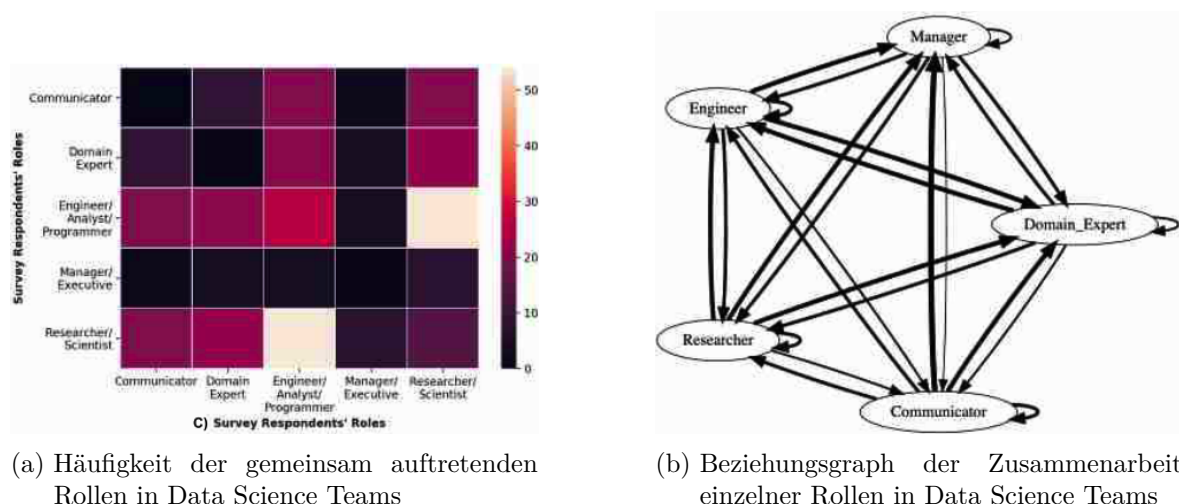


Abbildung 3.2: Rollen und deren Zusammenarbeit in Data Science Teams

Aus der Heatmap wird deutlich, dass die Kombination der Rollen *Researcher - Engineer* am Häufigsten auftritt. Dies resultiert vermutlich daraus, dass sich in der recht jungen Disziplin

¹⁰Vgl. Zhang et al., 2020, S. 7.

¹¹Vgl. Zhang et al., 2020, S. 7.

der Data Science noch wenig Standards etablierten, wodurch viele Konzepte in Projekten erstmalig zu entwickeln sind. Durch die Diagnose der Abbildung, in welcher die Häufigkeit von einzeln vorkommenden Rollenbesetzungen abgetragen ist, wird ersichtlich, dass die *Engineer* Rolle häufiger als alle anderen von einer Person ausgefüllt wird. Daraus lässt sich vermuten, dass der hohe Entwicklungsaufwand in Data Science Projekten die Fokussierung von Personal zur *Engineer* Rolle rechtfertigt. Als Person in der Rolle des Managers werden, mit Ausnahme der *Researcher* Rolle, noch seltener weitere Rollen ausgeübt, jedoch auch selten lediglich Aufgaben der eigenen Rolle übernommen. Begründbar erscheint dies durch die Annahme, dass das Management häufig Aufgaben außerhalb des direkten Geschehens im Data Science Projekt umfasst. Die Ausnahme der *Researcher* Rolle würde sich logisch durch den gemeinsamen hohen Bedarf an Arbeitserfahrung im Themenfeld begründen lassen.

Abbildung 3.2 (b) zeigt den Umfang der Zusammenarbeit zwischen den verschiedenen Rollen auf. Auffälligkeiten in der Grafik umfassen die Rolle des *Communicators* und des *Domain Experts*. Die Rolle des *Communicators* wird dabei vermutlich häufig sehr extrovertiert gestaltet, da hier besonders viel ausgehende Zusammenarbeit zu den anderen Rollen erkennbar ist, im Vergleich zu dessen Abhängigkeit. Etwas gegensätzlich dazu zeigt sich die Rolle des *Domain Experts*, welcher eine starke Abhängigkeit anderer Rollen abbildet, obwohl dabei wesentlich weniger technische Expertise zu erwarten ist.

In datengesteuerten Organisationen gilt es für eine Data Science Abteilung nicht nur innerhalb von sich selbst zusammenzuarbeiten, sondern auch z. B. Komponenten für maschinelles Lernen mit anderen Stakeholdern nutzbar zu gestalten. Dazu sind vielerlei Punkte in der Abstimmung beider Teams notwendig, wie bspw. Anforderungserhebung, Trainingsdaten und Modellintegration, dessen bewährte Praktiken nachfolgend detaillierter betrachtet werden. In der anfänglichen Phase der Anforderungserhebung ist es zum einen wichtig Data Scientists früh in die Definition der Produkthanforderungen einzubinden, jedoch zum anderen die Anforderungen an das zu entwickelnde Modell nicht losgelöst des Produkts zu betrachten.¹² Die Zusammenarbeit kann zusätzlich unterstützt werden, indem zum einen technische Schulungen für Stakeholder wie Kunden, Produktteams und Endnutzer eingerichtet werden und zum anderen das gemeinsame Verständnis der Anforderungen bestmöglich dokumentiert wird.¹³ Während der Abstimmung der Trainingsdaten mit anderen Abteilungen profitiert der Projektfortschritt, wenn Data Science Teams die Freiheit besitzen, Erwartungen bezüglich z. B. Datenqualität an Lieferanten zu stellen.¹⁴ Sollte aufgrund der Projekt- oder Organisationsgröße eine direkte Zusammenarbeit mit dem Datenlieferanten nicht umsetzbar sein, bleibt es notwendig, entsprechende Erwartungen in formalen Verträgen festzuhalten.¹⁵ Nach den bisherigen Abstimmung und der Entwicklung des Produkts bleibt der kontinuierliche Prozess des Betriebs. Die Qualitätssicherung spielt in dieser Phase eine signifikante Rolle und sollte geplant und mit Schulungen zu Themen wie *DevOps* und *MLOps* unterstützt werden.¹⁶

¹²Vgl. Nahar et al., 2022, S. 418.

¹³Vgl. Nahar et al., 2022, S. 418.

¹⁴Vgl. Nahar et al., 2022, S. 420.

¹⁵Vgl. Nahar et al., 2022, S. 420.

¹⁶Vgl. Nahar et al., 2022, S. 423.

4 Herausforderungen

Anschließend an die Beschreibung einer datengesteuerten Organisation, dessen Aufbau, Prozesse und Rollen, werden in diesem Kapitel die Herausforderungen zur Transformation in eine datengesteuerte Organisation thematisiert.

Bereits durch das äußere, sich schnell wechselnde kompetitive Umfeld einer Organisation, verkompliziert sich der Prozess datengesteuert z. B. Strategieentscheidungen in Organisation zu treffen.¹ Zusätzlich zu der Geschwindigkeit der Geschäftsumgebung, konnte durch eine Gartner Umfrage festgestellt werden, dass 65 % der Teilnehmenden im Zeitraum der letzten zwei Jahre (2021-2023) einen Zuwachs der Komplexität von Entscheidungen verzeichneten.² Damit ist das datengestützte Ableiten von Strategieentscheidungen ein besonders herausforderndes Feld aufgrund der hoch diversen Daten und der Vielzahl von Einflussvariablen.³ Diese äußeren Einflüsse wirken sich ebenfalls erschwerend auf interne managementbezogene und kulturelle Herausforderungen des Transformationsprozesses aus.⁴ Bestätigt wird dies durch die Forschungsergebnisse von Dalpiaz et al., welche 15 Interviews aus neun verschiedenen Softwareunternehmen auswerteten. Aus den Interviews konnten die folgenden in Abbildung 4.1 dargelegte Herausforderungen in der Transformation zu einer datengesteuerten Organisation identifiziert werden.⁵

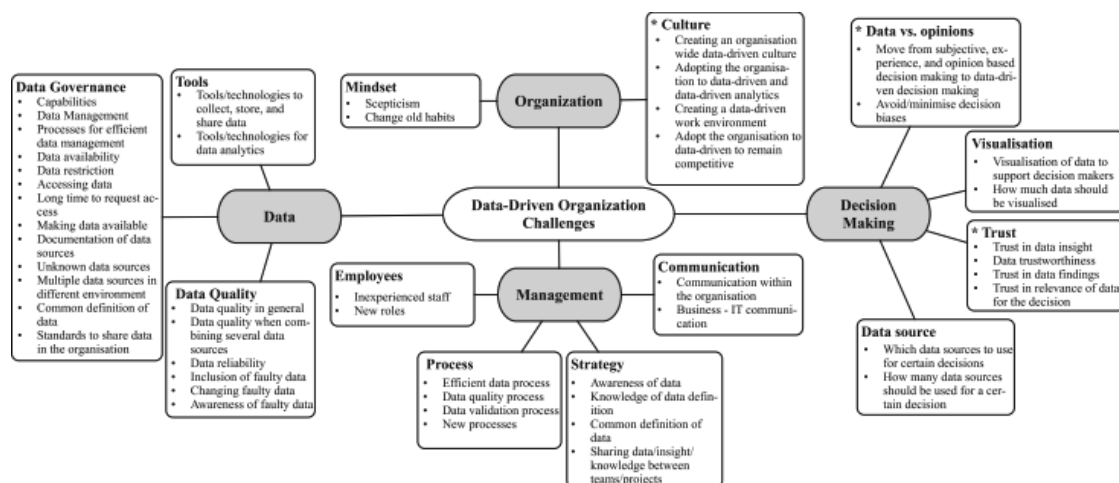


Abbildung 4.1: Herausforderungen einer datengesteuerten Organisation

Innerhalb der Grafik sind mittels * die drei wichtigsten Herausforderungen gekennzeichnet, welche alle 15 Teilnehmenden benannten: faktenbasierte Entscheidungen, Vertrauen und Kultur. Faktenbasierte Entscheidungen beschreiben die Absicht, subjektive, erfahrungsgestützte

¹Vgl. Pratt et al., 2023, S. 2.

²Vgl. Pratt et al., 2023, S. 65.

³Vgl. Pratt et al., 2023, S. 3.

⁴Vgl. Dalpiaz et al., 2020, S. 15.

⁵Dalpiaz et al., 2020, S. 9.

und meinungsbasierte Entscheidungen durch die Nutzung von Daten durchzuführen oder bisherige Prozesse zu unterstützen.⁶ Ein Einfluss auf die Umsetzung nimmt die zweite Herausforderung des Vertrauens. Für faktenbasierte Entscheidungen ist ein hohes Maß an Vertrauen gegenüber vielerlei Aspekte der Daten, entlang des Data Science Prozesses, notwendig. Diese Aspekte umfassen die Aufnahme relevanter Daten, die fehlerfreie Aufnahme der Daten, die fehlerlose Verarbeitung der Daten sowie die korrekte Ableitung von für die Entscheidung relevanten Erkenntnissen.⁷ Die Aufnahme relevanter Daten wird dadurch herausfordernd, dass in etwa 10 % der Daten insgesamt 90 % des strategischen Werts enthalten.⁸ Neben der Geschäftsleitung ist ein Vertrauen gegenüber Daten und dessen Auswertung auch als gelebte Kultur der Organisation notwendig.⁹ Eine gelebte datengesteuerte Kultur kann durch eine geringe Akzeptanz neuer Softwaretools und neuer Arbeitsmethoden verhindert werden.¹⁰ Eine notwendige Tool gestützte Arbeitsweise ist z. B. die Dokumentation der Zusammenarbeit innerhalb des Teams und mit anderen Abteilungen.¹¹ Innerhalb dieser Dokumentation können z. B. Lösungen für technische Herausforderungen von der Datenbeschaffung bis zum Modelleinsatz festgehalten werden.¹² Im Kontext von großen Datenumgebungen müssen technische Herausforderungen gelöst werden, welche zum einen die Aufnahme aller Daten betreffen.¹³ Zum anderen ist die Flexibilität zu gewährleisten, neue Datenquellen einfach anzubinden und Auswertungen schnell durchführen zu können.¹⁴ Durch Hinzunahme von Technologien, wie maschinelles Lernen zur Auswertung von Daten steigen ebenfalls die technischen Herausforderungen. Maschinelles Lernen bedarf Testsysteme zum Einsatz von Modellen, anpassbare Software zur Überwachung der Modelleistung und Möglichkeiten zur Erläuterung der Modellergebnisse.¹⁵

Im gesamten Verlauf der Datenbeschaffung, Modellentwicklung und Ableitung von neuen Optimierungsprozessen sind diverse Teams erforderlich, um vorurteilsbehaftete Datenverarbeitung zu vermeiden.¹⁶ Die Beschaffung von Arbeitskräften für ein solches diverses Team stellt eine weitere Herausforderung dar. Hierfür werden Bewerber mit datenbezogener Vergangenheit aus unterschiedlichen Branchen, oder Universitätsabsolventen und ein ausgearbeitetes Einarbeitungsprogramm benötigt.¹⁷

Diese Herausforderung allein, sowie in Kombination mit den anderen thematisierten Herausforderungen bedarf eines Rahmenprogramms zur Integration von Data Science in verschiedenste Institutionen.¹⁸

⁶Vgl. Dalpiaz et al., 2020, S. 9.

⁷Vgl. Dalpiaz et al., 2020, S. 10.

⁸Vgl. Pratt et al., 2023, S. 3.

⁹Vgl. Dalpiaz et al., 2020, S. 4.

¹⁰Vgl. Dalpiaz et al., 2020, S. 15.

¹¹Vgl. Zhang et al., 2020, S. 12.

¹²Vgl. Grossman und Siegel, 2014, S. 23.

¹³Vgl. Elgendy und Elragal, 2014, S. 217.

¹⁴Vgl. Elgendy und Elragal, 2014, S. 217.

¹⁵Vgl. Nahar et al., 2022, S. 1.

¹⁶Vgl. Zhang et al., 2020, S. 18.

¹⁷Vgl. Patil, 2011, S. 13.

¹⁸Vgl. Saltz und Grady, 2017, S. 1.

5 Integrationsprozesse

Nach den behandelten Herausforderungen werden in diesem Kapitel Möglichkeiten und Vorgehensmodelle zur Integration von Data Science in Organisationen thematisiert. Dazu werden insgesamt drei Modelle der Forschungsliteratur dargelegt.

5.1 CSPG Framework

Grossman et al. veröffentlichte 2014 das CSPG Framework zur Integration von Analytik, Domänenwissen und IT in Organisationen. *CSPG* steht repräsentativ als Abkürzung für die Komponenten *Culture*, *Staffing*, *Processes* und *Governance*. Der Aspekt der Kultur ist durch die Organisationsleitung umzusetzen, indem Verantwortung und Autorität für Datenbestände an eine Funktionsstelle übergeben wird. Die Aufnahme von Data Science Personal ist im CSPG Framework unausweichlich und durch den Analytik Leiter und die Geschäftsleitung durchzuführen. Dabei ist zu entscheiden, ob die analytische Funktion zentral, dezentral oder hybrid organisiert wird. Im dritten Aspekt des Frameworks sind die analytischen Prozesse in der Organisation aufzubauen. Ein Teil dieser Prozesse umfasst den Austausch von Daten zwischen Abteilungen und anderen Organisationen. Weitere Prozesse behandeln die Digitalisierung bestehender Inhalte, Produktanpassung zur Aufnahme von Daten und die Kombination von Datenbeständen. Die finale Komponente des Frameworks ist der Aufbau, die Verwaltung und die Weiterentwicklung der notwendigen Infrastruktur. Zur Bewältigung dieser Aufgabe sind die folgenden vier Bedingungen in der Organisation zu erfüllen:

- Langfristige Verpflichtung für Data Science und Sicherstellung des daraus entstehenden Geschäftswerts.
- Sichere und rechtlich unbedenkliche Umsetzung der Data Science Prozesse.
- Herstellen von Haftbarkeit, Transparenz und Rückverfolgbarkeit für Projektfinanzierung, Entwicklung und Ressourcen.
- Ressourcenbereitstellung für Daten-, Analyse- und Managementprozesse.

5.2 Design Parameters

Durch die Arbeit von Janine Adina Hagen et al. konnten die Gestaltungsparameter einer datengesteuerten Organisation ermittelt werden. Diese Parameter können als Richtlinien betrachtet werden, wie die eigene Organisation in verschiedenen Aspekten zu gestalten ist.

Subsystem/ component		Design dimension	Characteristic								
Social	Structure	Anchoring of data experts	Central			Hybrid			Decentral		
		Reporting line	Technology			Dual			Business		
		Horizontal linkage	Simplified examples	Meeting routines	Joint processes	Voluntary networks	Training	Integrator	Events		
		Collaboration initiative	Business team		Data and business team		Data team (business need-based)		Data team (data-based)		
		Collaboration mode	Prototyping			Structured backlog			Occasional use cases		
		Control mechanisms	Business impact	Back testing		Utilization		Data vs. human tournament		Transformation KPIs	
	Actors	Roles	Data-oriented			Hybrid			Business-oriented		
Technical	Tasks	Data tasks	Descriptive analysis	Predictive analysis	Prescriptive analysis	Tools	Infra-structure	Visualization	Dashboards		
		Business tasks	Process improvement		Decision-making	New insights		Products / services		Standardization	
	Technology	Data repository	United			Hybrid			Dispersed		

Abbildung 5.1: Gestaltungsparameter einer datengesteuerten Organisation

Die vorherige Tabelle zeigt die Gestaltungsparameter organisiert nach Komponenten, Dimensionen und konkreter Charakteristika.¹ Eine wichtige Komponente betrachtet die soziale Perspektive auf die Struktur und die Akteure der datengesteuerten Organisation. Die Struktur der Organisation kann durch die Parameter *Expertenorganisation*, *Berichtslinie*, *horizontale Verknüpfung*, *Kooperationsinitiative*, *Kooperationsmodus* und *Kontrollmechanismus* beeinflusst werden. Eine mögliche Ausprägung dieser Parameter umfasst z. B. zum einen dezentrale Data Science Experten, eine Berichtslinie zur Fachabteilung sowie regelmäßige Meetings und Events der Data Science Experten. Zum anderen werden z. B. durch die Datenexperten die Zusammenarbeiten mittels strukturiertem Backlog initiiert und anhand des Geschäftsmehrwerts evaluiert. In dieser Struktur könnten dann die Rollen z. B. hybrid, also datenorientiert sowie geschäftsorientiert gestaltet werden. Wird die technische Perspektive betrachtet, werden dessen Komponenten der Aufgaben und Technologien durch die Parameter *Datenaufgaben*, *Geschäftsaufgaben* und *Datenrepository* bestimmt. Beispielhaft könnte eine Organisation durch ihre Struktur und Akteure beschreibende, prädiktive und vorhersagende Analysen erstellen, um Prozesse zu verbessern und Entscheidungen zu unterstützen. Speicherorte für Daten und Software können z. B. hybrid für jedes Datenteam einer Fachabteilung und für den gemeinsamen Austausch mehrerer Datenteams eingerichtet werden.

5.3 Experiment Evolution Model

Ein weiteres Vorgehensmodell durch Fabijan et al. beschreibt den Transformationsprozess von Ad hoc Analysen zu skalierten kontrollierten Experimenten in Organisationen.² Das Experiment Evolution Modell verdeutlicht die Evolutionsphasen datengesteuerter Entwicklung in Unternehmen und Abteilungen. Zur Anwendung des Modells sind Voraussetzungen zu erfüllen, welche folgend thematisiert werden.

Eine Anwendung setzt insoweit Fähigkeiten der Data Science voraus, sodass Produktstatistiken evaluiert werden können. Die Fähigkeiten umfassen das Verständnis von Hypothesentests, Randomisierung, Stichprobengrößen und die Berechnung von Konfidenzintervallen. Die Kombination dieser Fähigkeiten mit Domänenwissen ermöglicht die Generierung und Evaluation erster Produkthypothesen. Als zweite Anforderung wird die Verfügbarkeit eines Zugangs zu

¹Vgl. Janine Adina Hagen und Thomas Hess, 2020, S. 5.

²Vgl. Fabijan et al., 2017, S. 5.

Daten der Produktverwendung vorausgesetzt.

Die im Modell abgebildete Evolution erfolgt in den drei Dimensionen Technisch, Organisatorisch und Geschäftlich. Jede Dimension folgt den vier Phasen *Krabbeln*, *Gehen*, *Rennen* und *Fliegen*. Zu Beginn der ersten Phase finden in der technischen Dimension der Aufbau von Logging-Systemen und erste manuelle produktspezifische Analysen von Data Science Teams statt. Die Geschäftsdimension beinhaltet die Definition von unternehmensweiten Evaluationskriterien in Verbindung mit den langfristigen Geschäftszielen. Während der Phase des *Gehens* werden in der technischen Dimension Metriken entwickelt, anhand dessen Eventdaten aggregiert betrachtet werden können. Diese Phase erzeugt den Bedarf einer übergreifenden Analyseplattform durch das Wachstum an Experimentarten und Anwendungen. In der Organisation bauen während dieser Phase Data Science Experten zunehmend Wissen zu einzelnen Produkten auf. Ebenfalls werden Daten und Vorgehensweisen weitreichend geteilt. Die Geschäftsdimension entwickelt sich durch Validierung bestehender und Entwicklung neuer Evaluationskriterien z. B. hinsichtlich der Datenqualität. Zur Phase des *Rennens* werden in der technischen Dimension Analyseplattformen skaliert und Experimente für mehrere Produktteams ausgeführt, um einen Fokus auf nahverwandte Metriken der Geschäftsziele zu setzen. Organisatorisch übernehmen Produktteams volle Verantwortung und für ihre Ergebnisse, werden aber noch fachlich von Data Science Experten begleitet. In der Geschäftsdimension werden bisherige Qualitäts- und Leistungskriterien verfeinert und kombiniert, um Ergebnisse zu Standardisieren und wenige klare Ziele zu fokussieren. Die letzte Phase des *Fliegens* fokussiert in der technischen Dimension die Standardisierung der Metrikentwicklung, Erweiterung der Plattformautomatisierung und Aufnahme von Daten zu jeder Produktänderung. Die Evolution der Organisation umfasst zu diesem Zeitpunkt die autonome Durchführung von Analysen durch Produktteams und Entwicklung von Plattformen durch zentrale Data Science Teams. In der Geschäftsdimension sind zu dieser Phase lediglich periodisch Änderungen an definierten Evaluationskriterien und deren Entwicklungsprozessen vorzunehmen.

Das Modell lässt sich zusammenfassend wie folgt in Abbildung 5.2 darstellen.³




	Category/ Phase	Crawl 	Walk 	Run 	Fly 
Technical Evolution	Technical focus of product dev. Activities 	(1) Logging of signals (2) Work on data quality issues (3) Manual analysis of experiments Transitioning from the debugging logs to a format that can be used for data-driven development.	(1) Setting-up a reliable pipeline (2) Creation of simple metrics Combining signals with analysis units. Four types of metrics are created: debug metrics (largest group), success metrics, guardrail metrics and data quality metrics.	(1) Learning experiments (2) Comprehensive metrics Creation of comprehensive set of metrics using the knowledge from the learning experiments.	(1) Standardized process for metric design and evaluation, and OEC improvement
	Experimentation platform complexity 	No experimentation platform An initial experiment can be coded manually (ad-hoc).	Platform is required 3 rd party platform can be used or internally developed. The following two features are required: • Power Analysis • Pre-Experiment A/A testing	New platform features The experimentation platform should be extended with the following features: • Alerting • Control of carry-over effect • Experiment iteration support	Advanced platform features The following features are needed: • Interaction control and detection • Near real-time detection and automatic shutdown of harmful experiments • Institutional memory
	Experimentation pervasiveness 	Generating management support Experimenting with e.g. design options for which it's not a priori clear which one is better. To generate management support to move to the next stage.	Experiment on individual feature level Broadening the types of experiments run on a limited set of features (design to performance, from performance to infrastructure experiments)	Expanding to (1) more features and (2) other products Experiment on most new features and most products.	Experiment with every minor change to portfolio Experiment with any change on all products in the portfolio. Even to e.g. small bug fixes on feature level.
Organisational Evolution	Engineering team self-sufficiency 	Limited understanding External Data Scientist knowledge is needed in order to set-up, execute and analyse a controlled experiment.	Creation and set-up of experiments Creating the experiment (instrumentation, A/A testing, assigning traffic) is managed by the local Experiment Owners. Data scientists responsible for the platform supervise Experiment Owners and correct errors.	Creation and execution of experiments Includes monitoring for bad experiments, making ramp-up and shut-down decisions, designing and deploying experiment-specific metrics.	Creation, execution and analyses of experiments Scorecards showing the experiment results are intuitive for interpretation and conclusion making.
	Experimentation team organization 	Standalone Fully centralized data science team. In product teams, however, no or very little data science skills. The standalone team needs to train the local product teams on experimentation. We introduce the role of Experiment Owner (EO).	Embedded Data science team that implemented the platform supports different product teams and their Experiment Owners. Product teams do not have their own data scientists that would analyse experiments independently.	Partnership Product teams hire their own data scientists that create a strong unity with business. Learning between the teams is limited to their communication.	Partnership Small data science teams in each of the product teams. Learnings from experiments are shared automatically across organization via the institutional memory features.
Business Evolution	Overall Evaluation Criteria (OEC) 	OEC is defined for the first set of experiments with a few key signals that will help ground expectations and evaluation of the experiment results.	OEC evolves from a few key signals to a structured set of metrics consisting of Success, Guardrail and Data Quality metrics. Debug metrics are not a part of OEC.	OEC is tailored with the findings from the learning experiments. Single metric as a weighted combination of others is desired.	OEC is stable, only periodic changes allowed (e.g. 1 per year). It is also used for setting the performance goals for teams within the organization.

Abbildung 5.2: Gestaltungsparameter einer datengesteuerten Organisation

³Vgl. Fabijan et al., 2017, S. 6.

6 Fazit

Ziel dieser Arbeit war die Beleuchtung der aktuellen Forschungsliteratur zur Integration von Data Science in Organisationen und Abteilungen. Dazu wurde zunächst ein Grundlagenverständnis der Data Science geschaffen und das Zielbild einer datengesteuerten Organisation thematisiert. Anschließend wurde im Detail auf die Herausforderungen der Integration eingegangen. Folgend wurden aus der Forschungsliteratur verschiedene Modelle zur Transformation in eine datengesteuerte Organisation und dessen Gestaltung dargelegt.

Durch die Hausarbeit konnte verdeutlicht werden, welche Auswirkung eine Integration der Data Science im Unternehmen erzeugen kann, jedoch auch, wie herausfordernd und aufwändig eine Transformation zur datengesteuerten Organisation ist. Als Disziplin, um aus Daten Mehrwerte zu generieren kann die Data Science in datengesteuerten Organisationen Prozesse verbessern, transparentere Entscheidungsgrundlagen schaffen und Innovationen fördern. Externe Faktoren, wie das sich schnell verändernde Umfeld und interne Herausforderungen, wie die Datenkultur, das Vertrauen in Daten und fehlende IT-Infrastruktur erschweren jedoch die Integration der Data Science. In der Forschung wurden daher verschiedene Modelle entwickelt, um die Integration phasenweise zu begleiten, oder das Zielbild in leichter umzusetzende Teilaspekte aufzugliedern.

Zusammenfassend lässt sich erkennen, dass die Integration der Data Science in Organisation eine wichtige Aufgabe für alle Teile der Gesellschaft wie Wirtschaft, Politik und Forschung geworden ist. Daher sollten alle Gesellschaftsbereiche mit Best Practices, Bildungsmaßnahmen und ausgiebiger Forschung zur Bewältigung dieser Aufgabe beitragen.

Literatur

- Dalpiaz, F., Zdravkovic, J., & Loucopoulos, P. (2020). *Research challenges in information science: 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23-25, 2020, Proceedings* (Bd. 385). Springer.
- Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. *Industrial Conference on Data Mining, 8557*, 214–227. https://doi.org/10.1007/978-3-319-08976-8_16
- Fabijan, A., Dmitriev, P., Olsson, H. H., & Bosch, J. (2017). The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale. *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 770–780. <https://doi.org/10.1109/ICSE.2017.76>
- Grossman, R., & Siegel, K. (2014). *Organizational Models for Big Data and Analytics*.
- Hupperz, M., Gür, I., Möller, F., & Otto, B. (2021). What is a Data-Driven Organization?
- Janine Adina Hagen & Thomas Hess. (2020). Linking Big Data and Business: Design Parameters of Data-Driven Organizations. https://www.researchgate.net/profile/janine-hagen/publication/343549422_linking_big_data_and_business_design_parameters_of_data-driven_organizations
- Marius Hupperz, Inan Gür, Frederik Möller & Boris Otto. (2021). What is a Data-Driven Organization? https://www.researchgate.net/profile/marius-hupperz/publication/351282206_what_is_a_data-driven_organization
- Moore, G. E. (1998). Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE, 86*(1), 82–85. <https://doi.org/10.1109/jproc.1998.658762>
- Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., Tariq, M. I., De-la-Hoz-Franco, E., & De-La-Hoz-Valdiris, E. (2022). Trends and Future Perspective Challenges in Big Data. In J.-S. Pan, V. E. Balas & C.-M. Chen (Hrsg.), *Advances in Intelligent Data Analysis and Applications* (S. 309–325). Springer Singapore.
- Nahar, N., Zhou, S., Lewis, G., & Kästner, C. (2022). Collaboration challenges in building ML-enabled systems. <https://doi.org/10.1145/3510003.3510209>
- Patil, D. J. (2011). *Building Data Science Teams*. O'Reilly Media, Inc.
- Pratt, L., Bisson, C., & Warin, T. (2023). Bringing advanced technology to strategic decision-making: The Decision Intelligence/Data Science (DI/DS) Integration framework. *Futures, 152*, 103217. <https://doi.org/10.1016/j.futures.2023.103217>

- Saltz, J. S., & Grady, N. W. (2017). The ambiguity of data science team roles and the need for a data science workforce framework. *2017 IEEE International Conference on Big Data (Big Data)*. <https://doi.org/10.1109/bigdata.2017.8258190>
- van der Aalst, W. (2016). Data Science in Action. In *Process Mining* (S. 3–23). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-49851-4_1
- Zhang, A. X., Muller, M., & Wang, D. (2020). How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1–23. <https://doi.org/10.1145/3392826>

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Köln, den 4. November 2023

Leon Henne