

Kurt Stockinger, Thilo Stadelmann, und Andreas Ruckstuhl

Zusammenfassung

Data Scientists sind gefragt: Laut Mc Kinsey Global Institute wird es in den nächsten Jahren allein in den USA einen Nachfrageüberschuss an 190.000 Data Scientists geben. Dieser sehr starke Nachfragetrend zeigt sich auch in Europa und im Speziellen in der Schweiz. Doch was verbirgt sich hinter einem Data Scientist und wie kann man sich zum Data Scientist ausbilden lassen?

In diesem Kapitel definieren wir die Begriffe Data Science und das zugehörige Berufsbild des Data Scientists. Danach analysieren wir drei typische Use Cases und zeigen auf, wie Data Science zur praktischen Anwendung kommt. Im letzten Teil des Kapitels berichten wir über unsere Erfahrungen aus dem schweizweit ersten Diploma of Advanced Studies (DAS) in Data Science, das an der ZHAW im Herbst 2014 erstmals gestartet ist.

Schlüsselwörter

Data Science • Data Warehousing • Machine Learning • Angewandte Statistik • Datenanalyse • Weiterbildung • Data Science Use Cases

Vollständig überarbeiteter und erweiterter Beitrag basierend auf „Data Science für Lehre, Forschung und Praxis“. In: HMD – Praxis der Wirtschaftsinformatik, HMD-Heft Nr. 298, 51 (4): 469–479, 2014.

K. Stockinger (✉) • T. Stadelmann • A. Ruckstuhl
Zürcher Hochschule für Angewandte Wissenschaften, Winterthur, Schweiz
E-Mail: Kurt.Stockinger@zhaw.ch; Thilo.Stadelmann@zhaw.ch; Andreas.Ruckstuhl@zhaw.ch

4.1 Data Science als Disziplin

Mit dem rasanten Aufkommen von Data Science als Disziplin geht die Genese des Berufsbildes des „Data Scientists“ einher (Loukides 2010). Beide Konzepte in ihrer heute populären Form entstanden dabei aus den Bedürfnissen der Wirtschaft heraus (Patil 2011). Eine wissenschaftliche Auseinandersetzung folgte etwas verzögert und nimmt aktuell an Fahrt auf (Brodie 2015a). Von Beginn an wurden sie jedoch begleitet von enormer medialer Beachtung bis hin zum Hype (z. B. Davenport und Patil 2012). Dies weckt Skepsis unter Fachleuten, sollte jedoch nicht den Blick für das reale Potenzial der Thematik trüben:

- Wirtschaftliches Potenzial – Das McKinsey Global Institute errechnet einen weltweiten Wert von ca. drei Billionen Dollar für die Nutzung von Open Data allein (Chui et al. 2014), zu dessen Realisierung bis zu 190.000 Data Scientists benötigt werden (Manyika et al. 2011).
- Gesellschaftliche Auswirkungen – Medizinische Versorgung, politische Meinungsbildung und die persönliche Freiheit werden durch Datenanalyse beeinflusst (Parekh 2015).
- Wissenschaftlicher Einfluss – Datenintensive Analyse als viertes Paradigma der Wissenschaft verspricht Durchbrüche von der Physik bis zu den Lebenswissenschaften (Hey et al. 2009).

Der Hype sagt: *„Data is the new oil!“*. Im Original setzt sich dieses Zitat fort mit *„[...] if unrefined, it cannot really be used. It has to be changed [...] to create a valuable entity that drives profitable activity“* (Humby 2006). Schon hier wird die Notwendigkeit der Arbeit des Data Scientists zur Realisierung dieses großen Potenzials angesprochen.

4.2 Definition: Data Scientists, Data Science und Data Products

4.2.1 Der Data Scientist

Geschichte

Data Science nach heutigem Verständnis¹ beginnt mit der Einführung des Begriffs „Data Scientist“ durch Patil und Hammerbacher während ihrer Arbeit bei LinkedIn respektive Facebook (Patil 2011). Sie empfinden, dass für die Mitarbeiter ihrer Teams, die über tief gehendes Ingenieurwissen verfügen und oft direkten Einfluss auf die Wirtschaftlichkeit der Kernprodukte im Unternehmen haben, eine neue Tätigkeitsbeschreibung notwendig sei: *„those who use both data and science to create something new“*.

¹Der Begriff „Data Science“ ist jedoch älter. Unter anderem hat William S. Cleveland 2001 in einem unveröffentlichten Artikel (vgl. Cleveland 2014) „Data Science“ als eigenständige Disziplin vorgeschlagen. Darin finden sich schon viele Elemente des heutigen Verständnisses.

Jones (2014) stellt in seinem Bericht anschaulich dar, wie es aus diesem Kerngedanken heraus sowie der Anforderung, dass Data Scientists ihre Ergebnisse und Einsichten auch selber kommunizieren können sollen, innerhalb kurzer Zeit zu einer inhaltlichen Explosion des Anforderungsprofils kommt: „...*business leaders see Data Scientists as a bridge that can finally align IT and Business*“. Auch auf technischer Seite gehen die Anforderungen über das ursprünglich geforderte „Deep Analytical Talent“ (Manyika et al. 2011) hinaus: Die Vorbereitung der Analyse, Data Curation genannt und bestehend aus dem Suchen, Zusammenstellen und Integrieren der möglicherweise heterogenen Datenquellen, macht ca. 80% der täglichen Arbeit eines Data Scientists aus (Brodie 2015b). Dies schliesst die Arbeit an und mit großen IT-Systemen ein.

Darüber hinaus trägt der Data Scientist hohe Verantwortung: Analysen wollen nicht nur vorbereitet, durchgeführt und kommuniziert werden; aufgrund des disruptiven Potenzials des datengetriebenen Paradigmas (Needham 2013) ist es angezeigt, deren inhärente Risiken explizit zu machen. Hierzu sind Analyseergebnisse auch mit Maßen über die erwartete Korrektheit, Vollständigkeit und Anwendbarkeit auszustatten (Brodie 2015b). Gerade in großen Datenbeständen und hohen Dimensionen versagt intuitives Verständnis für Zusammenhänge und macht Konfidenzbetrachtungen unumgänglich.² Eine informelle Umfrage unter allen ca. 190 Teilnehmern der SDS|2015 Konferenz³ ergab jedoch, dass von 80 praktizierenden Data Scientists nur etwa die Hälfte regelmässig solche Überlegungen anstellen – auch, da ihre Kunden nicht danach fragen.

Trends

Dem entgegen beobachten wir aktuell im Bereich der deutschsprachigen Wirtschaft, vor allem im Umfeld der Solution Provider für Big Data, folgende Trends:

- Die Arbeit des Data Scientist wird reduziert auf das Bedienen eines Tools im Sinne von „Self-Service BI“.
- Ergebnisse für komplexe und wissenschaftlich ungelöste Fragestellungen wie etwa Social Media Monitoring (Cieliebak et al. 2014) werden auf Knopfdruck versprochen.

In Beziehung gesetzt zu obiger Verantwortung offenbart diese Entwicklung Gefahren. Data Scientists sollten genau verstanden haben, welche Schlüsse die eingesetzten Methoden und Verfahren zulassen und welche nicht.

Gleichzeitig nehmen wir eine Zunahme an Data Scientist Positionen in der Schweiz war. Eine weitere informelle Umfrage am SDS|2015 ergab folgendes Bild: Circa 40% der Teilnehmer betrachten sich selber als praktizierende Data Scientists – jedoch trägt nur ein Viertel von ihnen den Titel „Data Scientist“ in der Arbeitsplatzbeschreibung. Zum Zeitpunkt der Planung der Vorgängerveranstaltung, 2013 waren schweizweit kaum 2 Data Scientists identifizierbar.

²Eine nicht unumstrittene Forderung – siehe die Abschaffung der Angabe von p-Werten in diesem Journal: <http://www.tandfonline.com/doi/pdf/10.1080/01973533.2015.1012991>

³SDS|2015 – The 2nd Swiss Workshop on Data Science (12. Juni 2015 in Winterthur), ist einer der Treffpunkte der Schweizer Data Science Community. <http://www.dlab.zhaw.ch/sds2015>

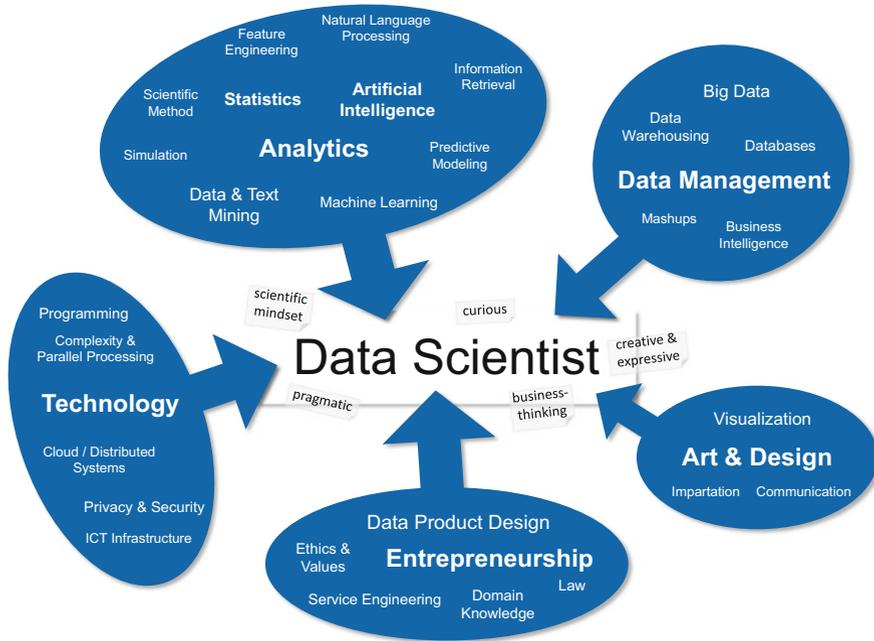


Abb. 4.1 Die im Vergleich zum Original leicht überarbeitete „Data Science Skill Set Map“ (Stadelmann et al. 2013)

Definition

Wie also kann das Berufsbild des Data Scientists aktuell definiert werden? Wir verwenden dafür die in Abb. 4.1 dargestellte Landkarte der zugeordneten Fertigkeiten und Eigenschaften des Data Scientists. Diese Landkarte dient zur Präzisierung der verwendeten Begriffe im Alltag unseres interdisziplinären Forschungslabors⁴ und ermöglicht einen schnellen Überblick über die Skillbereiche eines Data Scientists.

Die in den Ovalen zusammengefassten Gebiete auf dieser Karte entsprechen dabei wichtigen Kompetenzclustern, die sich der Data Scientist aus dem Repertoire teils mehrerer etablierter Teildisziplinen aneignet; die grauen Etiketten im Zentrum beschreiben Eigenschaften seiner Denk- und Arbeitsweise.

Im Einzelnen erfordert die zielgerichtete, analytische Arbeit an Datensätzen folgende Eigenschaften auf Seiten des Data Scientists:

- Kreativität, Neugier und wissenschaftliche Denkweise fördern neuartige Erkenntnisse zu Tage.
- Unternehmerisches Denken hält dabei ein klares Ziel vor Augen.

⁴Datalab – The ZHAW Data Science Laboratory. <http://dlab.zhaw.ch/>

- Pragmatismus sorgt für die notwendige Effizienz in einer komplexen Tool-Landschaft.

Diese Eigenschaften sind schwer trainierbar, aber wichtig für den praktischen Erfolg.

Im Folgenden gehen wir auf die einzelnen Kompetenzcluster im Profil des Data Scientists genauer ein:

Technologie und Datenmanagement. Der Umgang mit Daten ist so entscheidend, dass Datenmanagement-Fähigkeiten als eigenes Kompetenzgebiet auftauchen, auch (aber keineswegs nur) „at scale“ im Umfeld von Big Data. Doch auch andere technologische Fähigkeiten aus der Informatik und angrenzenden Gebieten sind in der Praxis des Data Scientists wichtig, allen voran das Programmieren – jedoch eher im Sinne von Scripting als der Entwicklung großer Softwaresysteme. Systemdesign spielt im Sinne des Zusammensetzens verschiedener (auch verteilter) Frameworks und Dienste eine Rolle.

Analytics. Analytische Fähigkeiten aus dem Bereich der Statistik, des maschinellen Lernens und der künstlichen Intelligenz zur Extraktion von Wissen aus Daten und zur Generierung von (Vorhersage-)Modellen sind die Kernfähigkeit des Data Scientists. Hierbei ist der unterschiedliche Zugang der Teildisziplinen, etwa von Statistikern und Informatikern bzgl. Modellierung (Breimann 2001), besonders wertvoll.⁵

Unternehmertum. Der Data Scientist hat nicht nur die Verantwortung zur Implementierung einer analytischen Lösung für ein gegebenes Problem. Er benötigt auch die Fähigkeit zum Stellen der richtigen Fragen bzgl. geschäftlichem Mehrwert sowie Folgen für Betrieb und Gesellschaft. Dies bedingt auch den Aufbau substanziellen Wissens aus der jeweiligen Fachdomäne sowie das Wahrnehmen ethischer Verantwortung: Viele Fragestellungen in Data Science berühren grundlegende Fragen des Datenschutzes und der Privatheit und sollten auch entsprechend rechtlich abgesichert sein.

Kommunikation und Design. Als Verantwortlicher für den gesamten analytischen Workflow kommuniziert der Data Scientist selbst seine Ergebnisse auf (Senior) Management Ebene. Dies benötigt neben adressatengerechter Kommunikation die Fähigkeit zur korrekten grafischen Aufbereitung komplexester Zusammenhänge via Informationsvisualisierung. Auch als Teil analytischer Lösungen und Services für den Kunden sind angemessene grafische Darstellungen bedeutend. Gleichzeitig spielt die grafische Aufbereitung von Daten unter dem Stichwort Visual Analytics bereits während der Datenexploration eine große Rolle.

⁵Breiman diskutiert den Unterschied zwischen modellgetriebenen Ansätzen und sogenanntem „algorithmischen“ Vorgehen. Der Unterschied liegt im Umfang der a priori Annahmen, welche den Raum möglicher Lösungen unterschiedlich einschränken bzw. formen. Beispielhaft sei die Annahme bestimmter Verteilungen der Daten vs. rein algorithmische Erfassung etwa mittels eines neuronalen Netzes genannt.

Schicht	Inhalt
Infrastruktur 	<ul style="list-style-type: none"> • Datenbanken • Cloud Computing • Big Data Technologien
Algorithmen 	<ul style="list-style-type: none"> • Data Mining, Statistik & Predictive Modeling • Maschinelles Lernen & Graphanalyse • Information Retrieval & Sprachverarbeitung • Business Intelligence & Visual Analytics • Data Warehousing & Entscheidungsunterstützung
Geschäft 	<ul style="list-style-type: none"> • Visualisierung & Kommunikation der Ergebnisse • Privatheit, Sicherheit & Ethik • Unternehmertum & Data Product Design

Abb. 4.2 Logischer Aufbau eines Data Science Curriculums

Der Weg zum Data Scientist

Ein erfahrener Data Scientist sollte etwa 80% dieser Kompetenzlandkarte abdecken, verteilt über alle fünf Ovale. Dies bedingt eine feste Verankerung als Experte in einem der vier Kompetenzcluster sowie gut abgestütztes Wissen und Fähigkeiten in wenigstens 2 weiteren, ohne dabei den Spagat zwischen technisch-analytischen und ökonomisch-kommunikativen Kompetenzen völlig zu vermeiden.

Die notwendigen Fähigkeiten können trainiert werden, gegeben eine Veranlagung zu quantitativen, komplexen und technischen Fragestellungen. Ein typischer Werdegang beginnt mit einem Studium etwa in Statistik, Informatik oder datenintensiven Wissenschaften, von dem aus Fähigkeiten in den anderen Bereichen durch interdisziplinäre Arbeit und Weiterbildung hinzugewonnen werden.

Abb. 4.2 skizziert die Inhalte eines Data Science Curriculums, aufgeteilt in Schichten nach Abstraktionsgrad vom Business Case. Die Inhalte des Geschäfts-Layers liegen nah an den Use Cases der Praxis, die vom Data Scientist nicht nur oberflächlich verstanden werden müssen. Dies spielt in die Auswahl der Verfahren aus dem Algorithmen-Layer hinein, abstrahiert aber weitgehend von der technischen Infrastruktur.

Aus oben diskutierten Gründen der Verantwortung für weitreichende Entscheidungen heraus scheint es uns wichtig zu betonen, die analytischen Aspekte ins Zentrum der Ausbildung zu stellen: Maschinelles Lernen, Statistik und deren zugrunde liegende Theorien müssen fest verankert sein, da sich aus ihnen heraus Machbarkeit und Folgenabschätzung ergeben.

In Abschn. 4.4 werden wir auf die Umsetzung dieses Konzepts eingehen.

4.2.2 Data Science als interdisziplinäre, angewandte Wissenschaft

Nachdem wir definiert haben, was ein Data Scientist können und wie er arbeiten sollte, wollen wir uns seiner Tätigkeit zusätzlich aus anderer Richtung nähern, nämlich über die Abgrenzung von Data Science als Disziplin seines Wirkens. Nach Brodie (Brodie 2015b) ist Data Science (oder „*Data-Intensive Analysis*“) die Anwendung der wissenschaftlichen Methode⁶ auf Daten. Wie grenzt sich diese Disziplin von ihrem Umfeld ab?

Data Science ist eine interdisziplinäre Wissenschaft, die Methoden zur Auswertung unterschiedlichster Arten von Daten mit verschiedensten Mitteln bündelt. Ausgehend von konkreten Fragestellungen wird ein Data Product entwickelt, d. h. eine neue Information oder ein neuer Service, der Wertschöpfung aus der Analyse bestehender Daten betreibt.

Es ist schwierig, sich mit Data Science auseinanderzusetzen und nicht allenthalben diese inhärente Interdisziplinarität wahrzunehmen. Gleichzeitig wird nirgends der Anspruch erhoben, diese Begriffe seien ihren Ursprungsdisziplinen zu enteignen und dieser neuen „Disziplin-in-Entstehung“ einzuverleiben.

Vielmehr ist Data Science eine einzigartige, also neue Mischung von Fertigkeiten aus Analytics, Engineering und Kommunikation, um ein spezifisches Ziel zu erreichen, nämlich die Erzeugung von einem (gesellschaftlichen oder betrieblichen) Mehrwert aus Daten. Als angewandte Wissenschaft lässt sie den Teildisziplinen ihren Wert und ist dennoch eigenständig und notwendig (siehe die Diskussion in (Provost und Fawcett 2013)).

Ob die Entwicklung dabei ähnlich verläuft wie die Bildung der Subdisziplin Data Mining, die bis heute etwa *in* Statistik- und Informatik-Curricula zu finden ist, oder analog zum *Herausschälen* der Informatik aus den Fachgebieten Elektrotechnik und Mathematik, ist noch nicht endgültig auszumachen. Wir beobachten, dass die Bündelung analytischen Wissens aller Fachgebiete in Forschungszentren und Ausbildungscurricula voranschreitet, während sich die prognostizierte Omnipräsenz analytischer Fragestellungen in Wirtschaft und Gesellschaft weiter entwickelt. Gleichzeitig sehen wir momentan eher interdisziplinäre Initiativen und Kooperationen anstatt einer grundlegenden Neuordnung der akademischen Landkarte bezüglich der Fachgebiete.

4.2.3 Data Products und wie sie entwickelt werden

In Erweiterung des explorativen Paradigmas im Data Mining plant der Data Scientist gezielt, was seiner Organisation einen Mehrwert verschaffen könnte. Entsprechend rückt das Data Product ins Zentrum des Analyseprozesses: Es ist das Ergebnis der wertgetriebenen Analyse (Loukides 2010) und kann dabei ein Service

⁶Experiment, Messung und Theoriebildung. In diesem Sinne wäre ein Data Scientist „...a person (professionally) involved in the conduct of data science“ (Brodie 2015b).

für Endkunden oder auch nur eine einzige Zahl als Entscheidungsunterstützung für interne Stakeholder sein; wesentlich ist, dass Mehrwert aus der Analyse von Daten geschöpft und realisiert wurde.

Was macht das Data Product aus? Siegel nennt 147 Beispiele erfolgreicher Data Products aus 9 Branchen (Siegel 2013). Zu den bekanntesten zählen sicher die Empfehlungsservices von Amazon und Netflix sowie diverse Kundenkartenprogramme, beispielsweise von Tesco. Doch Loukides weist darauf hin, dass, was auch immer mit den Daten „unter der Haube“ passiere, „...*the products aren't about the data; they're about enabling their users to do whatever they want, which most often has little to do with data*“ (Loukides 2011).

Entsprechend wichtig für die Entwicklung erfolgreicher Data Products erscheinen uns daher die Gedanken aus den Bereichen Service Engineering und klassische Produktentwicklung zu sein, welche mit dem „Value Proposition Design“ starten und vom Kunden aus denken (Osterwalder et al. 2014). Diese Ansätze fragen zuerst nach dem realisierbaren Wert des Produkts in spe, bevor die Analytik angestossen wird.

Einen ähnlichen Weg schlagen Howard und Kollegen mit dem „Drivetrain Approach“ vor (Howard et al. 2012): Zu Beginn ihres vierstufigen Prozesses steht die Definition des Ziels: Welches Kundenbedürfnis sollte idealerweise als nächstes adressiert werden? Im zweiten Schritt werden diejenigen Hebel identifiziert, mit deren Bewegung der Data Scientist die Erreichung dieses Ziels beeinflussen kann. Dies kann etwa eine neue analytische Idee sein, wie seinerzeit die Einführung des „Page Rank“ als Kriterium für die Güte von Suchresultaten bei Google.

Um die identifizierten Hebel in Bewegung setzen zu können, werden im dritten Schritt die notwendigen Datenquellen betrachtet. Diese müssen nicht nur bereits vorhandenen, internen Töpfen entstammen. In der Verknüpfung unternehmensinterner sowie externer Daten liegt manchmal der Schlüssel zum Beantworten analytischer Fragestellungen. Daher ist die Frage nach der Entwicklung von Data Products eng verknüpft mit Kenntnissen über den Datenmarkt sowie die Open Data Bewegung: Hier finden sich Möglichkeiten zum Aufstocken des verfügbaren eigenen Rohmaterials. Erst im vierten und letzten Prozessschritt wird schliesslich über analytische Modelle nachgedacht, denn deren Auswahl wird zu einem guten Teil durch das vorgegebene Ziel und dessen Rahmenbedingungen, die anzusetzenden Hebel sowie die verfügbaren Datenquellen (und -Mengen) bestimmt.

Es erstaunt uns, dass unserer Recherche nach bislang nur sehr wenige Ausbildungsprogramme für Data Product Design existieren. Im Syllabus eines der wenigen existierenden Kurse heißt es (Caffo 2015): „*The course will focus on the statistical fundamentals of creating a data product that can be used to tell a story about data to a mass audience*“. Anschließend stehen technologische Details zum Implementieren von Webanwendungen im Zentrum. Wir sehen hingegen einen starken Bedarf nach einem interdisziplinär ausgerichteten Kurs, der das „Data Storytelling“ nicht vergisst, jedoch den technisch ausgerichteten Data Scientists (siehe oben) das notwendige Businesswissen vermittelt, um erfolgreiche Produkte zu kreieren. In Abschn. 4.4 berichten über die erste Durchführung eines Kurses, der diese Gedanken aufnimmt.

4.3 Data Science Use Cases

In diesem Abschnitt stellen wir drei unterschiedliche Data Science Use Cases vor, die wir im Zuge von angewandten Forschungsprojekten mit Wirtschaftspartnern umgesetzt haben. Der erste Use Case (Markt Monitoring) basiert auf Technologien aus den Bereichen Data Warehousing und Machine Learning. Der zweite Use Case (Analytisches Customer Relationship Management) beruht auf dem Konzept des Customer Lifetime Value und Methoden aus dem predictive Modelling. Im dritten Use Case wird ein laufendes Projekt zur Entwicklung eines Systems für prädiktive Instandhaltung von Flugzeugkomponenten vorgestellt.

4.3.1 Markt Monitoring

Ausgangslage dieses Use Cases ist eine eCommerce Plattform mit Millionen von Produkten des täglichen Lebens. Ziel der Plattform ist es, den Usern gesündere Produkte bzw. Produkte aus nachhaltiger Produktion zu empfehlen. Im Unterschied zu typischen Recommender Systemen, die von Amazon oder Netflix bekannt sind, bleiben die User dieser eCommerce Plattform vollkommen anonym, da für das Durchsuchen der Produkte kein User Account notwendig ist. Ein weiterer Unterschied zu Amazon und Netflix ist, dass die User dieser Plattform nur in den seltensten Fällen ein Produkt kaufen, sondern sich lediglich über Produkteigenschaften informieren wollen. Somit stellt dieser Use Case eine andere und teilweise schwierigere Herausforderung dar, als die beiden bereits erwähnten Plattformen. Herkömmliche Ansätze verbinden Nutzerprofile mit dem positiven Feedback bestätigter Einkäufe. Da es jedoch weder Nutzerprofile noch bestätigte Einkäufe gibt, müssen andere Methoden herangezogen werden, um die Klickpfade auszuwerten.

Um die Kundenbedürfnisse besser zu verstehen, kann somit einerseits nur das Klickverhalten auf der eCommerce Webseite analysiert werden. Andererseits müssen auch die Eigenschaften der Produkte untersucht werden, sodass dem User bessere und nachhaltigere Produkte empfohlen werden können. Um diese Analysen durchzuführen, wurde zunächst ein *Data Warehouse* (DWH) erstellt, das die Produktinformation und die Klickpfade enthält (siehe Abb. 4.3).

Das Data Warehouse besteht aus den drei Schichten *Staging Area*, *Integration Layer* und *Enrichment Layer*. In der Staging Area werden zunächst die Daten der Produktdatenbank abgespeichert. Zusätzlich werden via Google Analytics statistische Information über Geschlecht und Altersgruppen der User hinzugefügt. Im Integration Layer werden die Daten homogenisiert, Duplikate entfernt und in einheitliches Datenmodell integriert. Da die Analyseergebnisse in einem Web-Portal mit vernünftigen Antwortzeiten dargestellt werden sollen, beinhaltet das Data Warehouse einen Enrichment Layer, in dem komplexe Berechnungen materialisiert und physisch reorganisiert werden. Der Enrichment Layer dient somit dazu, eine höchstmögliche Datenbank-Query-Performance zu erzielen.

Um die Kundenbedürfnisse zu analysieren, wurden unterschiedliche Machine Learning Ansätze verwendet. Abb. 4.4 zeigt eine schematische Darstellung einer

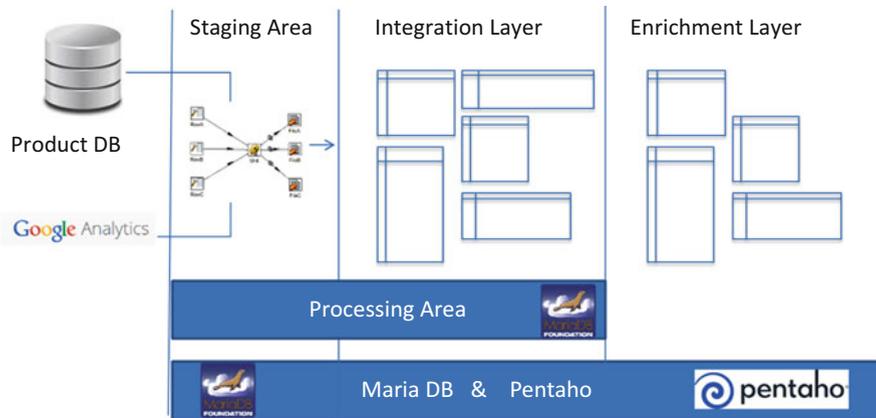


Abb. 4.3 Data Warehouse Architektur des Markt Monitoring

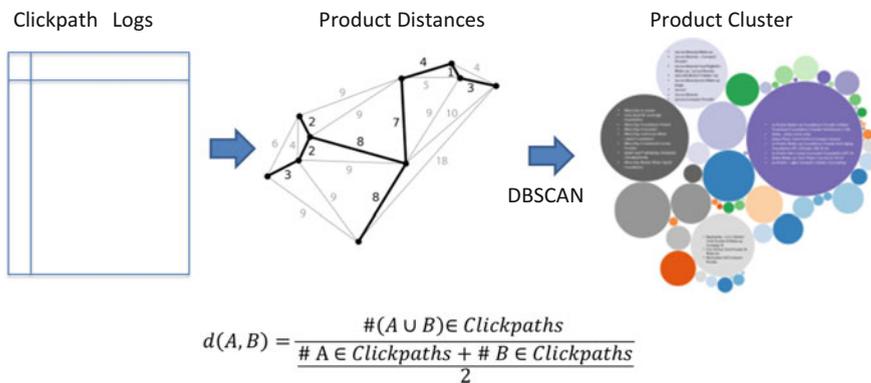


Abb. 4.4 Clusteranalyse von Produkten unterschiedlichster Klick-Pfade

Cluster-Analyse basierend auf den Klickpfaden der User sowie den Produkten, die sie angewählt haben. Mit Hilfe des Clustering-Algorithmus DBScan (Density-Based Spatial Clustering), konnten die ausgewählten Produkte in sinnvolle Cluster unterteilt werden, die es dann ermöglichen, entsprechend ähnliche Produkte zu finden. Um jedoch gesündere Produkte empfehlen zu können, müssen die gefundenen Cluster noch mit den Produktinhaltsstoffen kombiniert und entsprechend bewertet werden. Diese Bewertung kann vom End-User nach bestimmten Kriterien wie z. B. Zucker- oder Glutengehalt gewichtet und entsprechend priorisiert werden.

In unserem Projekt verwendeten wir ausschliesslich Open-Source-Technologien, da wir mit einem kleinen Start-Up zusammenarbeiteten und somit teure, kommerzielle Data Warehouse- und Analyse-Lösungen nicht in Frage kamen. Das Data Warehouse basiert auf MariaDB und Pentaho. Als Machine Learning Library wird Apache Mahout verwendet (vor allem die Algorithmen K-means, K-means mit Canopy und Fuzzy

K-Means). Da DBScan nicht in Mahout vorhanden ist, wurde dieser Algorithmus selbst in Java implementiert und optimiert.

Eine der wichtigsten Erkenntnisse dieses Projektes ist, dass ein gut entworfenes Data Warehouse eine wesentliche Grundvoraussetzung für komplexe Analysen ist. Durch das Integrieren und Bereinigen der Daten ermöglicht das Data Warehouse, Datenqualitätsprobleme zu erkennen und somit korrekte Analysen durchzuführen. Darüber hinaus ermöglicht ein physisch optimierter Enrichment Layer, dass End-User-Queries signifikant schneller ausgeführt werden können, als wenn die Queries direkt auf den Rohdaten abgesetzt werden. Eine der großen Herausforderungen war es, mehrere Tabellen mit 10^7 bis 10^8 Einträgen miteinander zu joinen, um möglichst effiziente Auswertungen zu machen. Hierfür mussten wir die Zugriffspfade und die End-User-Queries analysieren und entsprechende Datenbankindizes aufbauen, um einerseits die Abfragen zu beschleunigen und andererseits die Ladegeschwindigkeit des Data Warehouses möglichst wenig zu beeinträchtigen.

Ein weiterer wichtiger Punkt ist, möglichst früh mit einer explorativen Datenanalyse zu beginnen. Dies hilft einerseits, die Grundeigenschaften der Daten zu verstehen und somit erste Erkenntnisse zu gewinnen. Andererseits hilft es auch dabei, die Datenqualität besser zu verstehen und somit in einem iterativen Prozess das DWH zu erstellen und kontinuierlich zu erweitern bzw. die Query-Performance der Abfragen zu optimieren. Erst danach ist es angebracht, skalierende Machine Learning Algorithmen zu implementieren, um Millionen von Datensätzen in sinnvoller Zeit zu analysieren.

4.3.2 Analytisches Customer Relationship Management

Eine typische Aufgabe im Customer Relationship Management (CRM) ist die Selektion von geeigneten Kunden für Up-Selling-, Cross-Selling- oder Retentionsmaßnahmen. Traditionell wird dazu im analytischen CRM die Auswahl gemäß der Wahrscheinlichkeit, positiv auf die Maßnahme zu reagieren, oder äquivalent gemäß eines Scoring-Wertes getroffen. Allerdings müssen die auf diese Weise identifizierten Kunden nicht notwendigerweise diejenigen sein, von denen das Unternehmen am meisten profitiert. Deshalb fand in den letzten Jahren das Konzept des *Customer Lifetime Value (CLV)* zunehmend Beachtung. Die Selektion der Kunden beruht darin auf dem aktuellen (d. h. diskontierten) Wert aller zukünftigen Einnahmen, die durch die jeweiligen Kunden generiert werden. Während der traditionelle Ansatz dazu führt, dass die Anzahl der Kunden maximiert wird, führt der CLV-Ansatz zu einem finanziell ergiebigeren Portfolio von Kunden.

Das sind nicht nur aus betriebsökonomischer Sicht unterschiedliche Ansätze, sondern sie verlangen auch ein unterschiedliches Vorgehen in der Analyse. Im ersten Ansatz genügen geeignete Klassifikationsmethoden, die aufgrund der Kundeneigenschaften eine positive oder negative Selektion vornehmen. Der zweite Ansatz klingt sinnvoller, impliziert jedoch eine große analytische Herausforderung. Der CLV wird durch das künftige Verhalten der Kunden festgelegt. Wollen wir also einen CLV-Wert für einen Kunden bestimmen, muss das zukünftige Verhalten der

Kunden vorhergesagt werden. Eine Möglichkeit dafür ist die Verwendung geeigneter dynamischer Modelle, die an (umfangreichen) Daten aus der Vergangenheit kalibriert sind.

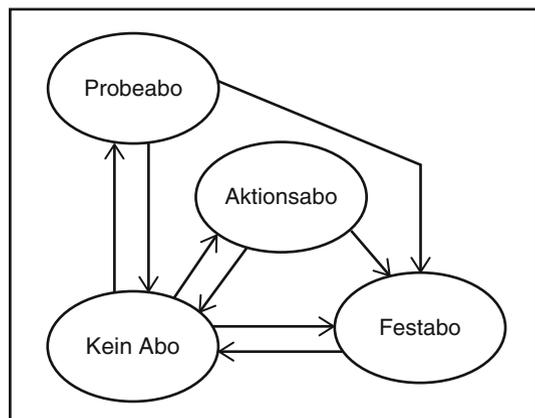
Bevor wir uns noch etwas detaillierter mit der dynamischen Modellierung befassen, sei hier auf das Verhältnis zwischen den drei Elementen Daten, Analytics und Einbettung des Produkts ins betriebliche Umfeld eingegangen. Im obigen Zusammenhang ist klar, dass die betriebsökonomische Einbettung der CRM-Aktion vorgibt, welche Eigenschaften oder Bedingungen das Endprodukt (oder das Data Product) erfüllen muss. Daraus lässt sich ableiten, welche Analysemethoden einsetzbar sind, und welche Daten zur Verfügung stehen müssten.

Umgekehrt wird im betrieblichen Umfeld oft auf Daten zurückgegriffen, die man sowieso zur Verfügung hat, d. h. die meistens aus anderen Gründen gesammelt wurden. Diese Gegensätze können im Projekt zu Spannungen führen und verlangen ein pragmatisches Vorgehen. Ein Data Science Projekt wird aus Sicht des Unternehmens dann erfolgreich sein, wenn alle drei Elemente, Daten, Analytics und betriebliche Einbettung des Data Products, zusammenpassen und eine zielführende Qualität haben. Weil das dritte Element von einer anderen Natur ist, wird es von Data Scientists gerne etwas vernachlässigt, ist aber für den Erfolg des Projekts unabdingbar.

Die dynamische Modellierung kann, wie Heitz, Ruckstuhl und Dettling zeigen, auf einem *Semi-Markov-Modell-Ansatz* aufbauen (Heitz et al. 2010). Die Grundlage dieses Ansatzes basiert auf der Unterteilung des Kundenverhaltens in mehrere disjunkte Zustände. Dies erlaubt eine differenzierte Modellierung der jeweiligen Verweildauer, der Übergangswahrscheinlichkeiten und der Verteilung der Umsätze.

In Abb. 4.5 hat das Beispiel für Zeitungsabonnemente vier verschiedene Zustände (bei anderen Beispielen können aber weit mehr Zustände nötig sein). Zu bemerken ist, dass eine Kundin oder ein Kunde nicht beliebig zwischen den Zuständen wechseln kann. Die Pfeile geben an, was möglich ist. Für jeden Zustand werden nun je drei zustandsspezifische Modelle benötigt:

Abb. 4.5 Zustände und erlaubte Übergänge für das Beispiel Zeitungsabonnemente



- ein *Vorhersagemodell für die Verweildauer* der Kundin im entsprechenden Zustand basierend auf discrete-duration models
- ein *Vorhersagemodell für die individuellen Übergangswahrscheinlichkeiten* bei einem Wechsel des Zustands (wir verwenden multinominale Logit-Modelle bei wenigen Merkmalen und das Random Forest Verfahren bei vielen Merkmalen)
- ein *Vorhersagemodell für die Umsätze*, die der Kunde generiert. Dies sind in unserem Fall jeweils zustandsspezifische Werte.

Diese Art der dynamischen Modellierung bedingt eine große Anzahl von Vorhersagemodellen und die Notwendigkeit von deren Kalibrierung. Im Weiteren beruhen diese Modelle auf unterschiedlichen Kundenmerkmalen, die sich üblicherweise aus der Benutzung der Produkte und sozio-demografischen Eigenschaften ergeben. Je nach Modellansatz muss eine Merkmalsselektion explizit vorgenommen werden. Werden viele Details der Produktnutzung festgehalten wie z. B. in der Telekommunikation, so müssen enorme Datenmengen verarbeitet und geeignet aufbereitet werden.

Bei Zeitungsabonnements ist die Datenlage viel spärlicher und man kämpft damit, überhaupt einen geeigneten Merkmalsatz zur Verfügung zu haben. In der letzten Zeit kam die Nachfrage auf, Merkmale, die auf Textdaten basieren, wie sie z. B. in Call-Centern entstehen, einzubeziehen. Unsere Erfahrung im Fall von Call-Center-Textdaten zeigt bisher, dass die (finanziellen) Vorteile den Aufwand für die immer wiederkehrende z. T. manuelle Textdatenbereinigung und das Textmining nicht decken.

Sind einmal alle Modelle für jeden Kunden kalibriert, so kann man die konkreten CLV-Werte explizit berechnen. Mit dem erstmaligen Aufsetzen und expliziten Berechnen des CLV-Ansatzes ist die Aufgabe jedoch lange nicht beendet. Üblicherweise müssen die Berechnungen periodisch (z. B. monatlich) wiederholt werden, um auch die Dynamik in der Kundschaft wie z. B. Verhaltensänderungen zu erfassen. Es zeigt sich, dass dabei eine vollständige Neuaufsetzung der Modelle nicht nötig ist, da sich in der Regel von einer Periode zur anderen die Situation nicht allzu sehr ändert, und dies auch viel zu aufwändig wäre. Deshalb werden nur die Parameter neu geschätzt, ohne Änderungen am eigentlichen Modell oder am benötigten Satz von Merkmalen vorzunehmen.

Über längere Zeit jedoch könnten sich dann aber doch Änderungen aufdrängen. Solche Bedürfnisse rufen nach einem automatischen System, welches die periodischen Resultate liefert und die Qualität der Resultate festhält im Sinne einer statistischen Prozessüberwachung. Ergeben sich deutliche Abweichungen in der Qualität der Resultate, muss das System entsprechende Warnungen ausgeben, sodass der Data Scientist die nötigen Modifikationen an den Vorhersagemodellen, respektive dem Merkmalsatz, vornehmen kann. Der Bau und der Betrieb eines solchen automatischen Überwachungssystems sind aufwendig. Weil die Ansätze inklusive des zur Verfügung stehenden Merkmalsatzes zudem oft (zu) kurzlebig sind, wird ein solches Überwachungssystem oft nur rudimentär implementiert.

Mehr zum oben beschriebenen Ansatz der dynamischen Modellierung findet sich in (Heitz et al. 2010) sowie (Heitz et al. 2011).

4.3.3 Predictive Maintenance

In letzter Zeit beschäftigen uns zunehmend Fragen aus dem Umfeld der prädiktiven oder zustandsbasierten Instandhaltung. Inzwischen sind einige Projekte gestartet worden und laufen. Obwohl wir noch keine abschließenden Resultate haben, möchten wir hier eines dieser laufenden Projekte vorstellen und aufzeigen, welche Herausforderungen bestehen.

Unser Wirtschaftspartner ist ein führender Anbieter von technischen Lösungen für Fluggesellschaften, unter anderem auch in der Instandhaltung von demontierbaren Flugzeugkomponenten (Pumpen, Computer, Ventile, Stellmotoren ...). Für die meisten Teile ist das Flugzeug so ausgelegt, dass es sicherheitstechnisch kein Risiko darstellt, die Komponenten bis zum Ausfall zu betreiben und den Fehler erst bei dessen Auftreten zu beheben; d.h. es wird eine reaktive Instandhaltungsstrategie gefahren. Sie hat zur Folge, dass die Materialplanung schwierig sowie aufwändig ist und ungeplante Betriebsunterbrechungen unvermeidlich sind. Dies kann dann zu Flugverspätungen und entsprechenden Kostenfolgen führen.

Das Ziel des Projekts ist es, neuartige Dienstleistungen im Bereich prädiktive Instandhaltung anbieten zu können. Dies bedingt, entsprechende Maintenance-Konzepte und Service-Produkte zu entwickeln, ein dazu passendes IT-System aufzubauen und auf die Service-Produkte ausgerichtete Analysekonzepte für Zustandsschätzung zu entwickeln. Weil man keine zusätzlichen Sensoren einbauen kann oder will, müssen die Analysen auf bereits vorhandenen Sensor- und Betriebsdaten der Flugzeugkomponenten beruhen. Aber auch schon so ist mit einem gewaltigen Datenfluss zu arbeiten. Der vorgesehene Daten- und Informationsfluss ist in Abb. 4.6 ersichtlich.

Erste Ergebnisse im Analytic-Teil zeigen, dass Störungen und Fehler von Komponenten mit geeigneten datenanalytischen Verfahren wie Hautkomponentenanalyse, auf robust geschätzten Kovarianzmatrizen beruhenden Mahalanobis-Distanzen, oder robusten nichtparametrischen Glättungsverfahren identifiziert werden können

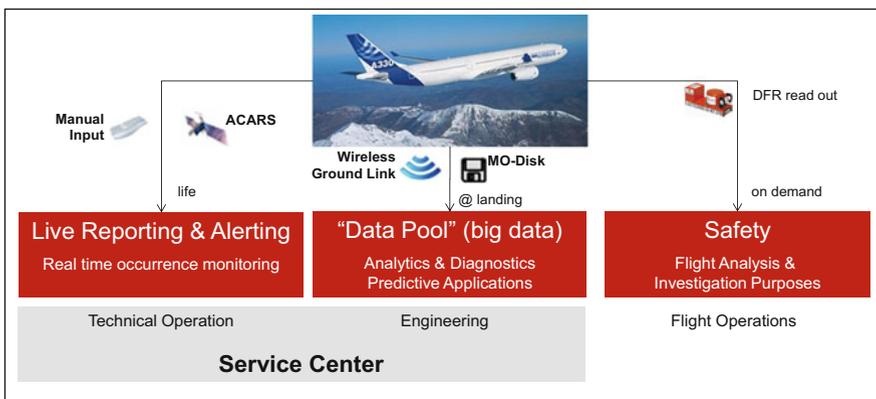
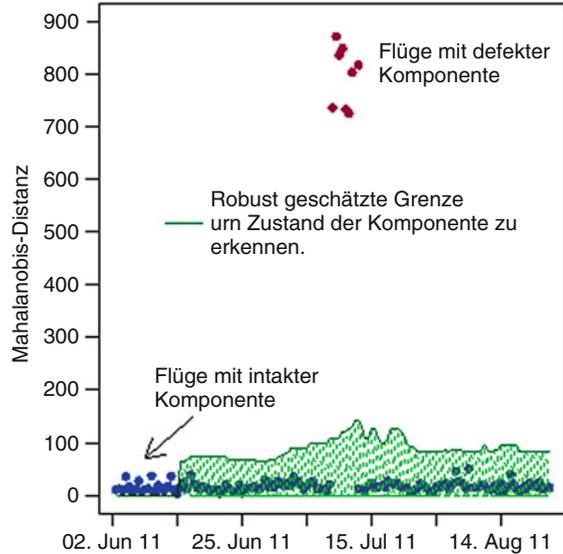


Abb. 4.6 Vorgesehener Daten- und Informationsfluss

Abb. 4.7 Identifizieren von defekten Komponenten mittels Mahalanobis-Distanzen



(vgl. Abb. 4.7). Dabei sind Komponentenausfälle teilweise mehrere Tage im Voraus in den Sensordaten zu erkennen. Ob das ausreicht, werden die Anforderungen aus dem Maintenance-Konzept, respektive aus den Service-Produkten zeigen.

4.4 Erfahrungen aus der Weiterbildung DAS Data Science

In diesem Kapitel berichten wir über unsere Erfahrungen bei der Durchführung des schweizweit ersten Diploma of Advanced Studies in Data Science (DAS). Insbesondere analysieren wir, welche Auswirkungen die Weiterbildung in Bezug auf neuerlernte Fähigkeiten, Jobprofil und Einsatzbereiche der Teilnehmer hat. Hierfür setzen wir einen Fragebogen ein, der das Jobprofil und die Skills der Kursteilnehmer vor und nach der Data Science Ausbildung sowie ihre Zukunftserwartungen erfasst.

4.4.1 Data Science Curriculum

Bevor wir uns die Analyseergebnisse im Detail ansehen, stellen wir kurz das Curriculum unseres DAS Data Science vor. Die gesamte Weiterbildung besteht aus den folgenden drei Certificats of Advanced Studies (CAS):

- **CAS Datenanalyse:** Dieser CAS besteht aus 5 Modulen und liefert vor allem die statistischen Grundlagen der Datenanalyse. Es werden Konzepte und Werkzeuge zur Beschreibung & Visualisierung von Daten behandelt und die in der Praxis wichtigen Methoden wie multiple Regression, Zeitreihenanalysen & Prognosen sowie Clustering & Klassifikation erarbeitet und vertieft.

- **CAS Information Engineering:** Dieser CAS besteht aus 4 Modulen und liefert vor allem die Informatikgrundlagen, die für einen Data Scientist wichtig sind. Die Module behandeln Scripting mit Python, Information Retrieval & Text Analytics, Datenbanken & Data Warehousing sowie Big Data.
- **CAS Data Science Applications:** Dieser CAS baut auf den beiden anderen CAS auf und vertieft das Wissen. Die Themen des aus vier Modulen bestehenden CAS sind Machine Learning, Big Data Visualisierung, Design & Entwicklung von Data Products sowie Datenschutz & Datensicherheit. Zusätzlich wird eine Projektarbeit mit Themen aus der Praxis absolviert.

Ein wesentliches Merkmal der Data Science Weiterbildung ist die Praxisrelevanz. Dies wird dadurch garantiert, dass die Dozierenden zusätzlich zu ihrer Hochschultätigkeit eine mehrjährige Berufserfahrung in der Wirtschaft haben und somit die Anforderungen der Wirtschaft mit den neuesten Erkenntnissen aus der Forschung gut in Einklang bringen können. Des Weiteren wird in der Weiterbildung großer Wert auf praktische Umsetzung und Implementierung gesetzt: Neben dem Besuch der Vorlesungen beschäftigen sich die Teilnehmer mit der konkreten Implementierung von Aufgaben in R, Python, SQL bzw. mit den entsprechenden Tools aus Data Warehousing, Information Retrieval und Big Data etc. Für weitere Informationen über den DAS Data Science verweisen wir auf die entsprechende Webseite⁷ der Hochschule.

4.4.2 Auswertung der Data Science Befragung

Nachdem wir das Data Science Curriculum kurz vorgestellt haben, widmen wir uns nun der Auswertung der ersten Kursdurchführung anhand der Umfragerückläufer der Teilnehmer. Insgesamt erhielten wir eine Rückmeldung von 25 der 30 Kursteilnehmer.

Abb. 4.8 zeigt, aus welchen Wirtschaftszweigen die Teilnehmer kommen. Auffallend ist, dass ein Großteil aus den Bereichen Beratung/Dienstleistung bzw. Versicherungswesen stammt. Dahinter reihen sich die Bereiche Verkehr, Softwareentwicklung, Finanzindustrie und Telekommunikation ein.

Abb. 4.9 zeigt die Berufsbezeichnung und Kenntnisse der Teilnehmenden vor der Data Science Weiterbildung. Hier fallen vor allem die beiden Berufsbezeichnungen Berater und BI Specialist auf. Die Teilnehmenden aus diesen beiden Bereichen haben fundierte Kenntnisse in Datenbanken und DWH Architekturen. Berater haben allerdings zusätzlich fundierte Kenntnisse in fortgeschrittener Analyse und Statistikpaketen, während BI Specialists sich durch Skills in BI Tools auszeichnen.

In Abb.4.10 sehen wir diejenigen Skills, deren Erwerb innerhalb der Weiterbildung die Teilnehmer als am wichtigsten empfunden haben: Programmierung in R,

⁷DAS Data Science der ZHAW: www.weiterbildung.zhaw.ch/de/school-of-engineering/programm/das-data-science.html

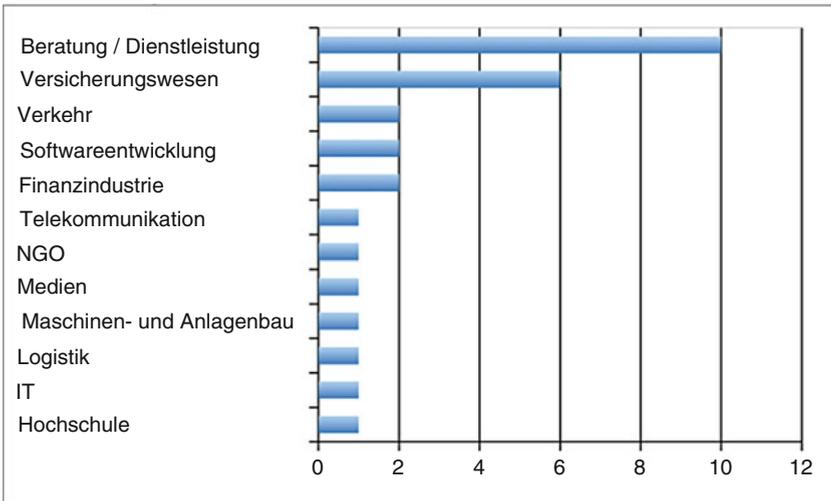


Abb. 4.8 Herkunftsbranchen der Teilnehmenden

Berufsbezeichnung	Kenntnisse													
	BI Tools	Big Data	Data Governance	Data Literacy	Datenbanken / SQL-Abfragen	Desk Research	Deskriptive Statistik	DWH Architekturen	Ethik	Fahrzeugkenntnisse (Bahn)	Fortgeschrittene Analysen	Programmiersprachen	Statistikpakete	Visualisierung
Berater	2	1	3	2	4	0	3	3	1	0	6	1	3	0
BI Specialist	5	0	0	1	5	0	0	4	0	0	0	0	0	0
Business Analyst	1	0	1	0	1	1	1	1	0	0	1	0	0	0
CFO	0	0	0	0	1	0	1	0	0	0	0	1	0	0
Data Miner	1	1	1	1	1	0	2	0	1	0	1	0	1	1
Datenanalytiker/in	2	0	0	1	2	0	2	0	0	0	0	0	1	1
DWH Consultant	1	0	0	0	1	0	0	1	0	0	0	0	0	0
IT-Projekt-Manager	1	0	1	1	1	0	1	1	0	0	1	0	1	0
Software-Entwickler	1	0	0	1	2	0	1	1	0	0	1	2	0	0
Stabstelle Marketing	0	0	0	0	0	0	1	0	0	0	1	0	1	0
Systemingenieur	0	0	0	1	0	0	0	0	0	1	0	1	0	0
Treuhänder	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Abb. 4.9 Berufsbezeichnung und Kenntnisse der Teilnehmenden vor der Data Science Weiterbildung. Mehrfachnennungen kommen vor, vor allem bei den Kenntnissen

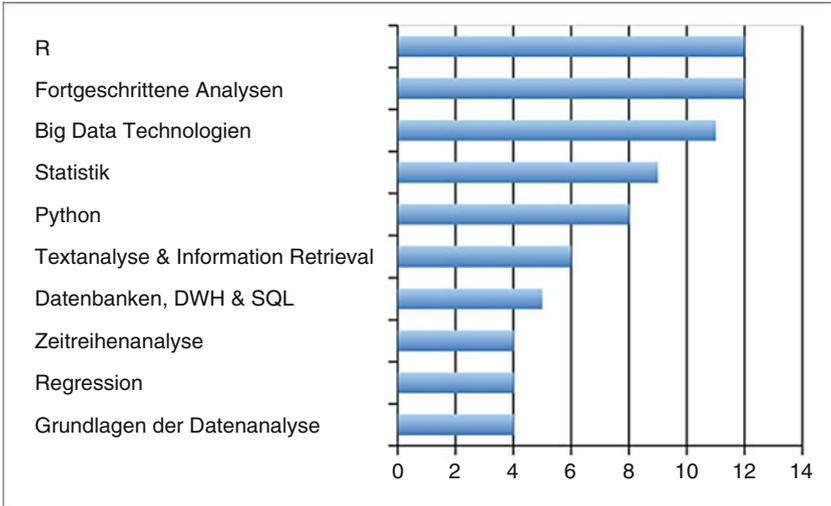


Abb. 4.10 Empfundene Wichtigkeit der Skills, die in der Data Science Ausbildung erworben wurden

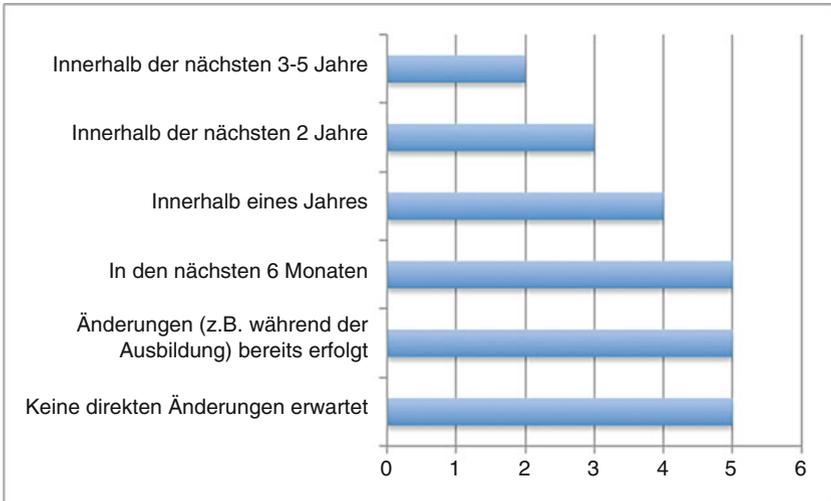


Abb. 4.11 Zeithorizont für erwartete Auswirkungen der Data Science Ausbildung

fortgeschrittene Analysen und Big Data Technologien, gefolgt von Statistik und Programmierung in Python.

Abb.4.11 zeigt den Zeithorizont, ab wann die Teilnehmer eine konkrete Auswirkung der Data Science Weiterbildung in ihrem Job-Alltag erwarten. Hier ist zu erkennen, dass ein Großteil der Teilnehmer bereits während der Ausbildung bzw.

innerhalb der nächsten 6 Monate eine Veränderung erwartet. Interessant ist auch, dass eine gleiche Anzahl an Teilnehmern keine direkten Änderungen erwartet. Bei genauerer Analyse konnte festgestellt werden, dass letztere Teilnehmergruppe vor allem an Data Science Methoden und Tools interessiert ist: Bereits als Lösungsarchitekten oder Führungskräfte tätig, möchten sie technologisch up-to-date bleiben.

Abb. 4.12 zeigt die antizipierte Job-Bezeichnung nach der Data Science Weiterbildung und vergleicht sie mit derjenigen vor der Weiterbildung. Wie zu erkennen ist, plant ein Großteil der Teilnehmer, als Berater oder BI Specialist zu arbeiten. In diesen beiden Fällen können wir keine Veränderung nach der Data Science Weiterbildung feststellen. Auffallend ist jedoch die neue Job-Bezeichnung des Data Scientists. Hier lässt sich erkennen, dass die Veränderungen aus den unterschiedlichsten Jobs resultieren.

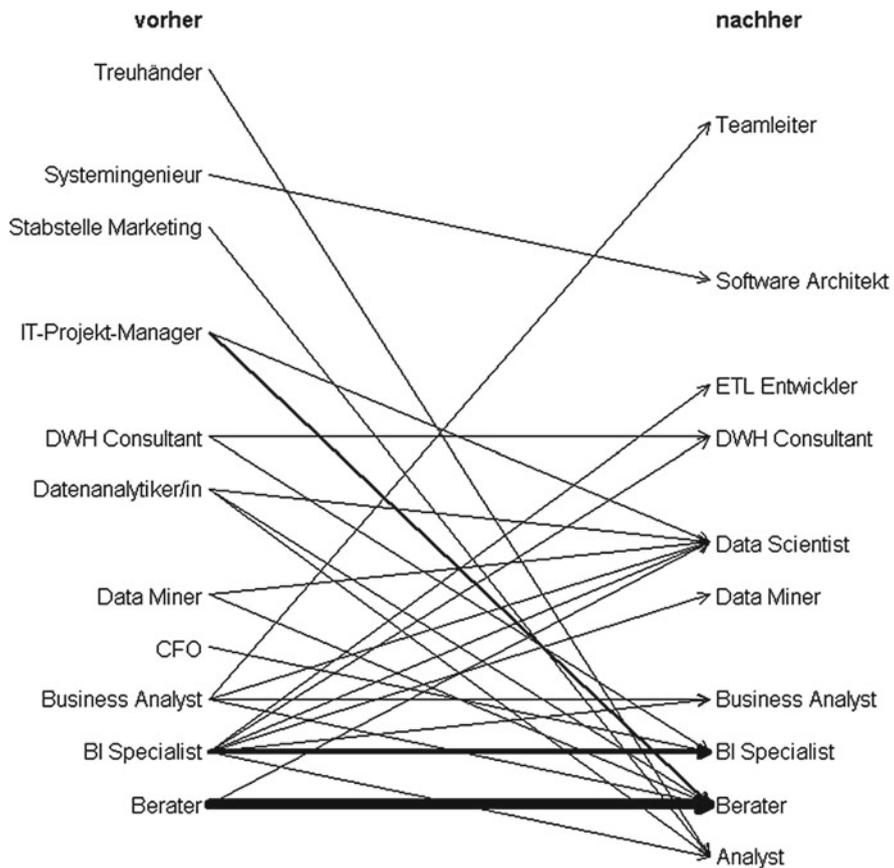


Abb. 4.12 Job-Bezeichnung vor und nach der Data Science Ausbildung. Die entsprechenden Entwicklungen sind mit Pfeilen sichtbar gemacht. Die Liniendicke ist proportional zu den Nennungen. Vereinzelt kommen Mehrfachnennungen vor

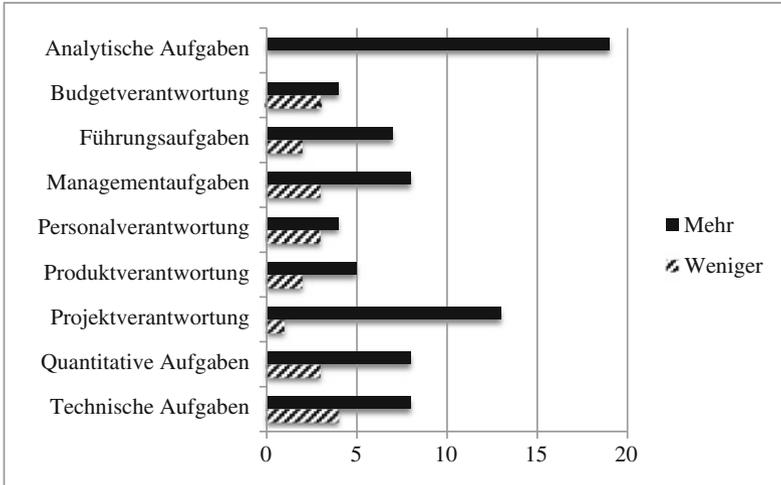


Abb. 4.13 Erwartete Veränderungen in Verantwortung und Aufgabenstellungen nach der Data Science Weiterbildung

Abb. 4.13 zeigt die erwarteten Veränderungen in Verantwortung und Aufgabenstellungen nach der Data Science Ausbildung. Ein Teil der Teilnehmer erwartet vor allem mehr Verantwortung bei analytischen, quantitativen und technischen Aufgaben als auch in Projekten. Andererseits lässt sich auch erkennen, dass andere Teilnehmer wiederum weniger Verantwortung bei technischen und quantitativen Aufgaben bzw. bei Budget-, Personal- und Managementfragen erwarten.

Abb. 4.14 zeigt den geplanten Einsatz der erworbenen Skills. Hier ist klar erkennbar, dass die Teilnehmer vor allem neue analytischen Aufgaben umsetzen wollen mit Schwerpunkt Marketing und Customer Analytics. Kundenberatung unter Zuhilfenahme eines Teams von Data Scientists bzw. das Verstehen und Analysieren von neuen Technologien sind weitere geplante Einsatzfelder für die neu erworbenen Data Science Skills.

Zusammenfassend lassen sich folgende **Erkenntnisse** gewinnen:

- Ein Großteil der Teilnehmer kam aus den datenintensiven Bereichen wie Data Warehousing und Business Intelligence und hatte bereits gute Kenntnisse in SQL und Data Warehousing. Diese Teilnehmer profitierten vor allem von den neu erworbenen Kenntnissen im Bereich tief gehender Analyse (vor allem mit R) und Big Data Technologien (siehe Abb. 4.10). Letztere Technologie wird oft als komplementär zum Data Warehousing angesehen und ist deshalb für Berater von Relevanz, die neue Technologie bewerten und in existierende Applikationslandschaften integrieren müssen.
- Die Analyse der erwarteten Veränderungen in Verantwortung und Aufgabenstellung zeigt zwei komplementäre Tendenzen. Ein Teil der Teilnehmer sieht die Weiterbildung als Sprungbrett für einen neuen Karrierepfad mit mehr Verantwortung im Sinne von Führung, während der andere Teil der Teilnehmer ein

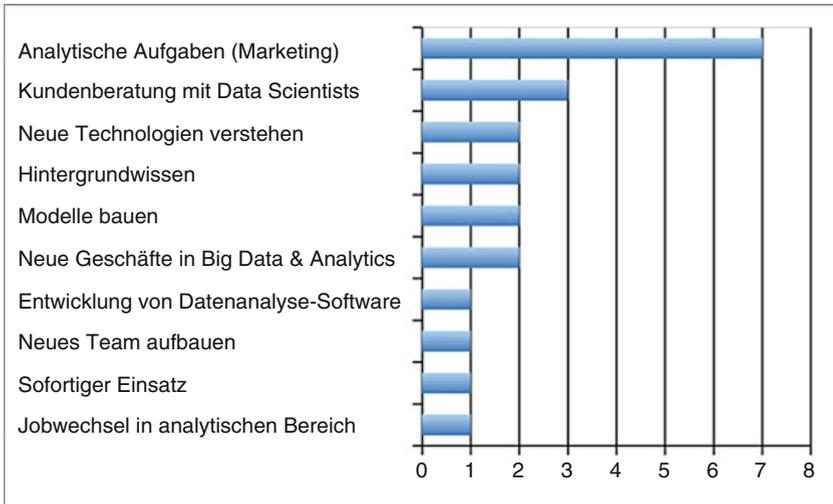


Abb. 4.14 Geplanter Einsatz der neu erworbenen Skills

ziemlich konträres Bild zeigt, nämlich einen Rückzug auf rein technische Aufgaben ohne „lästige“ Managementfunktionen.

- Wie aus den gewählten Themen für die Schlussarbeit zu schliessen ist, fokussiert sich die große Mehrheit der Teilnehmenden auf analytische Fragestellungen. Diese Fokussierung ist ganz im Einklang mit dem ersten Punkt, dass die Teilnehmenden im DAS vor allem von tiefergehenden analytischen Methoden und Big Data Technologien profitierten (siehe Abb. 4.10). Im Weiteren zeigt sich darin auch die große Kompetenzlücke „Analytics“ im betrieblichen Umfeld.

4.5 Ausblick

Wir betrachteten den Beruf des Data Scientists in seiner Entstehung, seiner Definition und Praxis sowie unter Ausbildungsgesichtspunkten. Folgender Punkt erscheint uns dabei aktuell betonenswert: Der „Scientist“ in „Data Scientist“ ist ernst zu nehmen.

Data Science ist eine gleichermaßen anspruchsvolle und verantwortungsvolle Tätigkeit aufgrund ihres allumfänglichen Potenzials zur teilweise disruptiven Veränderung von Wirtschaft und Gesellschaft. Aktuelle Tendenzen in der Industrie, Data Scientists gleichzusetzen mit Bedienern ausgefeilter Softwaretools, sind daher kritisch zu sehen. Die Aufgabe, Unternehmen und andere Organisationen mehr datengetrieben zu machen, sollte in den Händen gut ausgebildeter Fachleute mit viel Sachverstand und Weitblick liegen. Gleichzeitig werden, während mehr Data Scientists in die organisatorischen Strukturen der Unternehmen eingebunden werden, Leitungs- und Schnittstellenfunktionen im Umfeld der Data Scientists notwendig.

Aktuelle Trends zeigen, dass sich das Berufsbild des Data Scientists grob in zwei Stossrichtungen entwickeln wird: Die eine Stossrichtung ist eher im Bereich des

Managements von Data Science Aufgaben verankert. Die andere wird sich eher auf die technisch/methodischen Herausforderungen von Data Science fokussieren. Die zweite Stossrichtung ist allerdings so breit, dass sich vermutlich weitere Spezialisierungen ausbilden werden. Die Ansprüche werden auch so hoch sein, dass mindestens ein MSc als Ausbildung erforderlich sein wird. In jeden Fall verlangt die Arbeit im Data Science Bereich eine hohe Abstraktions-Kompetenz. Auch jemand in der Management-Stossrichtung muss die Möglichkeiten und Grenzen der Modelle und Algorithmen erkennen können. Entscheiden für einen für das Unternehmen nachhaltigen Einsatz von Data Science wird sein, dass sich die Data Scientists nicht selbstverliebt um die Methoden und Algorithmen kümmern, sondern Business Cases erkennen und den Einsatz von Methoden und Algorithmen zielgerichtet auf den Business Case lenken können.

Für die zukünftige Data Science Weiterbildung könnte dies bedeuten, mehr Wahlmöglichkeiten bzw. Spezialisierungen anzubieten. Beispielsweise könnte man sich einen Data Science Management Track und einen Data Science Technical Track vorstellen. Während der Technical Track Data Scientists wie in Abschn. 4.2.1 definiert ausbildet, würde der Management Track – zu Lasten technisch-analytischer Kompetenz – Technik-affine Manager von Data Scientists zum Ziel haben. Hier wäre es sinnvoll, verstärkt auf das Data Product und seine Einbettung in die Prozesslandschaft und Wertschöpfungskette des Unternehmens einzugehen. Im Technical Track wäre es sinnvoll, den Schwerpunkt noch vertiefender auf technischen Umsetzung bzw. Implementierung zu legen. Somit hätten Teilnehmer die Möglichkeit, sich entsprechend ihren Fähigkeiten und Zukunftsplänen weiterzubilden.

Literatur

- Breimann, L.: Statistical modeling: The two cultures. *Stat. Sci.* **16**(3), 199–309 (2001)
- Brodie, M.: Doubt and verify: Data science power tools. Blog Post. <http://www.kdnuggets.com/2015/07/doubt-verify-data-science-power-tools.html> (2015a). Zugegriffen im Juli 2015
- Brodie, M.: The emerging discipline of data science – Principles and techniques for data-intensive analysis. Keynote, SDS|2015, Winterthur, Schweiz. www.zhaw.ch/dlab/brodie (2015b). Zugegriffen im Juni 2015
- Caffo, B.: Developing Data Products. MOOC der Johns Hopkins University, Coursera. <https://www.coursetalk.com/providers/coursera/courses/developing-data-products> (2015). Zugegriffen im Mai 2015
- Chui, M., Farrell, D., Jackson, K.: How government can promote open data and help unleash over \$3 Trillion in economic value. http://www.mckinsey.com/insights/public_sector/how_government_can_promote_open_data (2014). Zugegriffen im April 2014
- Cieliebak, M., Dürr, O., Uzdilli, F.K.: Meta-Classifiers Easily Improve Commercial Sentiment Detection Tools. LREC. (2014, Mai)
- Cleveland, W.S.: Data science: An action plan for expanding the technical areas of the field of statistics. *Stat. Anal. Data Min.: The ASA Data Sci. J.* **7**(6), 414–417 (2014)
- Davenport, T.H., Patil, D.J.: Data scientist: The sexiest job of the 21st century. <http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1> (2012). Zugegriffen im Oktober 2012
- Heitz, Ch., Ruckstuhl, A., Dettling, M. Customer lifetime value under complex contract structures. In Morin, J.-H., Ralyt'e, J., Snene, M. (Hrsg.) IESS 2010, LNBIP 53, 276–281 (2010)
- Heitz, C., Dettling, M., Ruckstuhl, A.: Modelling customer lifetime value in contractual settings. *Int. J. Serv. Technol. Manag.* **16**(2) 172–190 (2011)
- Hey, T., Tansley, S., Tolle, K.: The forth paradigm, microsoft research. (2009, Oktober)

- Howard, J., Zwemer, M., Loukides, M.: *Designing Great Data Products*. O'Reilly Media, ISBN 978-1-449-33367-6 (2012, März)
- Humby, C.: Data is the new Oil!, ANA Senior marketer's summit, Kellogg School. http://ana.blogs.com/maestros/2006/11/data_is_the_new.html (2006). Zugegriffen im November 2006
- Jones, A.: Data science skills and business problems. Blog Post. <http://www.kdnuggets.com/2014/06/data-science-skills-business-problems.html> (2014). Zugegriffen im Juni 2014
- Loukides, M.: What is data science?. Blog Post. <http://radar.oreilly.com/2010/06/what-is-datascience.html> (2010). Zugegriffen im Juni 2010
- Loukides, M.: *The Evolution of Data Product*. O'Reilly Media, ISBN 978-1-449-31651-8 (2011, September)
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: *Big data: The next frontier for innovation, competition, and productivity*. Report. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (2011). Zugegriffen im Mai 2011
- Needham, J.: (2013) *Disruptive Possibilities – How Big Data Changes Everything*. O'Reilly Media, ISBN 978-1-449-36567-7 (2013, Februar)
- Osterwalder, A., Pigneur, Y., Bernarda, G., Smith, A.: *Value Proposition Design*. Wiley, Hoboken (2014)
- Parekh, D.: How big data will transform our economy and our lives in 2015. Blog Post. <http://techcrunch.com/2015/01/02/the-year-of-big-data-is-upon-us/> (2015). Zugegriffen im Januar 2015
- Patil, D.J.: Building data science teams. Blog Post. <http://radar.oreilly.com/2011/09/building-data-science-teams.html> (2011). Zugegriffen im September 2011
- Provost, F., Fawcett, T.: Data science and its relationship to big data and data-driven decision making. *Big Data* 1(1) (2013)
- Siegel, E.: *Predictive Analytics – The Power to Predict Who Will Click, Buy, Lie, or Die*. John Wiley & Sons, ISBN 978-1-118-35685-2 (2013)
- Stadelmann, T., Stockinger, K., Braschler, M., Cieliebak, M., Baudinot, G., Dürr, O., Ruckstuhl, A.: Applied data science in Europe – Challenges for academia in keeping up with a highly demanded topic. In: *European Computer Science Summit. ECSS 2013, Informatics Europe*, Amsterdam, August 2013
- Stockinger, K., Stadelmann, T.: *Data Science für Lehre, Forschung und Praxis. Praxis der Wirtschaftsinformatik, HMD*. 298 Springer, Heidelberg, S. 469–477 (2014)