Vorwort

Der Hype um Big Data und Data Analytics fällt in eine Zeit, in der es über das Internet einfacher denn je möglich ist, sich mit Themen zu beschäftigen, Wissen aufzubauen und zu vertiefen. Onlinekurse werden zu allen Facetten beider Themen zahlreich angeboten. Über Blogeinträge, Diskussionsforen oder Artikel können neueste Algorithmen und Technologien anhand von Beispielszenarien und -daten schnell erlernt und direkt angewandt werden. Programmiercode und ganze Bibliotheken werden aufwendig dokumentiert und frei über die Software Git bereitgestellt. Die Mühe und Zeit, die in die Erstellung der vielen frei zugänglichen Ressourcen und Dokumente geflossen sein muss, beeindruckt mich immer wieder auf ein Neues. Auch kommerzielle Plattformen wie Kaggle tragen über Wettbewerbe zu Analytics-Fragestellungen dazu bei, Data Scientists aus der ganzen Welt zusammenzuführen und den Austausch und die Entwicklung von Modellen und neuen Methoden zu fördern. Datenplattformen bieten Tausende realer Datensätze. Sogar der öffentliche Sektor stellt mehr und mehr Daten über eigens entwickelte Plattformen bereit. Es stehen also alle Mittel zur Verfügung, Data Analytics erfolgreich einzusetzen.

Es sind aber gerade die zahllosen Ressourcen, die es schwer machen, nach dem Einstieg in das Thema den Überblick zu behalten und zu entscheiden, welche Strategien und Vorgehensweisen sinnvoll sind. Informationen sind über viele Seiten verteilt und dadurch stark fragmentiert. Dazu kommt, dass Grundlagenkurse technisches Basiswissen sehr gut aufbereitet anbieten, häufig jedoch die Probleme vernachlässigen, die sich bei der Anwendung von Algorithmen in realen Use Cases ergeben. Durch wohl vorselektierte Daten scheinen Algorithmen fast von selbst perfekte Ergebnisse zu liefern. Vorverarbeitungen werden aufgrund des Fokus auf die Algorithmen häufig ganz außen vor gelassen. Es entsteht das Bild, dass Datenanalyse eigentlich nicht viel mehr als einen potenten Computer und «gute» Software verlangt. Der Übergang von Kursmaterialien zu realen Daten und Anwendungsszenarien vor allem in traditionellen Nicht-IT-Unternehmen, in denen Data Analytics erst eingeführt wird, ähnelt dann einem kalten Entzug.

Auf der anderen Seite entstehen neben vielen technischen Artikeln viele Veröffentlichungen dazu, wie Datenanalyseinitiativen, AI und Big Data in Unternehmen umgesetzt werden sollen. Dort werden teilweise recht reißerisch Dos and Don'ts beschrieben und Ratschläge erteilt, wie sich Misserfolge vermeiden lassen und wo schnelle Gewinne greifbar nahe sind. Die Empfehlungen bleiben dabei häufig oberflächlich. Aussagen wie «Fail fast, fail often» und «Startup-like» werden inflationär verwendet, ohne die Botschaften dieser Aussagen zu reflektieren. Personen, die (Analytics-) Projekte tatsächlich zum Erfolg führen wollen oder müssen, werden häufig nicht beachtet. Es scheint so, als käme es nur auf die richtigen Strategievorgaben an. Dann würden sich alle anderen Probleme von alleine lösen. Die Überraschung ist groß, wenn das nicht so einfach funktioniert und «Quick Wins» ausbleiben. Dann wird versucht, Probleme «organisatorisch» zu «pushen» anstatt sich die Zeit zu nehmen, die eigentlichen Hindernisse in den Fokus zu rücken und aus dem Weg zu räumen.

Die Erfahrung zeigt, dass gute Daten, nützliche Softwaretools und wegweisende Strategien zwar hilfreich sind, jedoch nicht ausreichen, um Analytics-Projekte langfristig erfolgreich umzusetzen. Das Buch bündelt die Grundsätze und Ansichten, die ich über die letzten Jahre in Analytics-Projekten in verschiedensten Situationen und Rollen gewonnen habe. Es soll dabei nicht nur Teilaspekte aus dem Themenfeld Analytics wie Methoden oder Tools vertiefen, sondern alle organisatorischen wie technischen Aspekte im Überblick zusammenführen. Dazu zählen Überlegungen zum Aufbau von Abteilungen und der Auswahl von Projekten genauso wie ein generelles Vorgehen, Algorithmen und Tools. Ich habe mit dem Buch nicht den Anspruch, Leser zu vollwertigen Data Scientists ausbilden zu wollen. Vielmehr ist das Ziel, Projektleitern, Teamleitern,

Experten, Fachabteilungen und anderen Stakeholdern genauso wie Data Scientists selbst ein gemeinsames Verständnis für Data Analytics in Unternehmen zu bieten sowie einige Handlungstipps an die Hand zu geben. Meine Erfahrung zeigt, dass eine gemeinsame Sprache und gegenseitiges Verständnis die eigentlichen zentralen Faktoren für erfolgreiche Analytics-Projekte und -Initiativen in Unternehmen sind. Dann werden Projekte realistisch angegangen und mit den richtigen Strategien, Daten, Tools und vor allem auch Experten erfolgreich umgesetzt.

Vielleicht untypisch für ein Buch enthält das vorliegende im Text und in den Referenzen eine Vielzahl von Links zu weiterführenden Ressourcen im Internet anstatt Standardwerke zu zitieren. Wie eingangs beschrieben, wird Wissen heute stark über das Internet verbreitet. Dem soll sich in diesem Buch nicht entgegengestellt werden. Im Gegenteil soll es ein Startpunkt sein und die Links sollen dazu beitragen, sich in einzelne Themen bei Interesse eigenständig detaillierter einzuarbeiten. Auch deshalb habe ich beschlossen, die Links zusammenzustellen und online verfügbar zu machen:

INTERNET

Alle im Buch enthaltenen Links zu weiterführenden Ressourcen werden zusammengefasst auch über die Adresse http://manufacturinganalytics.de bereitgestellt.

An dieser Stelle möchte ich es nicht versäumen, mich bei den Menschen zu bedanken, die den Entstehungsprozess entscheidend unterstützt haben. Der größte Dank gilt dabei meiner Frau Pavlina, die mir stets den Rücken frei hielt und moralische Unterstützung leistete, wenn ich mich das ein oder andere Mal fragte, wieso ich nach der Dissertation überhaupt noch mehr als zehn Seiten am Stück freiwillig schreibe. Meiner Tochter Ada, der ich dieses Buch widme, möchte ich für ihre Geduld danken. Darüber hinaus gilt mein Dank meiner Kollegin Julia Seitz, die sich aufopfernd durch den ersten Entwurf gekämpft und durch ihre Anmerkungen die Qualität sicherlich deutlich verbessert hat, und Alexander Piazza, der Algorithmen und Use Cases noch einmal Formel für Formel überprüft hat.

Nürnberg Dr. Johannes Kröckel

Inhaltsverzeichnis

1 Einführung

Wenn man Presse und Internet glaubt, befinden wir uns im goldenen Zeitalter von Analytics, Machine Learning und Big Data oder mittlerweile sogar KI (künstlicher Intelligenz). Alles scheint mit vielen Daten und Algorithmen möglich oder noch besser und schneller abbildbar, als es aktuell ohne algorithmische Hilfe der Fall ist. Jedes Jahr werden neue, bessere Algorithmen entwickelt, neue Werkzeuge zur Speicherung, Analyse und Visualisierung von noch größeren Datenmengen vorgestellt. Datenanalysten oder Data Scientists gelten als Mitglieder einer Profession, die auch als «Sexiest Job of the 21th Century» bezeichnet wird [1]. Forscher und Stars der Branche, die früher als Nerds galten, werden nun fast vergöttert.

In vielen, vor allem Online-nahen, Branchen werden Analytics-Teams seit Jahren erfolgreich eingesetzt. Sie helfen Onlineshops wie Amazon, Algorithmen zu entwickeln, die nach dem Eintippen der Internetadresse oder nach dem Klick auf Links im Hintergrund laufen. Die Modelle nutzen alle Daten, die über Kunden verfügbar sind, so dass diese bestmöglich beworben und zum Kauf angeregt werden. Internetdienste wie GoogleMail durchpflügen E-Mails ganz automatisch und ordnen diese bestimmten Kategorien zu oder erkennen Spam-Mails. Werbeflächen auf Webseiten werden höchstbietend versteigert – je nachdem, welche Informationen über den Aufrufer bekannt sind.

Auch traditionelle Branchen sehen nach einigem Zögern großes Potenzial in der Nutzung von bestehenden bzw. neu generierten Daten. Das Thema Datenanalyse in Industrieunternehmen erlebt geradezu einen Hype. Große Industrieunternehmen haben vor allem an Standorten wie München oder Berlin große Big-Data/Analytics-Einheiten aus der Taufe gehoben. Der Wunsch nach Nutzung der Ressource «Analytics» als neues Mittel zur Optimierung von bestehenden Prozessen und dadurch einhergehenden Kosteneinsparungen bzw. Ertragssteigerungen oder sogar zur Erschaffung neuer Geschäftsmodelle lockt Unternehmensleitungen in allen Branchen zu investieren. Projektpläne werden erstellt, Ziele definiert und Abteilungen aus dem Boden gestampft. Das führt mittlerweile dazu, dass Datenanalysten oder auch Data Engineers auf dem Arbeitsmarkt knapp werden und sich kaum mehr erfahrene Leute (günstig) finden lassen. HR und Fachabteilungen arbeiten sich durch zahllose Bewerbungen und Bewerbungsgespräche, um Teams zusammenzustellen.

Sind die Teams zusammengestellt, macht sich häufig früher oder später Ernüchterung breit. Ergebnisse können langsamer als erwartet oder überhaupt nicht realisiert werden, Einsparungen und Erträge durch Datenanalyse lassen sich schwer quantifizieren und noch weniger vorhersagen. Das traditionelle Vorgehen über Projekte mit strikt vordefinierten Zielen und Ergebnissen passt nicht ganz zu dem typischer Datenanalyseprojekte. Dann schwindet das Verständnis, «was Datenanalysten eigentlich tun», zugunsten der Frage, «was Datenanalysten eigentlich kosten». Dass die Schuld nicht unbedingt bei den Datenanalysten liegt, sondern häufig an der Reife des Unternehmens, digitale Projekte umzusetzen, wird dabei übersehen.

Gartner's Hypecyle of Emerging Technologies beschreibt die Stimmung rund um das Thema Datenanalyse gut: Er zeichnet die Themen im Jahr 2017 auf dem Höhepunkt des Hypes, also übersteigerter Erwartungen, und bevor die Ernüchterung und ein reifer Einsatz einsetzen [2]. Erste Fälle, bei denen diese Ernüchterung eingetreten ist, gibt es bereits. Trotz dieser entzaubernden, vielleicht etwas pessimistischen Bilanz lassen sich mithilfe der Analyse von Unternehmensdaten sowie Daten aus Maschinen, Sensoren und anderen vernetzten Geräten tatsächlich große Potenziale für Optimierung, Automatisierung und neue Geschäftsmodelle realisieren – allerdings nur dann, wenn abseits des Hypes realistische Ziele gesteckt und die Grundvoraussetzungen im Unternehmen geschaffen werden.

Hier soll dieses Buch ansetzen. Auf Basis der Erfahrungen des Autors in Unternehmen der Automobil- und Luftfahrtbranche sowie aus vielen Diskussionen nach Vorträgen auf Konferenzen und Messen sollen Denkanstöße und Handlungsanweisungen vermittelt werden, wie Datenanalysen erfolgreich(er) in produzierenden oder produktionsnahen Unternehmen umgesetzt werden können und welche Voraussetzungen dafür entscheidend sind. Das Buch adressiert dabei zwei Lesergruppen im Besonderen. Eine davon sind Entscheider, die kurz vor oder im Aufbau von Datenanalyseteams stehen oder diese im Unternehmen etablieren möchten. Ihnen soll das Buch ein Verständnis vermitteln, welche Gegebenheiten im Unternehmen und ihren Bereichen vorherrschen sollten, damit Datenanalyseteams erfolgreich arbeiten können bzw. was getan werden kann, um diese Gegebenheiten zu schaffen. Darüber hinaus soll anhand konkreter Beispielprojekte aufgezeigt werden, welche Potenziale Datenanalysen im Unternehmen bieten. Die zweite Gruppe sind Mitarbeiter dieser Datenanalyseteams, die vor der Aufgabe stehen, aus den Daten, die im Unternehmen vorhanden sind, zielgerichtet wirtschaftliche Potenziale zu heben. Ihnen soll das Buch Methoden, Ideen und Projektvorschläge an die Hand geben, um erste Projekte gezielt auszuwählen und zum Erfolg zu führen.

Der folgende Text ist in 6 Kapitel unterteilt. Kapitel 2 beschäftigt sich mit dem Aufbau und der Entwicklung von Datenanalyseteams. Hier sollen Fragen wie «Wie stelle ich ein erstes Team zusammen und wen benötige ich dafür?» geklärt werden. Dabei werden zwei Strategien sowie deren Vor- und Nachteile sowie Risiken diskutiert. Kapitel 3 beschreibt im Überblick, wie das Team von der Strukturierung und Einplanung neuer Analytics-Projekte bis zu deren Operationalisierung vorgehen sollte. Im Zentrum des Kapitels steht der bekannte CRISP-DM-Prozess, der um Praxiserfahrungen angepasst und erweitert wurde. Kapitel 4 bietet einen Überblick über die verschiedenen Charakteristika von Daten. Kapitel 5 stellt gängige Tools, Technologien und Methoden der Datenanalyse dar, um Analytics-Fragestellungen zu lösen. Es bildet damit einen Einstiegspunkt für Methoden der Datenanalyse, mit denen Analytics-Herausforderungen aus Produktion und Logistik angegangen werden können. Dieses Know-how wird in Kapitel 6 zur Anwendung gebracht. Hier werden typische Analytics-Projekte unter die Lupe genommen, Methoden vorgestellt und Handlungswege empfohlen.

In diesem Buch werden eine Vielzahl unterschiedlicher Themen knapp abgehandelt: vom Aufbau eines Teams über das Vorgehen in Analytics-Projekten bis hin zu Methoden und Algorithmen, mit denen sich Modelle erstellen lassen. Gerade den letzteren beiden Themen wurden bereits zahlreiche Fachbücher gewidmet. Im Gegensatz zu den Manuskripten, die Methoden und Vorgehensweisen technisch tief durchdringen, liegt der Fokus in diesem Buch darauf, einen Überblick über die vielen Facetten von Analytics zu bieten, diese miteinander zu verbinden und Handlungsempfehlungen zu geben, wie Analytics-Projekte begonnen werden können.

Das Buch soll die Angst vor Analytics-Projekten nehmen, gleichzeitig jedoch auch darauf hinweisen, dass Analytics-Projekte nicht eben schnell durchgezogen werden können. Die Projekte erfordern neben Daten und gut ausgewählten Experten Geduld, Struktur und ein agiles Mindset. Das lässt sich nur bedingt aus Onlinekursen und Büchern, sondern über Erfahrung lernen. Dennoch soll das Buch Unterstützung dabei bieten, nicht alle Erfahrungen selbst machen zu müssen.

Das Buch versteht sich daher auch als Aufschlag, sich mit dem Thema Datenanalyse in produzierenden und Logistikunternehmen zu beschäftigen, und bietet die Möglichkeit, einen gesamten Überblick über das Thema zu erhalten. Über die zahlreichen Links zu Internetquellen und Büchern soll eine Art erklärte Linkliste bereitgestellt werden, über die einzelne Themen tiefer betrachtet werden können. Am Ende gilt es festzustellen: Datenanalysten oder Data Scientists sind ganz normale Menschen, deren Passion die Betrachtung und Analyse von Daten und der Aufbau von Modellen ist. Sie im Unternehmen gut zu unterstützen bedeutet, gemeinsame Erfolge zu erzielen und das Unternehmen weiter zu bringen.

INTERNET

Das Buch enthält zahlreiche Links zu Webseiten mit Kursen, weiterführende Informationen zum Vertiefen von Themen oder Programmcode und Screenshots. Adressen ändern sich, Tools werden aktualisiert und es gibt schönere Aktivitäten als Webseitenadressen abzutippen. Daher werden alle Ressourcen auch über die Adresse http://manufacturinganalytics.de bereitgestellt.

2 Rollen und Teamstruktur

Dieses Kapitel befasst sich mit Themen rund um den Aufbau und die Entwicklung von Datenanalyseabteilungen. Neben dem Datenanalysten und/oder Data Scientist als Rolle werden weitere Positionen und deren Zusammenspiel vorgestellt. Außerdem wird beschrieben, wie ein Aufbau von Datenanalyseeinheiten vorangetrieben werden kann bzw. welche Vor- und Nachteile sich je nach Strategie ergeben. Ziel des Kapitels ist es, Entscheidern und Analysten Denkanstöße an die Hand zu geben, wie ein Analytics-Team strukturiert bzw. (weiter-) entwickelt und in das Unternehmen integriert werden kann.

2.1 Rollen

Datenanalysten und Data Scientists nehmen in Analytics-Projekten eine zentrale Rolle ein. Dennoch sind sie ohne die Hilfe von Datenarchitekten bzw. Data Engineers oder Business-Analysten häufig nur wenig schlagkräftig. Dabei müssen die genannten Rollen weder alle immer verfügbar sein noch unbedingt durch explizite Spezialisten besetzt werden, um mit Analysen zu beginnen. Häufig können zu Beginn breit ausgebildete Datenanalysten andere Rollen oder Doppelfunktionen in Projekten einnehmen.

Auch wenn nicht alle Rollen anfangs vorhanden sind, bedeutet das nicht, dass sie außer Acht gelassen werden können. Ein Mangel an Data Engineers führt unweigerlich dazu, dass Datenanalysten nur schwer oder mit viel manuellem Aufwand an Daten kommen und schnell frustriert sind. Eine Daumenregel besagt, dass für jeden Analysten oder Scientist mindestens ein Data Engineer verfügbar sein sollte. Wird dieses Verhältnis zum Nachteil von Data Engineers stark unterschritten, sind Datenanalysten gezwungen, die Aufgaben zu übernehmen. Wie später beschrieben wird, ist diese Situation aber maximal als vorübergehende Lösung geeignet. Umgekehrt bringen zu viele Data Engineers nichts, wenn niemand Daten anfordert. In der aktuellen Arbeitsmarktsituation ist ein solches Szenario aber eher unwahrscheinlich. Sind auf der anderen Seite nicht genügend Software Engineers im Unternehmen oder fehlen die Gegebenheiten, Modelle auszurollen, bleiben Lösungen in den Schubfächern. Abgesehen von Einmal-Analysen und Reports gilt dann: Was nicht ausgerollt bzw. in Produktion gesetzt wird, stiftet auch keinen Nutzen. Dadurch bekommen andere Abteilungen und Bereiche im Unternehmen den Eindruck, dass Analysten keine oder zu wenige Projekte vorantreiben, und das Management erhöht den Druck. Auch dann werden Data Engineers und Scientists frustriert sein, da ihre aufwendig entwickelten Pipelines und Modelle nicht zur Anwendung kommen. Benefits, die hinter den Lösungen stehen, werden nicht ausgeschöpft. Die Ergebnisse erscheinen schnell wertlos.

Dank Online-Angeboten wie Coursera, edX oder O'Reilly Safari

INTERNET
https://www.coursera.org/
https://www.edx.org/
https://www.safaribooksonline.com/

ist es leichter denn je, sich mit dem Thema Datenverarbeitung und -analyse zu beschäftigen und wertvolles Wissen aufzubauen. Viele der genannten Angebote sind zumindest zur Ansicht frei, so

dass Kosten nur dann anfallen, wenn Übungsaufgaben genutzt werden bzw. Zertifikate nach erfolgreichem Abschluss gewünscht sind. Die Qualität solcher Kurse ist hoch und die bereitgestellten Ressourcen sind vielseitig und praxisnah. Damit lässt sich im Unternehmen nicht nur durch neue Köpfe, sondern auch durch die Weiterbildung affiner Mitarbeiter neues Potenzial schöpfen. Das hat den Vorteil, dass diese Kollegen bereits notwendiges Domänenwissen mitbringen. Dennoch sollten ihnen fachlich und methodisch erfahrene Kollegen zur Seite gestellt werden.

Die Rollen Data Scientist und Datenanalyst werden häufig voneinander abgegrenzt und unterschiedlich definiert. [3]. Das Gleiche gilt für die beiden Rollen Data Architect und Data Engineer. Gerade in der Praxis verschwimmen diese Grenzen. Um Komplexität zu vermeiden, werden in diesem Buch die Begriffe Data Analyst und Data Scientist sowie Data Engineer und Data Architect zusammengefasst und synonym verwendet.

2.1.1 Datenanalyst / Data Scientist

Datenanalyse oder Data Science Jobs werden im Internet als «Sexiest Jobs of the 21st Century» gehypt und erwecken dadurch eine enorme Erwartungshaltung an die Personen, die diesen Professionen nachgehen. Dabei ranken sich gerade jetzt, wo das Thema einen Hype erlebt, Mythen um die Fähigkeiten, die ein Datenanalyst mitbringt, und die Methoden, die er verwendet. Es gibt aber auch Beschreibungen, die sich mehr an der Realität orientieren und versuchen, die verschiedenen Anforderungen an einen guten Datenanalysten zu formalisieren. Eine der wohl bekanntesten und einfachsten Darstellungen ist das Mengendiagramm von Drew Conway [4]. Die Darstellung wurde bereits im September 2010 in einem Blog-Eintrag vorgestellt und anschließend vielfach diskutiert und erweitert (siehe zum Beispiel [5] und [6]). In seiner ursprünglichen Form beschreibt es auf einfache Weise, welche Fähigkeiten ein Data Scientist mitbringen sollte (Bild 2.1).

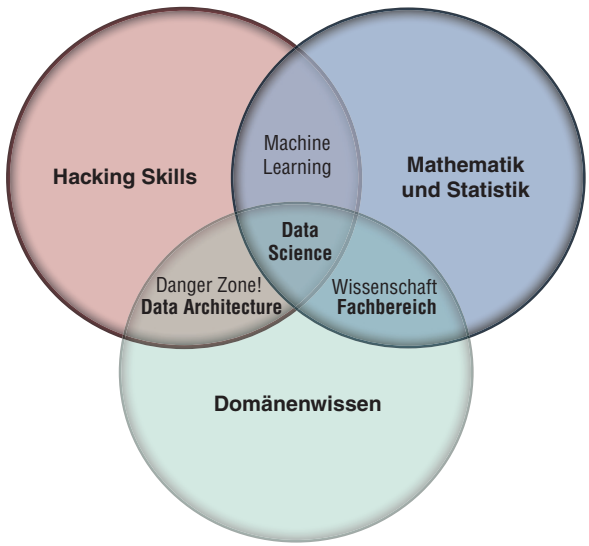

Bild 2.1 *Venn-Diagram nach Conway* [4]

In dem Diagramm sind drei Kernfähigkeiten eines Datenanalysten abgebildet. Mit «**Hacking Skills**» sind solide Kenntnisse von Programmiersprachen und -paradigmen gemeint. Das bedeutet, dass ein Data Scientist, obwohl er kein Software-Entwickler sein muss, dennoch Erfahrung

im Umgang mit Programmiersprachen mitbringen sollte. Das ist aus zwei Gründen wichtig. Programmiererfahrung schärft zum einen das Verständnis von Effizienz und Performance. So muss zum Beispiel nicht unbedingt das komplexeste Modell mit der höchsten Genauigkeit die beste Lösung sein. Stattdessen können auch Reaktionszeiten und andere Performancekennzahlen eine große Rolle bei der Auswahl der besten Methoden spielen. Zum anderen sind gerade in einer Zeit, in der immer mehr grafische Tools wie RapidMiner und KNIME die Nutzung von Analytics-Methoden unterstützen, solide Programmierkenntnisse in Sprachen wie R oder Python äußerst wertvoll, um eigene Methoden zu entwickeln bzw. bestehende anzupassen und zu verfeinern (siehe Abschnitt 5.2).

«**Mathematik und Statistik**» bilden zentrale Grundlagen eines jeden guten Datenanalysten. Die Anwendung dieses Wissens zieht sich von der ersten Betrachtung der Daten (z.B. mithilfe statistischer Mittel) bis hin zur finalen Modellbildung durch alle Phasen der Arbeit eines Analysten. Da auch heute noch ein Großteil der verwendeten Analytics-Modelle auf Mathematik und Statistik fußen, ist ein fundierter Einsatz von Methoden ohne das Verständnis zugrundeliegender Konzepte fast unmöglich. Darüber hinaus gibt es mittlerweile Sprachen wie SystemML,

INTERNET
https://systemml.apache.org/

die mathematische Beschreibungen von Algorithmen auch für etablierte Programmiersprachen wie R oder Python zulassen und eigenständig in performanten Programmcode überführen. Setzt sich dieser Trend fort, wird der Anteil notwendiger Programmierfähigkeiten bei der reinen Modellbildung zugunsten von mathematischen Kenntnissen geringer. Eine detailliertere Auseinandersetzung mit diesen Themen befindet sich in Abschnitt 3.5.

Die dritte Komponente «**Domänenwissen»** beschreibt Wissen und Erfahrung im Arbeitsumfeld, in dem sich der Datenanalyst befindet. Das sind Kenntnisse zu den Prozessen innerhalb des Unternehmens, aber vor allem auch zu den Produkten und deren Produktionsabläufen. Je nachdem, in welchem Bereich die Datenanalyse stattfinden soll, kann dabei Wissen über Controlling-, Verkaufs-, Einkaufsabläufe oder rund um die Herstellung und Beschaffenheit von Produkten relevant sein. Dass eine einzelne Person vor allem in großen Unternehmen keinen Überblick über alle (Produktions-) Prozesse haben kann, erklärt sich von selbst. Ein Datenanalyst sollte aber daran interessiert sein, eigenes Wissen rund um seine Analyseaufgaben aufzubauen, und sich nicht darauf verlassen, dieses ständig von den Fachbereichen zu erfragen. Je mehr Wissen und Erfahrung der Datenanalyst mitbringt, desto schneller kann er Problemstellungen erfassen, zu Ergebnissen gelangen, mögliche Fehler in Modellen erkennen und selbst neue Ideen entwickeln, wie bestehende Modelle erweitert oder neue Modelle entwickelt werden können.

Conway beschreibt in seinem Diagramm nicht nur die drei Domänen, sondern auch, welche Rollen sich ergeben, wenn nur zwei der drei Bereiche ausgeprägt sind. So sind Personen mit Programmiererfahrung und mathematisch-statistischem Wissen gut für die Entwicklung von Machine-Learning-Algorithmen geeignet. Das klingt nachvollziehbar, da diese Personen ihr mathematisches und statistisches Wissen einsetzen können, um neue Algorithmen zu entwickeln, die dann mithilfe der Programmiererfahrung in effiziente Codefragmente oder sogar eigene Bibliotheken umgesetzt werden. Die Kombination aus Mathematik und Statistik sowie ausgeprägtem Domänenwissen beschreibt Conway als traditionelle Wissenschaft. Personen mit diesen Fähigkeiten sind also sehr gut dafür geeignet, neue Methoden, die genau auf Domänenprobleme ausgerichtet sind, zu entwickeln, das heißt also, Spezialwissen bzw. -methoden für anwendungsgetriebene Modelle aufzubauen. Der dritte Überlappungsbereich aus Domänen- und Programmierwissen wird von Conway als «Danger Zone!» bezeichnet. Auch wenn das nicht zwingend zutreffen

muss, besteht aber genau in dieser Kombination an Fähigkeiten die größte Gefahr, dass vermeintliche Erkenntnisse geschaffen werden, die eigentlich keine sind. Gründe sind unter anderem falsche oder fehlerhafte Methoden und Modelle, Evaluationsmethoden oder Merkmale (vgl. Abschnitt 3.3). Das Wissen rund um Mathematik und Statistik hilft hier Methoden besser zu verstehen und eben diesen falschen Erkenntnissen vorzubeugen. Was für Datenanalysen schädlich sein kann, ist ein gutes Set an Fähigkeiten für Datenarchitekten bzw. Data Engineers. Daher sollten Personen mit diesen Fähigkeiten sich eher in Richtung dieser Rollen entwickeln (siehe Abschnitt 2.1.2).

Eine Fähigkeit, die Conway in seinem Venn-Diagramm nicht beschreibt, ist die Fähigkeit zu kommunizieren. Aus Erfahrung ist das wohl eine der zentralen Fähigkeiten, die ein Data Scientist mitbringen sollte. Seine Aufgaben erfordern es, mit allen in die Analyse involvierten Rollen zu interagieren. Er muss mit Datenarchitekten sprechen, um die richtigen Daten in der geeigneten Form und Qualität zu bekommen. Ebenso muss er sich mit Software-Entwicklern austauschen, um den Nutzen der entwickelten Modelle nicht in der Umsetzung zu verlieren. Schließlich muss der Datenanalyst auch mit Business-Analysten und Fachbereichen kommunizieren, um Ziele klar zu definieren, Ergebnisse vorzustellen und zu diskutieren. Gerade Fachbereiche bestehen aus Fachkräften, Ingenieuren und Führungskräften mit unterschiedlichen Interessen und Anforderungen. Jede dieser Gruppen will auf ihrer Ebene angesprochen werden und verstehen, worum es in den Projekten geht. Zusammenfassend muss der Datenanalyst die Zusammenarbeit mit den einzelnen Stakeholdern im Unternehmen für ein erfolgreiches Datenanalyseprojekt verstehen und meistern. Dies läuft über aktive Kommunikation, Storytelling, also das Beschreiben der Modelle und Ergebnisse im unternehmerischen Kontext, und Präsentation der Ansätze und Ergebnisse. Im Idealfall ist also das Set an Fähigkeiten eines Data Scientists die Schnittmenge aus allen drei Bereichen von Conway sowie ausgeprägten Kommunikationsfähigkeiten. In der Realität kommt es häufig vor, dass nicht alle Eigenschaften gleich gut ausgeprägt sind. Jungen oder branchenfremden Data Scientists fehlt das Domänenwissen. Quereinsteigern aus den Unternehmen fehlen unter Umständen mathematische oder Programmierkenntnisse. Daher ist eine gute Mischung und Besetzung der Rollen der Schlüssel für ein erfolgreiches Team.

Was die Unterscheidung zwischen Datenanalysten und Data Scientists anbelangt, so werden Datenanalysten vor allem als methodisch stark beschrieben. Sie haben also einen stärkeren Fokus auf mathematisch-statistisches und Domänenwissen. Data Scientists hingegen besitzen auch solide Hacking Skills, sind also versiert bei der Programmierung und Nutzung von progammiernahen Tools.

2.1.2 Data Engineer / Data Architect

Je nachdem, wen man fragt, bestehen 70 bis 90% der Arbeit in einem Datenanalyseprojekt aus Datenbeschaffung, -verständnis und -aufbereitung (siehe Abschnitt 3.3). Gerade beim Aufbau von großen Datenanalyseteams werden Daten häufig redundant angefragt, abgezogen und aufbereitet. Dies geschieht dann mit «Bordmitteln», also so, wie es schnell erledigt werden kann. Dabei entstehen zum Beispiel unvollständige Ordnerstrukturen mit CSV-Dateien auf lokalen Rechnern und dadurch Intransparenz, was die Verarbeitung und die Inhalte angeht. Data Governance und Security werden gänzlich außer Acht gelassen. Zwar lautet eines der Paradigmen heutiger Datenanalysen, schnell Modelle aufzubauen, zu evaluieren, um sie dann entweder zu verwerfen oder weiterzuentwickeln. Das gilt allerdings nur für die Erstellung von Modellen und nicht für die Sammlung und Zusammenführung von Daten. Ergebnisse sind nur dann reproduzierbar und nutzbar, wenn genau die gleiche Datenbasis verwendet wird und ein vollständiger Überblick über die vorhandenen bzw. verfügbaren Daten besteht. Es empfiehlt sich daher, eine Rolle oder ein Team zu schaffen, das sich mit diesen Themen zentral auseinandersetzt.

Genau hier kommt der Data Engineer bzw. Data Architect ins Spiel. Ziel dieser Rollen ist es, einen Überblick über alle oder zumindest einen abgegrenzten Teil der Datenquellen im Unternehmen zu erlangen und diese so zusammenzuführen und zu formalisieren, dass daraus im besten

Fall eine Art Data-as-a-Service-Angebot für die Datenanalysten entsteht. Sein Handwerkszeug sind Datenlandkarten, die detailliert beschreiben, wo sich welche Daten im Unternehmen befinden, und ETL-Tools (***E**xtract* = Extrahieren, ***T**ransform* = Transformieren und ***L**oad* = Laden), die speziell für das Extrahieren, Zusammenführen und Aufbereiten von Daten entwickelt wurden. Sie ermöglichen die Übertragung in übergeordnete Datenstrukturen. Im Idealfall entsteht so ein Data Lake oder Data Hub, in dem Daten aus allen benötigten Datenquellen zusammengeführt werden. Datenanalysten können sich so die Daten für Anwendungsfälle zusammenziehen (siehe Abschnitt 2.1.1). Data Engineers sollten daher beim Aufbau größerer Teams gleich zu Beginn eingeplant werden und gemeinsam mit den IT-Abteilungen an entsprechenden Konzepten arbeiten. Das hilft, Fehler in der Datenhaltung zu vermeiden und einheitliche Standards für die Ablage und Nutzung von Daten zu entwickeln und zu etablieren.

Auf dem Mengendiagramm von Conway (siehe Bild 2.1) bringen Data Engineers vor allem Programmierfähigkeiten bzw. allgemein eine hohe Technologieaffinität für Datenbanksysteme sowie Domänenwissen zum besseren Verständnis der Zusammenhänge mit. Darüber hinaus ist ein hohes Maß an Strukturiertheit notwendig – gerade zu Beginn, wenn viele Datenquellen kartografiert und vereinheitlicht werden sollen.

Wie eingangs kurz erwähnt, lässt sich darüber diskutieren, ob und welche Unterschiede es zwischen Data Engineers und Architects gibt. Datenarchitekten gehen dabei eher in Richtung datenübergreifende Architekturen und sind vor allem für die Strukturierung und Formalisierung von Datenquellen verantwortlich. Data Engineers hingegen haben ein größeres technisches Know-how im Umgang mit Datentechnologien und sind eher Macher. Ob und inwieweit die Rollen voneinander getrennt werden, hängt von der Größe und den Anforderungen im Unternehmen ab. In diesem Buch gibt es keine Unterscheidung.

2.1.3 Business Analyst / Fachbereich

Fachbereiche sind die Sponsoren von Analytics-Projekten. Sie haben geschäftsrelevante Ziele, die durch Datenanalysen bzw. deren Ergebnisse erreicht oder unterstützt werden sollen. Größere Unternehmen leisten sich darüber hinaus Business-Analysten, die mit den Fachbereichen in engem Kontakt stehen. Sie unterstützen die Kommunikation zwischen Fachbereichen und Analysten vor allem in der Projektanbahnung und besitzen detailliertes Domänenwissen, um neue Ideen gemeinsam mit den Fachbereichen zu entwickeln. Business-Analysten sollten im Idealfall neben solidem Domänenwissen ausreichende IT-Kenntnisse bzw. Grundlagen in der Datenanalyse mitbringen, um gegebenenfalls eine erste Vorselektion und Priorisierung von Themen vorzunehmen. Sie können Fachbereichen dabei helfen, Analytics Use Cases aufzunehmen und diese in einer Roadmap mit der Unterstützung von Datenanalysten nach Priorität und Machbarkeit strukturiert darzustellen. Diese Roadmap sollte auch Abhängigkeiten zu anderen, Nicht-Analytics-Projekten (zum Beispiel Montage von relevanten Sensoren), abbilden, deren Erfolg Voraussetzung für die Analytics-Projekte ist.

2.1.4 Software-Entwickler / Systemadministrator

Datenanalysen lassen sich grob in zwei Arten unterteilen: einmalige Analysen, die erstellt werden, um Erkenntnisse genau auf Basis eines zusammengetragenen Datensatzes durchzuführen, und kontinuierlich genutzte Modelle. Die erarbeiteten Ergebnisse einmaliger Analysen werden in Reports oder auf Folien zusammengefasst. Damit ist das Projekt abgeschlossen. Der größere Teil von Analysen soll jedoch in regelmäßigen Abständen oder wiederholt zum Beispiel auf Basis aktueller Daten durchgeführt werden. Das erfordert, dass die von Analysten entwickelten Modelle

operationalisiert und gegebenenfalls in bestehende Systeme integriert werden. Sollen zum Beispiel Dashboards erstellt werden, gilt es die erstellten Algorithmen und Verarbeitungsmethoden entweder in die Dashboarding-Tools zu integrieren oder getrennt durch andere Systeme im Hintergrund die Ergebnisse vorberechnen zu lassen. Sobald Analysemodelle in operationale Prozesse des Unternehmens integriert werden sollen, gilt es Software-Entwickler und Systemadministratoren einzubeziehen. Sie sorgen dafür, dass die Modelle entsprechend der vorliegenden IT-Infrastruktur in systemnahe Programmiersprachen übersetzt werden, Anforderungen an die Robustheit und Effizienz einhalten und andere unternehmenswichtige Systeme nicht stören. Unter Umständen müssen für die Anwendungsfälle auch neue Technologien aufgebaut und in die IT-Landschaft des Unternehmens integriert werden. Hier lohnt sich ein Blick in Abschnitt 5.2. Dort werden einige Technologien und Tools vorgestellt, die nicht nur bei der Erstellung der Analysemodelle unterstützen, sondern auch für deren Operationalisierung Basis sein können.

Der Austausch mit Entwicklern und IT-Abteilungen ist für Datenanalysten wichtig. Sie können wertvolle Hinweise liefern, unter welchen Gegebenheiten die Modelle später laufen müssen und welche Restriktionen auf technischer Seite vorliegen. Darüber hinaus funktioniert das reine Weitergeben von Code und Artefakten an Entwickler nicht. Vielmehr sollte an einer Art DevOps-Konzept gearbeitet werden. Das DevOps-Konzept besagt, dass derjenige, der die Modelle entwickelt, diese auch operational betreut. Dadurch werden Brüche, Verzögerungen und Verständnisprobleme vermieden. Diese können gerade auch dann entstehen, wenn Modelle schlecht dokumentiert sind oder die Kommunikation zwischen beiden Parteien schlecht läuft. Darüber hinaus reduzieren sich Entwicklungsaufwände, da Modelle von Anfang an in entsprechenden Programmiersprachen oder mit entsprechenden Vorgaben entwickelt werden. Außerdem lassen sich Anpassungen und die kontinuierliche Verbesserung und Weiterentwicklung der Modelle (sogenanntes Continuous Development) dynamischer umsetzen.

Das DevOps-Vorgehen lässt sich im Szenario zwischen Analysten und Entwicklern sicherlich nicht immer eins zu eins umsetzen. Allerdings fördert eine enge Zusammenarbeit zwischen Entwicklern, Administratoren und Datenanalysten das Verständnis für die Anforderungen und Ziele und somit eine effizientere Produktivsetzung von Modellen. Darüber hinaus können gemeinsam gesetzte Standards und Prozesse dazu beitragen, dass beide Seiten auf Wünsche und Veränderungen schnell reagieren können.

INTERNET

Weitere Information zum DevOps-Konzept gibt es unter

https://www.cloudcomputing-insider.de/devops-und-alle-ziehen-an-einem-strang-a-501139/index2.html oder
https://de.atlassian.com/devops

Bild 2.2 fasst die diskutierten Anknüpfungspunkte zusammen. Der Datenanalyst steht im Zentrum des Projekts und stimmt sich mit den anderen Rollen ab – je nachdem, in welcher Phase des Projekts er sich befindet. Dadurch erhält er Unterstützung, Anforderungen und Wissen zur Umsetzung des Projekts. Eine enge Vernetzung mit den anderen Rollen ist daher äußerst wertvoll, um Datenanalyseprojekte von der Beschreibung des Use Cases bis zur Operationalisierung von Analysemodellen erfolgreich zu gestalten. Eine detaillierte Beschreibung des Vorgehens in Analytics-Projekten folgt in Kapitel 3. Dort werden die Phasen Schritt für Schritt nachvollzogen.

Bild 2.2 *Zusammenarbeit zwischen den Rollen*

2.2 Teamaufbau

2.2.1 Teamstruktur

Der Aufbau von Analytics-Teams stellt Firmen vor einige Herausforderungen. Es ergeben sich Fragen, wie viele oder welche Datenanalysten benötigt werden, wo diese in der Unternehmensstruktur angesiedelt sein sollten und ob und wann weitere Rollen wie Data Engineers oder Business-Analysten besetzt werden sollten.

Die Antworten auf die Fragen hängen von Branche, Unternehmenskultur und Zielsetzung ab. Viele Unternehmen, die mittlerweile große Analytics-Einheiten besitzen, haben klein angefangen. Das heißt, es wurde eine Gruppe aus zwei bis fünf Datenanalysten und -architekten angestellt und entweder in eine entstehende Digitalisierungseinheit oder in die IT-Abteilung integriert. Ziel der Gruppe ist es dann, schnelle Erfolge zu realisieren, um zu zeigen, dass sich Datenanalysen im Unternehmen lohnen. Ist dies gelungen, werden die Abteilungen ausgebaut, um weitere Rollen erweitert und gegebenenfalls sogar fachlich unterteilt. Dieser Weg bietet im ersten Schritt ein vergleichsweise geringes finanzielles Risiko sowie geringe Management- und HR-Kosten. Gerade in Unternehmen, die mit Digitalisierungsthemen erst beginnen oder nicht allzu weit vorangeschritten sind, bietet dieses Vorgehen die Möglichkeit, sich an das Thema Analytics anzunähern und es in der Unternehmenskultur zu etablieren. Nachteilig ist hingegen die geringe Schlagkraft solcher kleinen Teams, da nur wenige Themen gleichzeitig abgearbeitet werden können und starke Managementunterstützung notwendig ist, um die Bereitschaft zur Zusammenarbeit mit den Analysten zu verstärken.

Andere Unternehmen vertrauen von vorneherein – vielleicht auch aufgrund des späten Einstiegs und hohen Marktdrucks – auf Erfolge aus Datenanalysen und bauen sofort große Teams mit verschiedenen Rollen auf. Der Vorteil liegt darin, dass die Sichtbarkeit im Unternehmen schnell vorhanden ist und viele Themen von Anfang an parallel angegangen werden können. Darüber hinaus müssen Strukturen nicht erst nach und nach mit dem Wachstum der Abteilung etabliert